การพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมาย โดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า

นายจักรินทร์ สันติรัตนภักดี

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาปรัชญาดุษฎีบัณฑิต
สาขาวิชาเทคโนโลยีสารสนเทศ
มหาวิทยาลัยเทคโนโลยีสุรนารี
ปีการศึกษา 2567

THE DEVELOPMENT OF A SEMANTIC-BASED IMAGE RETRIEVAL MODEL USING DEEP NEURAL NETWORK BY PRE-TRAINING APPROACH

CHAKKARIN SANTIRATTANAPHAKDI

A Thesis Submitted in Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy in Information Technology
Suranaree University of Technology
Academic Year 2024

การพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมาย
โดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า

มหาวิทยาลัยเทคโนโลยีสุรนารี อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาดุษฎีบัณฑิต

คณะกรรมการสอบวิทยานิพนธ์

(รองศาสตราจารย์ ดร.ชูพันธุ์ รัตนโภคา)
ประธานกรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.ศุภกฤษฎิ์ นิวัฒนากูล)
อาจารย์ที่ปรึกษา

(รองศาสตราจารย์ ดร.ธรา อั่งสกุล)
กรรมการ

(รองศาสตราจารย์ ดร.ศิรปัฐช์ บุญครอง)
กรรมการ

(อาจารย์ ดร.อรรคพล วงศ์กอบลาภ)
กรรมการ

(รองศาสตราจารย์ ดร.ยุพาพร รักสกุลพิวัฒน์)
รองอธิการบดีฝ่ายวิชาการและประกันคุณภาพ

(รองศาสตราจารย์ ดร.ธรา อั่งสกุล)
คณบดีสำนักวิชาศาสตร์และศิลป์ดิจิทัล

จักรินทร์ สันติรัตนภักดี : การพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายโดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า (THE DEVELOPMENT OF A SEMANTIC-BASED IMAGE RETRIEVAL MODEL USINGก DEEP NEURAL NETWORK BY PRE-TRAINING APPROACH)
อาจารย์ที่ปรึกษา : ผู้ช่วยศาสตราจารย์. ดร.ศุภกฤษฎิ์ นิวัฒนากูล, 226 หน้า.

คำสำคัญ: การค้นคืนรูปภาพเชิงความหมาย/การเรียนรู้ความหมายรูปภาพจากภาษาธรรมชาติ/โครงข่ายประสาทเทียมเชิงลึก

การวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายโดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า ประกอบด้วย 3 มอดูลหลัก ได้แก่ 1) มอดูลการสร้างชุดคำอธิบายรูปภาพ โดยประยุกต์ใช้ตัวแบบ CLIP ที่ถูกฝึกฝนล่วงหน้าสำหรับฝึกฝนตัวเข้ารหัสรูปภาพ และตัวเข้ารหัสข้อความอธิบายรูปภาพ เพื่อแยกความแตกต่างของแต่ละคลาสด้วยการวัดความคล้ายคลึงโคไซน์ตามหลักการเรียนรู้แบบคอนทราสต์ โดยคำนวณค่าความผิดพลาดจากการเปรียบเทียบระหว่างผลการพยากรณ์ป้ายกำกับกับผลการประเมินความหมายของรูปภาพโดยผู้เชี่ยวชาญ จากนั้นนำค่าการสูญเสียไปปรับปรุงพารามิเตอร์ด้วยตนเองเพื่อใช้ในการเรียนรู้ความคล้ายคลึงเชิงความหมายให้ใกล้เคียงกับการรับรู้มนุษย์มากที่สุด ก่อนจะเรียนรู้ซ้ำบนชุดข้อมูลที่กำหนดเอง และนำมาแปลงเป็นเวกเตอร์คุณลักษณะรูปภาพเพื่อจัดเก็บไว้ในชุดคำอธิบายรูปภาพ 2) มอดูลการประมวลผลข้อความค้นหาจากผู้ใช้ในรูปแบบภาษาธรรมชาติด้วยโมเดลภาษา DistilBERT ที่ถูกฝึกฝนล่วงหน้าสำหรับเข้ารหัสข้อความ เพื่อแปลงเป็นเวกเตอร์คุณลักษณะข้อความค้นหา 3) มอดูลการจับคู่เวกเตอร์คุณลักษณะรูปภาพกับเวกเตอร์คุณลักษณะข้อความค้นหาบนพื้นที่การฝังหลายรูปแบบ เพื่อเปรียบเทียบค่าความคล้ายคลึงของเวกเตอร์ ก่อนจะเรียงลำดับตามความเกี่ยวข้อง และแสดงเป็นผลลัพธ์การค้นคืนรูปภาพทั้งในบริบทของเนื้อหาและบริบทเชิงความหมายแก่ผู้ใช้

การประเมินประสิทธิภาพแบบจำลองการค้นคืนรูปภาพโดยผู้เชี่ยวชาญ 3 กลุ่ม ประกอบด้วยกลุ่มผู้เชี่ยวชาญด้านคอมพิวเตอร์และเทคโนโลยี กลุ่มผู้เชี่ยวชาญด้านการค้นคืนสารสนเทศ และกลุ่มผู้ใช้ทั่วไป กลุ่มละ 10 ท่าน รวมทั้งสิ้น 30 ท่าน ด้วยการวัดประสิทธิภาพการค้นคืนรูปภาพแบบไบนารี มีรายละเอียดดังนี้ ค่าความแม่นยำที่ k อันดับ พบว่า ผลลัพธ์จากการค้นคืนจำนวน 3 ลำดับแรกด้วยชื่อรูปภาพและหมวดหมู่ในลักษณะป้ายกำกับของรูปภาพ และข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพนั้นมีค่าความแม่นยำอยู่ในระดับดีมาก คิดเป็นร้อยละ 93.9 และ 86.4 ตามลำดับ อย่างไรก็ดี ความหมายเชิงคุณภาพนั้นขึ้นอยู่กับหลักการรับรู้ของมนุษย์ ดังนั้นประสบการณ์ของแต่ละบุคคลจึงส่งผลให้การประเมินนั้นแตกต่างกัน ส่งผลให้ค่าความแม่นยำ

จากข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพที่อยู่ในระดับดี คิดเป็นร้อยละ 83.0 เช่นเดียวกับ ค่าความครบถ้วนที่ k อันดับ พบว่า ผลลัพธ์จากการค้นคืนจำนวน 3 ลำดับแรก ภายใต้เงื่อนไขข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ในลักษณะป้ายกำกับของรูปภาพ และข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพนั้นมีค่าความครบถ้วนอยู่ในระดับดี ตรงกันข้ามกับข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพที่อยู่ในระดับปานกลาง อย่างไรก็ดี ค่าความครบถ้วนจะค่อย ๆ เพิ่มตามจำนวนผลลัพธ์ที่เพิ่มมากขึ้น เมื่อจำนวนผลลัพธ์เท่ากับ 10 คิดเป็นร้อยละ 81.8, 81.3 และ 80.2 ตามลำดับ นอกจากนั้น ค่าประสิทธิภาพโดยรวมที่ k อันดับ พบว่า ผลลัพธ์จากการค้นคืนจำนวน 3 ลำดับแรก ภายใต้เงื่อนไขข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ในลักษณะป้ายกำกับของรูปภาพ และข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพนั้นมีค่าความครบถ้วนอยู่ในระดับดี เช่นเดียวกับข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ โดยที่ค่าประสิทธิภาพโดยรวมมีค่าสูงขึ้น เมื่อทั้งค่าความแม่นยำในการค้นคืนรูปภาพ และค่าความครบถ้วนจากการค้นคืนรูปภาพมีค่าสูงขึ้นไปในทิศทางเดียวกัน ซึ่งเป็นสิ่งที่แสดงถึงประสิทธิภาพโดยรวมของแบบจำลองที่พัฒนาขึ้น เมื่อจำนวนผลลัพธ์เท่ากับ 10 คิดเป็น 86.0, 83.5 และ 81.0 ตามลำดับ

การประเมินประสิทธิภาพการค้นคืนรูปภาพดิจิทัลเชิงความหมายด้วยค่า NDCG ที่ k อันดับ เปรียบเทียบระหว่างการค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิมกับการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนัก พบว่า ผลลัพธ์จากการค้นคืนรูปภาพด้วยข้อความบรรยายรูปภาพในลักษณะป้ายกำกับของรูปภาพ โดยการค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิมเปรียบเทียบกับการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนัก มีค่า NDCG ที่ลำดับ 1, 3 และ 5 ไม่แตกต่างกันมากนัก เช่นเดียวกับผลลัพธ์จากการค้นคืนรูปภาพด้วยข้อความบรรยายรูปภาพเกี่ยวกับแนวคิดระดับสูงของรูปภาพ โดยการค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิม เปรียบเทียบกับการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนักนั้นมีค่า NDCG ที่ลำดับ 1, 3 และ 5 เพิ่มขึ้นจากเดิมเล็กน้อย ตรงกันข้ามกับ ผลลัพธ์จากการค้นคืนรูปภาพด้วยข้อความบรรยายรูปภาพที่อธิบายความหมายเชิงคุณภาพของรูปภาพ โดยการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนักนั้นมี ค่า NDCG ที่ลำดับ 1, 3 และ 5 เพิ่มขึ้นจากการค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิมถึงร้อยละ 21.50, 19.90 และ 22.80 ตามลำดับ

สำนักวิชาศาสตร์และศิลป์ดิจิทัล ลายมือชื่อนักศึกษา _____

ปีการศึกษา 2567 ลายมือชื่ออาจารย์ที่ปรึกษา _____

CHAKKARIN SANTIRATTANAPHAKDI : THE DEVELOPMENT OF A SEMANTIC-BASED IMAGE RETRIEVAL MODEL USING DEEP NEURAL NETWORK BY PRE-TRAINING APPROACH. THESIS ADVISOR : Asst. Prof. SUPHAKIT NIWATTANAKUL, Ph.D., 226 PP.

KEYWORD: Semantic-based image retrieval/Natural language supervision/ Deep neural network

This research aims to develop a semantic digital image retrieval model using a pre-trained deep neural network, comprising three main modules: 1) Image description set generate module: This module applies a pre-trained CLIP model to train an image encoder and a text encoder for image captions. It distinguishes between classes by measuring cosine similarity based on contrastive learning. The error is calculated by comparing predicted labels with expert semantic evaluations of the images. The loss is then used to manually adjust the weight parameters involved in learning semantic similarity to closely match human perception. Afterwards, the model is fine-tuned on a custom dataset, and image feature vectors are generated and stored in the image caption dataset. 2) Query processing module: This module processes natural language search queries using the pre-trained DistilBERT language model, which encodes queries into vector representations and 3) Vector Matching Module: This module matches image feature vectors with search query vectors in a multi-modal embedding space by comparing their similarity scores. The results are ranked by relevance and presented as image retrieval outputs, reflecting both content and semantic contexts to the user.

The retrieval performance was evaluated by 30 experts divided into three groups: 10 computer and technology specialists, 10 information retrieval experts, and 10 general users. Binary image retrieval performance was measured with the following results: Precision@k: For the top 3 retrieved results, queries based on image names and categories as image labels, as well as short queries concerning high-level image concepts, yielded very high precision rates of 93.9% and 86.4%, respectively. However, since qualitative semantic meaning depends on human

perception, individual experience caused variability in evaluation, resulting in a precision of 83.0% for queries describing qualitative image semantics. Recall@k: For the top 3 results, queries based on image names, category labels, and short high-level concept queries demonstrated good recall levels, whereas queries describing qualitative semantics showed moderate recall. Recall improved as the number of retrieved results increased, reaching 81.8%, 81.3%, and 80.2% at 10 results, respectively. F-measure@k: For the top 3 results, under queries based on image names, category labels, and short high-level concepts, effectiveness was good. Similarly, queries describing qualitative semantics also showed good effectiveness. Overall effectiveness increased as precision and recall improved together, reflecting the robust performance of the developed model. At 10 results, the effectiveness scores were 86.0%, 83.5%, and 81.0%, respectively.

The semantic image retrieval effectiveness was further evaluated using NDCG@k (Normalized Discounted Cumulative Gain), comparing retrieval results between the original CLIP model and the fine-tuned CLIP model. For image caption-based retrieval serving as image labels, NDCG scores at ranks 1, 3, and 5 were similar between the original and fine-tuned models. For retrieval with captions describing high-level image concepts, slight improvements in NDCG were observed at ranks 1, 3, and 5 with the fine-tuned model. In contrast, retrieval with captions describing qualitative semantic content showed significant NDCG increases of 21.50%, 19.90%, and 22.80% at ranks 1, 3, and 5, respectively, when using the fine-tuned CLIP model compared to the original.

Institute of Digital Arts and Science
Academic Year 2024

Student's Signature _____
Advisor's Signature S. Nimit.

กิตติกรรมประกาศ

ดุษฎีนิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดี ผู้วิจัยขอกราบขอบพระคุณทุกท่านที่มีส่วนร่วมในการให้คำปรึกษา แนะนำ สนับสนุนส่งเสริม และช่วยเหลือเป็นอย่างดี โดยเฉพาะอย่างยิ่ง ผู้ช่วยศาสตราจารย์ ดร.ศุภกฤษฎิ์ นิวัฒนากูล อาจารย์ที่ปรึกษาวิทยานิพนธ์ รวมถึงคณาจารย์ทุกท่านของสำนักวิชาศาสตร์และศิลป์ดิจิทัล และคณาจารย์สาขาวิชาอื่น ๆ ที่เกี่ยวข้องในหลักสูตรที่ให้ความสำคัญกับการพัฒนาผู้วิจัย โดยการผลักดันให้เรียนรู้ ศึกษาและวิจัยในประเด็นต่าง ๆ อย่างหลากหลาย อีกทั้งยังส่งเสริมแนวความคิดความสามารถ ประสบการณ์และความรู้ที่เป็นประโยชน์อย่างมหาศาลในการวิจัย เพื่อเพิ่มพูนความเชี่ยวชาญในการวิจัยด้าน การจำแนกรูปภาพเชิงความหมายโดยใช้โครงข่ายประสาทเทียมเชิงลึก การเรียนรู้ความหมายของรูปภาพด้วยข้อความภาษาธรรมชาติ การสกัดคุณลักษณะเชิงความจากการเรียนรู้ด้วยตนเอง และการค้นคืนรูปภาพเชิงความหมาย

ยิ่งไปกว่านั้นขอกราบขอบพระคุณรองศาสตราจารย์ ดร. ชูพันธุ์ รัตนโภคา ประธานกรรมการ ร่วมกับรองศาสตราจารย์ ดร.ธรา อั่งสกุล รองศาสตราจารย์ ดร. ศิรปัฐช์ บุญครอง และอาจารย์ ดร.อรรคพล วงศ์กอบลาภ คณะกรรมการสอบวิทยานิพนธ์ที่ให้คำแนะนำแนวทางในการดำเนินวิจัยของผู้วิจัยเป็นอย่างดียิ่ง

ขอขอบพระคุณมหาวิทยาลัยเทคโนโลยีสุรนารีที่ให้ทุนการศึกษาสำหรับผู้มีผลการเรียนดีเด่นที่เข้าศึกษาต่อหลักสูตรปริญญาเอก หลักสูตรปรัชญาดุษฎีบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ สำนักวิชาศาสตร์และศิลป์ดิจิทัล และขอขอบพระคุณมหาวิทยาลัยวงษ์ชวลิตกุลในฐานะต้นสังกัดที่ให้ทุนสนับสนุนการศึกษาในครั้งนี้

สุดท้ายนี้ ขอกราบขอบพระคุณบิดา มารดา และครอบครัวสันติรัตนภักดี อันเป็นที่รักและเคารพยิ่งที่ให้การอบรมเลี้ยงดู และส่งเสริมด้านการศึกษาเป็นอย่างดีตลอดมา

จักรินทร์ สันติรัตนภักดี

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญตาราง (ต่อ)

ตารางที่ **หน้า**

สารบัญตาราง (ต่อ)

ตารางที่ **หน้า**

สารบัญรูป

รูปที่ **หน้า**

สารบัญรูป (ต่อ)

รูปที่ **หน้า**

สารบัญรูป (ต่อ)

รูปที่ **หน้า**

บทที่ 1
บทนำ

1.1 ความสำคัญและที่มาของปัญหาการวิจัย

เทคโนโลยีการถ่ายภาพที่พัฒนาจากกล้องฟิล์มเป็นอุปกรณ์ถ่ายภาพดิจิทัล ประกอบกับมีราคาที่ถูกลง และสะดวกต่อการใช้งานในลักษณะอุปกรณ์พกพา ส่งผลให้เกิดรูปภาพจำนวนมาก ตามข้อมูลของเว็บไซต์ Phototutorial (Broz, 2024, www: 1) รายงานว่ามีจำนวนรูปภาพถึง 1.2 ล้านล้านภาพในปี 2021 และเพิ่มเป็น 1.72 ล้านล้านภาพในปี 2022 จำนวนนี้เพิ่มขึ้นเป็น 1.81 ล้านล้านภาพในปี 2023 และภายในปี 2025 จะมีรูปภาพมากกว่า 2 ล้านล้านภาพ อีกทั้งมีแนวโน้มที่จะเพิ่มมากขึ้นอีกในอนาคต โดยมีปัจจัยหลักอยู่ที่การพัฒนากล้องถ่ายภาพของสมาร์ทโฟน และพฤติกรรมการใช้เครือข่ายสังคมออนไลน์ในการแบ่งปันเรื่องราวและบันทึกความทรงจำ ส่งผลให้มีรูปภาพจำนวนมหาศาลบนอินเทอร์เน็ต ดังนั้นก่อนการนำไปใช้ประโยชน์จำเป็นต้องมีการจำแนกประเภทข้อมูลภาพ (Image classification) เพื่อแยกรูปภาพออกเป็นกลุ่มหรือคลาส (Class) ที่มีคุณลักษณะเป็นไปในทิศทางเดียวกัน (Canty, 2014)

หนึ่งในแนวคิดที่ถูกนำมาใช้อย่างแพร่หลายในการจำแนกประเภทข้อมูลภาพก็คือการเรียนรู้ของเครื่อง (Machine learning) (Aggarwal, 2015) เพื่อสร้างตัวจำแนกข้อมูล (Classifier) ด้วยโครงข่ายประสาทเทียม (Neural network classifier) (Wozniak, 2014) ที่เลียนแบบการทำงานของโครงข่ายประสาทของมนุษย์ (Vasilev, Slater, Spacagna, Roelants and Zocca, 2019) อย่างไรก็ดี โครงข่ายประสาทเทียมทั่วไปนั้นอาจทำงานได้ดีกับรูปภาพที่ไม่ซับซ้อนมากนัก เช่น รูปภาพที่มีวัตถุอยู่ตรงกลาง รูปภาพวัตถุในคลาสเดียวกัน หรือรูปภาพวัตถุที่มีคุณลักษณะเฉพาะตัวแตกต่างกันอย่างชัดเจน เนื่องจากมนุษย์เป็นผู้กำหนดคุณลักษณะในการเรียนรู้ ตรงกันข้ามหากรูปภาพมีความซับซ้อนจะยากต่อการกำหนดคุณลักษณะ และส่งผลให้ตัวจำแนกเรียนรู้ได้ไม่ดีนัก เป็นที่มาของการนำการเรียนรู้เชิงลึก (Deep learning) (Zhang, 2019) เข้ามาประยุกต์ใช้กับการจำแนกข้อมูลรูปภาพ ด้วยการใช้ตัวแบบที่ผ่านการฝึกฝนล่วงหน้า (Pre-trained model) โดยอาศัยเทคนิคการถ่ายโอนการเรียนรู้ (Transfer learning) บนชุดข้อมูลขนาดใหญ่ แล้วนำมาปรับแต่งพารามิเตอร์ (Fine-tuning) เพื่อให้สามารถเรียนรู้ซ้ำ (Retrain) บนชุดข้อมูลที่กำหนดเอง (Custom dataset) จึงสามารถใช้เพียงไม่กี่ร้อยภาพต่อคลาสในการฝึกตัวฝน ส่งผลให้ประหยัดเวลาในการฝึกฝนและประสิทธิภาพสูงกว่าตัวแบบที่เริ่มต้นการฝึกจากศูนย์ (Training from scratch) ทั้งนี้ กระบวนการเรียนรู้ของตัวแบบยังคงอาศัยกลไกการแพร่กระจายย้อนกลับ (Backpropagation) เพื่อนำค่าความสูญเสียกลับมาปรับปรุงค่าน้ำหนักของโครงข่ายประสาทเทียมในแต่ละชั้นอย่างเป็น

ลำดับจนกว่าจะได้ผลลัพธ์ที่เหมาะสมที่สุด โดยการวิจัยนี้ประยุกต์ใช้ตัวแบบ CLIP (Contrastive Language–Image Pre-training) ที่ถูกฝึกฝนล่วงหน้าจากคู่ข้อความและรูปภาพ (Radford et al., 2021) มาฝึกฝนตัวแบบให้เข้าใจความสัมพันธ์ระหว่างคำบรรยายภาพและรูปภาพให้ถูกต้องมากที่สุด ก่อนจะสร้างป้ายกำกับเชิงความหมายจากผลการพยากรณ์สำหรับการค้นคืนรูปภาพเชิงความหมาย เพื่อให้คอมพิวเตอร์สามารถเข้าใจรูปภาพได้ใกล้เคียงมนุษย์มากที่สุด (Tyagi, 2017)

แม้การค้นคืนรูปภาพเชิงเนื้อหาจะทำงานได้อย่างมีประสิทธิภาพด้วยการระบุแนวคิดระดับสูงที่เป็นรูปธรรม (Concrete concept) อย่างวัตถุ (Object) กิจกรรม (Activities) ฉาก (Scene) หรือเหตุการณ์ (Event) แต่ในความเป็นจริงพฤติกรรมผู้ใช้โดยมากจะค้นคืนรูปภาพโดยใช้ความสัมพันธ์ระหว่างแนวคิดระดับสูงที่เป็นรูปธรรม และแนวคิดที่เป็นนามธรรม (Abstract concept) อย่างอารมณ์และความรู้สึก (Mood and feelings) รวมเข้าด้วยกันเป็นข้อความค้นหาภาษาธรรมชาติ ในทางกลับกันคุณลักษณะที่ถูกแยกอัตโนมัติด้วยคอมพิวเตอร์มักเป็นแนวคิดระดับสูงที่เป็นรูปธรรมเท่านั้น ดังนั้นการค้นคืนรูปภาพจากเนื้อหาเชิงประจักษ์เพียงอย่างเดียวอาจส่งผลให้เกิดปัญหาช่องว่างความหมาย (Semantic-gap) (Liu and Song, 2011) ระหว่างความต้องการของผู้ใช้ที่อยู่ในลักษณะข้อความค้นหา ความหมายของรูปภาพ และคอมพิวเตอร์ ซึ่งโดยทั่วไปไม่มีการเชื่อมโยงโดยตรงระหว่างแนวคิดระดับสูงที่เป็นรูปธรรมและที่เป็นนามธรรมทำให้ผลลัพธ์จากการค้นคืนมีประสิทธิภาพไม่สูงเท่าที่ควร เป็นที่มาของงานวิจัยชิ้นนี้ที่มุ่งเชื่อมโยงระหว่างแนวคิดระดับสูงที่เป็นรูปธรรมและนามธรรมผ่านแบบจำลองการค้นคืนรูปภาพเชิงความหมาย (Semantic-based Image Retrieval: SBIR) โดยเปรียบเทียบระหว่างผลการพยากรณ์ความน่าจะเป็นของป้ายกำกับภาษาธรรมชาติกับผลการประเมินความหมายรูปภาพจากผู้เชี่ยวชาญ ก่อนจะนำไปปรับค่าน้ำหนักตามแนวคิดการแผ่กระจายย้อนหลัง เพื่อให้แบบจำลองเรียนรู้คุณลักษณะเชิงความหมายของรูปภาพให้ใกล้เคียงกับมนุษย์มากยิ่งขึ้น

ปัจจุบันงานวิจัยที่ประยุกต์ใช้เทคนิคต่าง ๆ เพื่อเพิ่มประสิทธิภาพการค้นคืนรูปภาพเชิงความหมายมีเป็นจำนวนมาก แต่ละงานวิจัยก็มีเทคนิคและวิธีการแตกต่างกันไป แต่เมื่อพิจารณาลงไปในรายละเอียดพบว่างานวิจัยส่วนใหญ่มุ่งใช้รูปภาพค้นหา (Image query) ในการค้นคืนรูปภาพ เนื่องจากระบบจะวิเคราะห์คุณลักษณะรูปภาพทั้งหมดเพื่อค้นคืนรูปภาพตามความเหมือนหรือคล้ายคลึงกับรูปภาพค้นหา จึงสามารถลดปัญหาช่องว่างความหมายลงได้ แต่ในความเป็นจริงพฤติกรรมผู้ใช้โดยมากจะค้นคืนรูปภาพโดยใช้ความสัมพันธ์ระหว่างคุณลักษณะระดับต่ำ ผสมผสานแนวคิดระดับสูงที่เป็นรูปธรรม และนามธรรมรวมเข้าด้วยกันเป็นข้อความค้นหาภาษาธรรมชาติ แต่ยังมีผู้ใช้จำนวนไม่น้อยที่ขาดความชำนาญในการใช้คำค้น ส่วนใหญ่มักใช้คำสำคัญ (Keyword) เพียง 1 หรือ 2 คำเป็นหลักในการค้นหาแต่ละครั้ง ประกอบกับบางอัลกอริทึมยังไม่เอื้อต่อการค้นหาด้วยข้อความในรูปแบบภาษาธรรมชาติ (Natural language) จึงอาจเกิดผลลัพธ์จำนวนมากทั้งที่ตรง

และไม่ตรงกับความต้องการของผู้ใช้ ส่งผลให้ผลลัพธ์จากการค้นคืนมีประสิทธิภาพไม่สูงเท่าที่ควร เนื่องจากไม่มีการแปลงเป็นแนวคิดเชิงความหมาย (High-level semantic) ในรูปแบบที่มนุษย์สามารถเข้าใจได้ อีกทั้งการค้นคืนโดยใช้รูปภาพค้นหายังไม่สะดวก และเป็นการผลักภาระให้กับผู้ใช้ในการหารูปภาพต้นฉบับอีกด้วย

งานวิจัยนี้มุ่งพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมาย เพื่อเชื่อมโยงระหว่างแนวคิดระดับสูงที่เป็นรูปธรรมและที่เป็นนามธรรมให้เป็นแนวคิดเชิงความหมายในรูปแบบที่มนุษย์สามารถเข้าใจได้โดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า อันจะสามารถลดช่องว่างความหมายระหว่างข้อความค้นหาภาษาธรรมชาติ ความหมายของรูปภาพ และการแปลความหมายของรูปภาพด้วยคอมพิวเตอร์ ตลอดจนช่วยสนับสนุนผู้ใช้ด้วยข้อความค้นหาภาษาธรรมชาติที่ยึดโยงกับความหมายของรูปภาพแทนที่จะยึดตามหลักไวยากรณ์ของภาษา

1.2 วัตถุประสงค์ของการวิจัย

1.2.1 เพื่อพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายโดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า

1.2.2 เพื่อประเมินประสิทธิภาพแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมาย

1.3 สมมติฐานการวิจัย

แบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมาย มีผลการประเมินประสิทธิภาพด้วยค่าความแม่นยำที่ k อันดับ (Precision@k) ค่าความครบถ้วนที่ k อันดับ (Recall@k) และค่าประสิทธิภาพโดยรวมที่ k อันดับ (F-measure@k) มากกว่าร้อยละ 80

1.4 ขอบเขตของการวิจัย

1.4.1 การพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายด้วยการเขียนโปรแกรมภาษาไพทอน โดยประยุกต์ใช้ตัวแบบ CLIP ที่ถูกฝึกฝนล่วงหน้ารุ่น ViT-B/32 เป็นโมเดลฐานบน Google Colaboratory

1.4.2 การประเมินประสิทธิภาพการค้นคืนรูปภาพแบบไบนารีด้วยมาตรวัดแบบไม่สนใจลำดับของผลลัพธ์ ประกอบด้วย ค่าความแม่นยำที่ k อันดับ ค่าความครบถ้วนที่ k อันดับ และค่าประสิทธิภาพโดยรวมที่ k อันดับ และมาตรวัดความเกี่ยวข้องแบบให้คะแนนด้วยค่า NDCG ที่ k อันดับ

1.4.3 การประเมินประสิทธิภาพการค้นคืนรูปภาพโดยผู้เชี่ยวชาญแบ่งออกเป็น 3 กลุ่ม ได้แก่ 1) กลุ่มผู้เชี่ยวชาญด้านคอมพิวเตอร์และเทคโนโลยี 2) กลุ่มผู้เชี่ยวชาญด้านการค้นคืนสารสนเทศ และ 3) กลุ่มผู้ใช้ทั่วไป กลุ่มละ 10 ท่าน รวมทั้งสิ้น 30 ท่าน

1.4.4 กำหนดคุณสมบัติผู้เชี่ยวชาญ ดังนี้ 1) จบการศึกษาระดับปริญญาโทขึ้นไป หรือมีความรู้ความเชี่ยวชาญเฉพาะด้านอันเป็นที่ประจักษ์ 2) มีประสบการณ์การใช้งานคอมพิวเตอร์และอินเทอร์เน็ตอย่างน้อย 10 ปีขึ้นไป 3) มีประสบการณ์การใช้งานคอมพิวเตอร์และสมาร์ทโฟนอย่างน้อย 10 ปีขึ้นไป 4) มีความสามารถในการให้ข้อความค้นหาในรูปแบบภาษาอังกฤษ และ 5) สามารถประเมินความถูกต้องของการค้นคืนรูปภาพจากข้อความค้นหาภาษาอังกฤษ

1.5 ข้อตกลงเบื้องต้น

1.5.1 กำหนดการสร้างข้อความบรรยายรูปภาพจากโมเดล LLaVA โดยใช้คำสั่ง (Prompt) เพื่อสร้างคำบรรยายรูปภาพในโดเมนอารมณ์พื้นฐานของมนุษย์ ประกอบด้วย ความรื่นเริง ความไว้วางใจ ความกลัว ความประหลาดใจ ความเศร้า ความรังเกียจ ความโกรธ และความคาดหวัง โดยในแต่ละโดเมนจะประกอบด้วยอารมณ์ย่อยโดเมนละ 5 อารมณ์

1.5.2 กำหนดคำบรรยายรูปภาพในรูปแบบภาษาอังกฤษจำนวน 5 รายการต่อ 1 รูปภาพ

1.5.3 กำหนดข้อความค้นหาในรูปแบบภาษาอังกฤษเท่านั้น โดยไม่มีข้อจำกัดความยาวของคำหรือประโยค ตลอดจนตัวอักษรพิมพ์เล็กพิมพ์ใหญ่ หรืออักขระพิเศษจะไม่มีผลในการค้นหา

1.5.4 กำหนดข้อความค้นหาภาษาอังกฤษภายใต้ 3 เงื่อนไข ได้แก่ 1) ข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ ในลักษณะป้ายกำกับของรูปภาพ 2) ข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพ และ 3) ข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ

15.5 กำหนดขอบเขตการค้นคืนรูปภาพบนชุดข้อมูล Flickr30k จำนวน 31,783 รูปภาพ และชุดข้อมูลที่ผู้วิจัยรวบรวมเองจำนวน 92,168 รูปภาพ รวมทั้งสิ้น 123,951 รูปภาพ

1.5.6 กำหนดขนาดของรูปภาพในการแสดงผลความกว้างและความยาวเท่ากับ 250 * 250 พิกเซล โดยเรียงลำดับผลลัพธ์ตามค่าความคล้ายคลึงจากมากไปน้อยจำนวน 20 รูปภาพ อย่างไรก็ดีจะไม่ปรากฏค่าคะแนนความคล้ายคลึงแก่ผู้ใช้ในขั้นตอนของการทำงาน

1.5.7 กำหนดการทดสอบแบบจำลองบน Google Colaboratory แบบมีค่าใช้จ่าย โดยใช้ GPU เป็นหน่วยประมวลผล และใช้หน่วยความจำขนาด 25 กิกะไบต์ในโหมด High-RAM

1.5.8 กำหนดการใช้งานรูปภาพอย่างคำนึงถึงสิทธิส่วนบุคคล โดยไม่ปรากฏใบหน้าของบุคคลหรือส่วนหนึ่งส่วนใดอย่างชัดเจนจนสามารถนำไปสู่การระบุอัตลักษณ์บุคคลได้

1.6 ประโยชน์ที่คาดว่าจะได้รับ

1.6.1 ได้แบบจำลองการค้นคืนรูปภาพดิจิทัลที่ลดข้อจำกัดช่องว่างความหมาย ภายใต้ข้อความค้นหาภาษาธรรมชาติที่ยึดโยงกับความหมายของรูปภาพแทนที่จะยึดความถูกต้องตามหลักไวยากรณ์ของภาษา

1.6.2 ได้ชุดข้อมูลคุณลักษณะเชิงความหมายของรูปภาพจากการเรียนรู้ของตัวแบบโครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้าจากการเรียนรู้ร่วมกันระหว่างรูปภาพและข้อความบรรยายรูปภาพ

1.6.3 ได้แนวทางการค้นคืนรูปภาพดิจิทัลเชิงความหมาย อันเป็นแนวทางในการพัฒนาการค้นคืนสารสนเทศในรูปแบบที่เกี่ยวข้องต่อไปในอนาคต

1.7 คำอธิบายศัพท์

1.7.1 รูปภาพดิจิทัล คือข้อมูลรูปภาพที่ถูกจัดเก็บในรูปแบบดิจิทัล โดยใช้เลขฐานสองในการกำหนดรายละเอียดแต่ละส่วนของภาพ ซึ่งสามารถเปลี่ยนแปลง แก้ไข ประมวลผล หรือถ่ายโอนได้โดยใช้ระบบคอมพิวเตอร์

1.7.2 โครงข่ายประสาทเทียมเชิงลึก คือพัฒนาการจากการเรียนรู้ของเครื่องในการสร้างแบบจำลองด้วยการประมวลผลข้อมูลหลายชั้นที่มีโครงสร้างที่ซับซ้อน

1.7.3 ตัวแบบที่ถูกฝึกฝนล่วงหน้า หมายถึง ตัวแบบที่ถูกฝึกฝนมาแล้วล่วงหน้า โดยใช้เทคนิคการถ่ายโอนการเรียนรู้ จากนั้นนำมาปรับปรุงพารามิเตอร์ แล้วจึงเรียนรู้ซ้ำบนชุดข้อมูลที่กำหนดเองเพื่อลดภาระการฝึกฝนแบบจำลองด้วยชุดข้อมูลจำนวนมาก

1.7.4 ตัวแบบ CLIP ที่ถูกฝึกฝนล่วงหน้า คือโครงข่ายประสาทเทียมเชิงลึกที่เรียนรู้แบบคอนทราสต์ระหว่างข้อความและรูปภาพ เพื่อเรียนรู้ความหมายของรูปภาพด้วยภาษาธรรมชาติ

1.7.5 การค้นคืนรูปภาพเชิงความหมาย คือการค้นคืนรูปภาพจากการแปลงคุณลักษณะของรูปภาพโดยเชื่อมโยงระหว่างแนวคิดระดับสูงที่เป็นรูปธรรมและนามธรรมเข้าด้วยกัน เพื่อค้นคืนรูปภาพด้วยเทคนิคต่าง ๆ ในการแปลความหมายที่เป็นนามธรรมที่ซ่อนอยู่ในรูปภาพ

บทที่ 2
ปริทัศน์วรรณกรรมและงานวิจัยที่เกี่ยวข้อง

การพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมาย โดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า การศึกษาครั้งนี้ผู้วิจัยได้ปริทัศน์วรรณกรรมและงานวิจัยที่เกี่ยวข้อง มีรายละเอียดดังนี้

2.1 การค้นคืนรูปภาพ
- 2.1.1 การค้นคืนรูปภาพด้วยข้อความ
- 2.1.2 การค้นคืนรูปภาพจากเนื้อหา
- 2.1.3 การค้นคืนรูปภาพเชิงความหมาย

2.2 เทคนิคการค้นคืนรูปภาพเชิงความหมาย
- 2.2.1 เทคนิคการค้นคืนรูปภาพหลายระดับ
- 2.2.2 เทคนิคการใช้แม่แบบเชิงความหมาย
- 2.2.3 เทคนิคการใช้ข้อความเพื่อเรียนรู้ความหมายของรูปภาพ
- 2.2.4 เทคนิคการตอบกลับของผู้ใช้
- 2.2.5 เทคนิคการใช้ออนโทโลยี
- 2.2.6 เทคนิคการจำแนกข้อมูลด้วยการเรียนรู้ของเครื่อง

2.3 โครงข่ายประสาทเทียมเชิงลึก

2.4 สถาปัตยกรรม ViT

2.5 การเรียนรู้ด้วยตนเอง

2.6 ตัวแบบ CLIP ที่ถูกฝึกฝนล่วงหน้า

2.7 การวัดประสิทธิภาพการค้นคืนรูปภาพแบบไบนารี
- 2.7.1 มาตรวัดแบบไม่สนใจลำดับของผลลัพธ์
- 2.7.2 มาตรวัดความเกี่ยวข้องแบบให้คะแนน

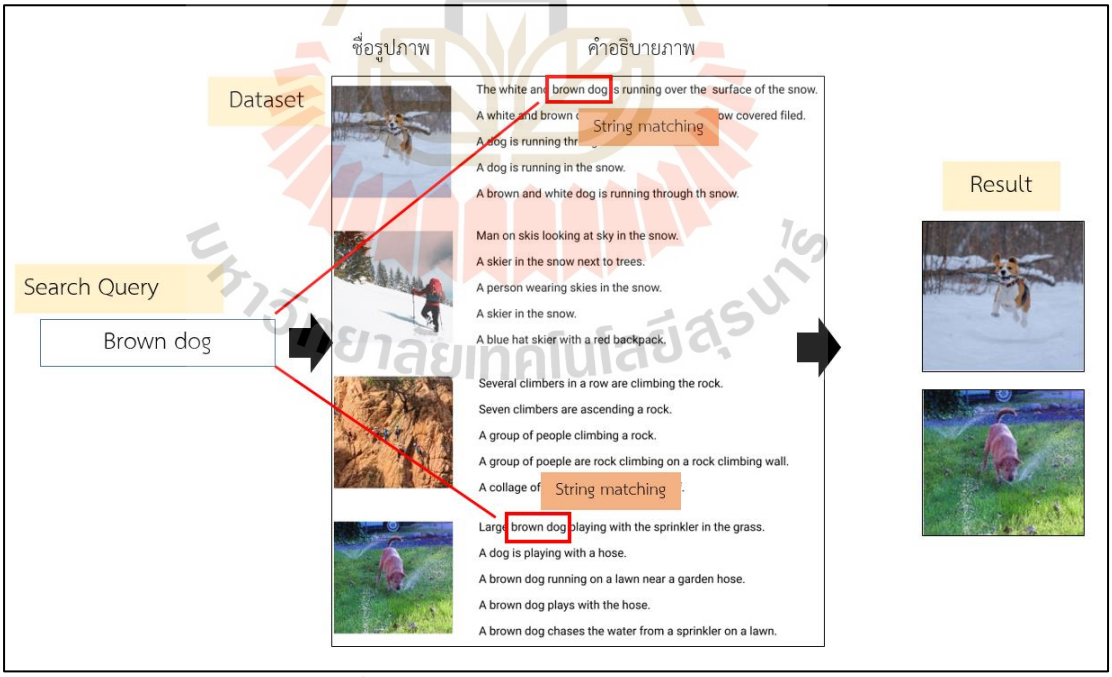
2.8 งานวิจัยที่เกี่ยวข้อง

2.1 การค้นคืนรูปภาพ

การค้นคืนรูปภาพเป็นหนึ่งในศาสตร์ด้านคอมพิวเตอร์วิทัศน์ที่มุ่งเปลี่ยนรูปภาพดิจิทัลให้กลายเป็นข้อมูลที่มีความหมาย เพื่อให้คอมพิวเตอร์สามารถเข้าใจรูปภาพได้ใกล้เคียงมนุษย์มากที่สุด (อารัมภ์ กาวีวงศ์, 2555) แบ่งตามลักษณะการทำงานได้ 3 รูปแบบ คือ

2.1.1 การค้นคืนรูปภาพด้วยข้อความ

การค้นคืนรูปภาพด้วยข้อความ (Text-Based Image Retrieval: TBIR) (Tyagi, 2017) การค้นคืนรูปภาพในยุคแรกจะใช้การเปรียบเทียบคำสำคัญ (Keyword) กับชื่อไฟล์รูปภาพหรือคำอธิบายภาพ (Image description) เมื่อพบข้อความที่ตรงกัน (String matching) (AbdElrazek, 2017) รูปภาพนั้นจะถูกดึงมาเป็นผลลัพธ์ เพื่อนำเสนอแก่ผู้ใช้ดังรูปที่ 2.1 การเปรียบเทียบระหว่างคำสำคัญกับคำอธิบายรูปภาพ แต่ปัญหาที่พบคือบางครั้งการตั้งชื่อไฟล์ไม่ตรงกับเนื้อหาของรูปภาพ เช่น ชื่อไฟล์เป็นตัวเลข 1.JPG จะทำให้รูปภาพเหล่านี้ไม่สามารถถูกค้นคืนได้ อีกทั้งการใส่คำอธิบายภาพโดยมนุษย์อาจไม่สามารถอธิบายได้ครอบคลุมเนื้อหาของรูปภาพได้ทั้งหมด หรือไม่สื่อความหมายที่แท้จริงของรูปภาพ ดังนั้นผลลัพธ์จากการค้นคืนรูปภาพด้วยข้อความจึงมักมีประสิทธิภาพต่ำ เนื่องจากมิได้วิเคราะห์ถึงแนวคิดระดับสูงของรูปภาพ

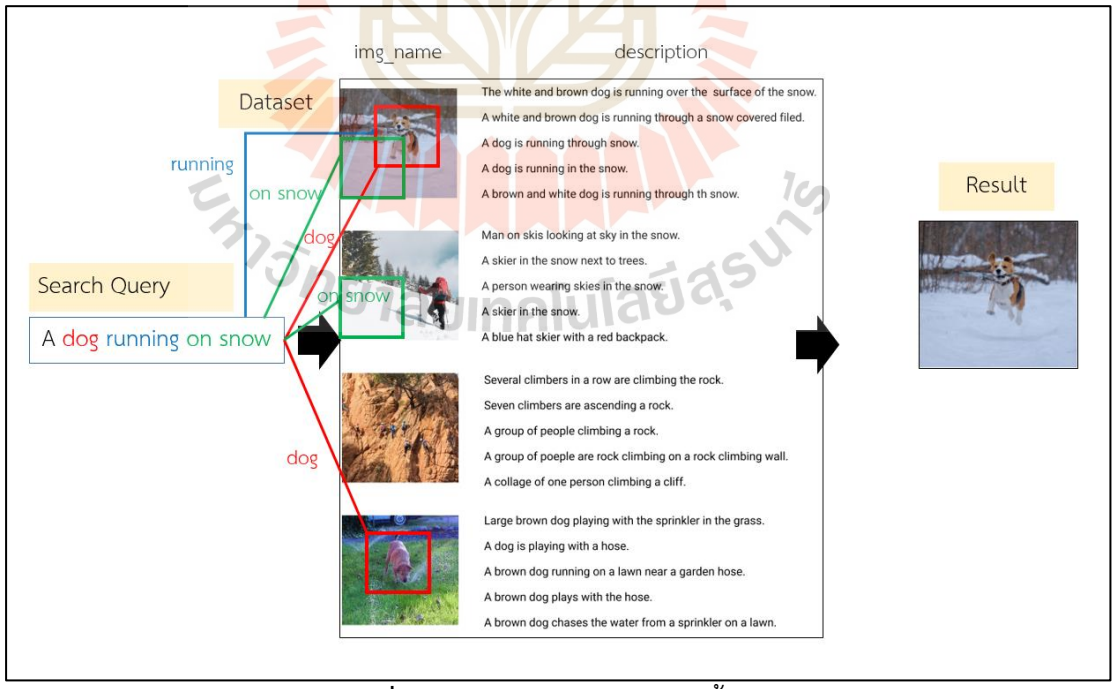


รูปที่ 2.1 การค้นคืนรูปภาพด้วยข้อความ

2.1.2 การค้นคืนรูปภาพจากเนื้อหา

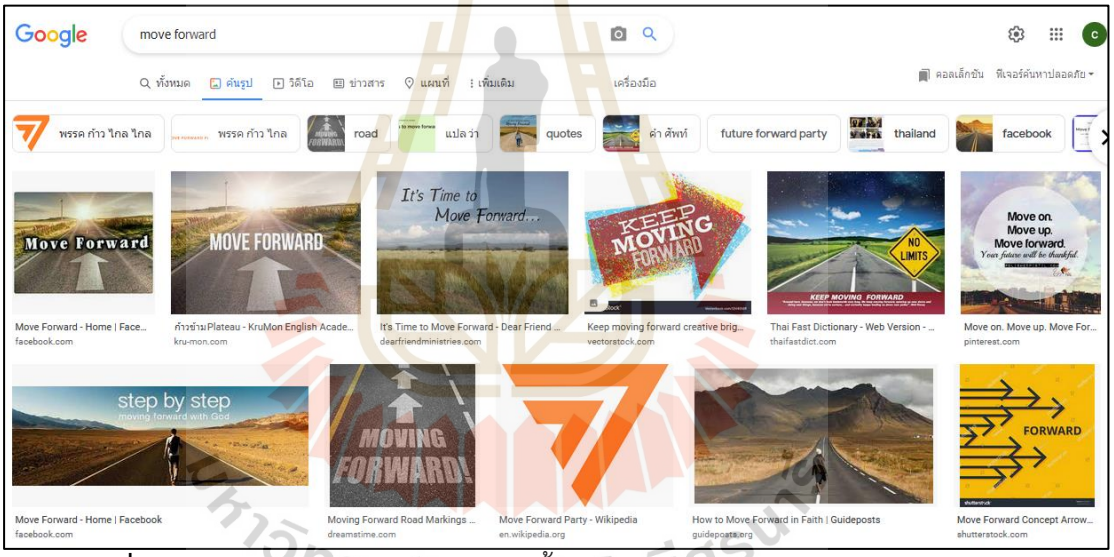
การค้นคืนรูปภาพจากเนื้อหา (Content-based Image Retrieval: CBIR) (Tyagi, 2017) คือการนำคุณลักษณะระดับต่ำ (Low-level features) ซึ่งเป็นข้อมูลที่ไม่มีความหมายในตัวเอง และไม่สื่อถึงความหมายใด ๆ ในรูปภาพมาเปรียบเทียบกับคำค้นหา ประกอบด้วยคุณลักษณะโกลบอล (Global features) ซึ่งเป็นคุณลักษณะแบบรวม ๆ ไม่เฉพาะเจาะจง จุดเด่นคือความง่าย ไม่ซับซ้อน และประมวลผลได้รวดเร็ว ประกอบด้วยคุณลักษณะสี คุณลักษณะรูปทรง คุณลักษณะพื้นผิว และคุณลักษณะตำแหน่งของวัตถุในภาพสองมิติ ใช้ร่วมกับคุณลักษณะโลคอล (Local features) ซึ่งจะเป็นคุณลักษณะเฉพาะของแต่ละส่วนในรูปภาพ ปัจจุบันมีเทคนิคที่ได้รับความนิยม ได้แก่ SIFT, SURF และ LBP เป็นต้น

ตัวอย่างการค้นคืนรูปภาพจากเนื้อหาจะเริ่มจากรับข้อความค้นหาจากผู้ใช้ (Search query) เช่น “a dog running on snow” ดังรูปที่ 2.2 ผ่านตัวเข้ารหัสข้อความ (Text encoder) เพื่อแปลงเป็นเวกเตอร์คุณลักษณะข้อความค้นหา แล้วนำมาเปรียบเทียบความคล้ึงกับเวกเตอร์คุณลักษณะรูปภาพในชุดข้อมูลที่ผ่านตัวเข้ารหัสรูปภาพ (Image encoder) บนพื้นที่การฝังหลายรูปแบบ (Multi-dimension embedding) พบว่า มีเพียงรูปภาพที่ 1 เท่านั้นที่มีค่าความคล้ายคลึงเกินค่าแบ่งและถูกค้นคืนออกมาเป็นผลลัพธ์



รูปที่ 2.2 การค้นคืนรูปภาพจากเนื้อหา

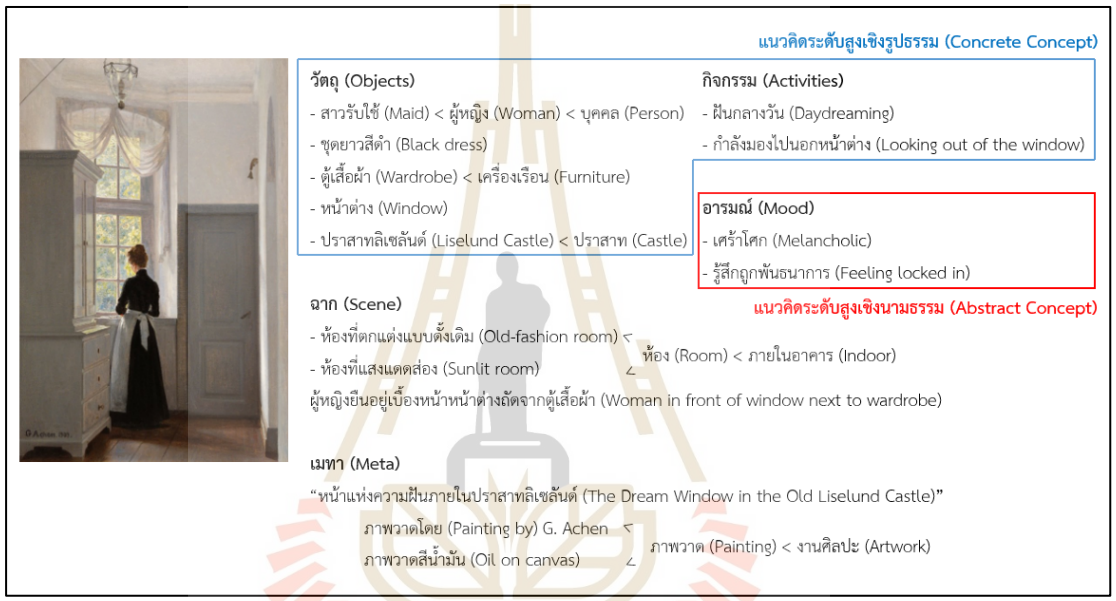
ปัจจุบันการค้นคืนรูปภาพจากเนื้อหาทำงานได้อย่างมีประสิทธิภาพ ดังรูปที่ 2.3 ผลลัพธ์การค้นคืนรูปภาพจากเนื้อหาบนเว็บไซต์ google.com ด้วยข้อความค้นหา “move forward” เป็นรูปภาพที่มีข้อความนั้นฝังอยู่ในภาพ หรือเป็นรูปภาพภายในเว็บเพจที่มีหัวข้อหรือเนื้อหาตรงกับข้อความค้นหาตามแนวคิดการใช้ข้อความที่อยู่รอบ ๆ ในการอธิบายความหมายของรูปภาพ ร่วมกับการรวบรวมข้อมูลด้วยเว็บครอว์เลอร์ (Crawling) การทำดัชนีในการค้นหา (Indexing) และการจัดอันดับผลการค้นหา (Ranking) มานำเสนอแก่ผู้ใช้ จะเห็นได้ว่าผลลัพธ์การค้นคืนรูปภาพจากเนื้อหา เช่น รูปภาพที่มีข้อความว่า Move Forward, Moving Forward หรือ Forward รวมถึงหัวข้อที่เกี่ยวข้องกับคำว่า Move Forward ซึ่งจะเห็นได้ว่าเป็นการค้นคืนจากเนื้อหาในรูปภาพ หรือข้อความในเว็บเพจที่มีข้อความเกี่ยวข้องกับคำค้นหาเท่านั้น แต่มิได้พิจารณาจากคุณลักษณะเชิงความหมายของรูปภาพเลย



รูปที่ 2.3 ผลลัพธ์การค้นคืนรูปภาพจากเนื้อหาด้วยข้อความค้นหา “move forward” บนเว็บไซต์ google.com

อย่างไรก็ดี “A picture is worth a thousand words” แสดงถึงเนื้อหาของรูปภาพสามารถสื่อความหมายได้หลากหลายจากความสัมพันธ์ระหว่างแนวคิดระดับสูงที่เป็นรูปธรรม (Concrete concept) อย่างวัตถุ กิจกรรม หรือเหตุการณ์ และแนวคิดระดับสูงที่เป็นนามธรรม (Abstract concept) อย่างอารมณ์และความรู้สึก (Barz and Denzler, 2020) ดังรูปที่ 2.4 ดังนั้นการค้นคืนรูปภาพจากเนื้อหาเชิงประจักษ์เพียงอย่างเดียวอาจส่งผลให้เกิดปัญหาช่องว่างความหมาย (Semantic-gap) ระหว่างความต้องการของผู้ใช้ที่อยู่ในลักษณะข้อความค้นหาภาษาธรรมชาติ ความหมายของรูปภาพ และคอมพิวเตอร์ เนื่องจากในความเป็นจริงพฤติกรรมผู้ใช้

ส่วนใหญ่จะมีแนวโน้มค้นคืนรูปภาพโดยใช้ความสัมพันธ์ระหว่างแนวคิดระดับสูงที่เป็นรูปธรรมและที่เป็นนามธรรม รวมเข้าด้วยกันเป็นข้อความค้นหา ในทางกลับกันคุณลักษณะที่ถูกแยกอัตโนมัติและแปลผลโดยใช้คอมพิวเตอร์มักเป็นแค่แนวคิดระดับสูงที่เป็นรูปธรรมเท่านั้น ซึ่งโดยทั่วไปไม่มีการเชื่อมโยงโดยตรงระหว่างแนวคิดระดับสูงที่เป็นรูปธรรมและที่เป็นนามธรรมให้เป็นแนวคิดเชิงความหมายในรูปแบบที่มนุษย์สามารถเข้าใจได้ ส่งผลให้การค้นคืนภาพไม่สามารถสะท้อนความต้องการของผู้ใช้ได้อย่างครบถ้วนในทุกมิติของความหมาย



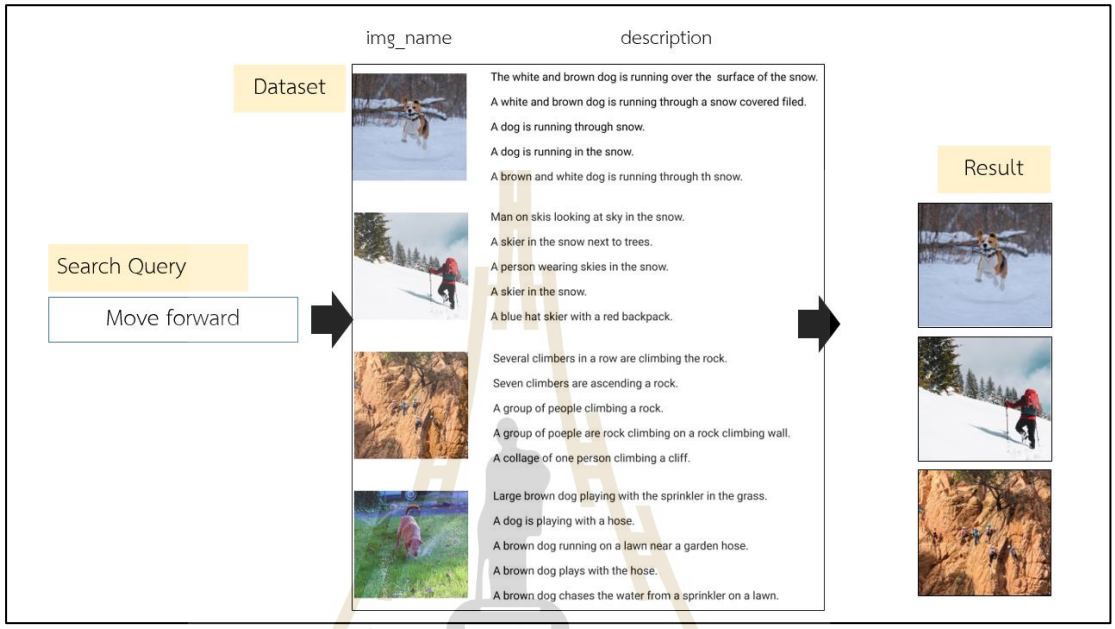
รูปที่ 2.4 ความสัมพันธ์ระหว่างแนวคิดระดับสูงที่เป็นรูปธรรมและที่เป็นนามธรรม
ที่มา: Barz and Denzler (2020)

2.1.3 การค้นคืนรูปภาพเชิงความหมาย

การค้นคืนรูปภาพเชิงความหมาย (Semantic-based Image Retrieval: SBIR) (Barz, 2020) คือการแปลงคุณลักษณะของรูปภาพโดยเชื่อมโยงระหว่างแนวคิดระดับสูงที่เป็นรูปธรรมและที่เป็นนามธรรมเข้าด้วยกัน เพื่อทำการค้นคืนรูปภาพ โดยใช้เทคนิคต่าง ๆ ในการแปลความหมายที่เป็นนามธรรมที่ซ่อนอยู่ในรูปภาพให้อยู่ในรูปแบบที่มนุษย์สามารถเข้าใจได้ เพื่อเพิ่มประสิทธิภาพการค้นคืนรูปภาพให้สูงขึ้น

ตัวอย่างข้อความค้นหาซึ่งอาจทำให้เกิดปัญหาช่องว่างความหมายในการค้นคืนรูปภาพ เช่น “Move Forward” ผ่านตัวเข้ารหัสข้อความ เพื่อแปลงเป็นเวกเตอร์คุณลักษณะข้อความค้นหา แล้วเปรียบเทียบความคล้ายคลึงกับคุณลักษณะรูปภาพที่ผ่านตัวเข้ารหัสรูปภาพอยู่ในรูปแบบเวกเตอร์คุณลักษณะรูปภาพที่อยู่ในชุดข้อมูล ก่อนจะนำมาวิเคราะห์หาคุณลักษณะเชิงความหมายของรูปภาพ

จากการเรียนรู้รูปภาพของตัวเข้ารหัสเชิงความหมาย (Semantic encoder) เมื่อเปรียบเทียบความคล้ายคลึงระหว่างเวกเตอร์ จะพบว่ามี 3 รูปภาพที่มีค่าความคล้ายคลึงเกินค่าเกณฑ์ที่กำหนด (Threshold) และถูกค้นคืนออกมาเป็นผลลัพธ์ดังรูปที่ 2.5



รูปที่ 2.5 การค้นคืนรูปภาพเชิงความหมาย

จากรูปที่ 2.5 จะเห็นได้ว่าผลลัพธ์ของกระบวนการค้นคืนรูปภาพเชิงความหมาย ได้แก่ รูปภาพที่ 1 สุนัขที่กำลังวิ่งไปข้างหน้า รูปภาพที่ 2 นักสกีที่กำลังมุ่งหน้าไปบนยอดเขาหิมะ และรูปภาพที่ 3 นักไต่เขากำลังมุ่งไปยังยอดเขานั้น ซึ่งแต่ละภาพนั้นมิได้มีคำว่า “Move Forward” ปรากฏอยู่ในรูปภาพ หรือในคำอธิบายรูปภาพเลย แต่ทั้ง 3 รูปภาพอาจมีความหมายแฝงว่า “Move Forward” ที่แปลว่ามุ่งไปข้างหน้า อันเป็นคุณลักษณะเชิงความหมายที่ตรงกับคุณลักษณะข้อความค้นหาจากผู้ใช้

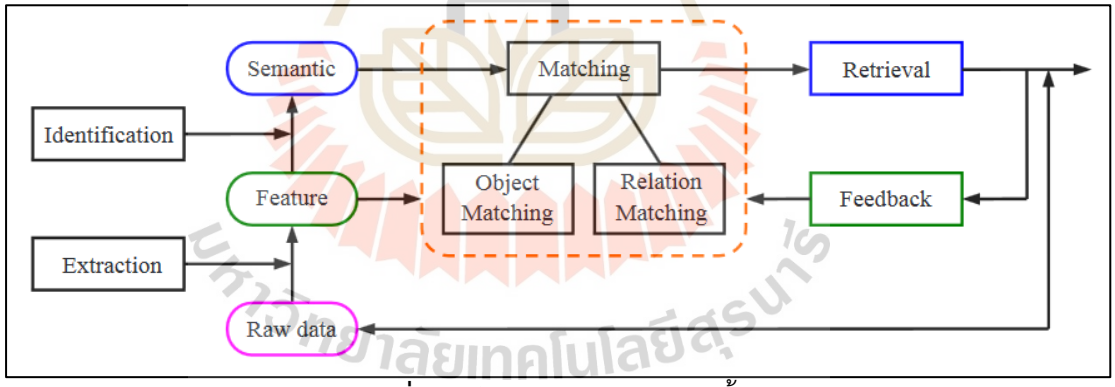
งานวิจัยนี้มุ่งพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมาย เพื่อเชื่อมโยงระหว่างแนวคิดระดับสูงที่เป็นรูปธรรมและที่เป็นนามธรรมให้เป็นแนวคิดเชิงความหมายในรูปแบบที่มนุษย์สามารถเข้าใจได้โดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า อันจะสามารถลดช่องว่างความหมายระหว่างความต้องการของผู้ใช้ที่อยู่ในลักษณะข้อความค้นหา ความหมายของรูปภาพ และคอมพิวเตอร์ และช่วยสนับสนุนผู้ใช้ด้วยข้อความค้นหาภาษาธรรมชาติที่ยึดโยงกับความหมายของรูปภาพแทนที่จะยึดตามหลักไวยากรณ์ของภาษา

2.2 เทคนิคการค้นคืนรูปภาพเชิงความหมาย

ปัจจุบันมีแนวคิดเกี่ยวกับการค้นคืนรูปภาพเชิงความหมายมากมาย โดยการประยุกต์แนวคิดหลายศาสตร์มาบูรณาการร่วมกัน เพื่อเชื่อมโยงระหว่างแนวคิดระดับสูงที่เป็นรูปธรรมและที่เป็นนามธรรมเข้าด้วยกัน สามารถจำแนกตามแนวคิดและวิธีการทำงานเป็น 6 กลุ่ม ได้แก่

2.2.1 เทคนิคการค้นคืนรูปภาพแบบหลายระดับ

รูปภาพแบบหลายระดับ (Multi-level image) คือการวิเคราะห์ความหมายของรูปภาพในรูปแบบลำดับชั้น (Hierarchical image analysis) (Ma et al., 2021) เพื่อจะอธิบายรูปภาพด้วยคำอธิบายที่เป็นลำดับขั้นที่ง่ายต่อการทำความเข้าใจเช่นเดียวกับการรับรู้ของมนุษย์ โดย Zhang (2007) นำเสนอโมเดลการแบ่งรูปภาพเป็น 3 ระดับชั้น ได้แก่ ระดับชั้นข้อมูลดิบ (Raw data layer) คือข้อมูลของรูปภาพในรูปแบบของพิกเซล ระดับชั้นคุณลักษณะ (Feature layer) เป็นลักษณะการจัดเรียงพิกเซลในรูปภาพ และระดับชั้นความหมาย (Semantic layer) ที่อธิบายความหมายของรูปภาพ และแสดงถึงส่วนประกอบของวัตถุหรือรายละเอียดภายในภาพ เพื่อให้สามารถแปลความหมายของรูปภาพได้ จะเห็นว่าในระดับชั้นต่าง ๆ เหล่านี้จะให้ข้อมูลของรูปภาพที่แตกต่างกัน ดังรูปที่ 2.6



รูปที่ 2.6 ข้อมูลรูปภาพ 3 ระดับชั้น

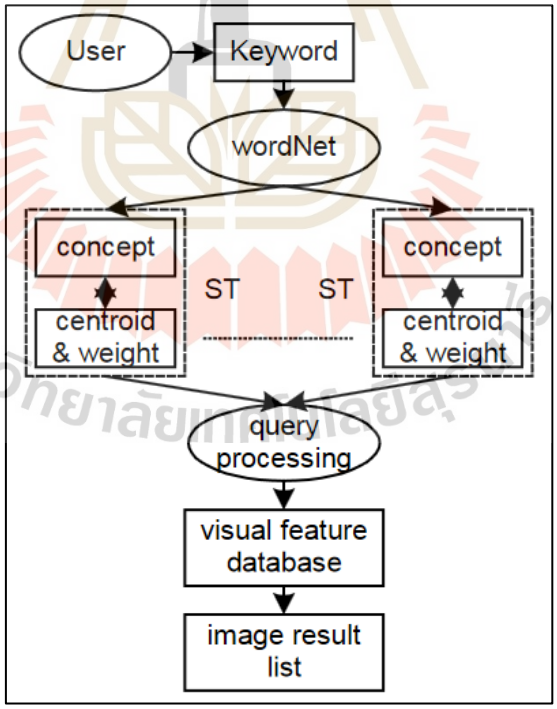
ที่มา: Zhang (2007)

จากรูปที่ 2.6 โมเดลในการนำเสนอการค้นคืนรูปภาพ 3 ระดับชั้น ประกอบ 2 ส่วนคือ 1) การหาว่าพื้นที่สำคัญในการให้ข้อมูลของวัตถุในภาพ (Extract meaningful region) เป็นขั้นตอนการหาพื้นที่ที่น่าสนใจในรูปภาพ โดยขั้นตอนนี้จะเกี่ยวข้องกับระดับชั้นข้อมูลดิบ และระดับชั้นคุณลักษณะ และ 2) การหาวัตถุที่สนใจในรูปภาพ (Identification of interesting objects) จากพื้นที่ที่ถูกระบุในขั้นตอนที่ 1 ระบบต้องระบุให้ได้ว่าวัตถุที่อยู่ในพื้นที่นั้นคืออะไร และวัตถุที่ค้นพบจะถูกนำไปใช้ในขั้นตอนการค้นคืนข้อมูลต่อไป

ปัจจุบันลำดับดับชั้นบนสุดคือระดับชั้นความหมาย แต่ในอนาคตอาจจะมีการนำเสนอชั้นข้อมูลอื่น ๆ มากกว่าวัตถุเพียงอย่างเดียว และสามารถนำไปสู่ผลสรุปเกี่ยวกับเหตุการณ์ต่าง ๆ เช่น ปฏิสัมพันธ์ระหว่างวัตถุต่าง ๆ ที่ตรวจพบในโดเมนของกีฬา หากพบคนอยู่เหนือคาน อาจจะสรุปได้ว่าในภาพนั้นกำลังเกิดเหตุการณ์นักกีฬากำลังกระโดดสูงอยู่ เป็นต้น

2.2.2 เทคนิคการใช้แม่แบบเชิงความหมาย

แม่แบบเชิงความหมาย (Semantic Template: ST) ที่ทำหน้าที่เชื่อมโยงระหว่างแนวคิดระดับสูง และคุณลักษณะของรูปภาพเข้าด้วยกัน เพื่อกำหนดเป็นตัวแทนของคุณลักษณะ (Representative feature) ของรูปภาพ โดย Miller et al. (1990) ได้นำแม่แบบของความหมายมาใช้ร่วมกับ WordNet เพื่อให้ง่ายต่อผู้ใช้งานมากยิ่งขึ้น หลักการทำงานคือเมื่อผู้ใช้ทำการระบุข้อคำถาม ซึ่งเป็นคำสำคัญเข้าไปในระบบ และทำการค้นหาเนื้อหาของคำค้นหาเหล่านั้นจาก WordNet และจะเชื่อมโยงแม่แบบของความหมายที่เกี่ยวข้องทั้งหมดเพื่อหารูปภาพที่เกี่ยวข้อง ดังรูปที่ 2.7



รูปที่ 2.7 การค้นคืนรูปภาพโดยใช้แม่แบบเชิงความหมายร่วมกับ WordNet
ที่มา: Miller et al. (1990)

ส่วนใหญ่งานวิจัยที่ใช้แม่แบบเชิงความหมายในการค้นคืนรูปภาพมักประยุกต์ใช้ร่วมกับออนโทโลยี เพื่อเชื่อมโยงหมวดหมู่รูปภาพแต่ละโดเมน อย่างไรก็ดี การใช้แม่แบบเชิงความหมายยังมีข้อจำกัดคือผู้ใช้ต้องมีความชำนาญหรือมีความรู้เกี่ยวกับคุณลักษณะของรูปภาพอย่างลึกซึ้ง ซึ่งยากต่อการเรียนรู้ของผู้ใช้งานปกติทั่วไป

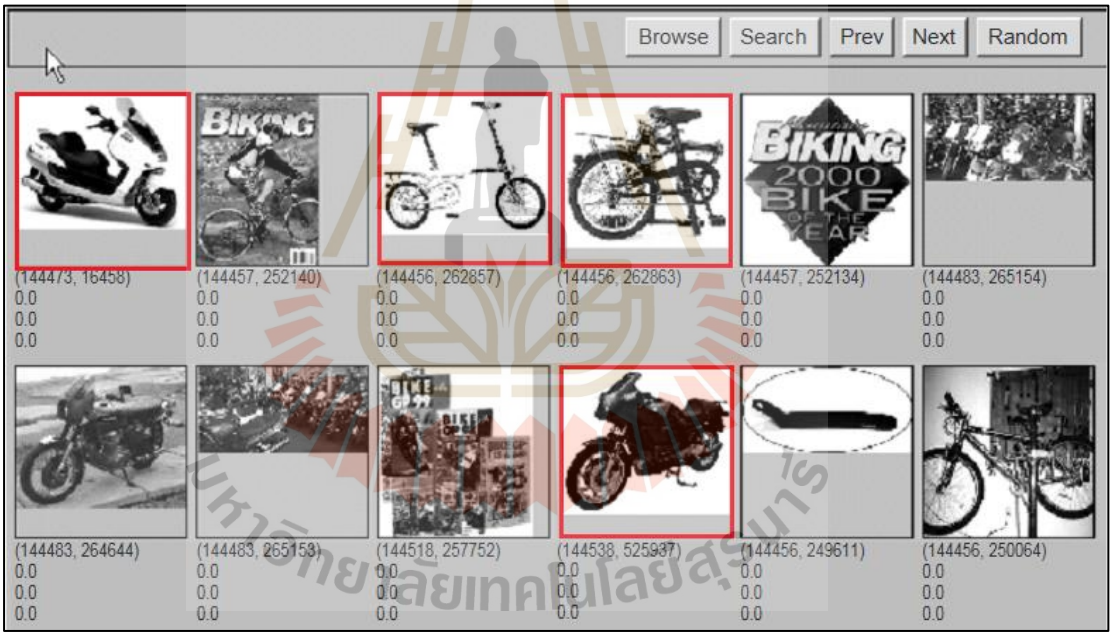
2.2.3 เทคนิคการใช้ข้อความเพื่อเรียนรู้ความหมายของรูปภาพ

การใช้ข้อความเพื่อเรียนรู้ความหมายของรูปภาพ (Using text to learn the image) เนื่องจากรูปภาพต่าง ๆ บนอินเทอร์เน็ตนั้นอาจจะมีข้อมูลบางอย่างที่ช่วยให้การให้ความหมายรูปภาพ เช่น แท็ก คำอธิบาย ไฮเปอร์ลิงก์ หรือส่วนหัวของเว็บเพจ เป็นต้น (Dong, Wang, Qiu and Sapiro, 2020) จึงมีแนวคิดการใช้ข้อความเพื่อเรียนรู้ความหมายของรูปภาพ บนแนวคิดการใช้ภาษาธรรมชาติในกระบวนการควบคุมหรือการฝึกฝนตัวแบบผ่านทางข้อความภาษาธรรมชาติ (Natural language supervision) ที่ง่ายต่อการทำความเข้าใจ เช่น การให้คำบรรยาย การจัดลำดับของข้อมูล หรือการกำหนดข้อกำหนดและเงื่อนไข เพื่อช่วยให้การสื่อสารระหว่างมนุษย์กับตัวแบบมีความกระชับและสื่อสารได้ง่ายขึ้น และยังช่วยให้ระบบเรียนรู้และปรับปรุงตามความต้องการของผู้ใช้ได้อย่างมีประสิทธิภาพมากขึ้นด้วย

แนวคิดการควบคุมด้วยภาษาธรรมชาตินั้นสัมพันธ์กับการเรียนรู้แบบคอนทราสต์ระหว่างรูปภาพและข้อความ (Contrastive learning for text-to-image) ในด้านการฝึกฝนตัวแบบให้เข้าใจความหมายของรูปภาพผ่านข้อความภาษาธรรมชาติ (Zhang, Koh, Baldridge, Lee, and Yang, 2020) โดยการควบคุมด้วยภาษาธรรมชาติเป็นกระบวนการที่ใช้ภาษาธรรมชาติเพื่อฝึกฝนและควบคุมตัวแบบด้วยข้อมูลและคำอธิบายในรูปแบบข้อความ ซึ่งสามารถปรับปรุงและเพิ่มความแม่นยำของตัวแบบได้ในหลาย ๆ รูปแบบ เช่น การเรียนรู้แบบแพตช์ (Patch learning) การขยายรายละเอียดของข้อมูล หรือการปรับแก้ข้อผิดพลาด ดังนั้นการควบคุมด้วยภาษาธรรมชาติจะช่วยให้ตัวแบบเข้าใจความหมายและปรับปรุงฐานความรู้ตามความต้องการของผู้ใช้ได้อย่างมีประสิทธิภาพ นอกจากนั้นการเรียนรู้แบบคอนทราสต์ระหว่างรูปภาพและข้อความเป็นกระบวนการที่ใช้ข้อมูลที่มีลักษณะความสัมพันธ์ระหว่างรูปภาพและข้อความเป็นชุดข้อมูลให้ตัวแบบเรียนรู้และเข้าใจความหมายของรูปภาพและข้อความในลักษณะเปรียบเทียบ มักใช้เพื่อเรียนรู้ลักษณะพิเศษของวัตถุหรือสถานการณ์ที่กำหนดโดยรูปภาพและข้อความที่มีเกี่ยวข้องกัน ซึ่งจะทำให้ตัวแบบเรียนรู้และเข้าใจความหมายของข้อมูลได้อย่างครอบคลุมได้เช่นเดียวกับการเรียนรู้ของมนุษย์

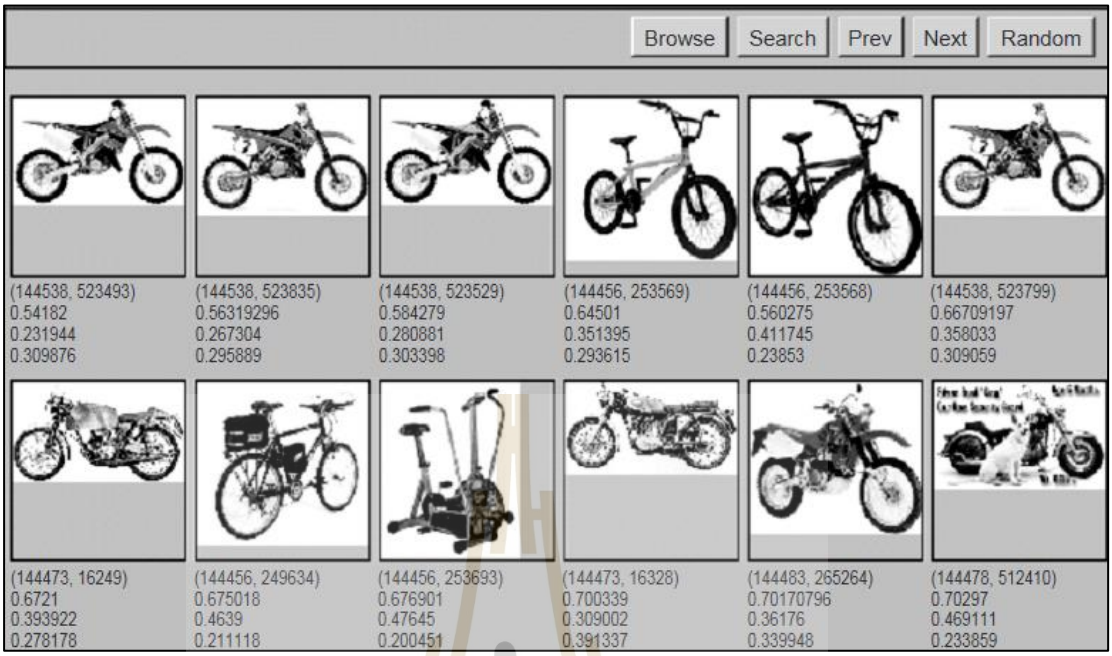
2.2.4 เทคนิคการตอบกลับของผู้ใช้

การตอบกลับของผู้ใช้ (Relevance feedback) เพื่อแก้ปัญหาของช่องว่างความหมายด้วยการเพิ่มผู้ใช้เข้าไปในกระบวนการค้นหาตามแนวคิดว่าผู้ใช้อาจไม่ทราบว่าตนเองต้องการค้นหารูปภาพอะไร (Manning, Raghavan and Schütze, 2008) จึงให้ผู้ใช้ป้อนข้อมูลย้อนกลับจากผลการค้นคืนแต่ละครั้งว่ามีข้อมูลใดบ้างที่ตรงกับสิ่งที่กำลังค้นหาอยู่จากชุดข้อมูลที่เป็นผลลัพธ์ในการค้นหาครั้งแรก เพื่อจะปรับปรุงผลลัพธ์ให้ตรงกับความต้องการของผู้ใช้มากขึ้นในครั้งต่อไป ดังรูปที่ 2.8 ผลลัพธ์เริ่มต้นสำหรับข้อความค้นหา เมื่อผู้ใช้ทำการเลือกผลลัพธ์ที่ 1, 2 และ 3 ในแถวบนและผลลัพธ์ที่ 4 ในแถวล่าง ระบบจะส่งข้อเสนอแนะนี้กลับไปประมวลผลเพื่อปรับปรุงผลลัพธ์ที่มีความเกี่ยวข้องกันมากขึ้นแก่ผู้ใช้ ดังรูปที่ 2.9 ซึ่งระบบจะทำกระบวนการดังกล่าวซ้ำกันไปเรื่อย ๆ จนกระทั่งผลลัพธ์ถูกต้องในระดับที่ผู้ใช้พอใจ



รูปที่ 2.8 ผลลัพธ์เริ่มต้นต่อผู้ใช้สำหรับข้อความค้นหาเกี่ยวกับจักรยาน

ที่มา: Manning, Raghavan and Schütze (2008)



รูปที่ 2.9 ผลลัพธ์หลังจากปรับปรุงความถูกต้องจากการตอบกลับของผู้ใช้

ที่มา: Manning, Raghavan and Schütze (2008)

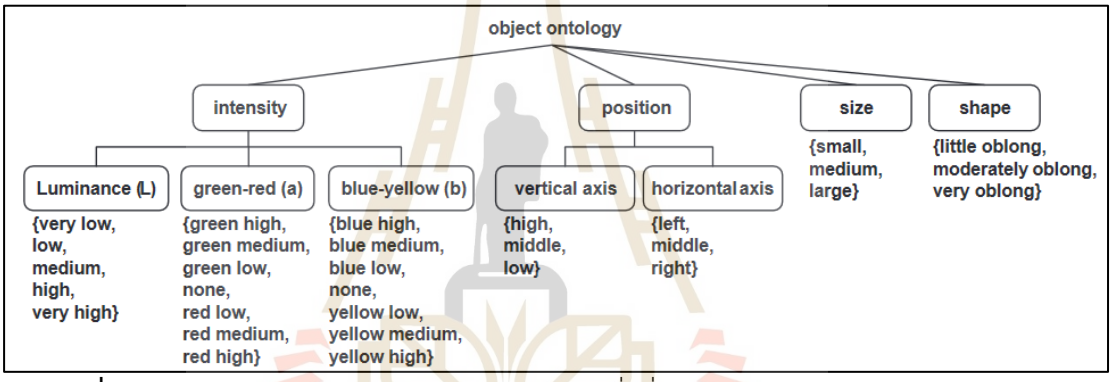
อย่างไรก็ตามเทคนิคการตอบกลับของผู้ใช้ อาจเป็นการเพิ่มภาระในการค้นหาให้กับผู้ใช้ เนื่องจากจำเป็นต้องให้ผู้ใช้ร่วมมือในการตอบกลับข้อคำถามให้กับระบบ เพื่อปรับปรุงผลลัพธ์ให้มีความถูกต้องมากขึ้น แต่วิธีการดังกล่าวอาจทำให้ผู้ใช้หมดความอดทน และเลิกใช้งานไปก่อน

2.2.5 เทคนิคการใช้ออนโทโลยี

ออนโทโลยี (Ontology) เป็นการพัฒนาภาษาเชิงความหมายที่แสดงโครงสร้างของแนวคิดที่บรรยายขอบเขตของความรู้เฉพาะเรื่อง มักถูกนำไปประยุกต์ใช้สร้างความเข้าใจพื้นฐานร่วมกันระหว่างผู้ใช้งานในโดเมนหนึ่ง ๆ (Davies, Studer and Warren, 2006) ประกอบด้วย 1) แนวคิด (Concept) หมายถึง ขอบเขตของความรู้เรื่องใดเรื่องหนึ่งในโดเมนที่สนใจ และสามารถอธิบายรายละเอียดได้ 2) คุณลักษณะ (Property) หมายถึง คุณสมบัติต่าง ๆ ที่นำมาใช้อธิบายแนวคิด 3) ความสัมพันธ์ (Relationship) หมายถึง รูปแบบของความสัมพันธ์กันระหว่างแนวคิด ได้แก่ 3.1) ความสัมพันธ์แบบลำดับชั้น (Subclass) คือ ความสัมพันธ์ที่มีคุณสมบัติการถ่ายทอดคุณสมบัติของแนวคิดแม่ไปยังแนวคิดลูก 3.2) ความสัมพันธ์แบบเป็นส่วนหนึ่ง (Part-of) คือ ความสัมพันธ์ที่หมายถึงการเป็นส่วนประกอบ 3.3) ความสัมพันธ์เชิงความหมาย (Syn-of) คือ ความสัมพันธ์ที่แสดงถึงแนวคิดที่มีความเหมือนเชิงความหมายต่อกัน 3.4) ความสัมพันธ์การเป็นตัวแทน (Instance-of) คือ ความสัมพันธ์ที่แสดงถึงการเป็นสมาชิกของแนวคิด

4) ข้อกำหนดในการสร้างความสัมพันธ์ (Axiom) หมายถึง เงื่อนไขกำหนดเฉพาะในการแปลงความสัมพันธ์ระหว่างแนวคิดกับคุณสมบัติ หรือแนวคิดกับแนวคิด เพื่อให้แปลงความหมายได้ถูกต้อง และ 5) ตัวอย่างข้อมูล (Instances) หมายถึง คำศัพท์ที่กำหนดความหมายไว้ในออนโทโลยี

การค้นคืนรูปภาพเชิงความหมายที่ใช้ออนโทโลยีจะพิจารณาจากความหมายของรูปภาพที่สัมพันธ์กัน แม้ว่าจะมีคุณลักษณะระดับต่ำที่แตกต่างกัน เพื่อกำหนดคุณลักษณะเชิงความหมายให้กับรูปภาพที่มีเนื้อหาซับซ้อนและไม่ทราบความหมายเชิงประจักษ์ โดย Mezaris, Kompatsiaris and Strintzis (2003) นำเสนอการแบ่งรูปภาพออกเป็น คุณลักษณะระดับต่ำที่อธิบายสี ตำแหน่ง ขนาด และรูปร่างของพื้นที่ เพื่อสร้างเป็นชุดออนโทโลยี (Object ontology) ตามคำจำกัดความเชิงคุณภาพของแนวคิดระดับสูงที่ผู้ใช้ค้นหา ดังรูปที่ 2.10



รูปที่ 2.10 ชุดออนโทโลยี แสดงคุณลักษณะระดับต่ำที่อธิบายสี ตำแหน่ง ขนาด และรูปร่าง
ที่มา: Mezaris, Kompatsiaris and Strintzis (2003)

ปัจจุบันมีหลายงานวิจัยที่นำแนวคิดของออนโทโลยีมาช่วยในการค้นคืนรูปภาพเชิงความหมายจากคำค้นหา และใช้รูปทรง สี และพื้นผิวเป็นหลักในการจำแนกประเภทในโดเมนที่ต้องการ ตลอดจนนำผสมผสานการใช้เหตุผล (Reasoning) แบบมัลติมีเดีย และการประมวลผลภาษาธรรมชาติตามโดเมนของออนโทโลยี ซึ่งช่วยให้สามารถกำหนดความหมายที่ซับซ้อนของรูปภาพได้มีประสิทธิภาพมากขึ้น

2.2.6 เทคนิคการจำแนกข้อมูลด้วยการเรียนรู้ของเครื่อง

การจำแนกข้อมูลด้วยการเรียนรู้ของเครื่อง (Machine learning classification) คือการได้มาซึ่งความหมายของรูปภาพต่าง ๆ โดยปกติจะต้องใช้เครื่องมือเพื่อวิเคราะห์ข้อมูลจากคุณลักษณะของรูปภาพ การเรียนรู้ของเครื่องที่จะช่วยวิเคราะห์และจำแนกคุณลักษณะของรูปภาพและแปลความหมายในระดับสูงขึ้น โดยการจำแนกข้อมูล (Classification) (Aggarwal, 2018) เป็นเทคนิคการสร้างแบบจำลองหรือตัวจำแนกข้อมูล (Classifier) เพื่อทำนายหมวดหมู่ของข้อมูล (Categories or class) โดยชุดข้อมูลที่เป็นอินพุตสำหรับสร้างตัวจำแนกข้อมูลเรียกว่าชุดข้อมูลฝึกสอน (Training data) หากในชุดข้อมูลมีแอททริบิวต์ (Attribute) หรือป้ายกำกับ (Label) หมวดหมู่ข้อมูลสำหรับจำแนกจะเรียกว่าการเรียนรู้แบบมีผู้สอน (Supervised learning) ที่สร้างตัวจำแนกข้อมูลจะถูกสอนหมวดหมู่ข้อมูลต่าง ๆ ตรงข้ามกับการเรียนรู้แบบไม่มีผู้สอน (Unsupervised learning or clustering) ที่จะไม่ทราบถึงหมวดหมู่ของข้อมูล

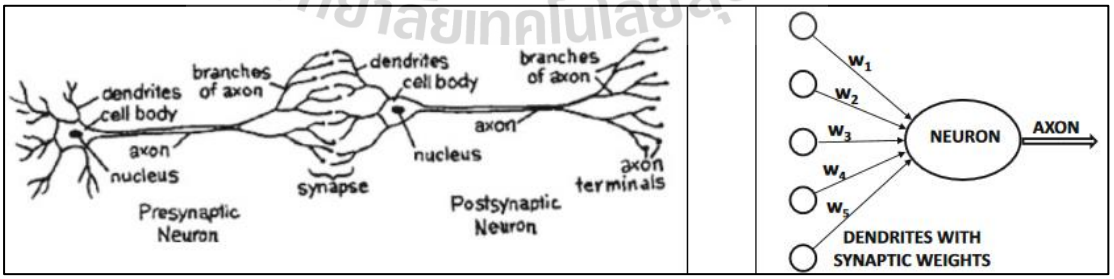
ปัจจุบันมีปหลายเทคนิคที่มักถูกนำมาใช้ในการสร้างตัวจำแนกข้อมูล (Wozniak, 2014) ประกอบด้วย 1) การจำแนกข้อมูลด้วยต้นไม้ตัดสินใจ (Decision tree classification) เป็นกระบวนการสร้างต้นไม้ขึ้นเพื่อใช้ในการตัดสินใจจากข้อมูลที่มีหมวดหมู่ข้อมูลระบุอยู่ 2) การจำแนกข้อมูลแบบเบย์เซียน (Bayesian classification) เป็นการสร้างตัวจำแนกข้อมูลโดยประยุกต์ใช้ทฤษฎีของเบย์ (Bayes' theorem) ด้วยค่าทางสถิติที่สามารถบ่งบอกถึงความน่าจะเป็นของข้อมูลเรคคอร์ดหนึ่งที่จะอยู่ในหมวดหมู่ของข้อมูลหนึ่ง ๆ ที่ช่วยให้สามารถจำแนกข้อมูลได้อย่างรวดเร็วและแม่นยำสูง 3) การจำแนกข้อมูลด้วยฐานกฎ (Rule-based classification) เป็นโมเดลที่จะแสดงผลด้วยเซตของกฎที่มีลักษณะแบบ if-thenโดยแต่ละกฎจะอยู่ในรูปของฟอร์ม 4) การจำแนกข้อมูลจากกฎความสัมพันธ์ของข้อมูล (Association rule classification) มีที่มาจากแนวคิดกฎความสัมพันธ์ของข้อมูลที่ประกอบด้วยรายการต่าง ๆ ที่ปรากฏร่วมกันบ่อย ๆ เพื่อใช้อธิบายถึงรูปแบบที่แอบแฝงอยู่ในชุดข้อมูล 5) การจำแนกข้อมูลแบบเพื่อนบ้านใกล้สุด k อันดับ (K-nearest neighbor classification) เป็นการเรียนรู้โดยเปรียบเทียบกันระหว่างข้อมูลที่ต้องการจำแนกทั้งหมดในชุดข้อมูลฝึกสอนที่มีลักษณะเหมือนกันหรือใกล้เคียงกันด้วยการพิจารณาข้อมูลแอททริบิวต์ต่าง ๆ โดยจะมีเรคคอร์ดหนึ่งที่ถูกกำหนดเป็นจุดหนึ่งในระนาบ พิจารณาจากระยะทางแบบยุคลิด (Euclidean distance) 6) การจำแนกข้อมูลด้วยซัพพอร์ทเวกเตอร์แมชชีน (Support vector machine classification) เป็นอัลกอริทึมในการวิเคราะห์ข้อมูลและจำแนกข้อมูลโดยอาศัยหลักการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลที่ถูกป้อนเข้าสู่กระบวนการฝึกฝนให้ระบบเรียนรู้ โดยเน้นไปยังการหาเส้นแบ่งแยกกลุ่มข้อมูลที่ดีที่สุด และ 7) การจำแนกข้อมูลด้วยโครงข่ายประสาทเทียม (Neural network classifier) ที่เลียนแบบการทำงานของโครงข่ายประสาทของมนุษย์ โดยเชื่อมต่อกันเป็นโหนดเรียงซ้อนกันหลายชั้น หรือเรียกว่า

โครงข่ายประสาทเทียมแบบหลายชั้น (Multi-layer perceptron) ค่าฟังก์ชันของแต่ละโหนดเกิดจากฝึกฝน โดยกระบวนการฝึกฝน เริ่มจากการระบุข้อมูลอินพุตเข้าโครงข่ายคำนวณค่าอินพุตคูณกับค่าน้ำหนักในโครงข่ายแบบป้อนไปข้างหน้า (Feedforward) แล้วปรับค่าน้ำหนักประสาท (Weight) และไบแอส (Bias) ตามกฎการเรียนรู้ เพื่อให้เอาต์พุตของโครงข่ายให้ค่าผลลัพธ์ใกล้เคียงเป้าหมายมากที่สุด แล้วส่งผ่านข้อมูลสู่ฟังก์ชันกระตุ้นและส่งออกเป็นเอาต์พุต เพื่อเปรียบเทียบกับค่าเป้าหมาย ซึ่งเป็นค่าผิดพลาด และจะถูกส่งย้อนกลับไปปรับค่าน้ำหนัก (Backpropagation algorithm) ให้พอดี (Fit) กับข้อมูลต่อไป ด้วยจุดเด่นคือมีความมั่นคงสูงต่อสิ่งรบกวน (Noise) ทำงานได้ดีกับข้อมูลอินพุตและเอาต์พุตที่มีค่าแบบไม่ต่อเนื่อง อีกทั้งสามารถจำแนกข้อมูลได้ทั้งแบบมีผู้สอนและไม่มีผู้สอน รวมถึงมีประสิทธิภาพในการจำแนกข้อมูล แม้พบความสัมพันธ์ระหว่างแอททริบิวต์กับหมวดหมู่ต่าง ๆ เพียงเล็กน้อยจากข้อดีดังกล่าวจึงถูกนำมาถูกประยุกต์ใช้อย่างแพร่หลาย

ปัจจุบันได้มีการพัฒนาไปเป็นการเรียนรู้เชิงลึก (Deep learning) ซึ่งมีการเรียงตัวของนิวรอนในแต่ละชั้นอย่างซับซ้อนและมีชั้นโครงข่ายลึกที่มากขึ้น เพื่อเพิ่มประสิทธิภาพในการเรียนรู้คุณลักษณะต่าง ๆ ได้อย่างหลากหลาย (Goodfellow et al., 2016) และได้รับการยอมรับว่ามีประสิทธิภาพในการวิเคราะห์และทำนายผลได้ดีกว่าการเรียนรู้ของเครื่องแบบเดิมเป็นอย่างมาก

2.3 โครงข่ายประสาทเทียมเชิงลึก

โครงข่ายประสาทเทียม (Neural network) (Wozniak, 2014) คือโมเดลทางคณิตศาสตร์สำหรับประมวลผลสารสนเทศด้วยการจำลองลักษณะการทำงานของเครือข่ายประสาทในสมองมนุษย์ที่ประกอบด้วยเซลล์ประสาท (Nerve cells) หรือเรียกว่านิวรอน (Neuron) ดังรูปที่ 2.11 ประกอบด้วย 4 ส่วนตามหลักชีววิทยา ได้แก่ โซมา เดนไดร์ท ซินแนปส์ และแอ็คซอน



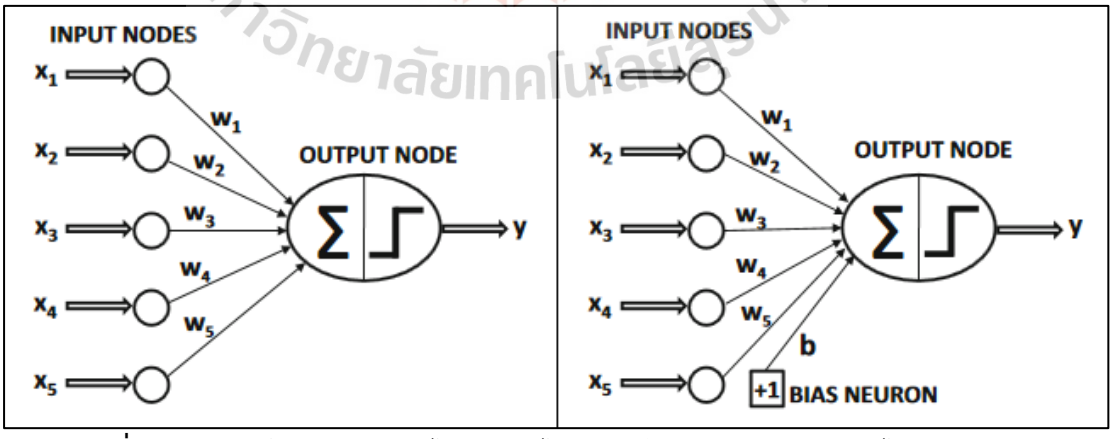
รูปที่ 2.11 แสดงระบบประสาทตามหลักชีววิทยา (ซ้าย) และโครงข่ายประสาทเทียม (ขวา)
ที่มา: Aggarwal (2018)

จากรูปที่ 2.11 การทำงานของระบบประสาทเกิดจากการเชื่อมต่อระหว่างเซลล์ประสาทหลายพันล้านเซลล์ แต่ละเซลล์ประกอบด้วยแขนงรับสัญญาณประสาทซึ่งเปรียบเสมือนหน่วยรับข้อมูลเข้าเรียกว่าเดนไดรท์ และส่วนปลายของเซลล์ประสาทซึ่งเป็นเสมือนหน่วยส่งข้อมูลออกของเซลล์เรียกว่าแอคซอน ซึ่งอาจทำให้เกิดได้ทั้งการกระตุ้นและยับยั้ง นอกจากนี้ วิธีการประมวลผลภายในแต่ละเซลล์ยังมีการขยายหรือลดขนาดของสัญญาณก่อนจะรวมกันเข้าสู่เซลล์ประสาท และหากสัญญาณรวมมีความแรงเกินระดับ (Threshold) ที่กำหนด เซลล์ประสาทก็จะส่งสัญญาณออกทางแอคซอนต่อไป เมื่อเปรียบเทียบกับส่วนประกอบของโครงข่ายประสาท ดังตารางที่ 2.1

ตารางที่ 2.1 ส่วนประกอบระบบประสาทตามหลักชีววิทยาเปรียบเทียบกับโครงข่ายประสาทเทียม

ระบบประสาทตามหลักชีววิทยา	โครงข่ายประสาทเทียม
โซมา (Soma)	นิวรอน (Neuron)
เดนไดร์ท (Dendrites)	อินพุต (Input)
ซินแนปส์ (Synapses)	ค่าน้ำหนัก (Weight)
แอ็คซอน (Axon)	เอาต์พุต (Output)

การเรียนรู้ของมนุษย์มีรากฐานมาจากการทำงานของระบบประสาท รวมทั้งการทำหน้าที่ด้านการคิดคำนึงต่าง ๆ โดยจำนวนนิวรอนที่มีอยู่ในสมองมนุษย์นั้นมีอยู่เป็นจำนวนมาก เปรียบได้กับเซลล์ประสาทที่เล็กที่สุดของโครงข่ายประสาทเทียม เรียกว่าเพอร์เซปตรอน (Perceptron) (Aggarwal, 2018) ดังรูปที่ 2.12



รูปที่ 2.12 เพอร์เซปตรอนแบบไม่กำหนดไบแอส (ซ้าย) และแบบกำหนดไบแอส (ขวา)
ที่มา: Aggarwal (2018)

จากรูปที่ 2.12 โครงสร้างของเพอร์เซปตรอนแต่ละโหนด เริ่มจากการป้อนอินพุตแทนด้วยสัญลักษณ์ ($\boldsymbol{x_{(n)}}$) ผ่านโหนดอินพุต เพื่อคูณกับค่าน้ำหนักของอินพุตแต่ละตัว ซึ่งแทนด้วย ($\boldsymbol{W_{(n)}}$) ตามกฎการเรียนรู้ แล้วปรับค่าน้ำหนัก (Weight) และไบแอส (Bias) ก่อนนำมารวมกันและส่งผ่านฟังก์ชันผลรวม (Summation function: $\boldsymbol{\Sigma}$) เพื่อให้เอาต์พุต ($\boldsymbol{y}$) ใกล้เคียงเป้าหมายมากที่สุด จากนั้นส่งไปยังฟังก์ชันกระตุ้น (Activation function) และส่งผลลัพธ์ผ่านโหนดเอาต์พุต ดังสมการที่ 2.1

$$0 = f\left(\sum_{i=1}^{n} w_i\, x_i + b\right) \tag{2.1}$$

เมื่อแทนค่าการทำงานของเพอร์เซปตรอนจากรูปที่ 2.5 ในสมการที่ 2.1 แบบกำหนดค่าไบแอส จะมีรายละเอียดดังนี้

$$0 = f\left((x_1 * w_1 + b) + (x_2 * w_2 + b) + (x_3 * w_3 + b) + (x_4 * w_4 + b) + (x_5 * w_5 + b)\right)$$

การเรียนรู้ของเพอร์เซปตรอนตั้งอยู่บนแนวคิดของการเรียนรู้แบบมีผู้สอน นั่นคือต้องมีชุดข้อมูลตัวอย่าง (Set of example) ที่อยู่ในรูปของ $\{P_1, t_1\}, \{P_2, t_2\}, \ldots., \{P_Q, t_Q\}$ เมื่อ P_Q เป็นเวกเตอร์ของอินพุตแต่ละตัวที่นำเข้าสู่โครงข่าย และ t_Q คือค่าของเอาต์พุตเป้าหมาย (Target output) ที่ต้องการของอินพุตตัวนั้น เพื่อให้โครงข่ายสามารถนำค่าของเอาต์พุตเป้าหมายนั้นไปใช้สำหรับการเปรียบเทียบกับค่าของเอาต์พุตที่เป็นผลลัพธ์จากโครงข่ายว่ามีความผิดพลาด (error) มากน้อยแค่ไหน และนำค่าความผิดพลาดดังกล่าวไปใช้ในขั้นตอนการปรับเปลี่ยนค่าน้ำหนักและไบแอส รวมถึงค่าพารามิเตอร์ต่าง ๆ ตามกฎการเรียนรู้ที่กำหนดไว้ในแต่ละโครงข่ายต่อไป โดยกฎการเรียนรู้ของเพอร์เซปตรอน ดังสมการที่ 2.2 และ 2.3

$$w^{new} = w^{old} + ep^t \tag{2.2}$$

$$b^{new} = b^{old} + e \tag{2.3}$$

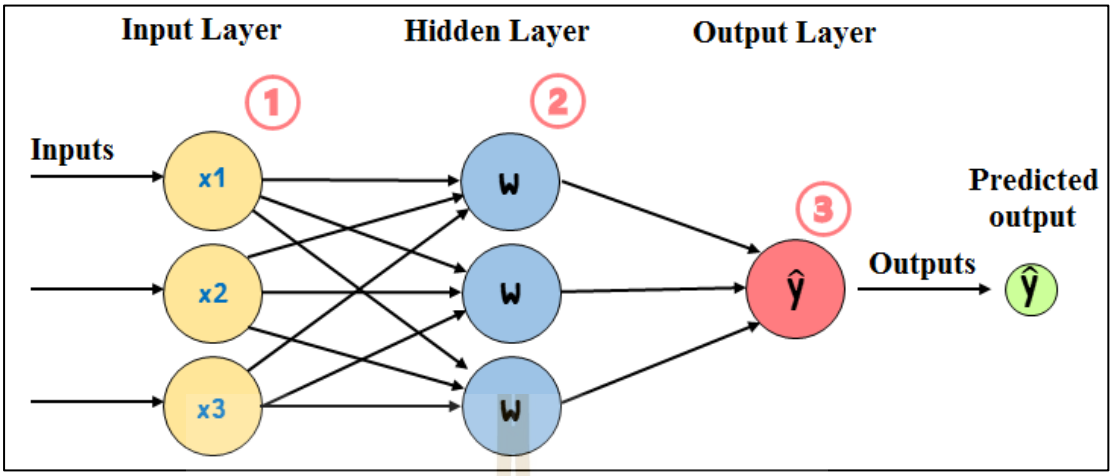
$$where, e = t - a$$

จากกฎการเรียนรู้จะเห็นว่าเป็นการนำค่าความผิดพลาด (e) มาใช้สำหรับการปรับค่าน้ำหนัก (w) และไบแอส (b)ซึ่งค่าความผิดพลาดมาจากการนำค่าเอาต์พุตเป้าหมาย (t)

ลบด้วยค่าเอาต์พุตที่ได้จากโครงข่าย (a) ซึ่งจะสังเกตได้ว่ากฎการเรียนรู้ของเพอร์เซปตรอนค่าของน้ำหนักจะไม่ถูกปรับ เมื่อค่าความผิดพลาดเป็นศูนย์ นั่นคือเมื่อโครงข่ายสามารถเรียนรู้อินพุตตัวนั้นแล้วได้ค่าของเอาต์พุตจากโครงข่ายเหมือนกับค่าของเอาต์พุตเป้าหมายหรือเอาต์พุตที่ต้องการจริง ๆ

อย่างไรก็ดี เนื่องจากคุณลักษณะของโครงข่ายเพอร์เซปตรอนแบบชั้นเดียวมีข้อจำกัดคือสามารถใช้จำแนกได้เฉพาะรูปแบบข้อมูลที่สามารถแบ่งแยกได้แบบเชิงเส้นเท่านั้น จึงเป็นที่มาของโครงข่ายเพอร์เซปตรอนแบบหลายชั้น (Multi-layer perceptron) ประกอบด้วย 1) แต่ละนิวรอนในโครงข่ายจะมีการนำฟังก์ชันกระตุ้นแบบไม่ใช่เชิงเส้น (Non-linear) มาใช้ 2) ชั้นซ่อนเร้นที่ประกอบด้วยนิวรอนที่ไม่ได้เป็นชั้นอินพุตหรือชั้นเอาต์พุต สามารถมีได้มากกว่า 1 ชั้น และ 3) โครงข่ายมีลักษณะการเชื่อมต่อกันสูง (High degree of connectivity) ด้วยค่าน้ำหนักของโครงข่าย ต่อมาได้มีการนำเสนอรูปแบบโครงสร้างการจัดวางของนิวรอน ภายในโครงข่ายเป็นองค์ประกอบสำคัญที่ทำให้คุณลักษณะต่าง ๆ ของโครงข่ายแตกต่างกันออกไป ตลอดจนการปรับเปลี่ยนค่าน้ำหนัก ไบแอส หรือแม้กระทั่งเงื่อนไขในการฝึกสอนของโครงข่ายเกิดเป็นสถาปัตยกรรมแบบป้อนไปข้างหน้า (Feedforward) ที่ได้รับความนิยมในเวลาต่อมา

โครงข่ายประสาทเทียมแบบป้อนไปข้างหน้า (Feedforward neural network) (Aggarwal, 2018) ที่มีลำดับในการคำนวณและส่งผ่านของข้อมูลไปในทิศทางเดียวกัน ประกอบด้วย 3 ชั้นหลัก คือ ชั้นอินพุต ชั้นซ่อนเร้น และชั้นเอาต์พุต โดยชั้นซ่อนเร้นสามารถมีได้มากกว่า 1 ชั้น และในแต่ละชั้นก็ไม่จำเป็นต้องมีจำนวนนิวรอนเท่ากัน ขึ้นอยู่กับการนำไปประยุกต์ใช้กับแต่ละปัญหา โดยในแต่ละชั้นจะมีนิวรอนจำนวนหนึ่งซึ่งไม่มีเส้นเชื่อมกันภายในชั้นเดียวกัน แต่จะมีเส้นเชื่อมกับนิวรอนตัวอื่น ๆ ที่อยู่ในลำดับชั้นที่ติดกัน โดยเอาต์พุตของนิวรอนในชั้นก่อนหน้าจะเป็นอินพุตของนิวรอนในชั้นต่อไปตามลำดับ ดังรูปที่ 2.13 การทำงานประกอบด้วย 3 ขั้นตอน คือ 1) การนำอินพุตเข้าสู่อินพุตเลเยอร์ 2) การคำนวณผลลัพธ์ของอินพุตแต่ละตัวจากการคูณกับค่าน้ำหนักที่กำหนดให้กับแต่ละนิวรอนในชั้นซ่อนเร้นแบบดอทโปรดัค (Dot product) และ 3) ผลลัพธ์จากการพยากรณ์ (Predicted output: $\hat{\boldsymbol{y}}$) จากเอาต์พุตเลเยอร์



รูปที่ 2.13 โครงข่ายประสาทเทียมแบบป้อนไปข้างหน้า

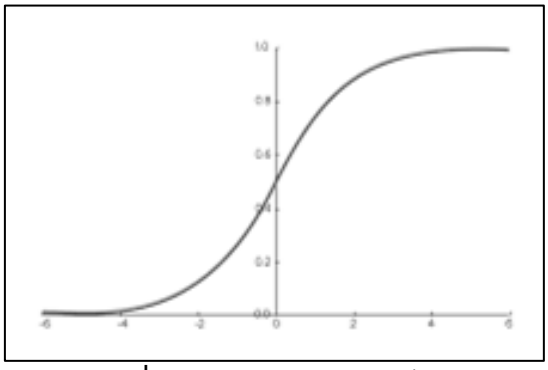
กำหนดสัญลักษณ์แทนการคำนวณไปข้างหน้า โดยให้ a_k^{l-1}แทนผลลัพธ์ของนิวรอนตัวที่ k ในลำดับที่ $l-1$ และ w_{jk}^l แทนน้ำหนักสำหรับนิวรอนตัวที่ j ในลำดับชั้น l ที่มีเส้นเชื่อมมาจากนิวรอนตัวที่ k ในระดับชั้นก่อนหน้า และb_j^l คือไบแอส นอกจากนี้ให้ g แทนฟังก์ชันกระตุ้นและให้ n แทนจำนวนนิวรอนในลำดับชั้นที่ $l-1$ สามารถแสดงการคำนวณ a_j^l ได้ดังสมการที่ 2.4 และ 2.5

$$z_j^i = \sum_k^n = 1\ w_{jk}^l a_k^l + b_j^l \tag{2.4}$$

$$a_j^l = g(z_j^l) \tag{2.5}$$

ฟังก์ชันกระตุ้นที่อยู่ในชั้นซ่อนเร้นเป็นตัวกำหนดว่าประสาทเทียมจะส่งผลลัพธ์ออกในลักษณะใด แบ่งเป็น 1) ฟังก์ชันกระตุ้นเชิงเส้น (Linear activation function) ที่เรียนรู้ความสัมพันธ์เชิงเส้นระหว่างอินพุตและเอาต์พุต จึงไม่สามารถหาคำตอบได้สำหรับบางกรณี และ 2) ฟังก์ชันกระตุ้นแบบไม่ใช่เชิงเส้น (Non-linear activation function) ซึ่งนิยมใช้กับการเรียนรู้เชิงลึกในปัจจุบัน (Ghosh, Sufian, Sultana, Chakrabarti, and De, 2019)

หนึ่งในฟังก์ชันกระตุ้นแบบไม่ใช่เชิงเส้นที่ได้รับความนิยมในปัจจุบันคือฟังก์ชันซอฟต์แม็กซ์ (Softmax function) ที่ให้ค่าผลลัพธ์ออกมาอยู่ในช่วง 0 ถึง 1 ดังรูปที่ 2.14



รูปที่ 2.14 ฟังก์ชันซอฟต์แม็กซ์

การคำนวณความน่าจะเป็นของป้ายกำกับ (Label probability) (Nwankpa, Ijomah, Gachagan and Marshall, 2018) โดยมักใช้ฟังก์ชันซอฟต์แม็กซ์ที่รับอินพุตเป็นเวกเตอร์ของจำนวนจริง (Logit) แล้วทำให้เป็นบรรทัดฐาน (Normalize) ที่มีหลักการคำนวณเพื่อให้คลาสที่มีการพยากรณ์ค่าความน่าจะเป็นต่ำจะถูกกดให้ต่ำลง ในทางกลับกันคลาสที่มีค่าความน่าจะเป็นสูงก็จะยิ่งถูกดันให้ใกล้เคียงกับ 1 มาแปลงให้เป็นมาตรฐานสำหรับการแจกแจงความน่าจะเป็นตามสัดส่วนของค่าเอาต์พุตแต่ละคลาสซึ่งเป็นค่ามากที่สุด (Max function) จากสมการที่ 2.6

$$f(x_i) = \frac{exp(x_i)}{\sum_j^n exp(x_j)} \tag{2.6}$$

โดย x_i คือ เวกเตอร์อินพุตของฟังก์ชันซอฟต์แม็กซ์

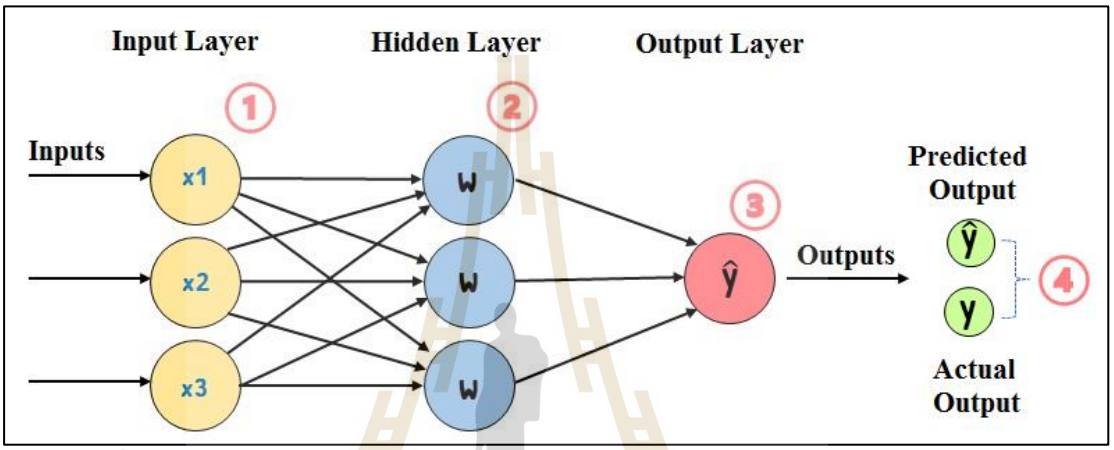
$exp(x_i)$ คือ ฟังก์ชันเอกซ์โพเนนเชียลมาตรฐานมีค่าคงที่เท่ากับ 2.71828 ยกกำลังด้วยเวกเตอร์อินพุต

$\sum_j exp(x_i)$ คือ การปรับค่าให้เป็นมาตรฐานโดยค่าเอาต์พุตทั้งหมดรวมกันเป็น 1 และแต่ละค่าอยู่ในช่วง (0, 1)

n คือ จำนวนคลาสทั้งหมดที่เป็นไปได้

ฟังก์ชันสูญเสีย (Loss function) (Nwankpa, Ijomah, Gachagan and Marshall, 2018) เปรียบเสมือนการตั้งเป้าหมายให้กับโครงข่ายประสาทเทียมเรียนรู้ ซึ่งมีหลายประเภทตามวัตถุประสงค์ในการทำงาน โดยทั่วไปแล้วจะมีค่าเพิ่มขึ้นหากผลลัพธ์ไม่ใช่คำตอบที่ถูกต้องตามการเรียนรู้ ตรงกันข้ามจะมีค่าน้อยหากได้ผลลัพธ์ตามที่ต้องการ

การทำงานของฟังก์ชันสูญเสียจะทำการปรับค่าน้ำหนักเพื่อที่ละลดค่าผลลัพธ์จากการเรียนรู้ของโครงข่ายประสาทเทียมด้วยการหาค่าความผิดพลาด (Loss) ที่ได้จากการเปรียบเทียบระหว่างผลลัพธ์จากตัวแบบ (Predicted output: $\hat{y}$) กับผลลัพธ์จริง (Actual output: y) ว่ามีผลการพยากรณ์ใกล้กับค่าจริงที่ต้องการมากน้อยแค่ไหนด้วยค่าการสูญเสียครอสเอนโทรปี (Cross-entropy loss) ดังรูปที่ 2.15



รูปที่ 2.15 การทำงานของฟังก์ชันสูญเสียในโครงข่ายประสาทเทียมแบบป้อนไปข้างหน้า

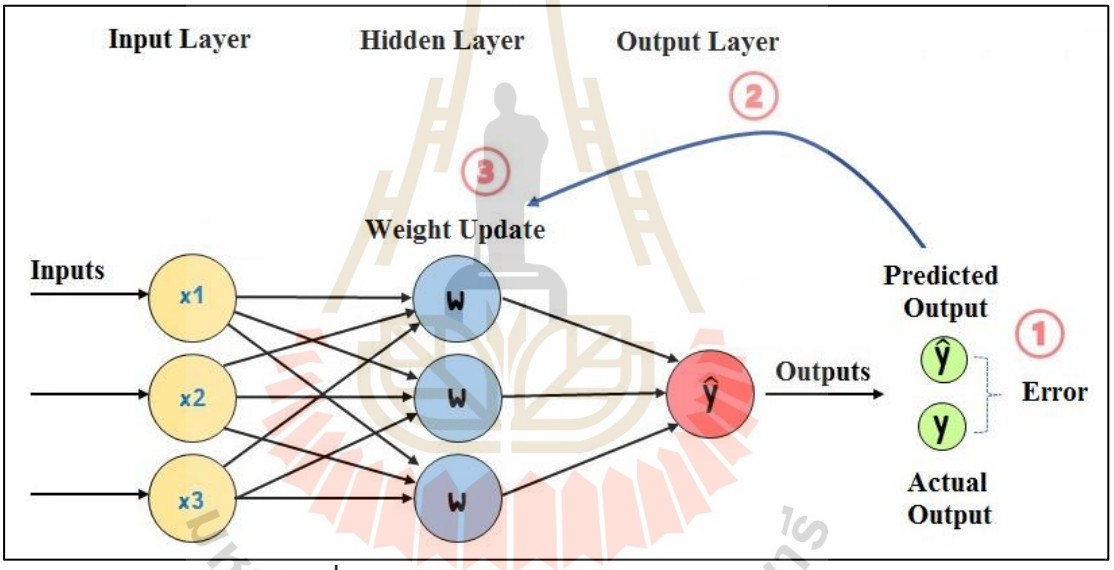
ค่าการสูญเสียครอสเอนโทรปี คือค่าที่เกิดจากการแจกแจงความน่าจะเป็นระหว่างผลลัพธ์จากการพยากรณ์กับผลลัพธ์จริง (True label) ในรูปแบบ One-hot encoding เพื่อจะให้ค่าความเชื่อมั่นที่ทุกคลาสรวมกันเท่ากับ 1 ต้องใช้งานร่วมกับฟังก์ชันซอฟต์แม็กซ์ในชั้นเอาต์พุตด้วยจากรูปที่ 2.5 การแปลงค่าให้เป็นมาตรฐานสำหรับการแจกแจงความน่าจะเป็นตามสัดส่วนของค่าเอาต์พุตจากฟังก์ชันซอฟต์แม็กซ์ด้วยการหาค่าการสูญเสียครอสเอนโทรปี ดังสมการที่ 2.7

$$H(p,q) = -\sum_{x\epsilon\ classes} p(x)\ log\ q\ (x) \tag{2.7}$$

โดย $p(x)$ คือ ผลลัพธ์จริงแบบ one-hot คลาสที่ถูกต้องจะมีค่าเท่ากับ 1 ตรงกันข้าม คลาสที่เหลือทั้งหมดจะเท่ากับ 0

$q(x)$ คือ ผลลัพธ์จากการพยากรณ์ค่าความน่าจะเป็นของตัวแบบที่เลเยอร์เอาต์พุตเป็นฟังก์ชันซอฟต์แม็กซ์

การทำงานของโครงข่ายประสาทเทียมเชิงลึกนั้น เมื่อมีอินพุตเข้ามาจะประมวลอัตโนมัติเพื่อหาข้อมูลตัวอย่างที่จำเป็นในการตรวจจับรูปแบบ หรือจัดหมวดหมู่ข้อมูล ความสามารถในการเรียนรู้คุณลักษณะอัตโนมัติ (Goodfellow, Bengio and Courville, 2016) โดยจะมีการหาค่าผิดพลาด (error) จากผลลัพธ์จากการพยากรณ์เปรียบเทียบกับผลลัพธ์จริงผ่านฟังก์ชันสูญเสียก่อนจะส่งค่ากลับเพื่อปรับค่าพารามิเตอร์ (Parameter) ประกอบด้วยค่าน้ำหนัก และไบแอสจากการหาอนุพันธ์ของฟังก์ชันสูญเสีย เพื่อให้ได้ค่าการสูญเสียที่ลดลง โดยเป้าหมายหลักของอัลกอริทึมเพิ่มประสิทธิภาพคือการทำให้ค่าการสูญเสียนั้นเคลื่อนที่ไปยังจุดต่ำสุดบนพื้นที่การสูญเสีย (Loss surface) และจะได้ผลลัพธ์เป็นค่าเกรเดียนต์เดสเซนท์ที่ถูกปรับปรุงตามค่าน้ำหนัก เรียกว่าการป้อนข้อมูลกลับ (Backpropagation) ดังรูปที่ 2.16



รูปที่ 2.16 โครงข่ายประสาทเทียมแบบย้อนกลับ

จะเห็นได้ว่าการหาค่าความผิดพลาดของโครงข่ายประสาทเทียมแบบป้อนไปข้างหน้าในชั้นสุดท้ายนั้นสามารถคำนวณค่าการสูญเสียครอสเอนโทรปีเปรียบเทียบระหว่างผลลัพธ์จากตัวแบบกับผลลัพธ์จริงว่ามีผลการพยากรณ์ใกล้กับค่าจริงมากน้อยเพียงใด แต่ในการหาค่าความผิดพลาดของโครงข่ายประสาทเทียมเพื่อใช้ในการเรียนรู้ของลำดับขั้นก่อนหน้านั้นไม่สามารถหาได้โดยตรง จึงต้องอาศัยวิธีการที่เรียกว่าการแพร่กระจาย ดังสมการที่ 2.8

$$\delta_j^l = \frac{\partial_j}{\partial z_j^l} = \frac{\partial_j}{\partial a_j^l}\frac{\partial a_j^l}{\partial z_j^l} = \frac{\partial j}{\partial a_j^l} g'(z_j^l) \tag{2.8}$$

โดย s คือ ค่าความผิดพลาดของโครงข่ายประสาทเทียมตัวที่ j ในลำดับชั้น l

j คือ ฟังก์ชันสูญเสีย

z คือ ค่าที่คำนวณได้ก่อนจะผ่านฟังก์ชันกระตุ้น g

สำหรับการหาค่า $\frac{\partial j}{\partial a_j^l}$ นั้น ในลำดับชั้นสุดท้ายสามารถคำนวณหาได้โดยตรงจากฟังก์ชันกระตุ้นที่เลือกใช้ ส่วนในลำดับชั้นก่อนหน้าจะต้องหาโดยวิธีแพร่กระจายย้อนกลับโดยจะทำคล้ายกับการป้อนไปข้างหน้าเพียงแต่กลับทิศกันเท่านั้น ดังสมการที่ 2.9

$$\frac{\partial j}{\partial a_j^l} = \sum_{k=1}^{m} \frac{\partial j}{\partial z_k^{l+1}} \frac{\partial z_k^{l+1}}{\partial a_j^l} = \sum_{k=1}^{m} \delta_k^{l+1} w_{kj}^{l+1} \quad (2.9)$$

โดย m คือ จำนวนโครงข่ายในลำดับชั้นที่ $l+1$

จากนั้นเมื่อคำนวณค่าความผิดพลาดของแต่ละระดับชั้นได้ก็สามารถหาค่าความผิดพลาดเทียบกับน้ำหนักและค่าไบแอสได้จากสมการที่ 2.10 และ 2.11

$$\frac{\partial j}{\partial w_{jk}^l} = \frac{\partial j}{\partial z_j^l} \frac{\partial z_j^l}{\partial w_{jk}^l} = \delta_j^l a_k^{l=1} \quad (2.10)$$

$$\frac{\partial j}{\partial b_j^l} = \frac{\partial j}{\partial z_j^l} \frac{\partial z_j^l}{\partial b_{jk}^l} = \delta_j^l \quad (2.11)$$

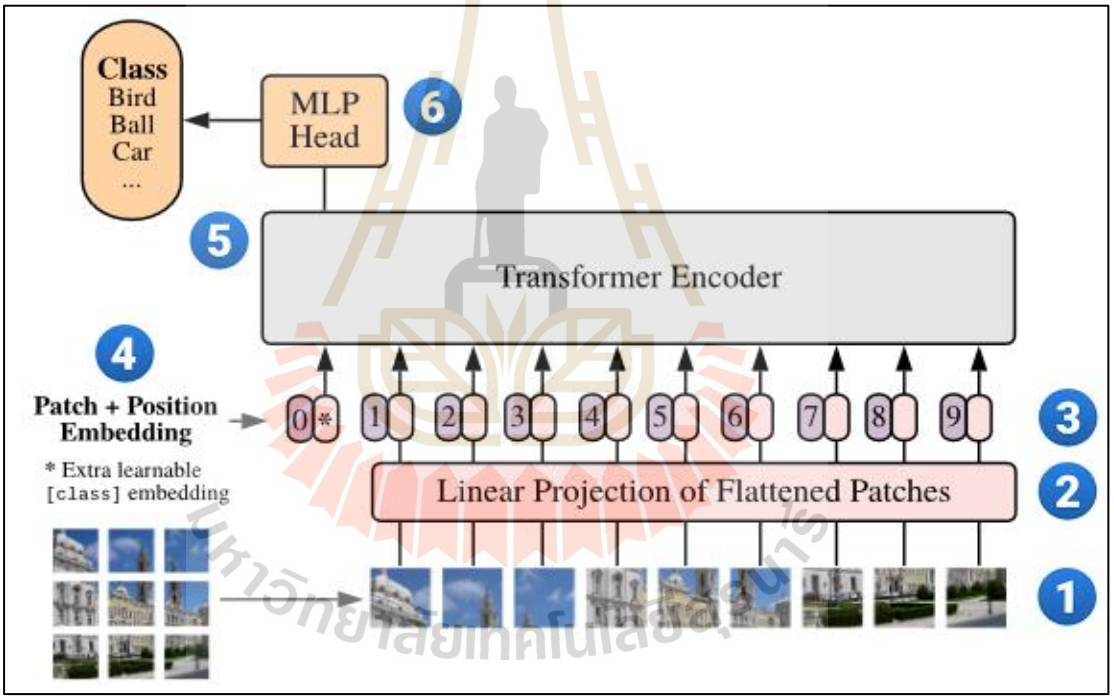
ดังนั้นการใช้สโตแคสติกเกรเดียนต์เดสเซนท์ (Stochastic Gradient Descent: SGD) เพื่อปรับปรุงค่าน้ำหนัก w ดังสมการที่ 2.12

$$w_{jk,t}^l = w_{jk,t-1}^l \alpha a_{k,t}^{l-1} \delta_{j,t}^l \quad (2.12)$$

ดังนั้นการหาค่าเกรเดียนต์ของแต่ละชั้น เพื่อหาค่าที่เหมาะสมที่สุดในการปรับแต่งค่าน้ำหนัก และเพิ่มประสิทธิภาพการปรับค่าอัตราการเรียนรู้ของแบบจำลองให้สามารถพยากรณ์ผลลัพธ์ออกมาได้อย่างแม่นยำจากค่าการสูญเสียที่มีค่าน้อยที่สุดหรือมีค่าความผิดพลาดน้อยที่สุดนั่นเอง

2.4 สถาปัตยกรรม ViT

สถาปัตยกรรม ViT (Vision in Transformer) คือนำสถาปัตยกรรม Transformer (Vaswani et al., 2017) ที่ใช้ในการประมวลผลข้อความมาประยุกต์ใช้สำหรับการจำแนกรูปภาพ ตามแนวคิดของดูข้อมูลภาพเป็นลำดับของแพตช์ (Patches) ขนาดคงที่ (Flattened) ซึ่งคล้ายกับวิธีประมวลผลคำในประโยค โดยแต่ละแพตช์จะถูกปรับให้เป็นแนวนอนและฝังเวกเตอร์แบบเชิงเส้น (Linearly embedded) แล้วประมวลผลตามลำดับด้วยตัวเข้ารหัสของ Transformer ร่วมกับการฝังตำแหน่งข้อความ (Positional embedding) เพื่อรักษาข้อมูลเชิงพื้นที่ (Spatial information) ตลอดจนการใช้เทคนิค Self-attention ที่จะช่วยให้สามารถเรียนรู้ความสัมพันธ์ระหว่างแพตช์คู่ใดก็ได้โดยไม่ต้องคำนึงถึงตำแหน่ง



รูปที่ 2.17 การทำงานของสถาปัตยกรรม ViT

ที่มา: Dosovitskiy et al., (2021)

จากรูปที่ 2.17 การจำแนกรูปภาพด้วยสถาปัตยกรรม ViT ในการสกัดคุณลักษณะรูปภาพประกอบด้วย 6 ขั้นตอน ได้แก่ 1) การแบ่งรูปภาพออกเป็นแพตช์ (Patches) ขนาด 16 * 16 ตามแนวคิดของ Transformer ที่จะแบ่งข้อความออกเป็นโทเค็น ดังนั้น ViT จะเปลี่ยนรูปภาพให้เป็นแพตช์ย่อย ๆ 2) การแปลงเทนเซอร์ (Tensor) ขนาด d * d * 3 ให้กลายเป็นเวกเตอร์ขนาด (d * d * 3) * 1 โดยที่ d คือ ขนาดของแพตช์ แล้วทำการฝังเวกเตอร์ (Vector embedding) เพื่อให้

ขนาดของเวกเตอร์เหมาะสมสำหรับนำเข้าสู่ตัวแบบ 3) การฝังข้อมูลตำแหน่ง (Positional embedding) จากการใช้ Muti-head self-attention เพื่อระบุตำแหน่งแพตช์จากการแบ่งรูปภาพไม่ให้สลับกัน อย่างไรก็ตาม การทำงานของ ViT นั้นมุ่งให้ความสนใจว่าแต่ละแพตช์มีความเชื่อมโยงกันอย่างไรมากกว่าลำดับของแพตช์ 4) การเพิ่มคลาสโทเค็นสำหรับจำแนกรูปภาพจำนวนเท่ากับคลาสที่ต้องการพยากรณ์มาบวกกับผลการฝังข้อมูลตำแหน่งแล้วเอาไปใส่ในตัวแบบพร้อมกับแพตช์ของรูปภาพ 5) การนำแพตช์และคลาสโทเค็นมาเชื่อมต่อกัน (Concatenate) แล้วส่งเข้าสู่ตัวเข้ารหัส Transformer และ 6) การจำแนกรูปภาพจากการวิเคราะห์บนโครงข่ายเพอร์เซปตรอนหลายชั้นผ่านการแปลงเป็นค่าความน่าจะเป็นตามสัดส่วนของผลลัพธ์สำหรับคลาสทั้งหมดด้วยฟังก์ชันซอฟต์แม็กซ์

ปัจจุบันมีการพัฒนาสถาปัตยกรรม ViT ในหลาย ๆ เวอร์ชัน (Dosovitskiy et al., 2021) เช่น ViT-Base, ViT-Large และ ViT-Huge ที่มีจำนวนพารามิเตอร์แตกต่างกันตามวัตถุประสงค์ในการทำงาน เกิดเป็นตัวแบบที่ได้รับความนิยม เช่น RN50, RN101, RN50x4, RN50x16, RN50x64, ViT-B/16, ViT-B/32, ViT-L/16, ViT-L/32 และ ViT-H/14 ดังตารางที่ 2.2

ตารางที่ 2.2 รายละเอียดตัวแบบ ViT แต่ละเวอร์ชัน

เวอร์ชัน	จำนวนชั้น	จำนวนชั้นซ่อนเร้น	ขนาดโครงข่ายประสาท	Heads	จำนวนพารามิเตอร์ (ล้าน)
ViT-base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

ที่มา: Dosovitskiy et al., (2021)

การทดสอบประสิทธิภาพของตัวแบบที่ถูกฝึกฝนล่วงหน้าบนชุดข้อมูล imagenet21k และปรับแต่งบนชุดข้อมูล imagenet2012 (Momin, www, 2021: 1) รายละเอียดดังตารางที่ 2.3 จะเห็นว่ายิ่งตัวแบบมีขนาดใหญ่ โดยพิจารณาจากจำนวนชั้นและจำนวนพารามิเตอร์นั้น จะส่งผลให้ใช้เวลาเรียนรู้มากขึ้นไปด้วย แม้จะดำเนินงานบนชุดข้อมูลเดียวกัน ดังนั้นงานวิจัยนี้จึงเลือกใช้ตัวแบบ ViT-B/32 เนื่องจากใช้เวลาในการเรียนรู้ต่ำที่สุดบนชุดข้อมูล imagenet21k และปรับแต่งอย่างละเอียดบนชุดข้อมูล imagenet2012 เท่ากันที่ 4.2 ชั่วโมง ในขณะที่ให้ความถูกต้องเท่ากันถึง 0.8179 อยู่ในระดับดีมาก และไม่แตกต่างจากตัวแบบอื่น ๆ ที่มีขนาดใหญ่กว่ามากนัก

ตารางที่ 2.3 การทดสอบประสิทธิภาพตัวแบบที่ถูกฝึกฝนล่วงหน้าบนชุดข้อมูล imagenet21k และปรับแต่งอย่างละเอียดบนชุดข้อมูล imagenet2012

ตัวแบบ	ชุดข้อมูล imagenet21k		ชุดข้อมูล imagenet2012	
	ค่าความถูกต้อง	เวลาที่ใช้ (ชม.)	ค่าความถูกต้อง	เวลาที่ใช้ (ชม.)
R50 + ViT-B/16	0.8505	25.9	0.8492	25.9
ViT-B/16	0.8462	19.9	0.8461	17.8
ViT-B/32	0.8179	4.2	0.8179	4.2
ViT-L/16	0.8507	59.2	0.8505	59.2
ViT-L/32	0.8122	14.7	0.8120	14.7

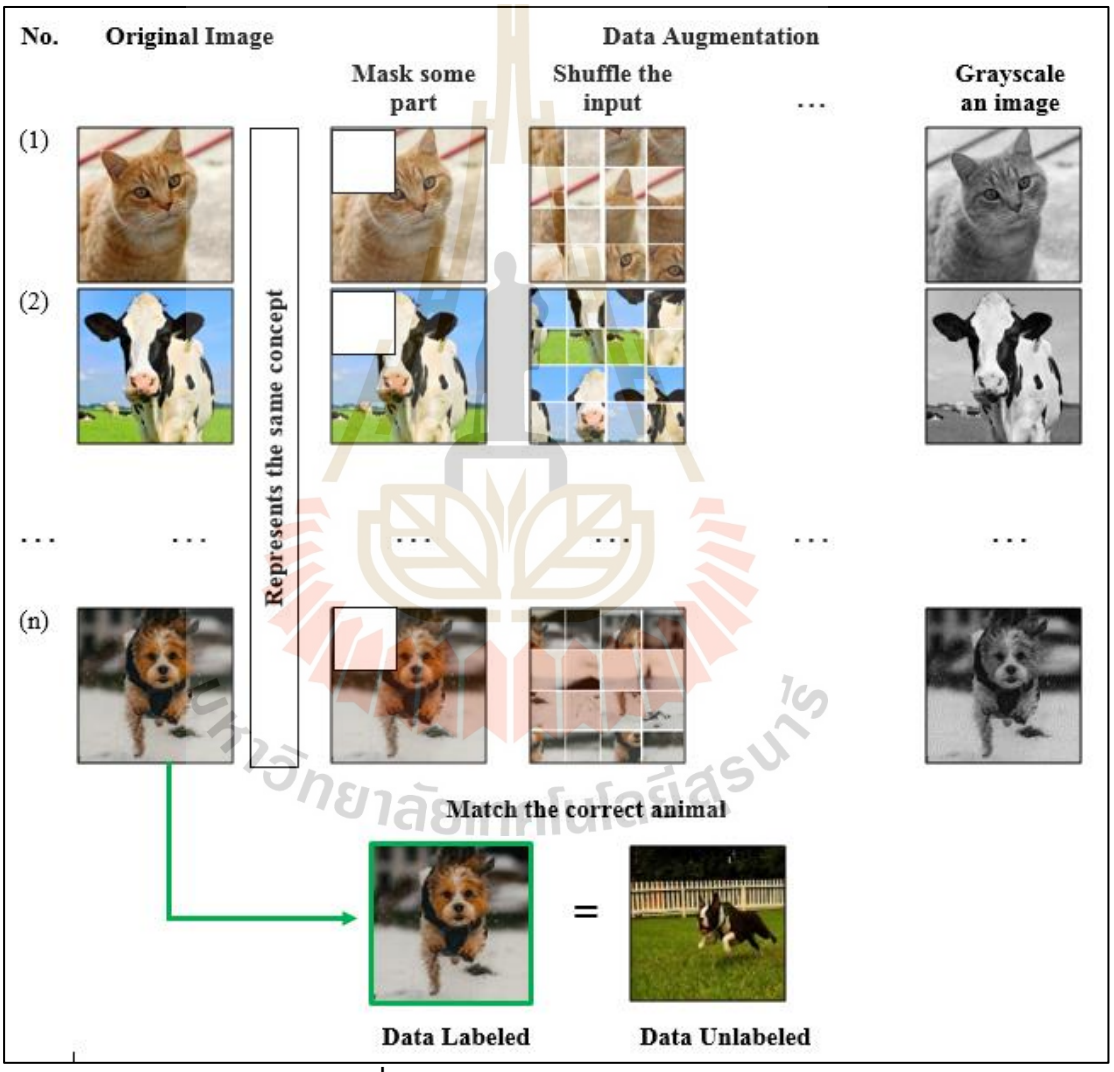
ที่มา: Dosovitskiy et al., (2021)

จากตารางที่ 2.3 โมเดล ViT-B/32 หรือ Vision in Transformer รุ่น Base ที่ใช้ขนาดแพตช์ 32 * 32 พิกเซล และมีจำนวนพารามิเตอร์ประมาณ 88 ล้านพารามิเตอร์ เป็นหนึ่งในสถาปัตยกรรมการเรียนรู้เชิงลึกที่นำแนวคิดของ Transformer ซึ่งประสบความสำเร็จอย่างสูงในงานประมวลผลภาษาธรรมชาติมาประยุกต์ใช้กับการประมวลผลภาพที่ถูกแบ่งออกเป็นแพตช์เล็ก ๆ โดยแต่ละแพตช์ขนาด 32 * 32 พิกเซล จากนั้นแต่ละแพตช์จะถูกแปลงเป็นเวกเตอร์เพื่อใช้เป็นโทเค็น (Token) ในการป้อนเข้าสู่โมเดลโดยไม่ต้องใช้คอนโวลูชัน (Convolution) เนื่องจาก โมเดลจะใช้กลไก Self-attention ที่อยู่ใน Multi-head attention เพื่อเรียนรู้ความสัมพันธ์ระหว่างแพตช์ทั้งหมดในรูปภาพ ทำให้สามารถจับบริบท (Context) ที่อยู่ห่างกันมาก ๆ ได้ตั้งแต่ชั้นแรกของโมเดล ส่งผลให้สามารถประมวลผลภาพในระดับโกลบอล (Global feature) ได้ดีกว่า โดยเฉพาะในงานที่ต้องการเข้าใจความหมายเชิงลึกของภาพ เนื่องจากให้ผลลัพธ์เป็นคุณลักษณะเชิงความหมายแบบมีบริบท (Contextual features) ซึ่งสามารถนำไปใช้ต่อในงานต่าง ๆ ได้อย่างมีประสิทธิภาพ เช่น การจำแนกประเภทภาพเชิงนามธรรมหรือการเข้าใจความหมายโดยรวม

2.5 การเรียนรู้ด้วยตนเอง

การเรียนรู้ด้วยตนเอง (Self-supervised learning) (Sawarkar, 2022) เป็นรูปแบบการเรียนรู้แบบไม่มีผู้สอนบนชุดข้อมูลแบบติดป้ายกำกับอัตโนมัติ (Automatically labeled) เนื่องจากแบบจำลองมีการเรียนรู้จากข้อมูลที่ไม่จำเป็นต้องมีป้ายกำกับล่วงหน้า แต่ต้องมีป้ายกำกับสำหรับการค้นหา และใช้ประโยชน์จากความสัมพันธ์ (Relations) หรือความเกี่ยวข้อง (Correlations) ระหว่างคุณสมบัติของข้อมูลอินพุตที่แตกต่างกัน โดยบังคับให้แบบจำลองเรียนรู้

การแสดงความหมายเกี่ยวกับข้อมูลด้วยการเรียนรู้ความสัมพันธ์ระหว่างรูปภาพกับข้อความภาษาธรรมชาติ แทนที่จะเรียนรู้ภายใต้คลาสที่จำกัดอย่างสุนัขหรือแมวเช่นเดิม เมื่อตัวแบบได้รับการฝึกฝนจากข้อความทั้งประโยคจะส่งผลให้ตัวแบบสามารถเรียนรู้สิ่งต่าง ๆ ได้มากขึ้น และสามารถค้นหารูปแบบความสัมพันธ์บางอย่างระหว่างรูปภาพกับข้อความด้วยการเรียนรู้ด้วยตนเอง จากนั้นชุดความรู้ (Knowledge) ที่เป็นผลลัพธ์จากการเรียนรู้จะถูกถ่ายโอนการเรียนรู้ (Transfer learning) ไปยังแบบจำลองเพื่อใช้สำหรับงานหลัก (Main task) ดังรูปที่ 2.18



รูปที่ 2.18 แนวคิดการเรียนรู้ด้วยตนเอง

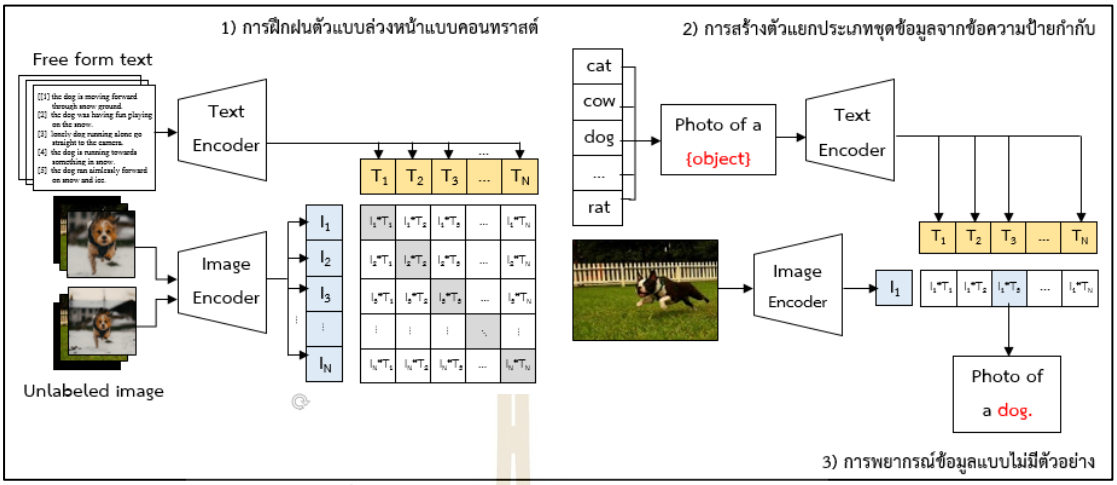
จากรูปที่ 2.18 แนวคิดการเรียนรู้ด้วยตนเองเพื่อทำความเข้าใจรูปภาพจากการเรียนรู้คุณลักษณะเฉพาะของแต่ละรูปภาพ (Represents the same concept) เช่น รูปภาพสุนัขต้นฉบับ (Original image) จากนั้นจะฝึกฝนตัวแบบให้เรียนรู้เพื่อจดจำรูปภาพสุนัขอื่น ๆ ที่มีรูปแบบหรือ

โครงสร้างที่เหมือนกันด้วยการดัดแปลงรูปภาพต้นฉบับ (Data augmentation) เช่น การมาสก์ปิดบังรูปภาพบางส่วน (Mask some part) การแบ่งรูปภาพออกเป็นส่วนย่อย ๆ แล้วสลับตำแหน่ง (Shuffle the input) และการปรับรูปภาพให้เป็นโทนสีเทา (Grayscale an image) เป็นต้น จะเห็นได้ว่าการเรียนรู้ด้วยตนเองมีลักษณะเดียวกับเรียนรู้ของเด็กที่แม้ยังไม่สามารถสื่อสารหรือเข้าใจภาษาได้ดูรูปสุนัข แล้วให้เลือกว่าภาพใดคล้ายคลึงมากที่สุดเมื่อเปรียบเทียบกับรูปภาพสัตว์ต่าง ๆ เป็นไปได้มากที่เด็กจะเลือกรูปภาพสุนัขได้อย่างถูกต้อง จะเห็นได้ว่าการเรียนรู้ด้วยตนเองนั้น ไม่จำเป็นต้องติดป้ายกำกับเพื่อแบ่งคลาสใด ๆ เหมือนเรียนรู้แบบมีผู้สอน (Supervised learning) ดังนั้นจึงถูกนำมาใช้กับการฝึกฝนแบบคอนทราสต์ เพื่อฝึกฝนแบบจำลองให้สามารถจัดหมวดหมู่วัตถุที่ไม่เคยเห็นมาก่อน (Zero-shot Prediction) คล้ายกับมนุษย์ที่สามารถจำแนกประเภทสิ่งของที่แตกต่างกันได้โดยที่ไม่ต้องมีการสอนหรือการเรียนรู้มาก่อน

งานวิจัยนี้ได้นำแนวคิดการเรียนรู้ด้วยตนเองมาใช้เพื่อให้แบบจำลองสามารถเรียนรู้และสร้างป้ายกำกับจากข้อมูลด้วยตนเอง โดยไม่จำเป็นต้องพึ่งพาการกำกับจากมนุษย์โดยตรง วิธีการนี้ช่วยลดข้อจำกัดในการเตรียมข้อมูลที่ต้องใช้แรงงานจำนวนมากในการจัดทำป้ายกำกับแบบดั้งเดิม อีกทั้งยังเปิดโอกาสให้สามารถนำข้อมูลขนาดใหญ่ที่มีอยู่แล้วมาใช้ในการฝึกตัวแบบได้อย่างมีประสิทธิภาพ โดยแบบจำลองจะเรียนรู้ความสัมพันธ์เชิงความหมายที่ซ่อนอยู่ในข้อมูล เช่น ความสอดคล้องระหว่างภาพและข้อความคำอธิบาย เพื่อสร้างตัวแทนข้อมูลเชิงความหมาย (Semantic representations) ด้วยตนเอง แนวทางนี้จึงเหมาะสมอย่างยิ่งสำหรับการประยุกต์ใช้ในบริบทที่ข้อมูลมีจำนวนมาก แต่ไม่มีการระบุป้ายกำกับอย่างเป็นทางการ ซึ่งส่งผลให้แบบจำลองสามารถนำไปประยุกต์ใช้ในงานที่หลากหลายได้อย่างยืดหยุ่นและครอบคลุมมากขึ้น

2.6 ตัวแบบ CLIP ที่ถูกฝึกฝนล่วงหน้า

ตัวแบบ CLIP (Contrastive Language–Image Pre-training) (Radford et al., 2021) เป็นตัวแบบโครงข่ายประสาทเทียมเชิงลึกที่ฝึกฝนล่วงหน้าจากรูปภาพและข้อความภาษาธรรมชาติบนอินเทอร์เน็ต 400 ล้านคู่ พัฒนาโดยบริษัทไม่แสวงหาผลกำไรชื่อ OpenAI ที่เป็นห้องปฏิบัติการวิจัยปัญญาประดิษฐ์ โดยได้รับการยอมรับว่า มีประสิทธิภาพสูงในการจำแนกข้อมูลแบบไม่มีตัวอย่างในการเรียนรู้ (Zero-shot classification) กรอบการทำงานของตัวแบบ CLIP ประกอบด้วย 3 ขั้นตอน ดังรูปที่ 2.19



รูปที่ 2.19 กรอบการทำงานของตัวแบบ CLIP

ที่มา: Radford et al. (2021)

1) การฝึกฝนล่วงหน้าแบบคอนทราสต์ (Contrastive pre-training) เพื่อฝึกฝนตัวเข้ารหัสรูปภาพ และตัวเข้ารหัสข้อความภาษาธรรมชาติบนพื้นที่การฝังหลายรูปแบบ มีรายละเอียดดังนี้

1.1) การเข้ารหัสรูปภาพ (Image encode) ฝึกฝนตัวเข้ารหัสรูปภาพแบบ ViT-B/32 สำหรับการสกัดคุณลักษณะรูปภาพตามแนวคิดของสถาปัตยกรรม Transformer ที่ทำงานกับข้อความ แต่นำมาปรับใช้กับรูปภาพ (Vision in Transformer) จากการเปรียบเทียบความสัมพันธ์ระหว่างแพตช์เพื่อคำนวณหาผังคุณลักษณะในการหาความสัมพันธ์กับบริเวณทั้งภาพ จึงมีความแม่นยำสูงกว่าหากชุดข้อมูลมีจำนวนมากพอ (Liu, Sun and Zhou, 2022) ผลลัพธ์จากการฝังรูปภาพ (Image embedding) จะแปลงเป็นเวกเตอร์คุณลักษณะรูปภาพ (Image feature vector)

1.2) การเข้ารหัสข้อความ (Text encode) ฝึกฝนตัวเข้ารหัสข้อความด้วย GPT-2 บนพื้นฐานสถาปัตยกรรม Transformer เพื่อเรียนรู้ความหมายโดยรวมของข้อความบรรยายรูปภาพแต่ละประโยค ผลลัพธ์จากการฝังข้อความ (Text embedding) แปลงเป็นเวกเตอร์คุณลักษณะข้อความ (Text feature vector)

1.3) การฝึกฝนตัวเข้ารหัสรูปภาพและตัวเข้ารหัสข้อความบนพื้นที่การฝังหลายรูปแบบด้วยการฝึกฝนล่วงหน้าแบบคอนทราสต์ระหว่างรูปภาพและข้อความภาษาธรรมชาติการเรียนรู้ระหว่างข้อความคำบรรยายรูปภาพด้วยตัวเข้ารหัสข้อความ (Text encoder) และการเรียนรู้รูปภาพด้วยตัวเข้ารหัสรูปภาพ (Image encoder) จะดำเนินการไปพร้อมกัน เพื่อพยายามหาค่าสูงที่สุด (Maximize) ของผลคูณผลคูณดอทยุคคลิดของเวกเตอร์รูปภาพ (I_x) คูณเวกเตอร์ข้อความ (T_x) ที่เป็นคู่กัน และหาค่าต่ำที่สุด (Minimize) ของเวกเตอร์รูปภาพ (I_x) และของเวกเตอร์ข้อความ (T_y) อื่น ๆ ที่ไม่ใช่คู่กันด้วยค่าความคล้ายคลึงโคไซน์ ดังสมการที่ 2.13

$$Cos\theta = \frac{A \,.\, B}{||A||\ ||B||} = \frac{\sum_{i=1}^{n} A_i B_1}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (2.13)$$

กำหนดให้ A และ B คือ เวกเตอร์

i คือ องค์ประกอบของเวกเตอร์ (Element of vector)

n คือ จำนวนขององค์ประกอบ (Number of element)

. คือ ผลคูณดอทยุคคลิด (Euclidean dot product)

|||| คือ ขนาดของเวกเตอร์ (Magnitude)

ดังนั้นกระบวนการเรียนรู้ CLIP จะพยายามจับคู่ภาพกับข้อความที่เกี่ยวข้องกัน และในขณะเดียวกันก็เรียนรู้ที่จะแยกแยะข้อความที่ไม่เกี่ยวข้องออกไป โดยไม่ต้องพึ่งพาป้ายกำกับที่มนุษย์กำหนดโดยตรง วิธีการนี้ช่วยให้ตัวแบบสามารถเข้าใจความสัมพันธ์เชิงความหมายระหว่างรูปภาพและข้อความได้อย่างลึกซึ้งยิ่งขึ้น ผลลัพธ์ของเวกเตอร์รูปภาพและเวกเตอร์ข้อความคู่ที่มีคุณลักษณะคล้ายคลึงกัน ให้เข้าใกล้กับ 1 มากยิ่งขึ้น ในทางกลับกัน ค่าความคล้ายคลึงโคไซน์ของเวกเตอร์รูปภาพและเวกเตอร์ข้อความคู่ที่คุณลักษณะแตกต่างกันก็จะลดลงจนเข้าใกล้กับ 0

2) การสร้างตัวจำแนกประเภทชุดข้อมูลจากป้ายกำกับ (Create dataset classifier from label text) โดยใช้ประโยชน์จากตัวแบบที่ถูกฝึกมาแล้วล่วงหน้าในการพยากรณ์ป้ายกำกับจากค่าน้ำหนักที่ได้จากการเรียนรู้ โดยฝังอ็อบเจกต์คลาสลงเป็นข้อความบรรยายรูปภาพ (Object classes into captions) เช่น “เป็นภาพของ {อ็อบเจกต์} (a photo of an {object})” หรือ “ตรงกลางของภาพคือ {อ็อบเจกต์} (a centered satellite photo of {object})” เป็นต้น

3) การพยากรณ์ข้อมูลแบบไม่มีตัวอย่าง (Use for zero-shot prediction) โดยตัวแบบไม่จำเป็นต้องมีชุดข้อมูลตัวอย่างในการพยากรณ์เพื่อติดป้ายกำกับ เมื่ออินพุตรูปภาพที่ต้องการจะจำแนกผ่านตัวเข้ารหัสรูปภาพ เพื่อสร้างเวกเตอร์ฝังตัว (Embedding vector) จากนั้นจะทำการวัดความคล้ายคลึงโคไซน์ระหว่างเวกเตอร์การฝังของรูปภาพและเวกเตอร์การฝังข้อความ โดยค่าที่มีความคล้ายคลึงโคไซน์สูงสุดจะเป็นค่าจากการพยากรณ์ จากรูปที่ 2.19 ผลการพยากรณ์ป้ายกำกับที่มีความคล้ายคลึงสูงสุดคือ “รูปภาพสุนัข (a photo of a dog)”

งานวิจัยนี้ประยุกต์ใช้ตัวแบบ CLIP ที่ถูกฝึกฝนล่วงหน้าบนแนวทางการเรียนรู้ด้วยตนเองแบบคอนทราสต์ในการฝึกฝนแบบจำลองจากรูปภาพกับข้อความบรรยายรูปภาพที่มีความหมายสอดคล้องกัน เพื่อเรียนรู้ความสัมพันธ์ที่เกี่ยวข้องกัน จากนั้นคำนวณความคล้ายคลึงกันระหว่างคู่รูปภาพกับป้ายกำกับ เพื่อจำแนกคุณลักษณะเชิงความหมายในรูปแบบที่มนุษย์สามารถเข้าใจได้

2.7 การวัดประสิทธิภาพการค้นคืนรูปภาพแบบไบนารี

การวัดประสิทธิภาพการค้นคืนรูปภาพแบบไบนารี (Binary relevance metrics) (Chaudhary, www, 2020: 1) ที่มักใช้ค้นคืนเนื้อหาบนเว็บไซต์ Google การค้นคืนวิดีโอบนเว็บไซต์ YouTube การค้นคืนผลิตภัณฑ์บนเว็บไซต์ Amazon หรือการค้นคืนบน Gmail เป็นต้น ด้วยการประเมินผลลัพธ์ที่เกี่ยวข้อง (Relevant) เทียบกับผลลัพธ์ที่ไม่เกี่ยวข้อง (Irrelevant) แล้วสรุปค่าออกมาในรูปแบบตาราง Confusion matrix

งานวิจัยมุ่งวัดประสิทธิภาพการค้นคืนรูปภาพแบบไบนารี ด้วยมาตรวัด 2 ประเภท ได้แก่ มาตรวัดแบบไม่สนใจลำดับของผลลัพธ์ และมาตรวัดความเกี่ยวข้องแบบให้คะแนน ซึ่งแต่ละประเภทมีลักษณะการใช้งานที่แตกต่างกันตามวัตถุประสงค์ในการประเมินผล มีรายละเอียดดังนี้

2.7.1 มาตรวัดแบบไม่สนใจลำดับของผลลัพธ์

มาตรวัดแบบไม่สนใจลำดับของผลลัพธ์ (Order-unaware metrics) เป็นมาตรวัดที่ไม่สนใจลำดับของรายการที่ถูกค้นคืนตามที่แบบจำลองจัดอันดับ ผลลัพธ์จะแสดงถึงรูปภาพที่ถูกค้นคืนนั้นถูกต้องหรือไม่ โดยไม่คำนึงถึงตำแหน่งที่รูปภาพนั้นปรากฏในลำดับ

ตัวอย่างผลลัพธ์จำนวน 5 รูปภาพจากข้อความค้นหา “the dog running on a grass” ดังตารางที่ 2.4 กำหนดให้หากผลลัพธ์ถูกต้องและครบถ้วนตามข้อความค้นหา เท่ากับ 1 คะแนน ตรงกันข้าม หากผลลัพธ์ไม่ถูกต้องหรือไม่ครบถ้วนตามข้อความค้นหา เท่ากับ 0 คะแนน

ตารางที่ 2.4 ผลลัพธ์ 5 รูปภาพจากการค้นคืนรูปภาพดิจิทัลที่ผ่านการประเมินความถูกต้อง

1	2	3	4	5
				
✓	✓	✓	✗	✓

1) ค่าความแม่นยำที่ k อันดับ (Precision@k) (Chaudhary, www, 2020: 1) เป็นสัดส่วนของรูปภาพที่ถูกต้องจากจำนวนรูปภาพทั้งหมดจากการค้นคืน ดังสมการที่ 2.14

$$Precision@k = \frac{\text{Number of relevant images in } k}{\text{Total number of images in } k} \quad (2.14)$$

จากรูปที่ 2.20 เมื่อแทนค่าในสมการที่ 2.14 การคำนวณหาค่าความแม่นยำของผลลัพธ์การค้นคืนรูปภาพแบบไม่สนใจลำดับของผลลัพธ์ เมื่อกำหนดเป็น k = 3 และ 5 ดังนี้

รูปที่ 1 ถูกต้อง
รูปที่ 2 ถูกต้อง
รูปที่ 3 ถูกต้อง $Precision@3 = \frac{3}{3} = \mathbf{1.00}$
รูปที่ 4 ไม่ถูกต้อง
รูปที่ 5 ถูกต้อง $Precision@5 = \frac{4}{5} = \mathbf{0.80}$

ดังนั้น เมื่อคำนวณหาค่าความแม่นยำของผลลัพธ์การค้นคืนรูปภาพที่ k = 3 และ 5 ค่าเท่ากับ 1.00 และ 0.80 ตามลำดับ

2) ค่าความครบถ้วนที่ k อันดับ (Recall@k) (Chaudhary, www, 2020: 1) เป็นสัดส่วนของรูปภาพที่ถูกต้องจากการค้นคืนจากจำนวนรูปภาพที่ถูกต้องทั้งหมดในชุดข้อมูล ดังสมการที่ 2.15

$$Recall@k = \frac{\text{Number of relevant images in } k}{\text{Total number of relevant item}} \quad (2.15)$$

จากรูปที่ 2.20 จะพบว่าจำนวนรูปภาพที่เกี่ยวข้องกับข้อความค้นหาทั้งหมดในชุดข้อมูลเท่ากับ 4 จาก 5 รูปภาพ เมื่อแทนค่าในจากสมการที่ 2.15 การคำนวณหาค่าความครบถ้วนของผลลัพธ์การค้นคืนรูปภาพแบบไม่สนใจลำดับของผลลัพธ์ เมื่อกำหนดเป็น k = 3 และ 5 ดังนี้

รูปที่ 1 ถูกต้อง
รูปที่ 2 ถูกต้อง
รูปที่ 3 ถูกต้อง $Recall@3 = \frac{3}{5} = \mathbf{0.60}$
รูปที่ 4 ไม่ถูกต้อง
รูปที่ 5 ถูกต้อง $Recall@5 = \frac{4}{5} = \mathbf{0.80}$

ดังนั้น เมื่อคำนวณหาค่าความครบถ้วนของผลลัพธ์การค้นคืนรูปภาพ k = 3 และ 5 มีค่าเท่ากับ 0.60 และ 0.80 ตามลำดับ

3) ค่าประสิทธิภาพโดยรวมที่ k อันดับ (F-measure@k) คือความสมดุลระหว่างความแม่นยำกับค่าความครบถ้วนที่เกิดจากการเปรียบเทียบกันระหว่างค่าความแม่นยำและค่าความครบถ้วนของแต่ละคลาสเป้าหมาย เพื่อเป็นมาตรวัดเดี่ยว (Single metric) ที่วัดความสามารถของโมเดลแทนค่าความแม่นยำและค่าความครบถ้วน เนื่องจากค่าความแม่นยำและค่าความครบถ้วนจะเป็นประโยชน์ต่อผู้ใช้ต่างกลุ่มกันไป โดยผู้ใช้จะให้ความสำคัญกับค่าความแม่นยำมากกว่าค่าความครบถ้วน เนื่องจากผู้ใช้ส่วนใหญ่ต้องการให้ข้อมูลที่เป็นผลลัพธ์ที่ตรงกับสิ่งที่ต้องการค้นหา แต่ในทางกลับกัน ผู้ที่พัฒนาระบบค้นหาข้อมูลจะเน้นไปที่ค่าความครบถ้วน เนื่องจากต้องการให้ระบบค้นหาและเลือกเอกสารที่ถูกต้องจากทั้งหมดให้ได้มากที่สุด ดังสมการที่ 2.16

$$F - measure@k = \frac{2 * (precision@k) * (recall@k)}{(precision@k) + (recall@k)} \quad (2.16)$$

ดังนั้น เมื่อคำนวณหาค่าประสิทธิภาพโดยรวม จากค่าความแม่นยำ k = 3 และ 5 มีค่าเท่ากับ 1.00 และ 0.80 และค่าความครบถ้วนที่ k = 3 และ 5 มีค่าเท่ากับ 0.60 และ 0.80 ตามลำดับ รายละเอียดดังตารางที่ 2.5

ตารางที่ 2.5 การคำนวณค่าประสิทธิภาพโดยรวมจากค่าความแม่นยำและค่าความครบถ้วน

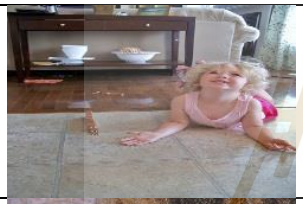
มาตรวัด	@1	@2	@3	@4	@5
Precision	1/1 = 1.00	2/2 = 1.00	3/3 = 1.00	3/4 = 0.75	4/5 = 0.80
Recall	1/5 = 0.20	2/5 = 0.40	3/5 = 0.60	3/5 = 0.60	4/5 = 0.80
F-measure	$\frac{2*(1/1)*(1/5)}{(1/1)+(1/5)}$ = 0.33	$\frac{2*(2/2)*(2/5)}{(2/2)+(2/5)}$ = 0.57	$\frac{2*(3/3)*(3/5)}{(3/3)+(3/5)}$ = 0.75	$\frac{2*(3/4)*(3/5)}{(3/4)+(3/5)}$ = 0.67	$\frac{2*(4/5)*(4/5)}{(4/5)+(4/5)}$ = 0.80

2.7.2 มาตรวัดความเกี่ยวข้องแบบให้คะแนน

มาตรวัดความเกี่ยวข้องแบบให้คะแนนด้วยค่า NDCG ที่ k อันดับ เนื่องจากแนวคิดระดับสูงที่เป็นนามธรรมนั้นยากที่จะประเมินผลแบบไบนารีเท่านั้น จึงเป็นที่มาของการใช้มาตรวัดความเกี่ยวข้องแบบให้คะแนนด้วยค่า NDCG (Normalized Discounted Cumulative Gain) ที่ k อันดับเข้ามาร่วมเป็นส่วนหนึ่งในการประเมินผลความหมายของแต่ละรูปภาพให้สอดคล้องกับหลักคอมพิวเตอร์วิทัศน์(Carnevali, www, 2023: 1) โดยเปรียบเทียบระหว่างการค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิมกับการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนัก

ค่า NDCG เป็นการวัดประสิทธิภาพผลลัพธ์การค้นคืนรูปภาพโดยใช้เกณฑ์ให้คะแนนกับรูปภาพที่เกี่ยวข้อง และให้ความสำคัญกับลำดับของรูปภาพ แบ่งระดับคะแนนความเกี่ยวข้อง (Judgment score) เป็น 5 ระดับ (5 Point scale) ตั้งแต่ 0 ถึง 4 โดย 0 คือรูปภาพไม่มีความเกี่ยวข้องกับข้อความค้นหาเลย จนถึง 4 คือรูปภาพมีความเกี่ยวข้องกับข้อความค้นหาครบถ้วน ยกตัวอย่างจากผลลัพธ์การค้นคืนจากข้อความค้นหา “a tabby cat on the floor” รายละเอียดดังตารางที่ 2.6

ตารางที่ 2.6 ตัวอย่างผลลัพธ์จำนวน 5 รูปภาพจากข้อความค้นหา “a tabby cat on the floor” ที่ผ่านการประเมินคะแนนความเกี่ยวข้อง

ลำดับ	รูปภาพจากการค้นคืน	คะแนนความเกี่ยวข้อง	คำอธิบาย
1)		1	รูปภาพเด็กผู้หญิงอยู่บนพื้น ถึงแม้ว่าจะไม่ใช่รูปภาพแมวลาย แต่มีความเกี่ยวข้องกับการค้นหาคือ on the floor
2)		4	แมวลายอยู่บนพื้น ตรงกับข้อความค้นหาทุกประการ
3)		4	แมวลายอยู่บนพื้น ตรงกับข้อความค้นหาทุกประการ
4)		2	เป็นแมวแต่เป็นแมวสีดำ และอยู่บนเก้าอี้ ไม่ได้อยู่บนโต๊ะ
5)		0	เป็นสุนัข และอยู่บนถนน ซึ่งไม่มีความเกี่ยวข้องใด ๆ เลยกับข้อความค้นหา

จากตารางที่ 2.6 เมื่อยกตัวอย่างผลลัพธ์จากการค้นคืนในลำดับที่ 1 และ 2 ที่ผ่านการประเมินคะแนนความเกี่ยวข้องด้วยค่า DCG (Discounted Cumulative Gain) ของผลลัพธ์จากการค้นคืน 2 ลำดับ ซึ่งสามารถคำนวณได้ด้วยจากสมการที่ 2.17

$$DCG@k = \sum_{k=1}^{K} \frac{rel_k}{log_2(1+K)} \quad (2.17)$$

โดย k คือ จำนวนผลลัพธ์จากการค้นคืน

rel_k คือ คะแนนความเกี่ยวข้องระหว่างผลลัพธ์จากการค้นคืนกับข้อความค้นหา

log_2 คือ ปัจจัยที่ทำให้คะแนนของผลลัพธ์จากการค้นคืนถูกลดทอนลงตามอัตราส่วน

เมื่อแทนผลลัพธ์จากการค้นคืนในลำดับที่ 1 และ 2 ในสมการที่ 2.17 เพื่อคำนวณค่า $_{original}DCG$ มีรายละเอียดดังนี้

$$_{original}DCG@2 = \frac{1}{log_2(1+1)} + \frac{4}{log_2(1+2)}$$

$$= 1 + 2.52$$

$$= 3.52$$

ดังนั้นเมื่อทำการสลับตำแหน่งผลลัพธ์ในลำดับที่ 1 และ 2 ของตารางที่ 2.6 แล้วแทนค่าในสมการที่ 2.17 อีกครั้ง เพื่อคำนวณค่า $_{swapped}DCG$ มีรายละเอียดดังนี้

$$_{swapped}DCG@2 = \frac{4}{log_2(1+1)} + \frac{1}{log_2(1+2)}$$

$$= 4 + 0.63$$

$$= 4.63$$

จากนั้นจึงคำนวณค่า IDCG@k คือผลลัพธ์ในอุดมคติของ DCG@k ซึ่งเป็นผลการค้นคืนรูปภาพที่ได้จัดเรียงตามลำดับความสำคัญทั้งหมดดังสมการที่ 2.18

$$IDCG@k = {}_{swapped}DCG@k + {}_{original}DCG@k \quad (2.18)$$

$$IDCG@k = \frac{4}{log_2(1+1)} + \frac{4}{log_2(1+2)}$$

$$= 4 + 2.52$$

$$= 6.52$$

ขั้นตอนสุดท้ายคือการคำนวณค่าประสิทธิผลของการค้นคืนจะถูกหักลดลงตามสัดส่วนของระดับความสำคัญของเอกสาร (Relevance grade) ที่ปรากฏในการจัดลำดับการค้นคืนด้วยค่า nDCG สามารถคำนวณได้ดังสมการที่ 2.19

$$nDCG@k = \frac{{}_{original}DCG@k}{ICDG@k} \quad (2.19)$$

$$nDCG@2 = \frac{2.52}{6.52}$$

$$= 0.39$$

จะเห็นได้ว่าการใช้ค่า nDCG นั้นมีข้อดีกว่าการวัดประสิทธิภาพด้วยค่าความแม่นยำเพียงอย่างเดียว เนื่องจากมีการคำนึงถึงการจัดเรียงลำดับของผลลัพธ์จากการค้นคืนได้อีกด้วย

งานวิจัยนี้มุ่งวัดประสิทธิภาพการค้นคืนรูปภาพแบบไบนารีด้วยมาตรวัดแบบไม่สนใจลำดับของผลลัพธ์ ประกอบด้วยค่าความแม่นยำที่ k อันดับ ค่าความครบถ้วนที่ k อันดับ และค่าประสิทธิภาพโดยรวมที่ k อันดับ ร่วมกับมาตรวัดความเกี่ยวข้องแบบให้คะแนนด้วยค่า NDCG ที่ k อันดับ เพื่อประเมินประสิทธิภาพแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายในทุกมิติ

2.8 งานวิจัยที่เกี่ยวข้อง

การปริทัศน์งานวิจัยที่เกี่ยวข้องกับการค้นคืนรูปภาพเชิงความหมาย เดิมเริ่มจากใช้คุณลักษณะระดับต่ำที่ถูกสกัดด้วยอัลกอริทึมต่าง ๆ ก่อนจะนำมาจัดเป็นหมวดหมู่เดียวกันด้วยโทนสี รูปร่างหรือรูปทรง (Manzoor et al., 2015) แต่อาจเกิดปัญหาเมื่อเจอกับภาพชายหาดและภาพชายทะเลที่ควรจัดอยู่ในหมวดหมู่เดียวกัน แต่ทั้งสองภาพมีคุณลักษณะของโทนสี และพื้นที่ที่แตกต่างกันอย่างชัดเจน จึงยากที่จะกำหนดให้อัลกอริทึมจัดให้หมวดหมู่เดียวกัน (Mai et al., 2017) ต่อมาเป็นการผสมผสานกันระหว่างคุณลักษณะสี รูปร่าง รูปทรง และพื้นผิวของรูปภาพ หรือมีการสกัดข้อมูลภาพเป็นฟีเจอร์แบ่งออกเป็นหลายระดับ (Kalimuthu and Krishnamurthi, 2015) แต่ในคุณลักษณะเชิงความหมายรูปภาพจะถูกติดป้ายกำกับและกำหนดคลาสที่เหมาะสมจากชุดข้อมูลที่ค้นหา (Mane, 2016) เพื่อให้สามารถวิเคราะห์ในรูปแบบที่ซับซ้อนได้มากขึ้น ก่อนจะพัฒนาเป็นการจับคู่ระหว่างคุณลักษณะระดับต่ำกับแนวคิดระดับสูง (Gupta and Singh, 2018) เพื่อค้นคืนรูปภาพจากฐานข้อมูลภาพขนาดใหญ่บนอินเทอร์เน็ต โดยจัดกลุ่มคุณลักษณะประเภทเดียวกันให้เป็นกลุ่มเพื่อจับคู่คุณลักษณะ (Pattern matching) (Potapov et al., 2018) อย่างไรก็ดีในความเป็นจริงแล้วการมองภาพของมนุษย์มักจะพิจารณาจากความหมายของรูปภาพเป็นหลักโดยที่ไม่จำเป็นต้องมีคุณลักษณะสีหรือรูปทรงแบบเดียวกันก็สามารถตัดสินว่าเป็นภาพชนิดเดียวกันได้ส่งผลให้การค้นคืนรูปภาพอาจไม่ได้ผลลัพธ์ที่ตรงกับความหมายในภาพอย่างแท้จริง

ปัญหาดังกล่าวจึงทำให้เกิดเป็นแนวคิดการแปลความหมายของภาพจากองค์ประกอบหรือวัตถุที่ปรากฏในภาพนั้น เพื่อให้ได้ผลลัพธ์ที่ตรงตามความต้องการมากที่สุด และแทนวัตถุนั้น ๆ ด้วยการแท็ก (Tag) หรือการให้ความหมายของวัตถุบนภาพเป็นชื่อวัตถุ หรือคำศัพท์ที่สอดคล้องกัน (Patel and Sampat, 2017) เพื่อสืบค้นข้อมูลแทนในลักษณะใช้ความสัมพันธ์ของคำหลัก (Synonym) เข้ามาใช้ในการค้นคืนข้อมูลภาพ เช่น “stone” มีความหมายเช่นเดียวกับ “rock” เป็นต้น ร่วมกับสร้างความสัมพันธ์ (Relationship) ของแท็กชื่อวัตถุบนภาพ เช่น การเชื่อมวัตถุด้วยคำต่าง ๆ เช่น “touch” “on” “top” เป็นต้น ซึ่งผลลัพธ์ที่ขึ้นกับคำศัพท์ที่ถูกแท็กไว้บนภาพยิ่งมีการแท็กบนภาพมากก็ยิ่งสามารถหาความหมายบนภาพได้มากขึ้น (Chen, Lu and Lin, 2020) แต่ในความเป็นจริงแล้วการแท็กข้อมูลบนภาพนั้นเป็นเพียงการฝังคำศัพท์ที่ต้องการบนภาพแต่ไม่ได้ให้ความหมายของภาพโดยรวม เช่นเดียวกับวิธีการที่นิยมใช้อยู่ในปัจจุบันอย่างการค้นคืนรูปภาพผ่านโปรแกรมค้นหาที่ผลลัพธ์จากการค้นคืนอาจเป็นรูปภาพที่มีข้อความค้นหานั้นฝังอยู่ในภาพหรืออาจเป็นรูปภาพภายในเว็บเพจที่มีหัวข้อหรือเนื้อหาที่ตรงกับข้อความค้นหาดังกล่าวตามแนวคิดการใช้ข้อความที่อยู่รอบ ๆ รูปภาพในการอธิบายความหมายของรูปภาพ (Chen, Trouve, Murakami and Fukuda, 2017) เข้ากับแนวคิดเชิงความหมายที่หลากหลายและการที่ใช้ฐานความรู้

รวมถึงใช้ผลตอบรับเชิงโต้ตอบจากเทคนิคการตอบกลับของผู้ใช้ เพื่อดึงผลลัพธ์ที่ผู้ใช้คาดหวัง โดยแก้ไขปัญหาช่องว่างทางความหมาย (Anandh, Mala and Babu, 2020)

ความหมายของภาพคือการนำวัตถุที่ปรากฏบนภาพมารวมกัน เพื่อวิเคราะห์ให้ได้ผลลัพธ์ที่แทนความหมายของภาพทั้งภาพ เป็นที่มาของการใช้ทฤษฎีการมองของมนุษย์ที่มีการพิจารณาจากตำแหน่งและขนาดของวัตถุที่เกิดขึ้น เพื่อนำมาใช้ในการแปลความหมายของภาพเรียกว่าโครงสร้างสเก็ตตรีตรอน (Zhang et al., 2020) ด้วยการวัดโครงสร้างกับภาพต้นฉบับที่มีการแท็ก แต่มักพบว่าค่าตำแหน่งในส่วนของภาพนั้นยังไม่เพียงพอต่อการนำมาใช้สำหรับการแปลความหมาย ภาพยังมีวิธีการที่ผสมผสานระหว่างการใช้แท็กคำหลักบนภาพกับการสร้างความสัมพันธ์ระหว่างวัตถุด้วยมัลติเปิลเคอร์แนล (Multiple kernel) (Zhao, Pei, Zheng and Tao, 2020) ของพื้นที่วัตถุบนภาพ ผลลัพธ์ที่ได้ส่วนใหญ่จึงเป็นกลุ่มของคำหลักที่ปรากฏชัดเจนบนภาพเท่านั้น จึงมีความพยายามใส่ข้อความ เพื่อบรรยายความหมายของภาพด้วยข้อมูลต่าง ๆ จนครบถ้วนในรูปแบบของใคร ทำอะไร ที่ไหน เมื่อไร ซึ่งบางครั้งเป็นข้อมูลที่นอกเหนือหรือเกินความจำเป็นทำให้เกิดความสับสนของผลลัพธ์ที่ได้มา จะเห็นว่าการใช้คำอธิบายภาพขึ้นอยู่กับวัตถุที่ผู้ใช้สนใจมีลักษณะเป็นแบบใด ดังนั้นแต่ละระดับของการให้คำอธิบายลงบนภาพนั้นมีความสำคัญ จึงมีการนำออนโทโลยีมาสร้างเป็นแนวทางสำหรับสร้างรูปแบบคำอธิบาย (Nhi, Le and Van, 2022) ซึ่งจะมีการกำหนดประเภทวัตถุ ความสัมพันธ์ระหว่างวัตถุ รวมทั้งความสัมพันธ์ระหว่างประเภท อย่างไรก็ตาม การวิเคราะห์คำบรรยายความหมายของภาพนั้น อาจจะไม่สามารถควบคุมการรู้จำวัตถุในบางกรณี เนื่องจากขนาดของวัตถุเล็กเกินไป อีกทั้งปริมาณคำศัพท์ที่มีความหมายคล้ายกันหรือใช้แทนกันนั้นยากในการกำหนดให้ครอบคลุม

แม้การศึกษางานวิจัยที่มุ่งประเมินสุนทรียภาพ (Aesthetics) และคุณภาพ (Technical quality) ของรูปภาพจะไม่ใช่เรื่องใหม่ โดย Talebi and Peyman (2017) นำเสนอ NIMA (Neural Image Assessment) ที่ใช้ตัวแบบการเรียนรู้เชิงลึกที่ผ่านการถ่ายโอนความรู้จากชุดข้อมูล ImageNet แล้วเรียนรู้เพิ่มเองด้วยรูปภาพจากฐานข้อมูล AVA (A Large-Scale Database for Aesthetic Visual Analysis) เพื่อประเมินรูปภาพในเชิงสุนทรียภาพ ร่วมกับฐานข้อมูล TID2013 (Tampere Image Database 2013) และฐานข้อมูล LIVE (LIVE In the Wild Image Quality Challenge Database) เพื่อประเมินรูปภาพในเชิงคุณภาพ เมื่อพิจารณาแท็กที่เกี่ยวข้องกับการประเมินเชิงสุนทรียภาพที่สอดคล้องกับวัตถุประสงค์ในงานวิจัยนี้ เมื่อพิจารณาในรายละเอียดพบว่า เป็นคุณลักษณะเชิงความหมาย เช่น อารมณ์ ปรัชญา หลักการ หรือแนวคิดทางทฤษฎี เป็นต้น อย่างไรก็ดี แม้ตัวแบบ NIMA จะมีประโยชน์ในการให้ความหมายรูปภาพ แต่ตัวแบบ NIMA ที่ถูกถ่ายโอนความรู้มานั้นเรียนรู้จากรูปภาพหลายหมวดหมู่ เช่น อาหาร ต้นไม้ ทิวทัศน์ สัตว์ หรือสถาปัตยกรรม เป็นต้น จึงได้รับการเรียนรู้จากคะแนนเชิงสุนทรียภาพหลายอย่างที่อาจจะไม่เกี่ยวข้อง

กับรูปภาพเชิงความหมายโดยตรง งานวิจัยนี้จึงนำแนวคิดบางส่วนมากำหนดแนวคิดที่เป็นนามธรรมของรูปภาพ เพื่อให้ตัวแบบมุ่งเน้นไปที่คุณลักษณะเชิงความหมายมากขึ้น

อย่างไรก็ดี ปัจจุบันมีการพัฒนาซอฟต์แวร์เพื่อแท็กข้อมูลรูปภาพอัตโนมัติด้วยคำศัพท์นั้นมีแบบแผนและโครงสร้างที่แน่นอน เรียกว่าแทนค่าความหมาย (Semantic representation) เพื่อทำให้คอมพิวเตอร์เข้าใจความหมายของรูปภาพ แต่การฝังแท็กอัตโนมัติลงบนภาพเพื่อสื่อความหมายนั้นมักพบปัญหา เนื่องจากมีเพียงข้อมูลหลักเท่านั้นที่คอมพิวเตอร์สามารถวิเคราะห์ได้จากภาพ แต่ข้อมูลที่เป็นองค์ประกอบอื่น ๆ ที่ช่วยสื่อความหมายของรูปภาพนั้นไม่สามารถวิเคราะห์จากรูปภาพโดยตรงได้ เป็นที่มาของการใช้เทคนิคการเรียนรู้ของเครื่องเข้ามาช่วยในการหาความหมายของภาพ ปัจจุบันพัฒนาเป็นการเรียนรู้เชิงลึกที่มีจุดเด่นในการเรียนรู้คุณลักษณะต่าง ๆ ได้อย่างหลายหลาย และใช้ตัวแปรจำนวนมากในการวิเคราะห์ เพื่อเพิ่มประสิทธิภาพในการวิเคราะห์ โดยเฉพาะโครงข่ายประสาทเทียมคอนโวลูชัน (Gkelios et al., 2021) เข้ามาเรียนรู้ความหมายของรูปภาพ จากลักษณะเด่นในขั้นตอนของการฝึกฝนที่จะทำการคัดเลือกคุณลักษณะอัตโนมัติ และทำซ้ำไปเรื่อย ๆ จนกว่าจะได้คุณลักษณะที่มีความสัมพันธ์กับผลลัพธ์มากที่สุด แต่ยังพบข้อจำกัดสำคัญคือความต้องการชุดข้อมูลจำนวนมหาศาลในการเรียนรู้ เพื่อปรับค่าน้ำหนัก และไบแอสของโมเดลให้มีค่าการสูญเสียน้อยลงและมีค่าความถูกต้องที่สูงขึ้น ในทางกลับกัน ก็มีผลกระทบที่ทำให้ต้องใช้ทรัพยากรต่าง ๆ ในการประมวลผลมากขึ้น ซึ่งหากเปรียบเทียบแล้วจะพบว่าแตกต่างกับความสามารถในการเรียนรู้ของมนุษย์ที่สามารถเรียนรู้สิ่งใหม่ ๆ ได้อย่างรวดเร็ว แม้จะมีข้อมูลที่จำกัด โดยอาศัยความรู้เดิมเป็นพื้นฐาน จึงเป็นที่มาของแนวคิดการเรียนรู้แบบ Zero-shot หมายถึงความสามารถของโมเดลในการจำแนกตัวอย่างใหม่ที่ไม่เคยเรียนรู้มาก่อน ซึ่งไม่มีอยู่ในข้อมูลการฝึกอบรม หนึ่งในตัวแบบที่ได้รับความนิยมคือ CLIP ที่เรียนรู้ความสัมพันธ์ระหว่างคำบรรยายภาพและรูปภาพที่มีความเกี่ยวข้องกันมากที่สุด จากนั้นคำนวณคะแนนความคล้ายคลึงกันระหว่างคู่รูปภาพกับป้ายกำกับ เพื่อจำแนกรูปภาพใหม่โดยใช้คะแนนความคล้ายคลึง

การประยุกต์ใช้งานตัวแบบ CLIP นั้นส่วนใหญ่จะเน้นไปที่การปรับปรุงพารามิเตอร์ของตัวแบบเพื่อใช้ในการจำแนกรูปภาพ โดยการสร้างตัวแยกประเภทชุดข้อมูลจากข้อความป้ายกำกับโดยฝังอ็อบเจกต์คลาสลงในข้อความบรรยายรูปภาพ ในขั้นตอนสุดท้าย เมื่อใส่รูปภาพเข้าไปในตัวแบบจะสร้างป้ายกำกับจากผลการพยากรณ์ข้อมูลแบบไม่มีตัวอย่าง อย่างไรก็ดี บางงานวิจัยที่นำมาประยุกต์ใช้กับการค้นคืนรูปภาพ (Bianchi et al., 2021) จากนั้น Le et al. (2022) ประยุกต์ใช้เพื่อการจำแนกรถยนต์ในมุมมองที่แตกต่างกันตามแนวคิดการเรียนรู้ด้วยตนเองระหว่างรูปภาพและข้อความ เพื่อติดป้ายกำกับรถยนต์ในรูปภาพ ประกอบด้วยสี รุ่น และทิศทาง เช่นเดียวกับงานวิจัยของ Xie et al. (2023) นำเสนอตัวแบบ RA-CLIP ที่นำมาปรับปรุงการเรียนรู้แบบคอนทราสต์แบบดั้งเดิม ซึ่งมีคำอธิบายข้อมูล เพื่อเพิ่มความแม่นยำในการจำแนกรูปภาพแบบละเอียด

จากที่กล่าวมาจะเห็นได้ว่า งานวิจัยส่วนใหญ่มุ่งใช้รูปภาพในการค้นคืนรูปภาพเชิงความหมาย เนื่องจากระบบจะวิเคราะห์คุณลักษณะรูปภาพเพื่อค้นคืนรูปภาพตามความเหมือนหรือคล้ายคลึงกับรูปภาพค้นหา จึงสามารถลดปัญหาช่องว่างความหมาย แต่ในความเป็นจริงพฤติกรรมผู้ใช้โดยมากจะค้นคืนรูปภาพโดยใช้ความสัมพันธ์ระหว่างคุณลักษณะระดับต่ำ ผสมผสานแนวคิดระดับสูงที่เป็นรูปธรรม และนามธรรมรวมเข้าด้วยกันเป็นข้อความค้นหาภาษาธรรมชาติที่อยู่ในลักษณะแนวคิดระดับสูง แต่ในทางกลับกันคุณลักษณะที่ถูกแยกอัตโนมัติโดยใช้คอมพิวเตอร์มักเป็นคุณลักษณะระดับต่ำเท่านั้น ซึ่งโดยทั่วไปไม่มีการเชื่อมโยงระหว่างคุณลักษณะระดับต่ำและแนวคิดระดับสูง ส่งผลให้ผลลัพธ์จากการค้นคืนมีประสิทธิภาพไม่สูงเท่าที่ควร อีกทั้งการค้นคืนโดยใช้รูปภาพค้นหายังไม่สะดวก และเป็นการผลักภาระให้กับผู้ใช้ในการหารูปภาพต้นฉบับอีกด้วย

งานวิจัยที่ประยุกต์ใช้เทคนิคต่าง ๆ เพื่อเพิ่มประสิทธิภาพการค้นคืนรูปภาพเชิงความหมายมีเป็นจำนวนมาก แต่ละงานวิจัยก็มีเทคนิคและวิธีการแตกต่างกันไป จากการปริทัศน์งานวิจัยที่เกี่ยวข้อง สามารถจำแนกอินพุตในการค้นหาได้ 3 รูปแบบ คือ 1) คำหลัก 2) รูปภาพค้นหา และ 3) ข้อความภาษาธรรมชาติ ร่วมกับการจำแนกเทคนิคการค้นคืนรูปภาพเชิงความหมายออกเป็น 7 กลุ่ม ได้แก่ 1) เทคนิคการค้นคืนรูปภาพหลายระดับ 2) เทคนิคการใช้แม่แบบเชิงความหมาย 3) เทคนิคการใช้ข้อความเพื่อเรียนรู้ความหมายของรูปภาพ 4) เทคนิคการตอบกลับของผู้ใช้ 5) เทคนิคการใช้ออนโทโลยี 6) เทคนิคการจัดกลุ่มด้วยการเรียนรู้ของเครื่อง และ 7) เทคนิคอื่น ๆ เช่น กราฟความสัมพันธ์ ฐานความรู้ หรือการคำนวณทางสถิติ เป็นต้น รายละเอียดดังตารางที่ 2.7

ตารางที่ 2.7 การปริทัศน์งานวิจัยที่เกี่ยวข้องกับเทคนิคการค้นคืนรูปภาพเชิงความหมาย

ผู้แต่ง	ชื่องานวิจัย	อินพุต*			เทคนิคการค้นคืนรูปภาพเชิงความหมาย**							รายละเอียดการวิจัยโดยรวม
		(1)	(2)	(3)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	
Manzoor et al. (2015)	Semantic Image Retrieval: An Ontology Based Approach	✓		✓	✓			✓				A domain-specific ontology is utilized for user-relevant retrieval, where users input concepts or images through a hybrid classification approach that incorporates shape, color, and texture features.
Kalimuthu and Krishnamurthi (2015)	Semantic-Based Facial Image-Retrieval System with Aid of Adaptive Particle Swarm Optimization and Squared Euclidian Distance		✓		✓						✓	This paper proposes a semantic-based facial image retrieval system that employs Adaptive Particle Swarm Optimization in conjunction with squared Euclidean distance to overcome limitations found in existing methods. The system operates through three key stages feature extraction, optimization, and retrieval utilizing both low-level and high-level features to improve accuracy and user-relevant results.
Mane (2016)	Semantic based image retrieval	✓			✓	✓						CBIR relies on low-level features such as color, texture, and shape to analyze, index, and retrieve images. These features are extracted directly from the visual content of images, enabling automated retrieval based on visual similarity. In contrast, semantic features involve labeling images with meaningful descriptors and assigning them to appropriate classes, allowing retrieval based on higher-level concepts rather than solely on visual characteristics.

ตารางที่ 2.7 การปริทัศน์งานวิจัยที่เกี่ยวข้องกับเทคนิคการค้นคืนรูปภาพเชิงความหมาย (ต่อ)

ผู้แต่ง	ชื่องานวิจัย	อินพุต*			เทคนิคการค้นคืนรูปภาพเชิงความหมาย**							รายละเอียดการวิจัยโดยรวม
		(1)	(2)	(3)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	
Mai et al. (2017)	Spatial-Semantic Image Search by Visual Feature Synthesis	✓	✓		✓	✓						The spatial-semantic image search allows users to specify both semantic and spatial constraints by placing concept text boxes on a 2D canvas. A convolutional neural network is trained to synthesize visual features that effectively capture these constraints, enabling more accurate and intuitive image retrieval based on both content and layout.
Gupta and Singh (2018)	A Framework for Semantic based Image Retrieval from Cyberspace by mapping low level features with high level semantics		✓			✓					✓	The proposed framework aims to develop software for complex system analysis with an emphasis on contextual information. The process involves mapping low-level image features to high-level semantic concepts, testing the effectiveness of the framework, and associating relevant metadata with existing images to enhance interpretability and retrieval accuracy.
Potapov et al. (2018)	Semantic Image Retrieval by Uniting Deep Neural Networks and Cognitive Architectures			✓		✓				✓		This semantic image retrieval approach integrates deep convolutional neural networks (DCNNs) for object detection with cognitive architectures for semantic analysis and query execution. The system, built upon YOLOv2 and OpenCog, enables users to perform queries that retrieve video frames containing specified objects arranged in defined spatial configurations.

ตารางที่ 2.7 การปริทัศน์งานวิจัยที่เกี่ยวข้องกับเทคนิคการค้นคืนรูปภาพเชิงความหมาย (ต่อ)

ผู้แต่ง	ชื่องานวิจัย	อินพุต*			เทคนิคการค้นคืนรูปภาพเชิงความหมาย**							รายละเอียดการวิจัยโดยรวม
		(1)	(2)	(3)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	
Patel and Sampat (2017)	Semantic image search using queries			✓				✓		✓		This study employs Long Short-Term Memory (LSTM) networks and conducts experiments on the Flickr8K dataset to enhance a model for generating natural language descriptions of images. By effectively capturing the semantic content of images, the proposed approach also contributes to improving image retrieval performance.
Chen, Lu and Lin (2020)	Ontology-based Dynamic Semantic Annotation for Social Image Retrieval	✓						✓				This paper presents the development of an automatic semantic image annotation model that leverages linked open data and ontology. The proposed model enables users to label images and identify underlying semantic intents, thereby improving retrieval accuracy and better addressing user information needs.
Chen, Trouve, Murakami and Fukuda (2017)	Semantic image retrieval for complex queries using a knowledge parser			✓			✓	✓			✓	This paper introduces the K-Parser, a tool designed to enhance text-based image retrieval through natural language queries. It extracts semantic representations from both image captions and user queries. Captions are stored as RDF triples, while queries are translated into SPARQL to enable precise and efficient retrieval.

ตารางที่ 2.7 การปริทัศน์งานวิจัยที่เกี่ยวข้องกับเทคนิคการค้นคืนรูปภาพเชิงความหมาย (ต่อ)

ผู้แต่ง	ชื่องานวิจัย	อินพุต*			เทคนิคการค้นคืนรูปภาพเชิงความหมาย**							รายละเอียดการวิจัยโดยรวม
		(1)	(2)	(3)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	
Anandh, Mala and Babu (2020)	Combined global and local semantic feature–based image retrieval analysis with interactive feedback		✓				✓	✓				This study identifies and applies suitable features for image retrieval, validating the results by comparison with existing systems. The use of a modified binary wavelet transform enhances similarity measurement, while interactive feedback mechanisms and efforts to address the semantic gap further improve retrieval performance.
Zhang et al., (2020)	Semantics-Guided Neural Networks for Efficient Skeleton-Based Human Action Recognition		✓							✓	✓	This paper introduces the Semantics-Guided Neural Network (SGN), which enhances feature representation by incorporating joint semantics, including joint type and frame index. The model employs hierarchical joint-level and frame-level modules to efficiently capture joint correlations and temporal dependencies across frames.
Zhao, Pei, Zheng and Tao (2020)	Online multi-core ranking model for image retrieval		✓							✓	✓	This article proposes an efficient online multi-core ranking model (OMKR), which is trained using triplet loss to minimize hard negative samples across multiple query dimensions and feature channels.

ตารางที่ 2.7 การปริทัศน์งานวิจัยที่เกี่ยวข้องกับเทคนิคการค้นคืนรูปภาพเชิงความหมาย (ต่อ)

ผู้แต่ง	ชื่องานวิจัย	อินพุต*			เทคนิคการค้นคืนรูปภาพเชิงความหมาย**							รายละเอียดการวิจัยโดยรวม
		(1)	(2)	(3)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	
Nhi, Le and Van (2022)	A Model of Semantic-Based Image Retrieval Using C-Tree and Neighbor Graph		✓					✓		✓	✓	This paper presents a semantic-based image retrieval system that leverages the combination of C-Tree and a neighbor graph to enhance retrieval accuracy. The k-NN algorithm categorizes similar images retrieved using Graph-CTree to generate visual words. A semi-automated ontology framework is then constructed from these visual words to facilitate semantic image retrieval.
Gkelios et al., (2021)	Deep convolutional features for image retrieval		✓							✓		This study introduces a procedure that utilizes the latest pre-trained CNN architectures, originally designed for image classification, to extract features for image retrieval. It investigates their effectiveness across various retrieval tasks without any optimization or exhaustive tuning. The results demonstrate that simple normalization of these pre-trained networks achieves performance comparable to state-of-the-art methods.

ตารางที่ 2.7 การปริทัศน์งานวิจัยที่เกี่ยวข้องกับเทคนิคการค้นคืนรูปภาพเชิงความหมาย (ต่อ)

ผู้แต่ง	ชื่องานวิจัย	อินพุต*			เทคนิคการค้นคืนรูปภาพเชิงความหมาย**							รายละเอียดการวิจัยโดยรวม
		(1)	(2)	(3)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	
Bianchi et al. (2021)	Contrastive Language–Image Pre-training for the Italian Language			✓			✓			✓		A multi-modal model learns from both images and text and is trained on vast amounts of English data, excelling in zero-shot tasks. This study focuses on adapting the model to the Italian language.
Le et al. (2022)	Tracked-Vehicle Retrieval by Natural Language Descriptions With Domain Adaptive Knowledge			✓				✓		✓		This paper develops a vehicle retrieval system designed to address domain bias arising from unseen scenarios and multi-view conditions. It employs CLIP for effective visual and textual representation learning and introduces a Domain Adaptive Training method that leverages labeled data to generate pseudo labels for new scenarios in the test set.
Xie et al. (2023)	RA-CLIP: Retrieval Augmented Contrastive Language-Image Pre-Training			✓						✓		This paper introduces Retrieval Augmented Contrastive Language-Image Pre-training (RA-CLIP), a method designed to enhance CLIP's performance without requiring extensive data. RA-CLIP utilizes online retrieval to augment embeddings by sampling relevant image-text pairs from a hold-out reference set, thereby enriching the learned representations.

ตารางที่ 2.7 การปริทัศน์งานวิจัยที่เกี่ยวข้องกับเทคนิคการค้นคืนรูปภาพเชิงความหมาย (ต่อ)

ผู้แต่ง	ชื่องานวิจัย	อินพุต*			เทคนิคการค้นคืนรูปภาพเชิงความหมาย**							รายละเอียดการวิจัยโดยรวม
		(1)	(2)	(3)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	
งานวิจัยนี้	The Development of a Semantic-based Images Retrieval Model using Deep Neural Network by Pre-Training Approach	✓		✓			✓			✓		Development of SBIR using CLIP Pre-Training Approach. First, contrastive training model from 400 random image and evaluation by expert for fine-tuning model, training model on flick30k dataset and custom dataset for semantic feature extraction encode to vector representation. Second, NPL search query processing and encode for vector representation. Last, vector matching between image vector and query vector by cosine similarity for retrieval to user.

* อินพุตในการค้นหาแบ่งได้ 3 รูปแบบ ประกอบด้วย

1) คำหลัก

2) รูปภาพ

3) ข้อความภาษาธรรมชาติ

** เทคนิคการค้นคืนรูปภาพเชิงความหมาย แบ่งได้ 7 กลุ่ม ประกอบด้วย

1) เทคนิคการค้นคืนรูปภาพหลายระดับ

2) เทคนิคการใช้แม่แบบเชิงความหมาย

3) เทคนิคการใช้ข้อความเพื่อเรียนรู้ความหมายของรูปภาพ

4) เทคนิคการตอบกลับของผู้ใช้

5) เทคนิคการใช้ออนโทโลยี

6) เทคนิคการจัดกลุ่มด้วยการเรียนรู้ของเครื่อง

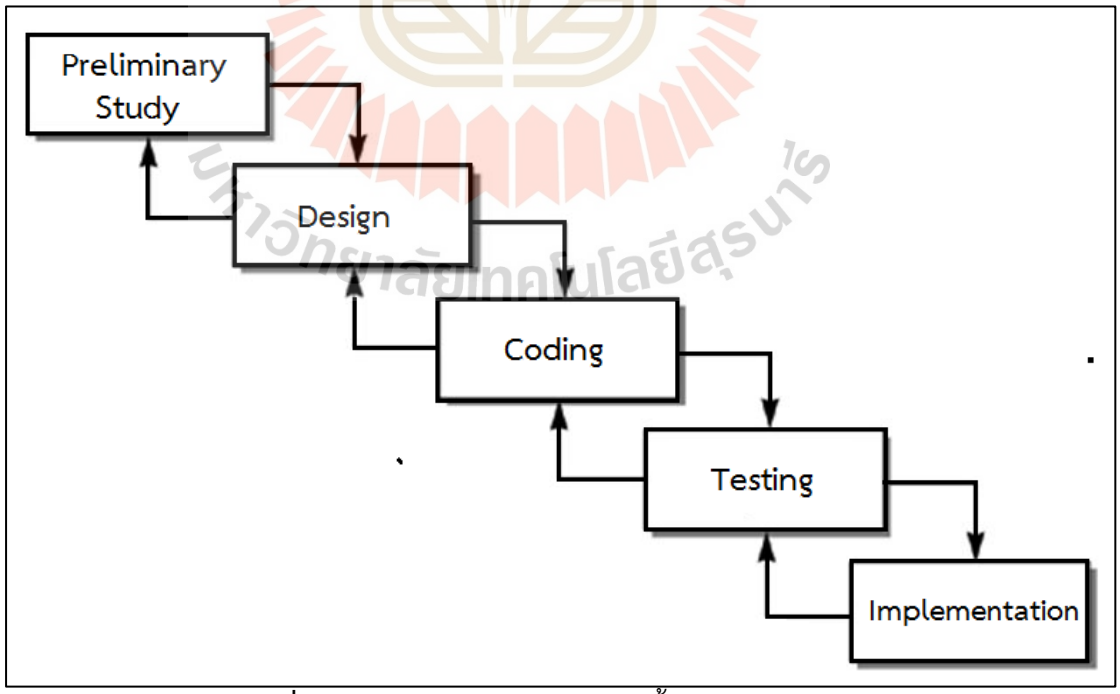
7) เทคนิคอื่น ๆ เช่น กราฟความสัมพันธ์ ฐานความรู้ หรือการคำนวณทางสถิติ เป็นต้น

บทที่ 3
วิธีดำเนินการวิจัย

การพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมาย โดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายโดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า และประเมินประสิทธิภาพแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมาย เนื้อหาในบทนี้ประกอบด้วยหัวข้อวิธีวิจัย เครื่องมือที่ใช้ในการวิจัย การรวบรวมข้อมูล และการวิเคราะห์ข้อมูล

3.1 วิธีวิจัย

งานวิจัยนี้ใช้วงจรการพัฒนาระบบ (System Development Life Cycle: SDLC) แบบน้ำตกแบบย้อนกลับได้ โดยปรับปรุงมาจาก Royce (1970) มาประยุกต์เป็นแนวทางการพัฒนาดังรูปที่ 3.1 ประกอบด้วย 5 ขั้นตอน ได้แก่ 1) การศึกษาข้อมูลเบื้องต้น 2) การออกแบบ 3) การเขียนโปรแกรม 4) การทดสอบ และ 5) การนำไปใช้



รูปที่ 3.1 วงจรการพัฒนาระบบแบบน้ำตกแบบย้อนกลับได้

3.1.1 การศึกษาข้อมูลเบื้องต้น

งานวิจัยนี้มีการศึกษาข้อมูลเบื้องต้น (Preliminary study) เพื่อกำหนดมาตรฐานสิ่งที่ส่งผลต่อการพัฒนาแบบจำลอง ประกอบด้วย

1) การศึกษาข้อมูลด้านฮาร์ดแวร์และซอฟต์แวร์ ด้วยเครื่องคอมพิวเตอร์ติดตั้งระบบปฏิบัติการไมโครซอฟต์วินโดวส์อีเลฟเวน (Microsoft Windows 11) หน่วยประมวลผลอินเทล คอร์ไอไฟว์-13400T (Intel® Core™ i5-13400T) ความเร็ว 4.4 กิกะเฮิรตซ์ (GHz) หน่วยความจำ DDR6 16 กิกะไบต์ และพื้นที่จัดเก็บข้อมูล 512 กิกะไบต์ เชื่อมต่ออินเทอร์เน็ตแบบไร้สาย ใช้งานผ่านเว็บไซต์ด้วยโปรแกรมเว็บเบราว์เซอร์ (Web browser)

2) การศึกษาข้อมูลสภาพแวดล้อมในการพัฒนาบน Google Colaboratory ซึ่งเป็นบริการแบบ Software as a Service (SaaS) โดยมีโฮสต์โปรแกรมเป็น Jupyter notebook ที่เป็นบริการอำนวยความสะดวกการพัฒนาซอฟต์แวร์บนคลาวด์ และสามารถดึงทรัพยากรจาก Google Cloud มาใช้ในการประมวลผลได้ โดยใช้บริการแบบมีค่าใช้จ่าย โดยสามารถใช้ GPU (Graphic Processing Unit) ซึ่งเป็นหน่วยประมวลผลเกี่ยวกับกราฟิกโดยเฉพาะ และสามารถใช้หน่วยความจำขนาด 25 กิกะไบต์ในโหมด High-RAM

3) การศึกษาข้อมูลสภาพแวดล้อมในการเรียนรู้ด้วยการเขียนโปรแกรมภาษาไพทอน (Python) โดยใช้ตัวแบบ CLIP ที่ถูกฝึกฝนไว้ล่วงหน้า รุ่น ViT-B/32 เป็นโมเดลฐาน (Baseline model) ในการทำงาน

3.1.2 การออกแบบ

การออกแบบ (Design) สถาปัตยกรรมแบบจำลองการค้นคืนรูปภาพเชิงความหมาย ประกอบด้วย 3 มอดูลหลัก ได้แก่ มอดูลการสร้างชุดคำอธิบายรูปภาพ มอดูลการประมวลผลข้อความค้นหา และมอดูลการจับคู่เวกเตอร์ มีรายละเอียดดังนี้

3.1.2.1 มอดูลการสร้างชุดคำอธิบายรูปภาพ (Image description set generate module) สำหรับใช้ในการค้นคืนรูปภาพจากการเปรียบเทียบความคล้ายคลึงเชิงความหมายระหว่างรูปภาพและข้อความค้นหา ประกอบด้วย 3 ขั้นตอน ได้แก่

3.1.2.1.1 การเตรียมข้อมูลก่อนการประมวลผล (Dataset pre-processing) งานวิจัยนี้กำหนดชุดข้อมูล Flickr30k ที่เป็นชุดข้อมูลรูปภาพที่รวบรวมจากเว็บไซต์ Flickr จำนวน 31,783 รูปภาพและชุดข้อมูลที่ผู้วิจัยรวบรวมเองจำนวน 92,168 รูปภาพรวมทั้งสิ้น 123,951 รูปภาพ คือประชากร จึงเลือกกลุ่มตัวอย่างที่เป็นไปตามโอกาสทางสถิติ (Probability sampling) ตามแนวคิดของ Yamane (1973) เพื่อประมาณค่าสัดส่วนประชากรที่ระดับความเชื่อมั่น 95% ดังสมการที่ 3.1

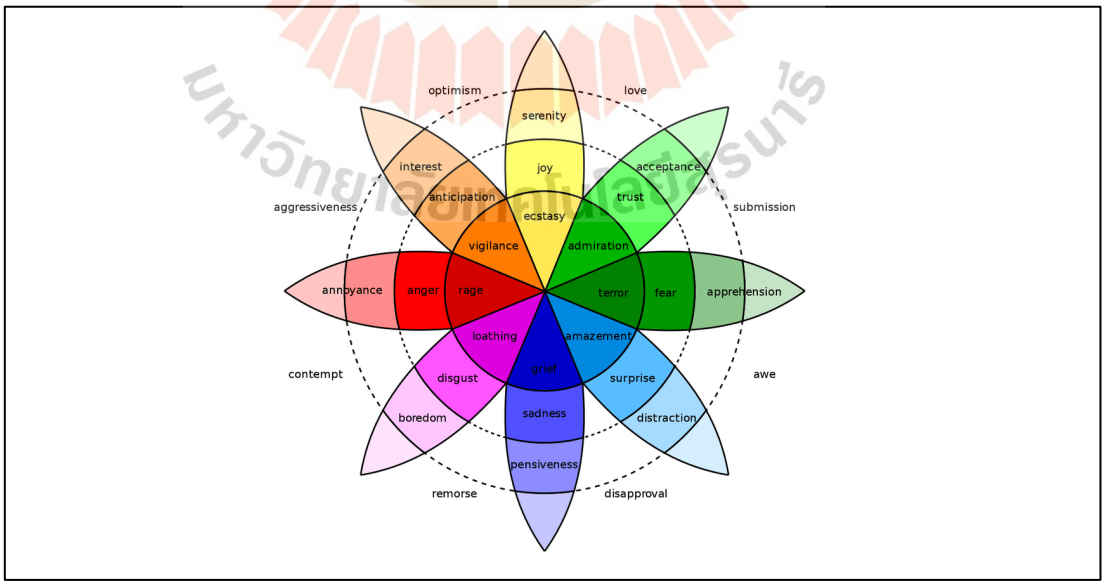
$$n = \frac{N}{1+(N)(e^2)} \quad (3.1)$$

กำหนดให้ n คือ ขนาดของรูปภาพที่สุ่มจากชุดข้อมูล

N คือ ขนาดของรูปภาพทั้งหมดในชุดข้อมูล

e คือ ค่าความคลาดเคลื่อนของการสุ่มตัวอย่างที่ยอมรับได้

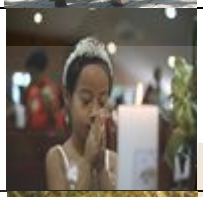
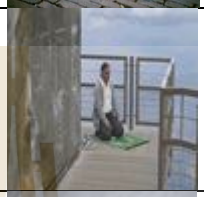
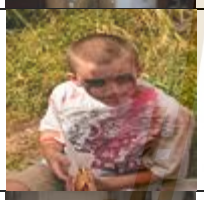
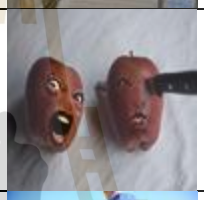
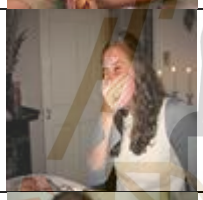
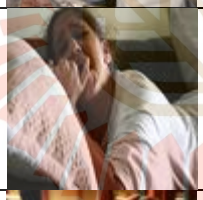
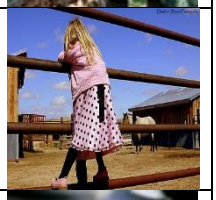
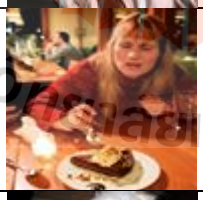
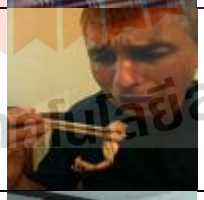
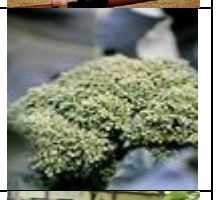
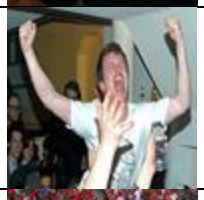
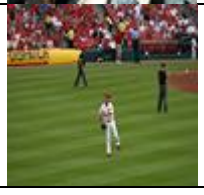
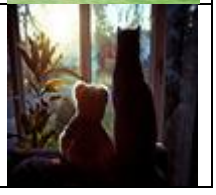
จากสมการที่ 3.1 พบว่า ขนาดของรูปภาพที่สุ่มจากชุดข้อมูลจะเท่ากับ 398.7133 รูปภาพ งานวิจัยนี้กำหนดจำนวนรูปภาพที่สุ่มจากชุดข้อมูลเท่ากับ 400 รูปภาพ แบ่งออกเป็น 8 โดเมน โดเมนละ 50 รูปภาพ รวมทั้งสิ้น 400 รูปภาพ ตามอารมณ์พื้นฐาน 8 รูปแบบ ประกอบด้วย ความรื่นเริง (Joy) ความไว้วางใจ (Trust) ความกลัว (Fear) ความประหลาดใจ (Surprise) ความเศร้า (Sadness) ความรังเกียจ (Disgust) ความโกรธ (Anger) และความคาดหวัง (Anticipation) ดังรูปที่ 3.2 แบ่งเป็นรูปภาพบนชุดข้อมูล Flickr30k จำนวน 100 รูปภาพ และอีก 300 รูปภาพจากชุดข้อมูลที่กำหนดเอง ตามสัดส่วนของรูปภาพในชุดข้อมูล ดังตารางที่ 3.1 เพื่อนำมาสร้างข้อความบรรยายรูปภาพที่อธิบายความหมายเชิงคุณภาพของรูปภาพ



รูปที่ 3.2 วงล้ออารมณ์

ที่มา: Plutchik (1980)

ตารางที่ 3.1 ตัวอย่างรูปภาพที่สุ่มจากชุดข้อมูล Flickr30k และชุดข้อมูลที่ผู้วิจัยรวบรวมเอง

โดเมน	รูปภาพ			
	1	2	...	50
1) ความรื่นเริง (Joy)			...	
2) ความวางใจ (Trust)				
3) ความกลัว (Fear)			...	
4) ความประหลาดใจ (Surprise)				
5) ความเศร้า (Sadness)			...	
6) ความรังเกียจ (Disgust)				
7) ความโกรธ (Anger)			...	
8) ความคาดหวัง (Anticipation)				

จากตารางที่ 3.1 งานวิจัยนี้ ได้สร้างข้อความบรรยายรูปภาพที่จำนวน 5 รายการต่อ 1 รูปภาพ โดยประยุกต์ใช้โมเดล LLaVA (Large Language and Vision Assistant) ซึ่งเป็นโมเดลที่ผสานความสามารถของการประมวลผลภาพและข้อความเข้าด้วยกัน เพื่อสร้างข้อความบรรยายรูปภาพเชิงความหมายภาษาอังกฤษตามอารมณ์พื้นฐานของมนุษย์ มีขั้นตอนดังนี้

1) การนำเข้าโมเดล LLaVA ที่ผ่านการฝึกฝนล่วงหน้าที่รองรับอินพุตแบบมัลติโมดัลพร้อมกันทั้งรูปภาพและข้อความ ซึ่งเป็นคุณสมบัติสำคัญที่ทำให้โมเดลสามารถสร้างข้อความบรรยายรูปภาพที่อิงจากความหมายของภาพได้อย่างมีนัยสำคัญ

2) การนำเข้าภาพและแปลงให้อยู่ในรูปแบบ Tensor ซึ่งเป็นรูปแบบที่โมเดลสามารถประมวลผลได้ โดยรวมถึงการปรับขนาดภาพ (Resize) และการทำนอมัลไลซ์ (Normalization) ตามค่าที่กำหนดไว้ของโมเดล LLaVA

3) การสร้างข้อความคำสั่ง (Prompt) ที่ใช้ร่วมกับภาพ เพื่อกระตุ้นให้โมเดลตอบสนองในบริบทที่ต้องการ เช่น การใช้คำสั่งสร้างคำบรรยายรูปภาพในโดเมนอารมณ์ของความรื่นเริง ประกอบด้วยอารมณ์ย่อย 5 ประเภท ได้แก่ “ecstasy”, “joy”, “serenity”, “optimism” และ “love” มากำหนดคำสั่ง ได้แก่ “Write a shortest ecstasy caption that goes along with this image”, “Write a shortest joy caption that goes along with this image”,“Write a shortest serenity caption that goes along with this image”, “Write a shortest optimism caption that goes along with this image” และ “Write a shortest love caption that goes along with this image” ตามลำดับ ซึ่งเป็นการกำหนดทิศทางของผลลัพธ์ให้สอดคล้องกับคุณลักษณะเชิงความหมายของภาพ โดยจะกระทำเช่นนี้ไปเรื่อย ๆ จนครบทั้ง 8 โดเมน

4) การส่งข้อมูลภาพที่แปลงเป็น Tensor และคำสั่งเข้าสู่โมเดลแบบเป็นคู่ เพื่อให้โมเดล LlaVA ทำการประมวลผลร่วมระหว่างข้อมูลสองประเภท และดำเนินการสร้างข้อความบรรยายรูปภาพออกมาโดยอัตโนมัติ ซึ่งเป็นผลลัพธ์ของกระบวนการสร้างผลลัพธ์ (Generate) ที่เกิดจากการเรียนรู้ร่วมกันระหว่างรูปภาพและคำสั่งในเชิงนามธรรม

5) การคืนผลลัพธ์ภาพพร้อมข้อความคำตอบกลับเป็นข้อความภาษาอังกฤษที่มีลักษณะเป็นคำบรรยายรูปภาพเชิงความหมาย โดยข้อความนี้จะถูกนำมาใช้เป็นตัวเลือกในกระบวนการประเมินของผู้เชี่ยวชาญ เพื่อพิจารณาว่าข้อความใดสื่อความหมายของรูปภาพได้ใกล้เคียงกับการรับรู้ของมนุษย์มากที่สุด

เมื่อเสร็จสิ้นกระบวนการสร้างข้อความบรรยายรูปภาพ 5 รายการในรูปแบบภาษาอังกฤษต่อ 1 รูปภาพ รวมทั้งสิ้น 2,000 รายการ ดังตารางที่ 3.2

ตารางที่ 3.2 ตัวอย่างรูปภาพที่สุ่มจากชุดข้อมูล Flickr30k และชุดข้อมูลที่ผู้วิจัยรวบรวมเอง ต่อข้อความบรรยายรูปภาพที่อธิบายความหมายเชิงคุณภาพของรูปภาพ

ลำดับ	ข้อความบรรยายรูปภาพที่อธิบายความหมายเชิงคุณภาพของรูปภาพ	รูปภาพ	โดเมน
1	(A) two women walking down a street, hugging each other, enjoying the moment together.		ความรื่นเริง (Joy)
	(B) two women walking arm in arm, enjoying each other's company.		
	(C) two women walking down a street, hugging each other, enjoying the moment.		
	(D) two women walking down a street, holding hands and smiling, enjoying each other's company.		
	(E) two women walking down the street, holding hands and sharing a love that transcends time.		
2	(A) a woman sits on the ground playing with a cat.		ความรื่นเริง (Joy)
	(B) a woman sitting on a brick walkway playing with a cat.		
	(C) a woman sitting on a brick street, petting a cat.		
	(D) a woman sitting on the ground petting a cat, surrounded by potted plants and a handbag.		
	(E) a woman sitting on the ground playing with a cat.		

ตารางที่ 3.2 ตัวอย่างรูปภาพที่สุ่มจากชุดข้อมูล Flickr30k และชุดข้อมูลที่ผู้วิจัยรวบรวมเอง ต่อข้อความบรรยายรูปภาพที่อธิบายความหมายเชิงคุณภาพของรูปภาพ (ต่อ)

ลำดับ	ข้อความบรรยายรูปภาพที่อธิบายความหมายเชิงคุณภาพของรูปภาพ	รูปภาพ	โดเมน
...	...	...	...
400	(A) a cat and a teddy bear sit together, watching the sunset.		ความคาดหวัง (Anticipation)
	(B) a cat and teddy bear sit together, looking out the window.		
	(C) a cat and a teddy bear sitting together by a window.		
	(D) a cat and a teddy bear sitting together, representing the bond between different species and the importance of companionship.		
	(E) a cat and a teddy bear sitting on a couch, the cat looking aggressive.		

จากตารางที่ 3.1 และตารางที่ 3.2 การกำหนดข้อความบรรยายรูปภาพจำนวน 5 ข้อความต่อ 1 รูปภาพเป็นแนวทางที่มีความเหมาะสมทั้งในเชิงวิชาการและในทางปฏิบัติ เนื่องจากภาพหนึ่งภาพสามารถตีความหมายได้หลากหลายขึ้นอยู่กับหลักการรับรู้ของมนุษย์ ซึ่งเกิดจากความรู้พื้นฐาน หรือประสบการณ์ของแต่ละบุคคลที่แตกต่างกัน ดังนั้นข้อความบรรยายรูปภาพที่หลากหลายจึงช่วยให้แบบจำลองสามารถเรียนรู้การเป็นตัวแทนเชิงความหมาย (Semantic representations) ที่ครอบคลุมและมีความยืดหยุ่นมากยิ่งขึ้นตามความหลากหลายของภาษาธรรมชาติ (Semantic diversity) (Young et al., 2014) อีกทั้งยังสอดคล้องกับแนวคิดสำคัญของโมเดล CLIP ที่พยายามเรียนรู้จากรูปภาพและข้อความบรรยายรูปภาพที่ตรงกัน ซึ่งการเพิ่มจำนวนคำบรรยายรูปภาพเป็นการเพิ่มจำนวนตัวอย่างต่ออินพุตหนึ่งหน่วย ส่งผลให้โมเดลสามารถเรียนรู้ความสัมพันธ์ข้ามโมดอล (Cross-modal alignment) ได้แม่นยำยิ่งขึ้น (Radford et al., 2021) นอกจากนี้ ชุดข้อมูลรูปภาพที่ใช้กันอย่างแพร่หลาย เช่น Flickr หรือ MS COCO ยังได้ออกแบบให้แต่ละรูปภาพมีข้อความบรรยายรูปภาพจำนวน 5 ประโยค ซึ่งแสดงถึงความพยายาม

ในการสร้างความหลากหลายเชิงภาษาศาสตร์ และเพื่อส่งเสริมให้แบบจำลองสามารถเข้าใจภาษาในระดับนามธรรมได้ดีขึ้น (Lin et al., 2014) และสามารถนำไปประยุกต์ใช้ในบริบทการค้นคืนรูปภาพได้อย่างมีประสิทธิภาพ

การเตรียมข้อมูลในงานวิจัยนี้แบ่งเป็นการเตรียมรูปภาพก่อนการประมวลผล และการเตรียมข้อความก่อนการประมวลผล มีรายละเอียดดังนี้

1) การสร้างรูปภาพใหม่จากการดัดแปลงรูปภาพต้นฉบับ (Image augmentation) (Mumuni and Mumuni, 2022) เนื่องจากประสิทธิภาพของการเรียนรู้เชิงลึกนั้นขึ้นกับปริมาณข้อมูลเป็นปัจจัยสำคัญอันดับหนึ่ง เกิดเป็นแนวคิดการปรับแต่งรูปภาพ เช่น การบิด ตัด หมุน เปลี่ยนสี ทำภาพให้มืดลงหรือสว่างขึ้น หรือใส่ตัวรบกวน (Noise) ลงไปให้ได้รูปภาพใหม่ออกมา (Xu, Yoon, Fuentes and Park, 2023) เพื่อให้แบบจำลองสามารถเรียนรู้เพื่อจดจำคุณลักษณะสำคัญของรูปภาพได้ดียิ่งขึ้น ในทางกลับกัน คุณลักษณะที่ไม่จำเป็นจะถูกนำมาเรียนรู้เพียงบางส่วน หรือไม่ถูกนำมาเรียนรู้เลย

2) การทำความสะอาดข้อความและการกำหนดมาตรฐานข้อความ (Text cleansing and text normalization) (Mitchell, 2018) แบ่งเป็นการเตรียมข้อความก่อนการประมวลผล เนื่องจากคำบรรยายที่ถูกสร้างขึ้นอาจมีรูปแบบที่ไม่ถูกต้องจากการบันทึกข้อมูล ดังรูปที่ 3.3 การทำความสะอาดข้อมูลเป็นกระบวนการตรวจสอบ สะสาง แก้ไข หรือจัดรูปแบบข้อมูลให้อยู่ในสภาพที่พร้อมใช้งาน รวมถึงลบเครื่องหมายต่าง ๆ และคัดกรองข้อมูลที่ไม่ถูกต้องหรือไม่จำเป็นออกจากชุดข้อมูล เพื่อให้ชุดข้อมูลพร้อมนำไปวิเคราะห์และใช้ประโยชน์

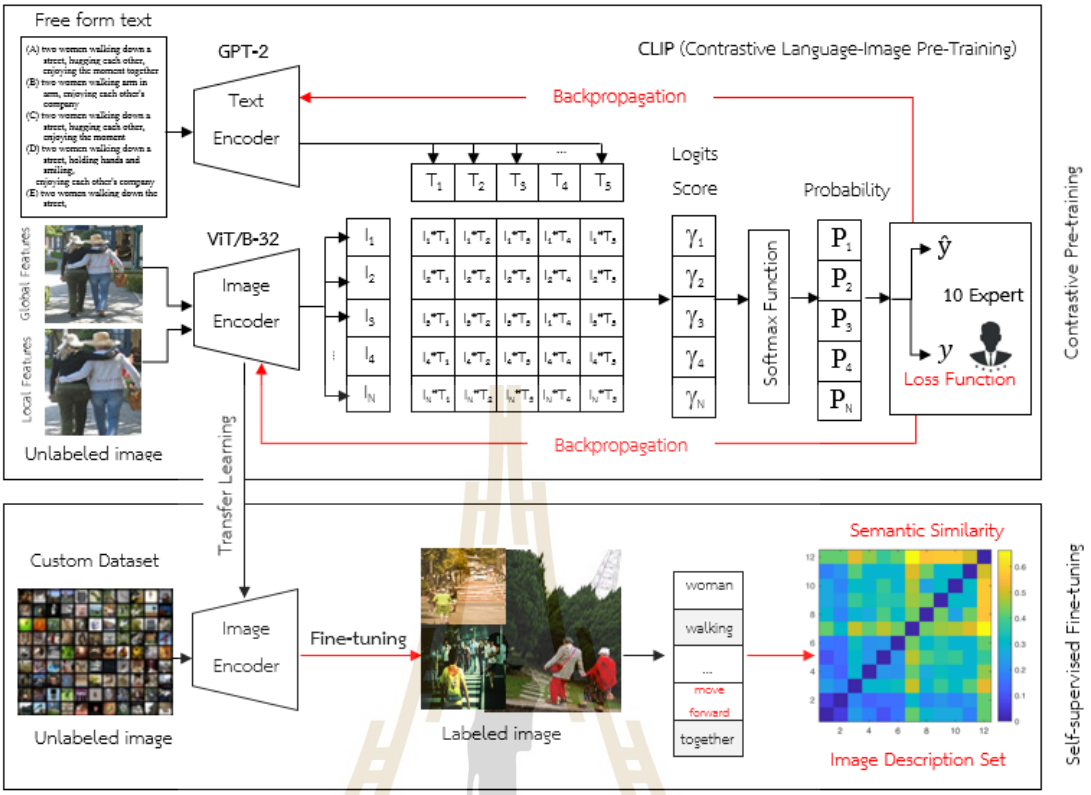
3.1.2.1.2 การฝึกฝนล่วงหน้าแบบคอนทราสต์ระหว่างรูปภาพและข้อความ (Image-text contrastive pre-training) งานวิจัยนี้ใช้ตัวแบบ CLIP เป็นโมเดลฐาน ดังรูปที่ 3.3 มีรายละเอียดดังนี้

1) การฝึกฝนตัวเข้ารหัสรูปภาพด้วยตัวแบบ ViT-B/32 ที่มีจำนวน 12 ชั้น ชั้นซ่อนเร้นจำนวน 168 ชั้น ขนาดโครงข่ายประสาทเท่ากับ 3,072 จำนวน 12 Heads และ 86 ล้านพารามิเตอร์ จากการเปรียบเทียบความสัมพันธ์ระหว่างแพตช์เพื่อคำนวณหาความสัมพันธ์กับพื้นที่ทั้งหมดในภาพ ผลลัพธ์จากการฝังรูปภาพ (Image embedding) จะแปลงเป็นเวกเตอร์คุณลักษณะรูปภาพ (Image feature vector) ขนาด 2,048

2) การฝึกฝนตัวเข้ารหัสข้อความด้วยตัวแบบ GPT-2 บนพื้นฐานสถาปัตยกรรม Transformer โดยการเรียกใช้ CLIPTokenizer ที่ใช้เข้ารหัสข้อความ และ CLIPTextModel ที่ใช้ฝังข้อความบนพื้นที่ฝังหลายรูปแบบ เพื่อเรียนรู้ความหมายโดยรวมของข้อความบรรยายรูปภาพแต่ละประโยค ผลลัพธ์จากการฝังข้อความ (Text embedding) แปลงเป็นเวกเตอร์คุณลักษณะข้อความ (Text feature vector) ขนาด 768

3) การเรียนรู้แบบคอนทราสต์บนพื้นที่การฝังหลายรูปแบบ (Contrastive learning on multi-modal embedding space) (Baltrusaitis, Ahuja and Morency, 2019) เนื่องจากเวกเตอร์มีขนาดต่างกัน จึงต้องเรียกใช้ Projection head (Weers, Shankar, Katharopoulos, Yang and Gunte, 2023) ในการเปลี่ยนขนาดเวกเตอร์บนพื้นที่การฝังให้มีมิติเท่ากับ 256 เพื่อฝึกฝนตัวเข้ารหัสรูปภาพ และตัวเข้ารหัสข้อความด้วยค่าความคล้ายคลึงโคไซน์ (Cosine similarity) (Zizka, Darena and Svoboda, 2020) จากผลคูณดอทยุคคลิดของเวกเตอร์แต่ละคู่ โดยหากเป็นเวกเตอร์คู่เดียวกันจะมีค่าความคล้ายคลึงสูง ตรงกันข้ามหากเป็นเวกเตอร์คู่ที่แตกต่างกันจะมีค่าความคล้ายคลึงต่ำด้วย

ตัวอย่างการคำนวณโดยยกตัวอย่างเป็นอาร์เรย์ 1 มิติ เพื่อง่ายต่อการทำความเข้าใจ เช่น รูปภาพ แทนด้วย I_1 ประกอบด้วยตัวเลขชุด I_1 {0.45, 0.32, 0.46, 0.23} และข้อความบรรยายรูปภาพในลำดับที่ (1) “two women walking down a street, hugging each other, enjoying the moment together” ลำดับที่ (2) “two women walking arm in arm, enjoying each other’s company” ลำดับที่ (3) “two women walking down a street, hugging each other, enjoying the moment” ลำดับที่ (4) “two women walking down a street, holding hands and smiling, enjoying each other’s company” และสุดท้ายในลำดับที่ (5) “two women walking down the street, holding hands and sharing a love that transcends time” แทนด้วย T_1, T_2, T_3, T_4 และ T_5 ดังรูปที่ 3.3



รูปที่ 3.3 แบบจำลองการค้นคืนรูปภาพเชิงความหมาย โดยฝึกฝนล่วงหน้าแบบคอนทราสต์

จากรูปที่ 3.3 อาร์เรย์ 1 มิติ แทนด้วย T_1, T_2, T_3, T_4 และ T_5 ประกอบด้วยชุดตัวเลข T_1 {0.24, 0.46, 0.44, 0.35}, T_2 {-0.66, 0.69, 0.53, 0.22}, T_3 {0.82, 0.25, 0.98, 0.99}, T_4 {0.22, 0.29, 0.35, 0.67} และ T_5 {-0.22, 0.64, 0.37, 0.78} ตามลำดับ แทนค่าในสมการที่ 3.2 มีรายละเอียดดังนี้

$$sim(I_1, T_1) = \frac{(I_{1,1}*T_{1,1})+(I_{1,2}*T_{1,2})+(I_{1,3}*T_{1,3})+(I_{1,4}*T_{1,4})}{\sqrt{I_1^2+I_2^2+I_3^2+I_4^2}*\sqrt{T_1^2+T_2^2+T_3^2+T_4^2}} \quad (3.2)$$

$$= \frac{(0.45*0.24)\ +\ (0.32*0.46)\ +}{\sqrt{0.45^2+0.32^2+0.46^2+0.23^2}\ *}$$
$$\frac{(0.46*0.44)\ +\ (\ 0.23*0.35)}{\sqrt{0.24^2+0.46^2+0.44^2+0.35^2}}$$

$$= \frac{0.54}{0.58} = 0.93$$

จากสมการที่ 3.2 พบว่า ค่าความคล้ายคลึงโคไซน์ระหว่างรูปภาพ I_1 กับข้อความ T_1 จะเท่ากับ 0.93 จากนั้นนำ I_1 คำนวณต่อกับข้อความบรรยายรูปภาพลำดับ T_2 ถึง T_5 จะได้ผลลัพธ์เท่ากับ 0.26, 0.91, 0.80 และ 0.55 ตามลำดับ

4) การคำนวณความน่าจะเป็นของป้ายกำกับ (Label probability calculator by softmax function) เนื่องจากเอาต์พุตจากการฝึกฝนล่วงหน้าแบบคอนทราสต์นั้นจะอยู่ในรูปแบบค่า Logits score ที่เกิดจากค่าความคล้ายคลึงของเวกเตอร์ จึงต้องคำนวณโดยใช้ฟังก์ชันซอฟต์แม็กซ์มาแปลงให้เป็นมาตรฐานการแจกแจงความน่าจะเป็นตามสัดส่วนของค่าเอาต์พุตแต่ละคลาส จะมีผลรวมเท่ากับ 1 ดังสมการที่ 3.3

$$f(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \tag{3.3}$$

จากรูปที่ 3.3 ตัวอย่างค่า Logits score ที่เกิดจากค่าความคล้ายคลึงระหว่างเวกเตอร์คุณลักษณะรูปภาพ I_1 กับเวกเตอร์คุณลักษณะข้อความบรรยายรูปภาพลำดับที่ T_1 ถึง T_5 มีค่าเท่ากับ 0.93, 0.26, 0.91, 0.80 และ 0.55 ตามลำดับ จากนั้นกำหนดเป็นค่าเฉลี่ยเลขคณิตถ่วงน้ำหนักของแต่ละข้อมูล โดยค่าน้ำหนักจะสะท้อนถึงความสำคัญของข้อมูลแต่ละตัวที่นำมาคำนวณ ซึ่งจะใช้ในกรณีที่ข้อมูลแต่ละตัวมีความสำคัญไม่เท่ากัน เมื่อแทนค่าเวกเตอร์อินพุตของคำบรรยายรูปภาพลำดับที่ T_1 ที่เท่ากับ 0.93 ลงในสมการที่ 3.3 มีรายละเอียดดังนี้

$$f(0.93) = \frac{\exp(0.93)}{\exp(0.93)+\exp(0.26)+\exp(0.91)+\exp(0.80)+\exp(0.55)}$$

$$= \frac{2.5345}{2.5345 + 1.2969 + 2.4843 + 2.2255 + 1.7333}$$

$$= 0.2467$$

ผลการแปลงค่าให้เป็นมาตรฐานสำหรับการแจกแจงความน่าจะเป็นตามสัดส่วนของค่าเอาต์พุตของคำบรรยายรูปภาพลำดับที่ T_1 กำหนดทศนิยม 4 ตำแหน่งจะเท่ากับ 0.2467 จากนั้นคำนวณต่อกับข้อความบรรยายรูปภาพลำดับ T_2 ถึง T_5 จะได้ผลลัพธ์เท่ากับ 0.1262, 0.2418, 0.2166 และ 0.1687 ตามลำดับ ซึ่งมีผลรวมเท่ากับ 1 ดังตารางที่ 3.3

ตารางที่ 3.3 ผลการพยากรณ์ความน่าจะเป็นของป้ายกำกับภาษาธรรมชาติ

ข้อความบรรยายรูปภาพที่อธิบายความหมายเชิงคุณภาพของรูปภาพ	รูปภาพ	ผลการพยากรณ์ความน่าจะเป็น
(A) two women walking down a street, hugging each other, enjoying the moment together		0.2467
(B) two women walking arm in arm, enjoying each other's company.		0.1262
(C) two women walking down a street, hugging each other, enjoying the moment		0.2418
(D) two women walking down a street, holding hands and smiling, enjoying each other's company		0.2166
(E) two women walking down the street, holding hands and sharing a love that transcends time		0.1687

3.1.2.1.3 การปรับแต่งค่าน้ำหนักโครงข่ายด้วยตนเอง (Manually set weight parameters) เพื่อติดป้ายกำกับอัตโนมัติสำหรับการค้นหา และจากการเรียนรู้คุณลักษณะเชิงความหมายจากความสัมพันธ์ หรือความเกี่ยวข้องระหว่างรูปภาพและข้อความบรรยายรูปภาพในรูปแบบภาษาธรรมชาติ จากนั้นจะถูกถ่ายโอนความรู้ไปยังแบบจำลองเพื่อใช้เรียนรู้ซ้ำบนชุดข้อมูลที่กำหนดเอง เพื่อติดป้ายกำกับข้อมูลให้กับรูปภาพสำหรับการค้นคืนรูปภาพเชิงความหมาย มีรายละเอียดดังนี้

1) การคำนวณค่าถ่วงน้ำหนักเฉลี่ย และสร้างค่าน้ำหนักใหม่โดยอิงผู้เชี่ยวชาญ (Weighted mean calculation and expert-guided weight reconstruction) มาปรับปรุงการอัลกอริทึมการเรียนรู้จากการหาค่าความผิดพลาดที่ได้จากการเปรียบเทียบระหว่างผลลัพธ์จากตัวแบบกับผลลัพธ์ที่ถูกประเมินความหมายโดยผู้เชี่ยวชาญด้วยการหาค่าการสูญเสียจากฟังก์ชันการสูญเสีย (Loss function) จากนั้นนำค่าการสูญเสียมาใช้กับฟังก์ชันเพิ่มประสิทธิภาพ (Optimize function) สำหรับปรับค่าพารามิเตอร์ที่ใช้ในการเรียนรู้ของตัวแบบ

กระบวนการพัฒนาโมเดลโครงข่ายประสาทเทียมนั้นการตั้งค่าน้ำหนักของพารามิเตอร์ภายในโมเดลเป็นหนึ่งในกระบวนการสำคัญที่มีผลต่อประสิทธิภาพของ

การเรียนรู้ การตั้งค่าน้ำหนักโดยอัตโนมัติโดยใช้การเรียนรู้ผ่านการทำซ้ำ (Iterative learning) อย่างไรก็ตาม ในบางบริบท การกำหนดค่าน้ำหนักโครงข่ายด้วยตนเอง (Manually set weights) (Liu and Motoda, 2012) กลับมีความจำเป็นและเหมาะสมมากกว่า โดยเฉพาะอย่างยิ่งในบริบทที่ต้องการควบคุมพฤติกรรมของโมเดลให้ตรงตามความต้องการเฉพาะหรือมีข้อจำกัดของข้อมูล

การตั้งค่าน้ำหนักด้วยตนเองมักใช้เพื่อควบคุมลักษณะเฉพาะของโมเดลตามวัตถุประสงค์ในการพัฒนา โดยงานวิจัยนี้ให้น้ำหนักมากกับฟังก์ชันการสูญเสียด้านการจำแนก (Classification loss) และให้น้ำหนักน้อยกับฟังก์ชันการสูญเสียด้านตำแหน่ง (Localization loss) เพื่อเน้นความแม่นยำของการจำแนกเชิงความหมายเหนือกว่าการหาตำแหน่ง (Zhao et al., 2019) อีกทั้งการกำหนดน้ำหนักด้วยตนเองมีบทบาทสำคัญในการแก้ไขปัญหาของข้อมูลไม่มีป้ายกำกับ ซึ่งเป็นปัญหาที่พบได้บ่อยในการวิเคราะห์ข้อมูลเชิงความหมาย เนื่องจากการจัดเตรียมชุดข้อมูลที่มีป้ายกำกับอย่างถูกต้องมีค่าใช้จ่ายสูงและใช้เวลานาน การปรับน้ำหนักด้วยตนเองจึงช่วยชี้นำกระบวนการเรียนรู้ของโมเดลให้สามารถจับความสัมพันธ์เชิงความหมายได้ดีขึ้น แม้ไม่มีตัวอย่างที่มีป้ายกำกับอย่างเพียงพอ

นอกจากนี้ ในกรณีที่มีการใช้โมเดลที่ผ่านการฝึกมาแล้วล่วงหน้า และนำมาใช้งานใหม่ในบริบทที่แตกต่างกันนั้น การกำหนดน้ำหนักให้กับชั้นที่เพิ่มเติมเข้ามาหรือการปรับค่าที่จำเป็นด้วยตนเองเป็นวิธีการที่ช่วยปรับแต่งโมเดลให้เข้ากับข้อมูลใหม่ได้อย่างรวดเร็วและมีประสิทธิภาพ โดยไม่ต้องฝึกใหม่ทั้งระบบ (Pan and Yang, 2010) ตลอดจนการตั้งค่าน้ำหนักด้วยตนเองยังมักใช้ในงานวิจัยเชิงทดลองหรือสถานการณ์ที่ต้องการการควบคุมที่แม่นยำ เช่น การศึกษาผลของการปรับค่าพารามิเตอร์ต่อการเรียนรู้ความหมายเชิงนามธรรมของโมเดล หรือการจำลองสถานการณ์เฉพาะทาง ซึ่งยากที่จะทำได้ด้วยกระบวนการเรียนรู้อัตโนมัติทั่วไป

การออกแบบกระบวนการประเมินโดยเลือกข้อความบรรยายรูปภาพที่เหมาะสมที่สุดเพียงหนึ่งรายการ จากทั้งหมด 5 ตัวเลือกที่แตกต่างกัน มีเป้าหมายเพื่อประเมินว่าข้อความใดมีความสอดคล้องกับความหมายเชิงคุณภาพของรูปภาพมากที่สุด การเลือกใช้แนวทางนี้ สะท้อนถึงหลักการที่สำคัญในการพัฒนาระบบปัญญาประดิษฐ์ที่ได้รับการยอมรับอย่างกว้างขวางในระดับสากลคือการทดสอบทัวริง (Turing test) (Turing, 1950) ที่มีวัตถุประสงค์เพื่อทดสอบว่าปัญญาประดิษฐ์สามารถสร้างปฏิสัมพันธ์ในระดับที่ใกล้เคียงกับการรับรู้ของมนุษย์เพียงใด ดังนั้นการให้ผู้ใช้ประเมินข้อความบรรยายรูปภาพจึงเป็นกลไกสำคัญในการพัฒนาตัวแบบที่สามารถค้นคืนรูปภาพเชิงความหมายที่ไม่เพียงแต่อิงความแม่นยำเชิงตัวเลข แต่ยังครอบคลุมในบริบทของการเข้าใจความหมาย ซึ่งเป็นประเด็นสำคัญของการค้นคืนรูปภาพในปัจจุบัน การเลือกใช้แนวทางนี้แทนที่จะกำหนดเพียงตัวเลือกเดียวต่อภาพ จึงสะท้อนให้เห็นถึงหลักการที่สำคัญในด้านการประเมินผลของการค้นคืนรูปภาพเชิงความหมาย ซึ่งต้องอาศัยการตัดสินแบบสัมพัทธ์ (Relative judgment)

แทนการตัดสินแบบสัมบูรณ์ (Absolute judgment) ซึ่งเป็นการเรียนรู้จากสถานการณ์จำลองที่โมเดลอาจเจอกับรูปภาพที่มีคุณลักษณะเชิงความหมายที่ซับซ้อน หรือมีความหมายใกล้เคียงกันมาก และยากที่จะจำแนกด้วยการประเมินแบบถูกหรือผิดเท่านั้น

อย่างไรก็ดี แนวคิดระดับสูงที่เป็นนามธรรมนั้นยากที่จะประเมินผลเพียงค่าความแม่นยำเท่านั้น จึงเป็นที่มาของการให้ผู้เชี่ยวชาญเข้ามาร่วมเป็นส่วนหนึ่งในการประเมินผลความหมายของแต่ละรูปภาพให้สอดคล้องกับหลักคอมพิวเตอร์วิทัศน์ โดยผู้วิจัยนำรูปภาพจากการสุ่มจาก 8 โดเมน ได้แก่ ความรื่นเริง ความวางใจ ความกลัว ความประหลาดใจ ความเศร้า ความรังเกียจ ความโกรธ และความคาดหวัง โดเมนละ 50 รูปภาพ รวมทั้งสิ้นจำนวน 400 รูปภาพ มาสร้างเป็นชุดคำถามแบบเลือกคำตอบที่ถูกต้องเพียงข้อเดียว เพื่อให้ตัวแบบเรียนรู้คุณลักษณะเชิงความหมายบนรูปภาพได้หลากหลาย และเพื่อหลีกเลี่ยงการรู้จำรสนิยม เนื่องจากหากมีผู้เชี่ยวชาญให้คะแนนเพียงท่านเดียว แล้วเอาคะแนนเหล่านั้นไปให้ตัวแบบเรียนรู้ อาจเกิดปัญหาที่ตัวแบบจะเรียนรู้รสนิยมของผู้เชี่ยวชาญท่านนั้นในการให้คะแนนรูปภาพ ซึ่งการยึดรสนิยมของผู้เชี่ยวชาญเพียงหนึ่งคนนั้นไม่เป็นผลดี เพราะผู้ใช้งานก็ย่อมมีรสนิยมต่างกันออกไป ดังนั้นจึงต้องการคะแนนที่ไม่ลำเอียง งานวิจัยนี้กำหนดให้ผู้เชี่ยวชาญทั้ง 10 ท่านให้ประเมินผลความหมายรูปภาพจากข้อความที่อธิบายความหมายเชิงคุณภาพของรูปภาพ กำหนด 1 รูปภาพต่อ 5 ข้อความบรรยาย แล้วเอาคะแนนไปถ่วงน้ำหนักเพื่อเป็นตัวแทนความหมายของรูปภาพนั้น ดังตารางที่ 3.4

ตารางที่ 3.4 ตัวอย่างการเปรียบเทียบผลการประเมินความหมายของรูปภาพจากผู้เชี่ยวชาญท่านที่ 1 กับผลการพยากรณ์ความน่าจะเป็นของป้ายกำกับภาษาธรรมชาติ

ข้อความบรรยายรูปภาพ ที่อธิบายความหมายเชิงคุณภาพของรูปภาพ	**รูปภาพ**	**ผลการพยากรณ์ความน่าจะเป็น**
☑ (A) two women walking down a street, hugging each other, enjoying the moment together		0.2467
☐ (B) two women walking arm in arm, enjoying each other's company		0.1262
☐ (C) two women walking down a street, hugging each other, enjoying the moment		0.2418

ตารางที่ 3.4 ตัวอย่างการเปรียบเทียบผลการประเมินความหมายของรูปภาพจากผู้เชี่ยวชาญท่านที่ 1 กับผลการพยากรณ์ความน่าจะเป็นของป้ายกำกับภาษาธรรมชาติ (ต่อ)

ข้อความบรรยายรูปภาพ ที่อธิบายความหมายเชิงคุณภาพของรูปภาพ	รูปภาพ	ผลการพยากรณ์ความน่าจะเป็น
☐ (A) two women walking down a street, holding hands and smiling, enjoying each other's company		0.2166
☐ (B) two women walking down the street, holding hands and sharing a love that transcends time		0.1687

จากตารางที่ 3.4 ผู้วิจัยตรวจสอบคุณภาพของข้อความบรรยายรูปภาพทั้งหมดให้อยู่ในระดับนามธรรม อีกทั้งมีความหมายโดยรวมไม่แตกต่างกันมากนักและไม่ปรากฏความสุดโต่งที่ชัดเจน ซึ่งสอดคล้องกับจุดมุ่งหมายของการศึกษาที่ต้องการจำลองสถานการณ์การรับรู้ในโลกจริง (Realistic perceptual interpretation) ที่ข้อความแต่ละชุดมีความเป็นไปได้ในการอธิบายภาพได้ใกล้เคียงกัน ดังนั้นการออกแบบให้ผู้เชี่ยวชาญเลือกเพียงหนึ่งข้อความที่เหมาะสมที่สุด จึงมิใช่การระบุสิ่งที่ถูกต้องทางวัตถุ (Objective correctness) แต่เป็นการวัดความสอดคล้องเชิงความหมายเชิงอัตวิสัย (Subjective semantic alignment) ซึ่งถือเป็นตัวชี้วัดสำคัญต่อการประเมินประสิทธิภาพของโมเดลที่เรียนรู้คุณลักษณะเชิงความหมายของรูปภาพจากข้อความที่เป็นนามธรรม

การเปรียบเทียบกับผลการพยากรณ์ป้ายกำกับจากตัวแบบด้วยการหาค่าการสูญเสียครอสเอนโทรปี เพื่อวัดค่าความน่าจะเป็นของป้ายกำกับ ว่ามีผลการพยากรณ์ใกล้กับค่าจริงที่ต้องการมากน้อยแค่ไหน (Error measurement) ด้วยฟังก์ชันการสูญเสียดังสมการที่ 3.4

$$H(p,q) = -\sum_{x\epsilon\, classes} p(x) \log q\,(x) \qquad (3.4)$$

จากตารางที่ 3.4 ตัวอย่างป้ายกำกับที่ผู้เชี่ยวชาญท่านที่ 1 เลือกคือลำดับที่ 1 ถือว่าเป็นป้ายกำกับจริง เนื่องจากมีผลการพยากรณ์ความน่าจะเป็นสูงที่สุด ดังนั้นตัวเลือกลำดับที่ 2 ถึง 5 จะเท่ากับ 0 ทั้งหมด แทนค่า p[1, 0, 0, 0, 0] ในทางกลับกัน ผลการพยากรณ์ความน่าจะเป็นของป้ายกำกับจากฟังก์ชันซอฟต์แม็กซ์ในตัวเลือกที่ 1 ถึง 5 เท่ากับ 0.2467,

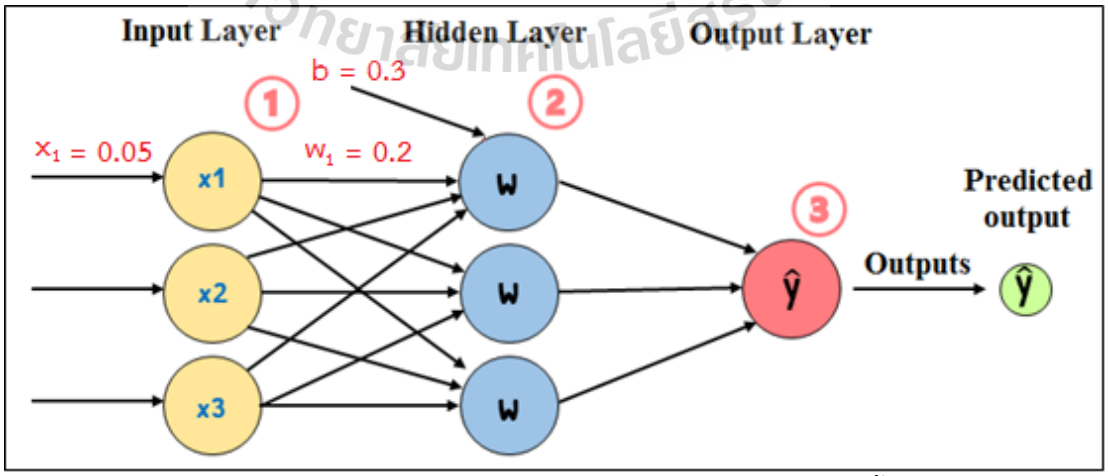
0.1262, 0.2418, 0.2166 และ 0.1687 ตามลำดับ โดยแทนค่าใน q[0.2467, 0.1262, 0.2418, 0.2166, 0.1687] เมื่อแทนค่าในสมการที่ 3.4 มีรายละเอียดดังนี้

$$H(p,q) = -[\,1 * log_2(0.2467) + 0 * log_2(0.1262) + 0 * log_2(0.2418) + 0 * log_2(0.2166) + 0 * log_2(0.1687)]$$

$$H(p,q) = 2.0192$$

ดังนั้นผลการเปรียบเทียบระหว่างผลการประเมินความหมายของรูปภาพจากผู้เชี่ยวชาญคือตัวเลือกที่ 1 ตรงกับผลการพยากรณ์ความน่าจะเป็นของป้ายกำกับที่สูงที่สุดในตัวเลือกที่ 1 ส่งผลให้ค่าครอสเอนโทรปี เท่ากับ 2.0192 โดยจะอยู่ระหว่าง 0 ถึงไม่จำกัด ซึ่ง 0 คือไม่มีค่าการสูญเสียเลย สื่อถึงความมั่นใจในผลการพยากรณ์ หากค่าครอสเอนโทรปีมีค่าน้อยแสดงถึงตัวแบบพยากรณ์ความน่าจะเป็นของป้ายกำกับถูกต้องตรงกับผลประเมินจากผู้เชี่ยวชาญได้อย่างมั่นใจ ตรงกันข้าม แม้ผลการพยากรณ์ความน่าจะเป็นของป้ายกำกับถูกต้องแต่มีค่าความน่าจะเป็นน้อย ค่าครอสเอนโทรปีก็จะมีค่ามาก เช่นเดียวกับผลพยากรณ์ความน่าจะเป็นของป้ายกำกับที่ไม่ถูกต้อง แม้จะมีค่าความน่าจะเป็นสูงก็จะทำให้ค่าครอสเอนโทรปีมากตามไปด้วย

2) การปรับแต่งค่าน้ำหนักโครงข่ายตามแนวคิดการแพร่กระจายย้อนกลับ (Update weights using backpropagation algorithm) เพื่อปรับแต่งค่าน้ำหนักให้เป็นตัวแทนของข้อมูลได้เที่ยงตรงยิ่งขึ้น อธิบายจากโครงข่ายประสาทเทียมแบบป้อนไปข้างหน้าอย่างง่าย ประกอบด้วยชั้นอินพุต ชั้นซ่อนเร้น และชั้นเอาต์พุต ดังรูปที่ 3.4



รูปที่ 3.4 โครงข่ายประสาทเทียมแบบป้อนไปข้างหน้าที่ถูกกำหนดอินพุต ค่าน้ำหนัก และไบแอส

จากรูปที่ 3.4 กำหนดให้อินพุตที่ X_1 ค่าน้ำหนักที่ W_1 และไบแอสที่ b เท่ากับ 0.05, 0.2 และ 0.3 ตามลำดับ ตัวอย่างกระบวนการปรับค่าถ่วงน้ำหนักแบบย้อนกลับ (Backward propagation) เพื่อปรับค่าน้ำหนัก W_1 จากการหาอนุพันธ์ของค่าการสูญเสีย (Loss: L) เทียบกับ W_1 หรือความชัน (Gradient) ของ Error ที่ W_1 จากตารางที่ 3.3 กำหนดให้ผลลัพธ์เป็นจริง (Actual output: y) เท่ากับ 1 และผลลัพธ์จากตัวแบบ (Predicted output: $\hat{y}$) ของข้อความบรรยายรูปภาพในลำดับที่ 1 เท่ากับ 0.2467 เมื่อแทนค่าดังสมการที่ 3.5

$$Error\_at\_W_1 = \frac{\partial L}{\partial W_1} \tag{3.5}$$

แต่เนื่องจากไม่มี W_1 ใน L จึงต้องใช้กฎลูกโซ่ (Chain rule) สำหรับการหาอนุพันธ์ของฟังก์ชันคอมโพสิต ดังสมการที่ 3.6

$$\frac{\partial L}{\partial W_1} = \frac{\partial L}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial z} * \frac{\partial z}{\partial W_1} \tag{3.6}$$

เมื่อหาค่าในแต่ละ Term ดังสมการที่ 3.7 และ 3.8

$$\frac{\partial L}{\partial \hat{y}} = \frac{\partial((y-\hat{y})^2)}{\partial \hat{y}} = -2(y-\hat{y}) \tag{3.7}$$

$$= -2(1-0.2467)$$

$$= -1.5066$$

เช่นเดียวกับ

$$\frac{\partial \hat{y}}{\partial z} = \hat{y}\,(1-\hat{y}) \tag{3.8}$$

$$= 0.2467(1-0.2467)$$

$$= 0.1858$$

เมื่อคำนวณหาค่าอินพุตที่ X_1 ดังสมการที่ 3.9

$$\frac{\partial z}{\partial w_1} = \frac{\partial(w_1x_1 + w_2x_2 + b)}{\partial w_1} = x_1 \quad (3.9)$$

$$= 0.05$$

ดังนั้น

$$Error_{at_{W_1}} = (-1.5066)(0.1858)(0.05)$$

$$= -0.014$$

เมื่อกำหนดอัตราการเรียนรู้ (Learning rate) เท่ากับ 0.5 จะสามารถปรับค่าน้ำหนักที่ W_1 ได้ดังนี้

$$Update\ W1 = W1 - Learning\_Rate * Error\_at\_W1$$

$$= 0.2 - (0.5)(-0.014)$$

$$= 0.2070$$

การประเมินความหมายของรูปภาพบนชุดคำถามแบบเลือกคำตอบที่ถูกต้องเพียงข้อเดียวโดยให้ผู้เชี่ยวชาญ 10 ท่านประเมินความหมายรูปภาพแบบเป็นอิสระต่อกัน แล้วเก็บคำตอบไว้ในชุดคำตอบสำหรับเปรียบเทียบกับผลการพยากรณ์ความน่าจะเป็นของป้ายกำกับจากตัวแบบด้วยการคำนวณหาค่าการสูญเสีย เพื่อนำไปปรับปรุงพารามิเตอร์ตามแนวคิดการแผ่กระจายย้อนหลัง

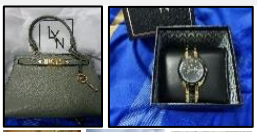
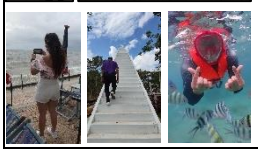
อย่างไรก็ดี การฝึกฝนโมเดลทั่วไปมักอาศัยการเรียนรู้จากข้อมูลจำนวนมากซ้ำหลายรอบ (epoch) ซึ่งหมายถึงหนึ่งรอบเต็มของการนำข้อมูลฝึกทั้งหมดเข้าสู่ระบบโมเดล อย่างไรก็ตาม ในบางกรณี โดยเฉพาะในการพัฒนาโมเดลต้นแบบ อาจมีความจำเป็นต้องกำหนดให้ปรับค่าน้ำหนักเพียงหนึ่งรอบ ซึ่งเป็นแนวทางที่ตั้งใจเพื่อให้ได้ผลลัพธ์ภายใต้เงื่อนไขเฉพาะเจาะจง โดยการปรับค่าน้ำหนักเพียงรอบเดียวเหมาะสมกับการทดลองที่ต้องการประเมินประสิทธิภาพเบื้องต้นของโมเดล โดยเฉพาะเมื่อน้ำหนักของโมเดลถูกกำหนดขึ้นด้วยตนเอง (Manual weights) ว่าที่กำหนดไว้สามารถสร้างผลลัพธ์ที่ตรงกับวัตถุประสงค์หรือไม่ โดยไม่ต้องอาศัย

การเรียนรู้ซ้ำ ซึ่งอาจเปลี่ยนแปลงได้จากโครงสร้างหรือถูกรบกวนจากจำนวน epoch ที่แตกต่างกัน (Goodfellow et al., 2016) อีกทั้งการงดใช้ epoch เป็นกลยุทธ์เพื่อลดตัวแปรที่อาจส่งผลต่อผลลัพธ์การทดลอง เช่นในการเปรียบเทียบโครงสร้างโมเดลหลากหลายแบบภายใต้ชุดข้อมูลเดียวกัน หรือต้องการให้ทุกโมเดลเริ่มต้นจากเงื่อนไขเดียวกัน และวัดผลหลังจากกระบวนการเดียวกัน อีกทั้งการปรับค่าน้ำหนักเพียงรอบเดียวสามารถลดการใช้ทรัพยากรลงได้อย่างมีนัยสำคัญ โดยเฉพาะอย่างยิ่ง เมื่อทำการทดลองเบื้องต้นในระบบที่มีข้อจำกัดของต้นทุน และเวลาในการประมวลผล

นอกจากนี้ การใช้โมเดลเพื่อการสกัดคุณลักษณะเชิงความหมาย การปรับค่าน้ำหนักเพียงรอบเดียว จะช่วยให้ผลลัพธ์สะท้อนคุณสมบัติโครงสร้างเดิมได้ชัดเจนมากกว่า จึงเป็นที่ของงานวิจัยนี้ที่กำหนดรอบการเรียนรู้เพียงรอบเดียว แต่จะดำเนินการซ้ำจนครบตามจำนวนโหนดของตัวแบบ CLIP ที่ถูกฝึกฝนล่วงหน้า

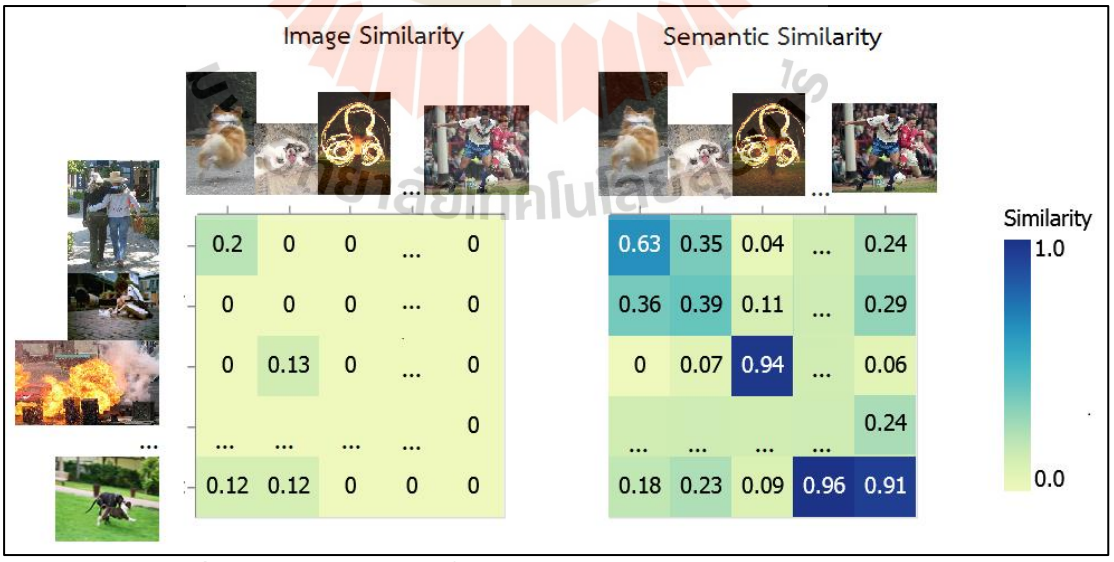
3) การเรียนรู้ซ้ำบนชุดข้อมูลที่กำหนดเอง (Retrain network on custom dataset) เนื่องจากการฝึกฝนแบบจำลองตั้งแต่เริ่มต้น (Training from scratch) จำเป็นต้องใช้ปริมาณข้อมูลจำนวนมากเพื่อให้ได้แบบจำลองที่มีความสามารถในการใช้งานทั่วไป (Generalization) นอกจากนั้น ยังมีข้อจำกัดด้านเวลาและทรัพยากรในการประมวลผล จึงทำให้เทคนิคการถ่ายโอนการเรียนรู้ (Transfer learning) ได้รับความนิยม โดยเป็นกระบวนการนำความรู้จากแบบจำลองที่ได้รับการฝึกฝนล่วงหน้ามาใช้ร่วมกับแบบจำลองใหม่ ผ่านการใช้ค่าน้ำหนักที่ได้รับการปรับแต่งด้วยตนเอง เพื่อช่วยลดระยะเวลาในการฝึกฝน ทั้งนี้ จำเป็นต้องมีการปรับค่าพารามิเตอร์ให้เหมาะสมกับคุณลักษณะเชิงความหมายลักษณะ ก่อนนำมาใช้ในการเรียนรู้ซ้ำบนชุดข้อมูลที่กำหนดเอง เพื่อติดป้ายกำกับเชิงความหมายให้กับรูปภาพจากการเรียนรู้ด้วยตนเองบนข้อมูล Flickr30k จำนวน 31,783 รูปภาพและชุดข้อมูลที่ผู้วิจัยรวบรวมเองจำนวน 92,168 รูปภาพ รวมทั้งสิ้น 123,951 รูปภาพ ดังตารางที่ 3.5

ตารางที่ 3.5 ตัวอย่างรูปภาพจากชุดข้อมูลที่ผู้วิจัยรวบรวมเอง

ภาพบุคคล	ภาพสัตว์เลี้ยง	ภาพสิ่งของ	ภาพทิวทัศน์และสิ่งปลูกสร้าง
			
			

งานวิจัยนี้ให้ความสำคัญกับประเด็นลิขสิทธิ์ของรูปภาพจึงใช้ชุดข้อมูลรูปภาพที่ผู้วิจัยรวบรวมเองมากกว่าร้อยละ 75 ในการเรียนรู้ของตัวแบบ ตลอดจนคำนึงถึงสิทธิส่วนบุคคลที่ปรากฏในภาพ จึงหลีกเลี่ยงการใช้รูปภาพที่ปรากฏใบหน้าของบุคคลหรือส่วนหนึ่งส่วนใดอย่างชัดเจนที่สามารถนำไปสู่การระบุอัตลักษณ์บุคคลได้เป็นเงื่อนไขในการเลือกรูปภาพเพื่อเป็นชุดข้อมูลที่กำหนดเอง

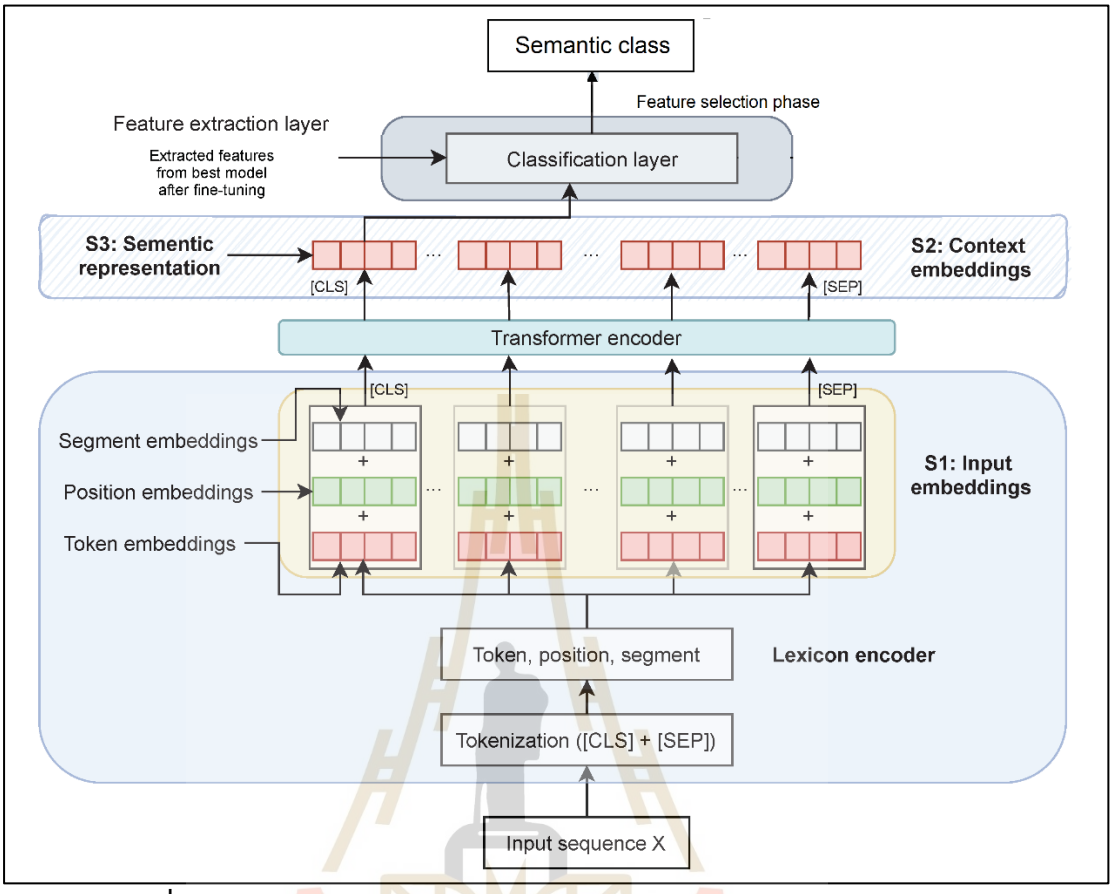
4) การเรียนรู้ความคล้ายคลึงเชิงความหมายด้วยกระบวนการเรียนรู้ด้วยตนเอง โดยบังคับให้แบบจำลองเรียนรู้การแสดงความหมายเกี่ยวกับข้อมูล เปรียบเทียบกับแนวคิดระดับสูงในรูปแบบนามธรรม เพื่ออธิบายว่าแต่ละรูปภาพซ้อนทับกันในเชิงความหมายมากน้อยเพียงใด มีค่าตั้งแต่ 0 ถึง 1 ดังรูปที่ 3.5 ตัวอย่างภาพ “ผู้หญิงสองคนกำลังเดินไปข้างหน้า” เปรียบเทียบกับภาพ “สุนัขกำลังวิ่งไปข้างหน้า” หากพิจารณาในเชิงความคล้ายคลึงของภาพ (Image similarity) จะพบว่าทั้งสองภาพมีลักษณะที่แตกต่างกันอย่างชัดเจน ไม่ว่าจะเป็นวัตถุหลักอย่างมนุษย์กับสุนัข หรือองค์ประกอบของภาพ ตลอดจนสีและพื้นหลัง ส่งผลให้ค่าความคล้ายคลึงของภาพอยู่ในระดับต่ำ ในทางตรงกันข้าม เมื่อพิจารณาด้วยความคล้ายคลึงเชิงความหมาย (Semantic similarity) จะพบว่าทั้งสองภาพมีความใกล้เคียงกันในระดับสูง เนื่องจากทั้งภาพของผู้หญิงและสุนัขต่างแสดงถึงพฤติกรรมของ “move forward” ซึ่งสามารถจัดให้อยู่ในกลุ่มเดียวกับ “moving forward” หรือ “forward motion” ด้วยเหตุนี้การค้นคืนรูปภาพที่มีประสิทธิภาพต้องสามารถประเมินให้ภาพทั้งสองมีความสัมพันธ์กันในเชิงความหมาย แม้จะต่างกันในเชิงรูปลักษณ์



รูปที่ 3.5 ตัวอย่างรูปภาพที่ผ่านการเรียนรู้ความคล้ายคลึงเชิงความหมาย

3.1.2.2 มอดูลการประมวลผลข้อความค้นหา (Query processing module) จากผู้ใช้ด้วยการประมวลผลภาษาธรรมชาติ คือการแปลความหมายของคอมพิวเตอร์โดยอาศัยระบบองค์ความรู้ (Knowledge base system) มาช่วยแปลความหมายของคำสั่งต่าง ๆ และตอบสนองต่อผู้ใช้งานด้วยการใช้ภาษาเช่นเดียวกับที่มนุษย์ใช้สื่อสารโดยไม่สนใจรูปแบบไวยากรณ์หรือโครงสร้างของภาษามากนัก งานวิจัยนี้ใช้ตัวแบบภาษา DistilBERT (Sanh et al., 2019) ที่ถูกฝึกฝนล่วงหน้าบนพื้นฐานการทำงานจากตัวแบบ BERT แต่พารามิเตอร์น้อยกว่าจึงประมวลผลได้รวดเร็วกว่าและใช้ทรัพยากรเครื่องน้อยกว่า ในขณะที่รักษาประสิทธิภาพการทำงานของ BERT ไว้ได้มากกว่าร้อยละ 97 ตามเกณฑ์มาตรฐานความเข้าใจภาษาของ GLUE ด้วยเทคนิคที่เรียกว่า Distillation เพื่อลดขนาดของตัวแบบภาษา BERT ให้มีขนาดเล็กลง โดยเก็บสิ่งที่สำคัญสำหรับการเข้าใจประโยคหรือข้อความ และลบสิ่งที่ไม่สำคัญ ซึ่งจะทำให้ตัวแบบมีขนาดเล็กลง แต่ยังคงความสามารถในการเข้าใจและสร้างความหมายของประโยคหรือข้อความได้ดีเช่นเดิม

การทำงานของตัวแบบ DistillBERT เริ่มจากนำเข้าข้อความค้นหาผ่านตัวเข้ารหัสคลังคำ (Lexicon encoder) เพื่อแยกข้อความออกเป็นโทเค็น (Token) ตำแหน่ง (Position) และตัวแบ่ง (Segment) จากนั้นนำเข้าไปยังขั้นตอนการแปลงไปเป็นเวกเตอร์ (Weers, Shankar, Katharopoulos, Yang and Gunte, 2023)โดยเพิ่มโทเค็นพิเศษ 2 ตัว คือ [CLS] เพื่อใช้แทนตำแหน่งเริ่มต้นของข้อมูล และ [SEP] สำหรับคั้นกลางระหว่างประโยค และกำหนดขอบเขตระหว่างประโยคที่ 1 และ ประโยคที่ 2 แล้วจึงผ่าน Embedding 3 ตัว ได้แก่ 1) Token embedding ที่เป็น Trainable parameters ซึ่งระหว่างการฝึกฝนตัวแบบจะค่อย ๆ ปรับตัวเองเพื่อแทนข้อมูลในระดับคำของแต่ละโทเค็น 2) Positional embedding ซึ่งเป็นเทคนิคเดียวกับที่ใช้ใน Transformer เพื่อเพิ่มข้อมูลเกี่ยวกับตำแหน่งของคำต่าง ๆ ให้กับโมเดล และ 3) Segment embedding เป็น Trainable parameters เช่นเดียวกัน แต่เพิ่มเข้ามาเพื่อช่วยให้โมเดลสามารถแยกแยะระหว่างประโยคที่ 1 และประโยคที่ 2 โดยจะนำผลลัพธ์จากทั้ง 3 Embedding มารวมกัน จากนั้นป้อนรหัสโทเค็น และสร้างมาสก์ระบุความสนใจ (Attention masks) ที่เป็นเทนเซอร์ขนาดเท่ากับรหัสโทเค็น ประกอบด้วยเลข 1 ที่แปลว่าโทเค็นในตำแหน่งนั้น ๆ จำเป็นต้องพิจารณา ตรงกันข้ามกับเลข 0 ที่จะไม่ถูกพิจารณาบนชั้นความสนใจ และจะถูกส่งไปคำนวณในชั้นต่อไป ดังรูปที่ 3.6 การประมวลผลข้อความทั้งประโยคจาก [CLS] ซึ่งถูกใส่เป็นโทเค็นแรกของทุก ๆ ประโยคก่อนเริ่มประมวลผล หากต้องการ Fine-tuning เพื่อคำนวณหาคะแนนความคล้ายคลึงเชิงความหมายระหว่างประโยคจึงต้องนำเวกเตอร์ที่เป็นผลลัพธ์ของ [CLS] ไปใส่ตัวจำแนกเพื่อคำนวณคำตอบ



รูปที่ 3.6 กระบวนการประมวลผลข้อความค้นหาด้วยตัวแบบภาษา DistilBERT

จากรูปที่ 3.6 กระบวนการ Feature extraction เป็นการฝึกฝนโมเดล 2 รอบ ในลักษณะเดียวกับการ Fine-tuning แต่ในรอบที่ 2 นั้นจะไม่ปรับปรุงผลลัพธ์ทั้งโมเดล แต่จะปรับปรุงเฉพาะในส่วนตัวจำแนกที่เพิ่มเข้ามา เนื่องจากจะนำข้อมูลอินพุตทุกตัวไปแปลงเป็นเวกเตอร์ก่อน แล้วนำฝึกฝนตัวจำแนก โดยไม่ต้องไปแก้ไข Pre-trained model ทั้งนี้การทำ Feature extraction จำเป็นต้องทดสอบสร้างคุณลักษณะหลาย ๆ รูปแบบ เพื่อเรียนรู้ความหมายโดยรวมของข้อความค้นหาแต่ละประโยค ผลลัพธ์คือเวกเตอร์ตัวแทนคุณลักษณะข้อความค้นหา (Query representation vector) ขนาด 768 สำหรับใช้ในการค้นคืนรูปภาพต่อไป

3.1.2.3 มอดูลการจับคู่เวกเตอร์ (Vector matching module) เกิดจากการคำนวณความคล้ายคลึงของเวกเตอร์โดยคำนวณออกมาเป็นตัวเลขเรียงลำดับความคล้ายคลึงจากมากไปน้อย กระบวนการทำงานเริ่มจากนำเวกเตอร์ตัวแทนรูปภาพที่ผ่านการฝัง (Image vector embedded) จากตัวเข้ารหัสรูปภาพ และเวกเตอร์ตัวแทนข้อความค้นหาที่ผ่านการฝัง (Query vector embedded) จากตัวเข้ารหัสข้อความค้นหามาคำนวณค่าความคล้ายคลึงในพื้นที่เวกเตอร์

(Vector space model) เวกเตอร์ทั้งหมดจะถูกเปรียบเทียบในรูปแบบของเส้นตรงที่ลากจากจุดโคออดิเนต (0, 0) ไปยังจุดของเวกเตอร์นั้น ๆ บนพื้นที่เวกเตอร์ มีรายละเอียดดังนี้

1) การคำนวณค่าความคล้ายคลึงของเวกเตอร์ (Vector similarity) ระหว่างรูปภาพ และข้อความค้นหา เนื่องจากเวกเตอร์ตัวแทนรูปภาพจากชุดคำอธิบายรูปภาพนั้นอาจมีจำนวนมากขึ้นอยู่กับขนาดและความซับซ้อนของรูปภาพจากการสกัดคุณลักษณะ เช่นเดียวกับเวกเตอร์ตัวแทนข้อความค้นหาจากผู้ใช้ที่มีขนาดความยาวของข้อความที่แตกต่างกัน จึงทำให้เกิดเวกเตอร์ขนาดที่แตกต่างกัน ดังนั้นงานวิจัยนี้จะแปลงขนาดของเวกเตอร์ ให้มีขนาดเท่ากันที่ 256 บนพื้นที่เวกเตอร์ เพื่อให้สามารถเปรียบเทียบกันได้

เมื่อตัวแบบสามารถเรียนรู้และเข้าใจคุณลักษณะต่าง ๆ ที่ใช้ในการพิจารณาความหมายของรูปภาพจากชุดคำอธิบายรูปภาพจากการเปรียบเทียบระหว่างเวกเตอร์คุณลักษณะรูปภาพกับเวกเตอร์คุณลักษณะข้อความค้นหาโดยตรง ซึ่งสามารถทำได้หลายวิธี เช่น Pearson correlation และ Cosine vector similarity อย่างไรก็ตามทั้งสองวิธีนี้ ไม่เหมาะกับการเปรียบเทียบข้อมูลเวกเตอร์ที่มีขนาดเล็ก ดังนั้นงานวิจัยนี้จึงเลือกใช้วิธีการวัดค่าความห่างระหว่างเวกเตอร์คุณลักษณะรูปภาพกับเวกเตอร์คุณลักษณะข้อความค้นหาแทน จากนั้นแปลงค่าความห่างไปเป็นค่าความคล้ายคลึงอีกครั้งด้วยวิธี Euclidean distance ดังสมการที่ 3.10 ตัวอย่างการแปลงค่าความแตกต่างกลับมาเป็นค่าความคล้ายคลึง สามารถอธิบายได้ดังสมการที่ 3.11

$$d(u,m) = \sqrt{\sum_{j=1}^{n}(c_j - a_j)^2} \tag{3.10}$$

$$sim(u,m) = 1 / (1 + d(u,m)) \tag{3.11}$$

กำหนดให้ n คือ จำนวนคุณลักษณะต่าง ๆ ทั้งหมดที่ถูกนำมาเปรียบเทียบระหว่างคุณลักษณะข้อความค้นหา (u) และคุณลักษณะเชิงความหมายของรูปภาพ (m, c_J)

u คือ เวกเตอร์คุณลักษณะข้อความค้นหา

m คือ เวกเตอร์คุณลักษณะเชิงความหมายของรูปภาพ

a_J คือ เกณฑ์ที่ j ของคุณลักษณะเชิงความหมายของรูปภาพ และคุณลักษณะที่ j ของคุณลักษณะข้อความค้นหา

ตัวอย่างการทำนายแบบหลายเกณฑ์ (Multi-criteria matching approach) ตัวอย่างรูปภาพจากตารางที่ 3.3 กำหนดคุณลักษณะเชิงความหมายของรูปภาพ (m) ในรูปแบบเวกเตอร์ ได้แก่ enjoy, move forward และ together เท่ากับ 30, 20 และ 10 ตามลำดับ ในทางกลับกัน คุณลักษณะข้อความค้นหา (u) คือ “moving forward” เมื่อแปลงเป็นรูปแบบเวกเตอร์ เท่ากับ 20 และ 30 ตามลำดับ เมื่อแทนค่าในสมการที่ 3.10 จะได้ผลลัพธ์เท่ากับ 10

$$d(u,m) = \sqrt{(30-20-10)^2 + (20-30)^2}$$

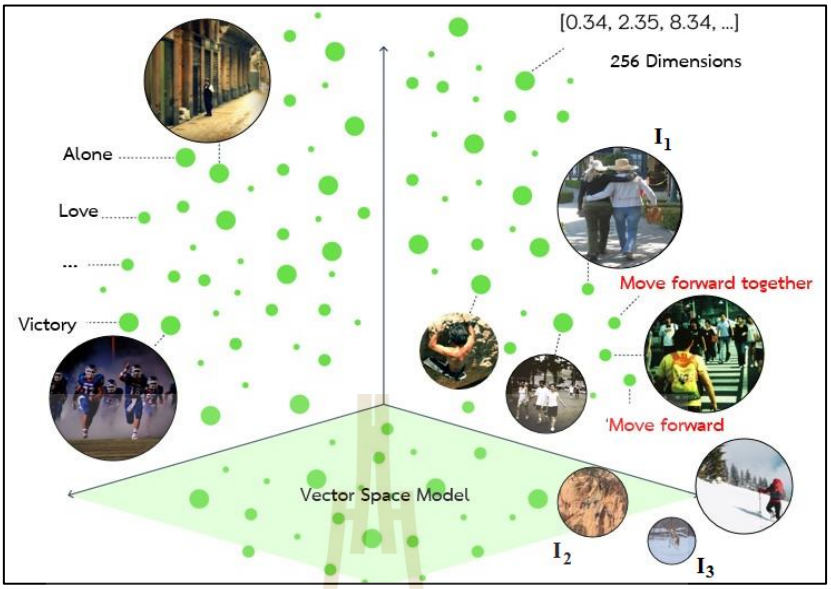
$$= 10.00$$

เมื่อแปลงค่าความแตกต่างกลับมาเป็นค่าความคล้ายคลึงดังสมการที่ 3.11 จะได้ผลลัพธ์เท่ากับ 0.0909 โดยระดับความคล้ายคลึงจะอยู่ในช่วง 0 ถึง 1 โดยหากใกล้เคียงกับ 0 หมายถึงไม่มีความคล้ายคลึงกันเลย ตรงกันข้าม หากเข้าใกล้ 1 แสดงว่ามีความคล้ายคลึงกันมากที่สุด

$$sim(u,m) = 1 / (1 + 10.00)$$

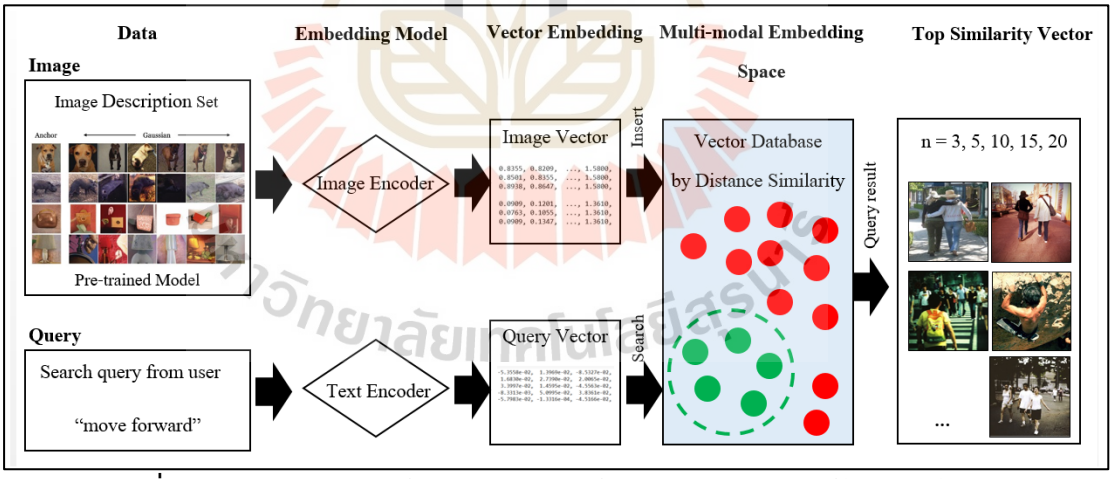
$$= 0.0909$$

ตัวอย่างการพิจารณาระยะทางระหว่างเวกเตอร์จากข้อความค้นหา “move forward” นั้นมีความสัมพันธ์โดดเด่นกับรูปภาพผู้หญิงสองคนที่เดินไปข้างหน้าในตำแหน่ง i_1 จึงส่งผลให้มีค่าน้ำหนักสูง ในทางกลับกันข้อความค้นหานั้นยังมีความสัมพันธ์กับรูปภาพนักปีนเขากำลังไต่หน้าผาขึ้นสู่ยอดเขาในตำแหน่ง i_2 และรูปภาพสุนัขกำลังวิ่งไปข้างหน้าในตำแหน่งที่ i_3 ด้วยแต่ในคนละกรณี ซึ่งไม่ใช่คุณลักษณะหลักที่มีความสำคัญเป็นศูนย์กลางในรูปภาพ เนื่องจาก “move forward” นั้นมักใช้กับการมุ่งไปข้างหน้า แต่รูปภาพ i_2 นั้นอาจคล้ายคลึงกับคำว่า “climbing” มากกว่า เช่นเดียวกับรูปภาพ i_3 อาจคล้ายคลึงกับคำว่า “running” มากกว่า ดังนั้นรูปภาพผู้หญิงสองคนที่เดินไปข้างหน้าจึงมีความคล้ายคลึงกับข้อความค้นหามากที่สุด เช่นเดียวกับรูปภาพบุคคลเดินไปข้างหน้าอื่น ๆ แต่อาจจะมีความคล้ายคลึงมากขึ้นหากข้อความค้นหาเป็น “move forward together” ที่คุณลักษณะเป็นบุคคลมากกว่า 1 เดินไปด้วยกันในภาพดังรูปที่ 3.7



รูปที่ 3.7 การพิจารณาระยะทางระหว่างเวกเตอร์ตัวแทนรูปภาพและเวกเตอร์ตัวแทนข้อความค้นหา

2) การเรียงลำดับค่าความคล้ายคลึง (Similarity sorting) จากการเปรียบเทียบระหว่างเวกเตอร์คุณลักษณะข้อความค้นหา เรียงลำดับค่าความคล้ายคลึงจากมากไปน้อย



รูปที่ 3.8 การจับคู่เวกเตอร์ ระหว่างเวกเตอร์รูปภาพ และเวกเตอร์ข้อความค้นหา

3.1.3 การเขียนโปรแกรม

การเขียนโปรแกรม (Coding) การพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายโดยประยุกต์ใช้ตัวแบบ CLIP ที่ถูกฝึกฝนล่วงหน้า เพื่อเป็นแกนหลักในการทำงานด้วยการเขียนโปรแกรมภาษาไพทอน และแสดงผลการทำงานผ่าน Google Colaboratory ที่เป็นบริการโฮสต์โปรแกรม Jupyter Notebook โดยบริษัทกูเกิล สำหรับเขียนและเรียกใช้ภาษาไพทอนจากเดิมที่ทำงานบนคอมพิวเตอร์มาให้บริการบนคลาวด์ โดยแพลตฟอร์มนี้ช่วยอำนวยความสะดวกในการเขียน เรียกใช้งาน และการทดสอบแบบอินเทอร์แอกทีฟ (Interactive) โดยไม่จำเป็นต้องติดตั้งโปรแกรมหรือใช้ทรัพยากรของคอมพิวเตอร์ส่วนบุคคล จากการใช้งาน GPU ที่จัดสรรโดย Google สำหรับการประมวลผลโมเดลขนาดใหญ่ ทำให้กระบวนการพัฒนาแบบจำลองมีความยืดหยุ่นและรวดเร็วยิ่งขึ้น อีกทั้งยังสามารถบันทึกผลลัพธ์และก่อนจะเผยแพร่เป็นสาธารณะแก่ผู้สนใจต่อไป

3.1.4 การทดสอบ

การทดสอบ (Testing) การประเมินประสิทธิภาพการค้นคืนรูปภาพดิจิทัลภายใต้ข้อความค้นหาภาษาอังกฤษ 3 เงื่อนไข ได้แก่ 1) ข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ในลักษณะป้ายกำกับของรูปภาพ 2) ข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพ และ 3) ข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ เงื่อนไขละ 10 รายการ รวมทั้งสิ้น 30 รายการด้วยมาตรวัดแบบไม่สนใจลำดับของผลลัพธ์ ประกอบด้วยค่าความแม่นยำที่ k อันดับ ค่าความครบถ้วนที่ k อันดับ และค่าประสิทธิภาพโดยรวมที่ k อันดับ บนชุดข้อมูล Flickr30k จำนวน 31,783 รูปภาพ และชุดข้อมูลที่ผู้วิจัยรวบรวมเองจำนวน 92,168 รูปภาพ รวมทั้งสิ้น 123,951 รูปภาพ โดยกำหนดจำนวนรูปภาพในการแสดงผลแต่ละครั้งเท่ากับ 20 รูปภาพ ร่วมกับมาตรวัดความเกี่ยวข้องแบบให้คะแนนด้วยค่า NDCG ที่ k อันดับ โดยผู้วิจัยคัดเลือกข้อความค้นหาภาษาอังกฤษที่มีค่าความแม่นยำที่ 5 ลำดับแรกสูงที่สุด จากข้อความค้นหาทั้ง 3 เงื่อนไข เงื่อนไขละ 100 รายการ รวมทั้งสิ้น 300 รายการ ระบุข้อความค้นหาภาษาอังกฤษผ่าน Google Colaboratory เมื่อกระบวนการค้นคืนรูปภาพเสร็จสิ้นจะแสดงผลลัพธ์จำนวน 5 รูปภาพ

ดังนั้นเพื่อเป็นการสะท้อนถึงความพร้อมของแบบจำลองต่อการใช้งานในวงกว้าง โดยเฉพาะในบริบทที่ผู้ใช้งานแต่ละคนมีเงื่อนไขที่แตกต่างกัน จึงเป็นที่มาของการให้ผู้เชี่ยวชาญเข้ามามีบทบาทในการประเมินประสิทธิภาพ เพื่อเพิ่มความน่าเชื่อถือและลดข้อผิดพลาดที่อาจเกิดขึ้นจากกระบวนการอัตโนมัติหรือข้อจำกัดของแบบจำลองในการจำแนกข้อมูลเชิงความหมายในสถานการณ์จริงได้อย่างมีประสิทธิภาพและสอดคล้องกับความต้องการของผู้ใช้งาน นอกจากนั้นการประเมินจากผู้เชี่ยวชาญหลายคนยังช่วยลดความคลาดเคลื่อนของการประเมินที่อาจเกิดขึ้นจากการพิจารณาโดยผู้เชี่ยวชาญเพียงท่านเดียว ทำให้ผลการประเมินมีความน่าเชื่อถือมากขึ้น งานวิจัยนี้สิ่งที่ผู้วิจัยจะต้องคำนึงถึงในการเลือกกลุ่มผู้เชี่ยวชาญ ตามเกณฑ์ในการคัดเลือกผู้เชี่ยวชาญ (Experts)

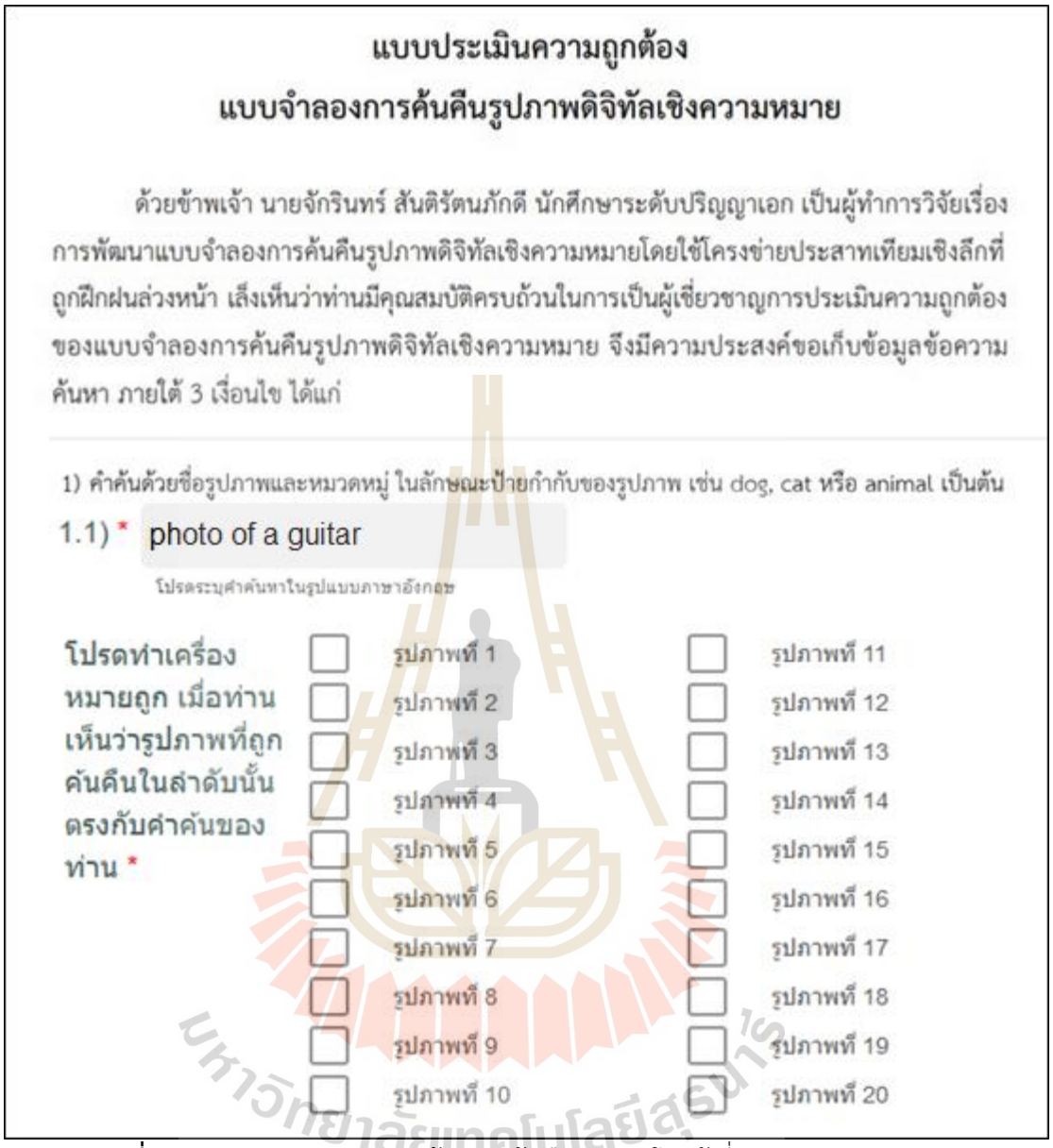
หรือผู้มีประสบการณ์ (Specialist) หรือผู้ทรงคุณวุฒิ (อรนุช ศรีสะอาด, 2538) กำหนดให้ผู้เชี่ยวชาญต้องมีคุณสมบัติดังเงื่อนไขต่อไปนี้ ได้แก่ 1) จบการศึกษาระดับปริญญาโทขึ้นไป หรือมีความรู้ความเชี่ยวชาญเฉพาะด้านอันเป็นที่ประจักษ์ 2) มีประสบการณ์การใช้งานคอมพิวเตอร์และอินเทอร์เน็ตอย่างน้อย 10 ปีขึ้นไป 3) มีประสบการณ์การใช้งานคอมพิวเตอร์และสมาร์ทโฟนอย่างน้อย 10 ปีขึ้นไป 4) มีความสามารถในการให้ข้อความค้นหาในรูปแบบภาษาอังกฤษ และ 5) สามารถประเมินความถูกต้องของการค้นคืนรูปภาพจากข้อความค้นหาภาษาอังกฤษได้ รวมถึงคำนึงถึงการป้องกันผลประโยชน์ทับซ้อน (Conflict of interest) โดยคัดเลือกผู้เชี่ยวชาญที่ไม่มีส่วนได้ส่วนเสียกับงานวิจัย ไม่เป็นผู้ที่ได้รับผลประโยชน์ส่วนตัวหรือผลประโยชน์ทางวิชาชีพ ไม่เป็นผู้มีความสัมพันธ์ส่วนตัวกับผู้วิจัย เช่น เป็นบุคคลในครอบครัวหรือผู้ใกล้ชิด ไม่เป็นผู้ได้รับผลประโยชน์ที่มีมูลค่าสูงจากโครงร่างการวิจัยในทางตรงหรือทางอ้อม และไม่เป็นผู้ร่วมวิจัยหรือให้คำปรึกษาในการวิจัย (Office of General Counsel the California State University, 2021) โดยแบ่งผู้เชี่ยวชาญออกเป็น 3 กลุ่ม ได้แก่ 1) กลุ่มผู้เชี่ยวชาญด้านคอมพิวเตอร์และเทคโนโลยี 2) กลุ่มผู้เชี่ยวชาญด้านการค้นคืนสารสนเทศ และ 3) กลุ่มผู้ใช้ทั่วไป กลุ่มละ 10 ท่าน รวมทั้งสิ้น 30 ท่าน

3.1.5 การนำไปใช้งาน

การนำไปใช้งาน (Implementation) แบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายที่ผ่านการทดสอบและประเมินผลเรียบร้อยแล้ว และเผยแพร่ในลักษณะบทความวิจัยสู่สาธารณะผ่านวารสารวิชาการที่มีมาตรฐานระดับชาติและนานาชาติ เพื่อนำเสนอแนวทางการค้นคืนสารสนเทศที่เกี่ยวข้องต่อไปในอนาคต เพื่อให้นักวิจัยและผู้สนใจสามารถเข้าถึงองค์ความรู้และแนวทางที่พัฒนาไว้ได้อย่างทั่วถึง อันจะนำไปสู่การต่อยอดองค์ความรู้ในด้านการค้นคืนสารสนเทศ การประมวลผลภาพ และปัญญาประดิษฐ์ในอนาคต ทั้งในเชิงทฤษฎีและเชิงประยุกต์ ซึ่งถือเป็นการขยายผลเชิงนโยบายและการพัฒนาองค์ความรู้ในวงกว้าง ทั้งยังสนับสนุนการนำไปประยุกต์ในสถานการณ์จริงต่อไป

3.2 เครื่องมือที่ใช้ในการวิจัย

แบบประเมินความถูกต้องการค้นคืนรูปภาพโดยผู้เชี่ยวชาญที่สร้างและประเมินผ่านแบบฟอร์มออนไลน์ด้วย Jotform โดยขอความอนุเคราะห์จากผู้เชี่ยวชาญแต่ละท่านระบุข้อความค้นหา ภายใต้ 3 เงื่อนไข ได้แก่ 1) ข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ ในลักษณะป้ายกำกับของรูปภาพ 2) ข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพ และ 3) ข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ อย่างละ 10 รายการ รวมทั้งสิ้น 30 รายการ ดังรูปที่ 3.9



รูปที่ 3.9 แบบประเมินความถูกต้องการค้นคืนรูปภาพโดยผู้เชี่ยวชาญผ่าน Jotform

3.3 การเก็บรวบรวมข้อมูล

เนื่องจากความหมายของรูปภาพจะอยู่ในลักษณะนามธรรมที่ยากจะวัดผลความถูกต้อง จึงเป็นที่มาของการให้ผู้เชี่ยวชาญ เข้ามาร่วมเป็นส่วนหนึ่งในการประเมินความถูกต้องของผลลัพธ์จากการค้นคืนตามทฤษฎีโครงสร้างทางสติปัญญาของกิลฟอร์ด (Guilford's structure of the intellect model) (Guilford, 1967) ในมิติที่ 1 กระบวนการคิด (Operation) แบบเอกนัย (Convergent production) หมายถึง ความสามารถทางสมองของบุคคลในการคิดหาคำตอบที่ดีที่สุด หรือหาเกณฑ์หรือหาบทสรุปได้สมเหตุสมผลที่สุด เมื่อได้รับหรือเห็นสิ่งเร้าในมิติที่ 2 ด้านเนื้อหา (Content)

เป็นรูปภาพ (Figural) และวัดผลความสามารถในการมองเห็น (Visual) เป็นผลในมิติที่ 3 ผลการคิด (Production) จากกระบวนการคิดที่กระทำกับเนื้อหาให้สอดคล้องกับหลักคอมพิวเตอร์วิทัศน์ที่เป็นศาสตร์ที่ใช้ในการเข้าใจว่าคอมพิวเตอร์สามารถเข้าใจรูปภาพหรือวิดีโอในรูปแบบเดียวกับการมองเห็นของมนุษย์

งานวิจัยนี้รวบรวมผลการประเมินจากผู้เชี่ยวชาญด้วยการทดสอบแบบกล่องดำ (Blackbox testing) โดยขอความอนุเคราะห์จากผู้เชี่ยวชาญแต่ละท่านระบุข้อความค้นหาผ่านแบบฟอร์มออนไลน์ด้วย Jotform จากนั้นผู้วิจัยนำเข้าข้อความค้นหาไปยังตัวแบบที่พัฒนาขึ้นบน Google Colaboratory ทีละข้อความ เมื่อกระบวนการเสร็จสิ้นจะปรากฏผลลัพธ์เป็นรูปภาพที่มีความเกี่ยวข้องกันมากที่สุดมาแสดงเป็นผลลัพธ์การค้นคืนรูปภาพ โดยให้ผู้เชี่ยวชาญประเมินความถูกต้องของรูปภาพทีละรูปภาพผ่านแบบประเมินความถูกต้องทีละรายการ จนครบทั้ง 30 รายการต่อผู้เชี่ยวชาญ 1 ท่าน และดำเนินการต่อจนครบทั้ง 30 ท่าน จะมีข้อมูลข้อความค้นหาภายใต้ 3 เงื่อนไข ได้แก่ 1) ข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ ในลักษณะป้ายกำกับของรูปภาพ 2) ข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพ และ 3) ข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ อย่างละ 300 รายการ รวมทั้งสิ้น 900 รายการ

อย่างไรก็ดี มีผู้เชี่ยวชาญจำนวนไม่น้อยที่ไม่สามารถกำหนดข้อความค้นหาที่เหมาะสม หรืออาจขาดความหลากหลายในการสร้างข้อความค้นหา เนื่องจากเป็นการค้นคืนรูปภาพที่ไม่ได้มาจากความต้องการของผู้ใช้จริง ๆ ส่งผลให้เกิดผลลัพธ์รูปภาพจำนวนมากจากค้นคืนทั้งที่ตรงกับความต้องการและไม่ตรงกับความต้องการ ดังนั้นจึงอำนวยความสะดวกให้กับผู้เชี่ยวชาญด้วยการสร้างข้อความค้นหาจาก ChatGPT (Chatbot Generative Pre-trained Transformer) (Liu et al., 2023) ให้แต่งประโยคภาษาอังกฤษเพื่อเป็นแนวทางในการสร้างข้อความค้นหา โดยกำหนด Prompt หมายถึง ชุดคำสั่งที่อินพุตเข้าไปเพื่อเริ่มการสนทนาหรือสร้างการตอบกลับจากแบบจำลอง สำหรับกำหนดบริบท และขอบเขตของเอาต์พุตที่เกิดจากการตอบสนองต่อชุดคำสั่งที่ส่งเข้าไป มีรายละเอียดดังนี้ 1) ข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ ในลักษณะป้ายกำกับของรูปภาพ กำหนด Prompt เป็น write 500 English classify classes with labels, e.g. photo of a {object} 2) ข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพ กำหนด Prompt เป็น write 500 English sentences about actions or events or activities of a {object} และ 3) ข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ กำหนด Prompt เป็น write 500 English sentences about emotions or feelings of a {object} ดังตารางที่ 3.6

ตารางที่ 3.6 ตัวอย่างข้อความค้นหาภายใต้ 3 เงื่อนไข ที่ถูกสร้างจาก ChatGPT

เงื่อนไขที่ 1	เงื่อนไขที่ 2	เงื่อนไขที่ 3
1.1) photo of a dog	2.1) the dog playing on ground	3.1) the dog wagged its tail happily
1.2) photo of a cat	2.2) the cat pounced on a mouse	3.2) the cat purred contentedly while being petted
1.3) photo of a car	2.3) the car sped down the highway	3.3) the car revved its engine eagerly
...	...	...
1.500) photo of a tradition	2.500) we enjoy our family traditions every year	3.500) family gathers to celebrate cultural traditions with love and joy

งานวิจัยนี้จึงสุ่มประโยคภาษาอังกฤษที่ครอบคลุมกับแนวคิดระดับสูงที่เป็นรูปธรรม ได้แก่ วัตถุ กิจกรรม ฉาก หรือเหตุการณ์ และที่เป็นนามธรรม ได้แก่ อารมณ์และความรู้สึก จำนวน 10 หมวดหมู่ หมวดหมู่ละ 10 รายการ เพื่อลดอคติจากการสร้างข้อความค้นหาที่อาจเอื้อต่อผลการประเมิน ประกอบด้วยหมวดคนสัตว์สิ่งของ หมวดอาหารและเครื่องดื่ม หมวดการท่องเที่ยว หมวดการศึกษา หมวดสุขภาพ หมวดธรรมชาติ หมวดเทคโนโลยี หมวดศิลปะและการบันเทิง หมวดธุรกิจและการเงิน และหมวดสังคมและวัฒนธรรม รวมทั้งสิ้น 100 รายการ ดังตารางที่ 3.7 เพื่อนำมาเป็นแนวทางให้กับผู้เชี่ยวชาญในการสร้างข้อความค้นหา ซึ่งผู้เชี่ยวชาญสามารถเลือกใช้ข้อความค้นหาดังกล่าวหรือไม่ก็ได้ หรือแก้ไขเพียงบางส่วนได้ตามความต้องการ

ตารางที่ 3.7 คำศัพท์จากการสุ่มและตัวอย่างข้อความค้นหาที่สร้างจากคำศัพท์ด้วย ChatGPT

คำศัพท์	เงื่อนไขที่ 1	เงื่อนไขที่ 2	เงื่อนไขที่ 3
1. dog	photo of a dog	the dog playing on ground	the dog wagged its tail happily
2. baby	photo of a baby	the baby sleeps peacefully in the bed	she experienced a profound sense of joy as she held her newborn baby in her arms
3. man	photo of a man	the man wears a blue suit	the man smiles warmly as he plays catch with his dog
4. dog	photo of a dog	the dog playing on ground	the dog wagged its tail happily
5. baby	photo of a baby	the baby sleeps peacefully in the bed	she experienced a profound sense of joy as she held her newborn baby in her arms
6. man	photo of a man	the man wears a blue suit	the man smiles warmly as he plays catch with his dog
7. woman	photo of a woman	the woman smiles brightly	the woman laughs joyfully while chatting with her friends
8. guitar	photo of a guitar	the man playing the guitar	the guitarist strums passionately, lost in the melody
...	...	...	...
98. religion	photo of religion	many people find comfort in their religion	her faith in her religion gives her strength during difficult times
99. friend	photo of a friend	a true friend is always there for you	a true friend is always there to lend a helping hand when needed
100. tradition	photo of tradition	we enjoy our family traditions every year	family gathers to celebrate cultural traditions with love and joy

3.4 การวิเคราะห์ข้อมูล

การวิเคราะห์ผลการค้นคืนรูปภาพแบบไบนารีด้วยโปรแกรมสำเร็จรูปทางสถิติ เพื่อประเมินประสิทธิภาพแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายในทุกมิติ ประกอบด้วย

1) มาตรวัดแบบไม่สนใจลำดับของผลลัพธ์ เป็นมาตรวัดที่แสดงถึงรูปภาพที่ถูกค้นคืนนั้นถูกต้องหรือไม่ โดยไม่คำนึงถึงตำแหน่งที่รูปภาพนั้นปรากฏในลำดับ ประกอบด้วย

1.1) ค่าความแม่นยำที่ k อันดับ (Precision@k) เป็นมาตรวัดที่ใช้ประเมินประสิทธิภาพของแบบจำลอง โดยพิจารณาจากสัดส่วนของรายการที่เกี่ยวข้อง ในลำดับผลลัพธ์สูงสุด k รายการจากการค้นคืน โดยค่าดังกล่าวมีค่าอยู่ในช่วงระหว่าง 0 ถึง 1 หากค่าเข้าใกล้ 1 แสดงถึงแบบจำลองมีความสามารถในการค้นคืนรูปภาพที่เกี่ยวข้องได้แม่นยำยิ่งขึ้น

1.2) ค่าความครบถ้วนที่ k อันดับ (Recall@k) เป็นค่าวัดความสามารถของแบบจำลองโมเดลในการค้นคืนรูปภาพที่เกี่ยวข้องทั้งหมดออกมา โดยพิจารณาว่าในจำนวนรายการที่เกี่ยวข้องทั้งหมด มีสัดส่วนเท่าใดที่ปรากฏในผลลัพธ์ k อันดับแรก

1.3) ค่าประสิทธิภาพโดยรวมที่ k อันดับ (F-measure@k) เป็นค่าเฉลี่ยฮาร์โมนิกของที่ใช้เพื่อประเมินความสมดุลระหว่างค่าความแม่นยำและค่าความครบถ้วน

2) มาตรวัดความเกี่ยวข้องแบบให้คะแนนด้วยค่า NDCG ที่ k อันดับ ที่ให้ความสำคัญกับลำดับของรายการที่เกี่ยวข้องที่ปรากฏในลำดับต้น ๆ จะได้รับน้ำหนักมากกว่ารายการที่อยู่ในลำดับหลัง โดยค่าจะถูกปรับให้อยู่ในช่วง 0 ถึง 1 เพื่อความสะดวกในการเปรียบเทียบระหว่างการค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิมกับการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนัก

การตีความผลการวิเคราะห์ข้อมูลในรูปของร้อยละให้ง่ายต่อการทำความเข้าใจ งานวิจัยนี้ใช้เกณฑ์การแปลผลตามระดับคะแนน 5 ระดับตามแนวทางของณัฏฐ์ธนัน ทองดี (2553) โดยพิจารณาความเหมาะสมในการวิเคราะห์ข้อมูลทางสถิติในงานวิจัยเชิงปริมาณ ดังตารางที่ 3.8

ตารางที่ 3.8 การตีความผลการวิเคราะห์ข้อมูลในรูปของร้อยละ ตามระดับคะแนน 5 ระดับ

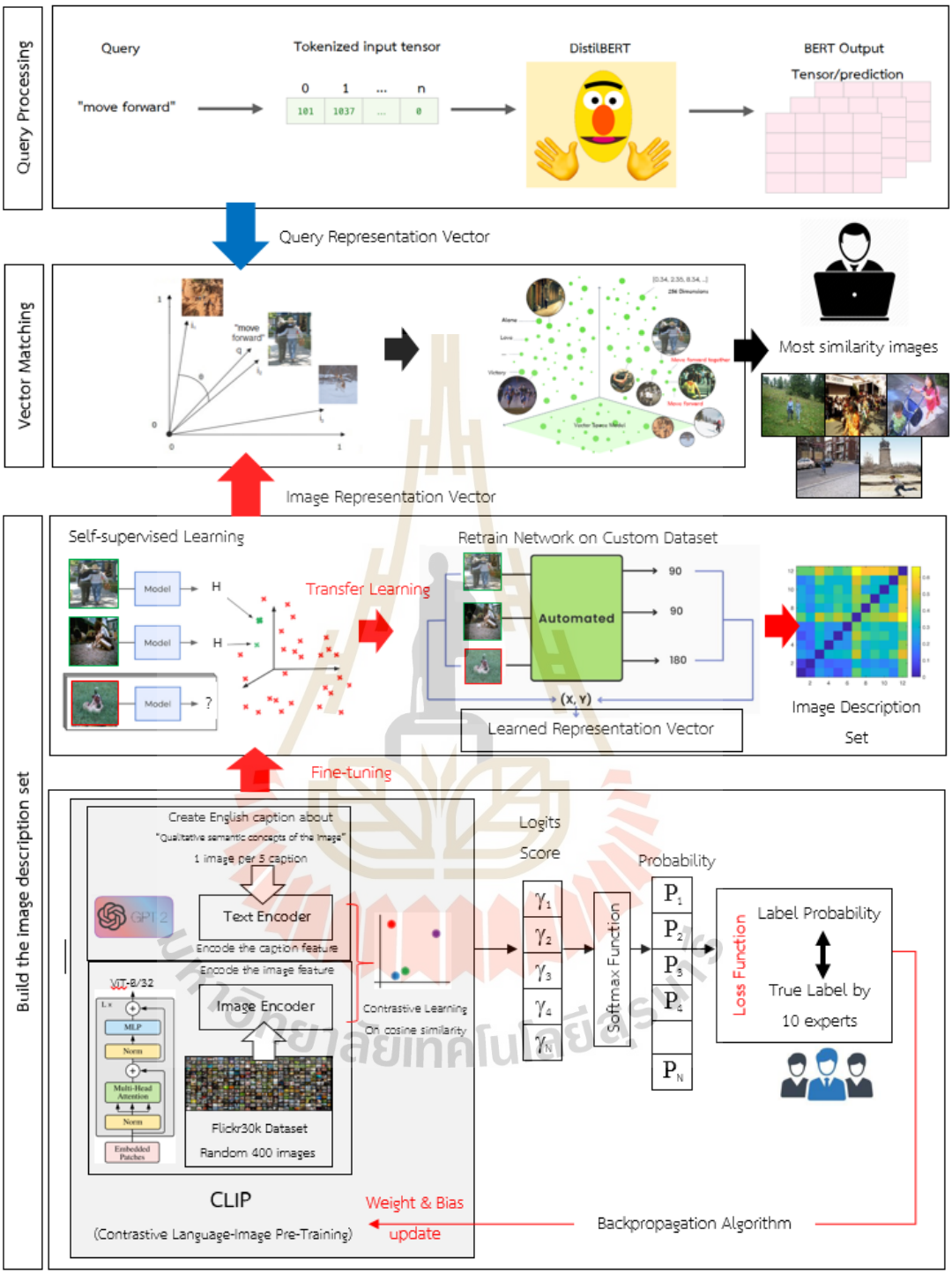
ช่วงร้อยละ (%)	ระดับการแปลผล
81 – 100	ดีมาก
61 – 80	ดี
41 – 60	ปานกลาง
21 – 40	น้อย
0 – 20	น้อยที่สุด

บทที่ 4
ผลการวิจัยและการอภิปรายผล

การพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมาย โดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า แบ่งผลการวิจัยและการอภิปรายผลออกเป็น 2 ส่วน ตามวัตถุประสงค์ ได้แก่ 1) เพื่อออกแบบและพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายโดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า และ 2) เพื่อประเมินประสิทธิภาพแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมาย มีรายละเอียดดังนี้

4.1 ผลการพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมาย โดยประยุกต์โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า

งานวิจัยนี้สุ่มรูปภาพจำนวน 400 รูปภาพแบ่งเป็น 100 รูปภาพจากชุดข้อมูล Flickr30k และอีก 300 รูปภาพจากชุดข้อมูลกำหนดเอง ในการเรียนรู้คุณลักษณะรูปภาพจากข้อความภาษาธรรมชาติด้วยตัวแบบ CLIP ที่ถูกฝึกฝนล่วงหน้า เพื่อพยากรณ์ป้ายกำกับภาษาธรรมชาติกำหนดให้ 1 รูปภาพต่อ 5 ข้อความที่ผ่านการแจกแจงความน่าจะเป็นตามสัดส่วนของค่าเอาต์พุตจากฟังก์ชันซอฟต์แม็กซ์ แล้วเปรียบเทียบกับผลการประเมินความหมายของรูปภาพจากผู้เชี่ยวชาญ 10 ท่าน ผ่านฟังก์ชันสูญเสีย จากนั้นนำไปปรับค่าน้ำหนักการเรียนรู้ของโครงข่ายประสาทเทียมตามแนวคิดการแพร่กระจายย้อนกลับ เพื่อให้ตัวแบบสามารถจำแนกคุณลักษณะเชิงความหมายได้ใกล้เคียงกับมนุษย์มากที่สุด แล้วจึงถ่ายโอนความรู้เพื่อเรียนรู้ซ้ำอีกครั้งบนชุดข้อมูลที่กำหนดเอง และแปลงเป็นเวกเตอร์จัดเก็บไว้ในชุดคำอธิบายรูปภาพสำหรับเปรียบเทียบความคล้ายคลึงกับเวกเตอร์ข้อความค้นหาจากผู้ใช้ ก่อนจะเรียงลำดับตามความเกี่ยวข้องและนำเสนอเป็นผลลัพธ์การค้นคืนรูปภาพแก่ผู้ใช้ ดังรูปที่ 4.1 ประกอบด้วย 3 มอดูล ได้แก่ มอดูลการสร้างชุดคำอธิบายรูปภาพ มอดูลการประมวลผลข้อความค้นหา และมอดูลการจับคู่เวกเตอร์ มีรายละเอียดดังนี้

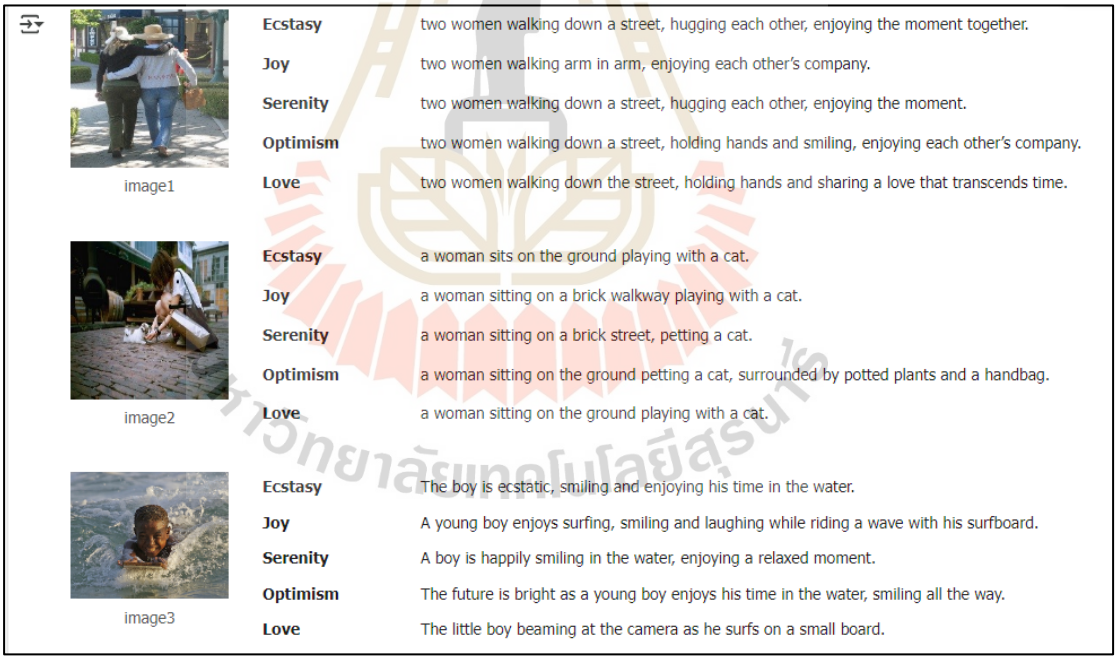


รูปที่ 4.1 การออกแบบและพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมาย โดยประยุกต์โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า

4.1.1 ผลการพัฒนามอดูลการสร้างชุดคำอธิบายรูปภาพ

การพัฒนามอดูลการสร้างชุดคำอธิบายรูปภาพสำหรับใช้ในการค้นคืนรูปภาพจากการเปรียบเทียบความคล้ายคลึงเชิงความหมาย ประกอบด้วย 3 มอดูล ได้แก่

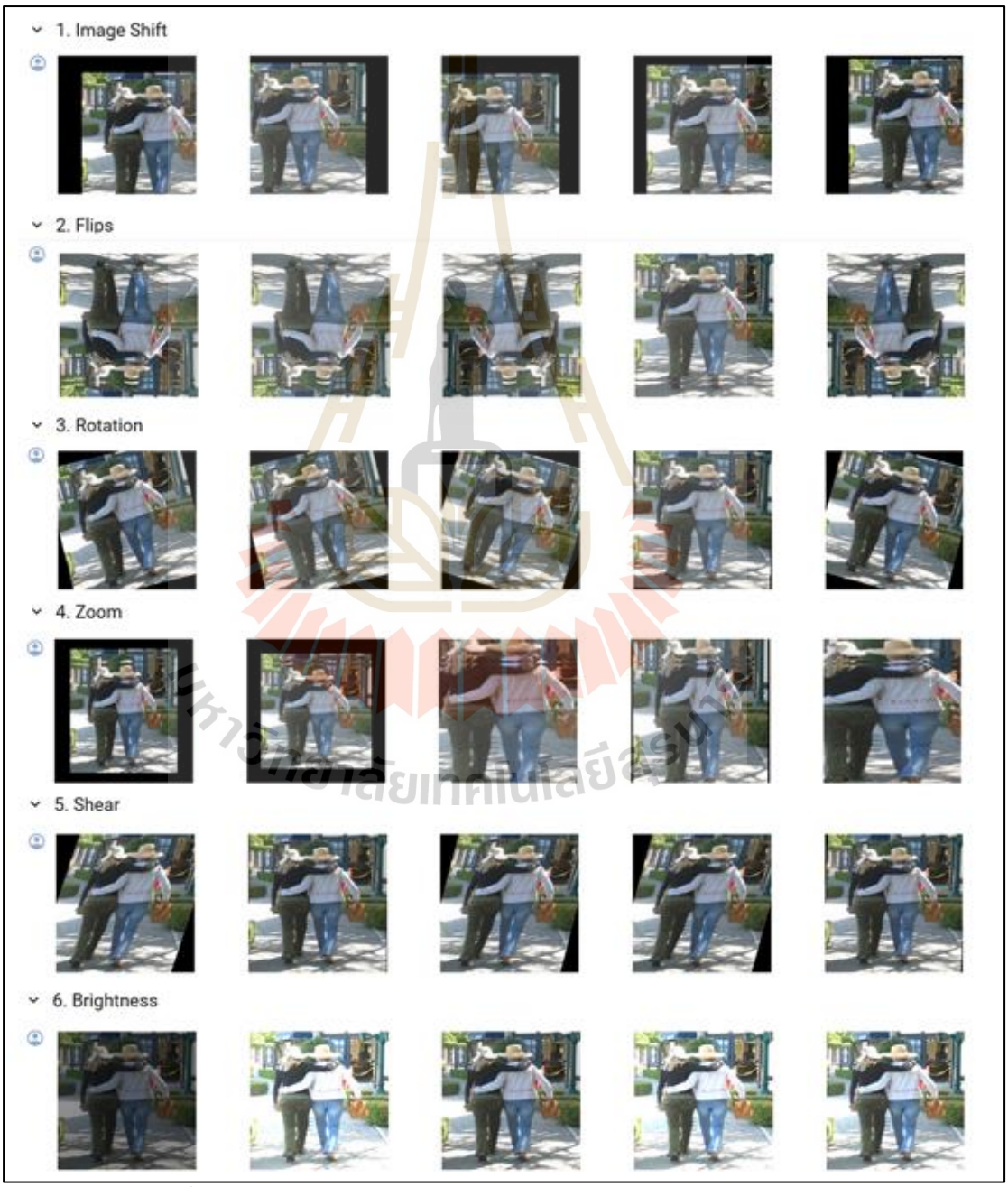
4.1.1.1 ผลการเตรียมข้อมูลก่อนการประมวลผล งานวิจัยนี้สุ่มรูปภาพจำนวน 400 รูปภาพ สำหรับฝึกฝนให้คอมพิวเตอร์เรียนรู้จากรูปภาพและข้อความภาษาธรรมชาติ แบ่งเป็นรูปภาพบนชุดข้อมูล Flickr30k จำนวน 100 รูปภาพ และอีก 300 รูปภาพจากชุดข้อมูลที่ผู้วิจัยรวบรวมเองตามสัดส่วนของจำนวนรูปภาพที่ใช้ในงานวิจัย มาสร้างข้อความบรรยายรูปภาพในรูปแบบภาษาอังกฤษ กำหนดให้ 1 รูปภาพต่อ 5 ข้อความบรรยายรูปภาพ ซึ่งเป็นผลลัพธ์จากโมเดล LLaVA ที่ได้รับการตั้งค่าคำสั่ง (Prompt) โดยใช้คีย์เวิร์ดเกี่ยวกับอารมณ์ (Emotion-related keywords) ที่แตกต่างกันในแต่ละโดเมน เพื่อสร้างคำบรรยายรูปภาพสำหรับใช้ประเมินว่าข้อความใดมีความสอดคล้องกับความหมายเชิงคุณภาพของรูปภาพมากที่สุดในระดับที่ใกล้เคียงกับการรับรู้ของมนุษย์ ดังรูปที่ 4.2



รูปที่ 4.2 ผลการสร้างข้อความบรรยายรูปภาพในรูปแบบภาษาอังกฤษจากโมเดล LLaVA ที่ได้รับการตั้งค่าคำสั่ง โดยใช้คีย์เวิร์ดเกี่ยวกับอารมณ์

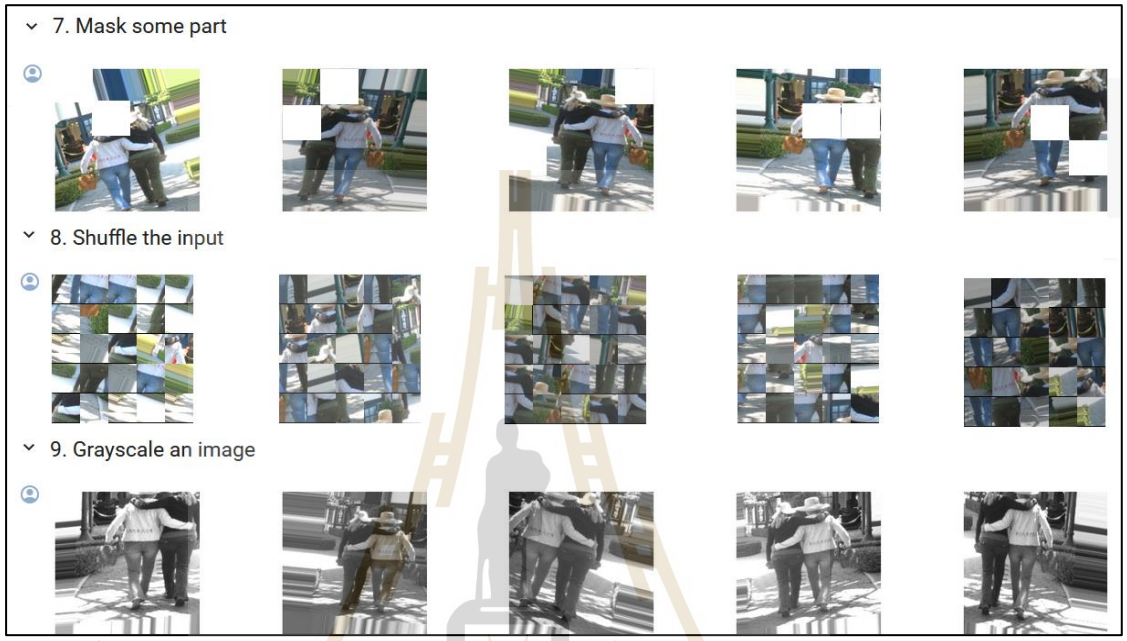
เมื่อนำรูปภาพจำนวน 400 รูปภาพจากการสุ่ม มาสร้างข้อความบรรยายรูปภาพ กำหนดจำนวน 5 ข้อความต่อหนึ่งรูปภาพ รวมทั้งสิ้น 2,000 รายการผ่านการเตรียมข้อมูลก่อนการประมวลผล มีรายละเอียดดังนี้

1) ผลการสร้างรูปภาพใหม่จากการดัดแปลงรูปภาพต้นฉบับ เพื่อให้แบบจำลองจดจำสิ่งที่ไม่จำเป็นได้ยากขึ้น และจดจำคุณลักษณะสำคัญของรูปภาพได้ดียิ่งขึ้น ประกอบด้วยการเลื่อนภาพ (Image shift) การพลิกภาพ (Image flips) การหมุนภาพ (Image rotate) การขยายและย่อภาพ (Image zoom) การบิดภาพ (Image shear) การปรับความสว่างภาพ (Image brightness) กำหนดรูปแบบละ 5 รูปภาพ ดังรูปที่ 4.3



รูปที่ 4.3 ตัวอย่างการสร้างรูปภาพใหม่จากการดัดแปลงรูปภาพต้นฉบับ

จากนั้นผสมผสานกันเพื่อสร้างรูปภาพใหม่ด้วยการมาสก์ปิดบังรูปภาพบางส่วน (Mask some part) การแบ่งรูปภาพออกเป็นส่วนย่อย ๆ แล้วสลับตำแหน่ง (Shuffle the input) และการปรับรูปภาพให้เป็นโทนสีเทา (Grayscale an image) ดังรูปที่ 4.4



รูปที่ 4.4 การผสมผสานกันการสร้างรูปภาพใหม่ เพื่อให้ตัวเข้ารหัสเรียนรู้คุณลักษณะรูปภาพ

2) ผลการทำความสะอาดข้อความและการกำหนดมาตรฐานข้อความ ประกอบด้วย การตัดเครื่องหมายวรรคตอนและการแปลงข้อความทั้งหมดให้เป็นตัวพิมพ์เล็ก ซึ่งสอดคล้องกับแนวโน้มการกำจัดคำหยุดที่มีจำนวนคำลดลงเรื่อย ๆ จนในปัจจุบันการเรียนรู้เชิงลึกไม่มีการกำจัดคำหยุดเลย เนื่องจากถือว่าคำทั้งหมดมีผลต่อการกำหนดคุณลักษณะของรูปภาพ

A	B	C
001.jpg	[1]	two women walking down a street, hugging each other, enjoying the moment together
001.jpg	[2]	two women walking arm in arm, enjoying each other's company
001.jpg	[3]	two women walking down a street, hugging each other, enjoying the moment
001.jpg	[4]	two women walking down a street, holding hands and smiling, enjoying each other's company
001.jpg	[5]	two women walking down the street, holding hands and sharing a love that transcends time
002.jpg	[1]	a woman sits on the ground playing with a cat
002.jpg	[2]	a woman sitting on a brick walkway playing with a cat
002.jpg	[3]	a woman sitting on a brick street, petting a cat
002.jpg	[4]	a woman sitting on the ground petting a cat, surrounded by potted plants and a handbag
002.jpg	[5]	a woman sitting on the ground playing with a cat
...	...	...
400.jpg	[1]	a cat and a teddy bear sit together, watching the sunset
400.jpg	[2]	a cat and teddy bear sit together, looking out the window
400.jpg	[3]	a cat and a teddy bear sitting together by a window
400.jpg	[4]	a cat and a teddy bear sitting together, representing the bond between different species and the importance of companionship
400.jpg	[5]	a cat and a teddy bear sitting on a couch, the cat looking aggressive

รูปที่ 4.5 คำบรรยายรูปภาพที่ผ่านการทำความสะอาดข้อความและการกำหนดมาตรฐานข้อความ

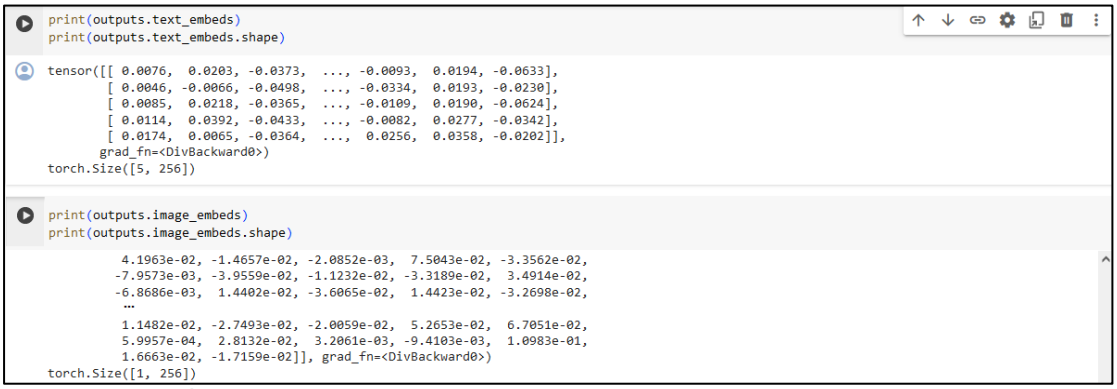
จากรูปที่ 4.5 ก่อนจะแยกข้อความออกเป็น 3 ส่วน ประกอบด้วยชื่อไฟล์รูปภาพในคอลัมน์ A ลำดับข้อความต่อรูปภาพในคอลัมน์ B และข้อความบรรยายรูปภาพในคอลัมน์ C ดังรูปที่ 4.5

2) ผลการฝึกฝนล่วงหน้าแบบคอนทราสต์ระหว่างรูปภาพและข้อความ โดยใช้ตัวแบบ CLIP ที่ถูกฝึกฝนล่วงหน้าบนชุดข้อมูลขนาดใหญ่จึงสามารถถ่ายโอนความรู้ไปใช้ได้เลยโดยไม่จำเป็นต้องฝึกฝนตัวแบบใหม่ตั้งแต่ต้น เพียงปรับแต่งตัวแบบเพิ่มเติมด้วยชุดข้อมูลที่เฉพาะเจาะจงในงานที่ต้องการ เนื่องจากตัวแบบได้เรียนรู้คุณลักษณะที่สำคัญของรูปภาพเอาไว้เป็นโมเดลฐาน เพื่อลดภาระการฝึกฝนตัวแบบจากเดิมที่ต้องการข้อมูลจำนวนมากในการเรียนรู้ อีกทั้งมีประสิทธิภาพดีกว่าตัวแบบที่ถูกสร้างขึ้นมาเองตั้งแต่ต้น มีรายละเอียดดังนี้

2.1) ผลการฝึกฝนตัวเข้ารหัสรูปภาพ ด้วยตัวแบบ ViT-B/32 สำหรับการสกัดคุณลักษณะรูปภาพตามแนวคิดของสถาปัตยกรรม Transformer มาปรับใช้กับรูปภาพ (Vision in Transformer) จากการเปรียบเทียบความสัมพันธ์ระหว่างแพตซ์ เพื่อคำนวณความสัมพันธ์กับพื้นที่ทั้งหมดในภาพ ผลลัพธ์จากการฝังรูปภาพ (Image embedding) จะแปลงเป็นเวกเตอร์คุณลักษณะรูปภาพ (Image feature vector) ขนาด 2,048

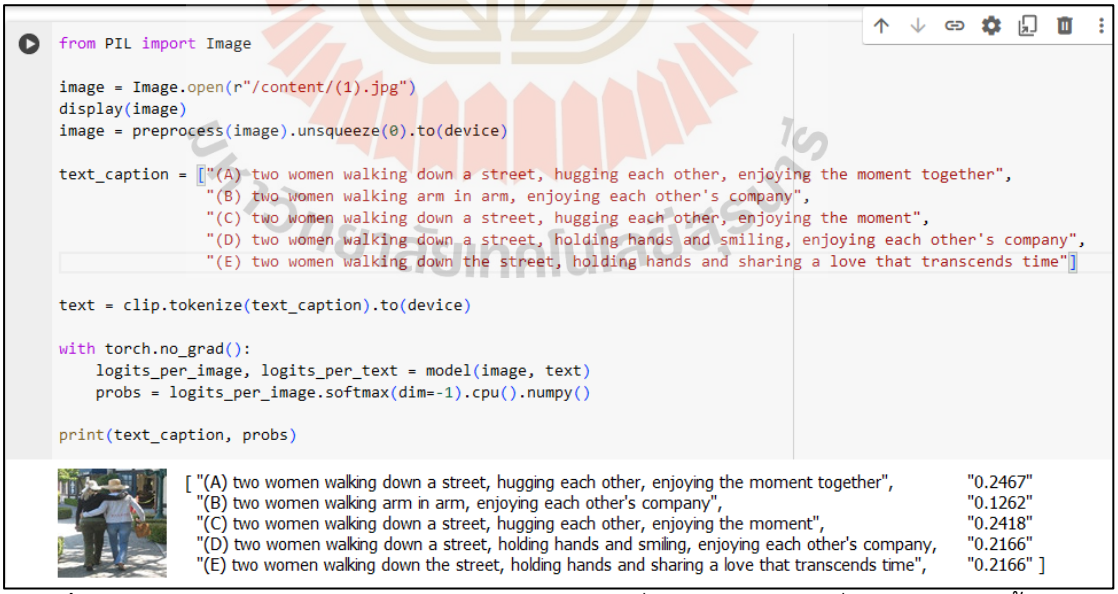
2.2) ผลการฝึกฝนตัวเข้ารหัสข้อความด้วยตัวแบบ GPT-2 บนพื้นฐานสถาปัตยกรรม Transformer โดยการเรียกใช้ CLIPTokenizer ที่ใช้เข้ารหัสข้อความและ CLIPTextModel ที่ใช้ฝังข้อความบนพื้นที่ฝังหลายรูปแบบ เพื่อเรียนรู้ความหมายโดยรวมของข้อความบรรยายรูปภาพแต่ละประโยค ผลลัพธ์จากการฝังข้อความ (Text embedding) แปลงเป็นเวกเตอร์คุณลักษณะข้อความ (Text feature vector) ขนาด 768

2.3) ผลการเรียนรู้แบบคอนทราสต์บนพื้นที่การฝังหลายรูปแบบ เนื่องจากเวกเตอร์คุณลักษณะข้อความและเวกเตอร์คุณลักษณะรูปภาพมีขนาดต่างกัน จึงต้องเรียกใช้ Projection head ในการเปลี่ยนขนาดเวกเตอร์บนพื้นที่การฝังให้มีมิติเท่ากับ 256 เพื่อฝึกฝนร่วมกันระหว่างตัวเข้ารหัสรูปภาพ และตัวเข้ารหัสข้อความด้วยการพยายามขยับเวกเตอร์ตัวแทนด้วยค่าความคล้ายคลึงโคไซน์ (Cosine similarity) จากผลคูณดอทยุคคลิดของเวกเตอร์แต่ละคู่ หากเป็นเวกเตอร์คู่เดียวกันจะมีค่าความคล้ายคลึงสูง ตรงกันข้ามหากเป็นเวกเตอร์คู่ที่แตกต่างกันจะมีค่าความคล้ายคลึงต่ำด้วย ดังรูปที่ 4.6 เพื่อให้ตัวแบบเรียนรู้คุณลักษณะเชิงความหมาย (Output feature) โดยหากเป็นภาพใกล้เคียงกันก็จะมีค่าใกล้เคียงกัน แต่หากเป็นภาพที่ต่างกัน ก็จะพยายามแยกให้อยู่ไกลกันออกไป และภาพที่ใกล้เคียงกันเกิดจากการสร้างรูปภาพใหม่จากการดัดแปลงรูปภาพต้นฉบับให้มีความต่างออกไป หรืออาจจะเรียนรู้ภาพคนละมุมมอง ส่งผลให้ตัวแบบสามารถเรียนรู้ที่จะสกัดคุณลักษณะที่จะยังคงให้ผลลัพธ์เป็นกลุ่มเดียวกันแม้จะเป็นคนละมุมมอง



รูปที่ 4.6 การเรียนรู้แบบคอนทราสต์ระหว่างเวกเตอร์คุณลักษณะข้อความ และเวกเตอร์คุณลักษณะรูปภาพบนพื้นที่การฝังหลายรูปแบบ

ผลการคำนวณความน่าจะเป็นของป้ายกำกับ เนื่องจากเอาต์พุตจากการเรียนรู้แบบคอนทราสต์นั้นจะอยู่ในรูปแบบค่า Logits score ที่เกิดจากค่าความคล้ายคลึงของเวกเตอร์ ซึ่งไม่ใช่การแจกแจงความน่าจะเป็นที่ถูกต้อง จึงต้องใช้ฟังก์ชันซอฟต์แม็กซ์มาแปลงให้เป็นมาตรฐานตามสัดส่วนของค่าเอาต์พุตแต่ละคลาส โดยมีผลรวมเท่ากับ 1 ดังรูปที่ 4.7 ค่าความน่าจะเป็นของป้ายกำกับภาษาธรรมชาติที่กำหนดเป็นค่าเฉลี่ยเลขคณิตถ่วงน้ำหนัก เพื่อนำมากำหนดเป็นผลลัพธ์จากการพยากรณ์ (Predicted output: $\hat{y}$)



รูปที่ 4.7 ความน่าจะเป็นของป้ายกำกับภาษาธรรมชาติที่กำหนดเป็นค่าเฉลี่ยเลขคณิตถ่วงน้ำหนัก

3) ผลการปรับแต่งค่าน้ำหนักโครงข่ายด้วยตนเอง ในการเรียนรู้ของตัวแบบเพื่อเปรียบเทียบความคล้ายคลึงเชิงความหมาย มีรายละเอียดดังนี้

3.1) ผลการคำนวณค่าถ่วงน้ำหนักเฉลี่ย และสร้างค่าน้ำหนักใหม่โดยอิงผู้เชี่ยวชาญจากการเปรียบเทียบระหว่างผลลัพธ์จากตัวแบบกับผลลัพธ์ที่ถูกประเมินความหมายโดยผู้เชี่ยวชาญด้วยการหาค่าการสูญเสีย เพื่อย้อนกลับไปปรับแต่งค่าน้ำหนักด้วยตนเอง (Custom update) โดยตัวอย่างค่าถ่วงน้ำหนักเฉลี่ยจากผู้เชี่ยวชาญทั้ง 10 ท่านดังตารางที่ 4.1

ตารางที่ 4.1 การคำนวณค่าถ่วงน้ำหนักเฉลี่ยจากผลการประเมินความรูปภาพโดยผู้เชี่ยวชาญจำนวน 10 ท่านสำหรับนำไปปรับค่าน้ำหนักแบบย้อนกลับ

ผู้เชี่ยวชาญ (ท่านที่)	ตัวเลือก	ค่าน้ำหนัก	Error	$\frac{\partial L}{\partial \hat{y}}$	$\frac{\partial z}{\partial W_1}$	Error at W_1	Update W_1
1	A	0.2467	2.0192	-1.5066	0.1858	-0.0140	0.2070
2	A	0.2467	2.0192	-1.5066	0.1858	-0.0140	0.2070
3	B	0.1262	-2.9862	-1.7476	0.2156	-0.0188	0.2094
4	A	0.2467	2.0192	-1.5066	0.1858	-0.0140	0.2070
5	C	0.2418	-2.0481	-1.5164	0.1870	-0.0142	0.2071
6	C	0.2418	-2.0481	-1.5164	0.1870	-0.0142	0.2071
7	A	0.2467	2.0192	-1.5066	0.1858	-0.0140	0.2070
8	D	0.2166	-2.2069	-1.5668	0.1933	-0.0151	0.2076
9	D	0.2166	-2.2069	-1.5668	0.1933	-0.0151	0.2076
10	A	0.2467	2.0192	-1.5066	0.1858	-0.0140	0.2070
ค่าถ่วงน้ำหนักเฉลี่ย							**0.2074**

จากตาราง 4.1 เพื่อง่ายต่อการอธิบายรายละเอียด งานวิจัยนี้ยกตัวอย่างค่าถ่วงน้ำหนักเริ่มต้น โดยกำหนดให้เท่ากับ 0.2 เมื่อผ่านคำนวณเป็นค่าถ่วงน้ำหนักเฉลี่ยจะปรับค่าน้ำหนักที่ W_1 จากเดิม 0.2 เป็น 0.2074 สำหรับนำไปปรับปรุงค่าน้ำหนักโครงข่ายด้วยตนเอง เริ่มจากการโหลดโมเดล openai/clip-vit-base-patch32 ที่ผ่านการฝึกฝนแล้วจากไลบรารี transformers เพื่อให้ได้โครงสร้างและค่าน้ำหนักเริ่มต้นของโมเดล จากนั้นทำการเรียกค่าน้ำหนักทั้งหมดจากเลเยอร์ visual_projection ซึ่งเป็นเลเยอร์สำคัญสำหรับการสกัดคุณลักษณะรูปภาพ

3.2) ผลการปรับแต่งค่าน้ำหนักโครงข่ายตามแนวคิดการแพร่กระจายย้อนกลับ เนื่องจากค่าน้ำหนักใหม่ที่สร้างโดยอ้างอิงผู้เชี่ยวชาญกับจำนวนพารามิเตอร์ในตัวแบบ CLIP ที่ถูกฝึกฝนล่วงหน้ามีจำนวนไม่เท่ากัน งานวิจัยนี้จึงการนำค่าน้ำหนักจากการคำนวณค่าการสูญเสียระหว่างป้ายกำกับภาษาธรรมชาติจากการพยากรณ์ของตัวแบบเปรียบเทียบกับผลการประเมินจากผู้เชี่ยวชาญจำนวน 400 ค่า มาใช้เป็นฐานน้ำหนัก (base_weights) ในการสร้างชุดน้ำหนักใหม่ โดยแปลงข้อมูลให้อยู่ในรูปแบบรายการ (List) เพื่อความสะดวกในการจัดการ จากนั้นเติมค่าที่เหลือจนครบจำนวนพารามิเตอร์ทั้งหมดของเลเยอร์ (512 * 768 = 393,216 ค่า) ด้วยค่าที่สุ่มตามการแจกแจงแบบปกติ (Normal distribution) โดยใช้ค่าเฉลี่ย (Mean) และส่วนเบี่ยงเบนมาตรฐาน (Standard deviation) จากชุด base_weights เดิม เพื่อให้ค่าที่เติมเข้ามามีลักษณะใกล้เคียงและสอดคล้องกับข้อมูลเดิม ก่อนจะแปลงชุดน้ำหนักดังกล่าวเป็นเทนเซอร์ของ PyTorch และปรับรูปทรง (Reshape) ให้สอดคล้องกับขนาดของเลเยอร์ที่ต้องการ เพื่ออัปเดตน้ำหนักชุดใหม่ที่สร้างขึ้นในเลเยอร์ visual_projection ของโมเดลโดยตรงด้วยการใช้ชุดข้อมูลน้ำหนักจริงควบคู่กับการเติมค่าน้ำหนักที่สุ่มอย่างมีการควบคุม ส่งผลให้รักษาคุณลักษณะสำคัญของโมเดลและลดความเสี่ยงของการเปลี่ยนแปลงน้ำหนักอย่างผิดธรรมชาติ ดังรูปที่ 4.8 ก่อนที่จะทำการบันทึกโมเดล CLIP ที่ผ่านการปรับค่าน้ำหนักแล้วเป็น CLIP_trained ลงใน GitHub repository เพื่อความสะดวกในการเรียกใช้งานภายหลัง

รูปที่ 4.8 การปรับแต่งค่าน้ำหนักเองตามแนวคิดการแพร่กระจายย้อนกลับ

จากรูปที่ 4.8 ผลลัพธ์ของกระบวนการสร้างและปรับปรุงค่าน้ำหนักในเลเยอร์ visual_projection ของโมเดล CLIP ประกอบด้วย 3 ส่วนคือ 1) ค่าน้ำหนักเดิม (Base weights) แสดงค่าน้ำหนักจำนวน 5 แถวแรกจากทั้งหมด 400 ค่า ซึ่งได้มาจากการคำนวณค่าสูญเสีย

(Loss) ระหว่างค่าที่ได้จากโมเดลกับผลการประเมินจากผู้เชี่ยวชาญ โดยค่าน้ำหนักเหล่านี้มีทั้งค่าบวกและลบ แสดงถึงทิศทางและขนาดของความสัมพันธ์ระหว่างข้อความกับภาพในมิติต่าง ๆ ซึ่งถือเป็นความรู้เชิงมนุษย์ (Human knowledge)ซึ่งเกิดจากประสบการณ์ การเรียนรู้ การใช้ภาษา ความเข้าใจบริบท และความสามารถในการตีความนัยสำคัญของข้อมูลในเชิงลึกที่ฝังไว้ในตัวแบบเบื้องต้น 2) ค่าน้ำหนักใหม่ที่ถูกเติม (Sampled weights) เป็นค่าน้ำหนักที่ถูกสุ่มขึ้นโดยใช้การแจกแจงแบบปกติ โดยมีการคำนวณค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานจากฐานน้ำหนัก (base_weights) เพื่อให้ค่าน้ำหนักที่สุ่มขึ้นใหม่มีลักษณะทางสถิติใกล้เคียงกับข้อมูลเดิม วิธีนี้ช่วยควบคุมไม่ให้ค่าน้ำหนักใหม่เบี่ยงเบนไปจากบริบทที่ผู้เชี่ยวชาญประเมินไว้มากเกินไป และ 3) ค่าน้ำหนักที่เปลี่ยนแปลงแล้ว (Final weights) เป็นผลลัพธ์หลังจากรวมค่าน้ำหนักเดิมจำนวน 400 ค่า และค่าน้ำหนักที่สุ่มขึ้นใหม่จนครบจำนวนที่ต้องการจำนวน 393,216 ค่า แล้วทำการปรับให้อยู่ในรูปแบบเทนเซอร์พร้อมใช้งานจริงในโมเดล โดยผลลัพธ์นี้คือค่าน้ำหนักสุดท้ายที่จะถูกนำไปอัปเดตในเลเยอร์ visual_projection เพื่อเพิ่มประสิทธิภาพของโมเดลในการจับคู่คุณลักษณะรูปภาพเชิงความกับข้อความค้นหา

3.3) ผลการเรียนรู้ซ้ำบนชุดข้อมูลที่กำหนดเอง กระบวนการสกัดคุณลักษณะเชิงความหมาย (Semantic features) ประกอบด้วยรูปภาพจากชุดข้อมูล Flickr30k จำนวน 31,783 รูปภาพ และชุดข้อมูลที่ผู้วิจัยรวบรวมเองจำนวน 92,168 รูปภาพ รวมทั้งสิ้น 123,951 รูปภาพ เมื่อผ่านการแบ่งข้อมูลออกเป็นกลุ่มย่อยหรือแบตช์ (Batch) กำหนดให้แต่ละแบตช์ประกอบด้วยรูปภาพจำนวน 16 รูป ซึ่งช่วยให้สามารถประมวลผลข้อมูลได้อย่างมีประสิทธิภาพมากขึ้น และยังสามารถจัดเก็บผลลัพธ์ของแต่ละแบตช์ได้อย่างเป็นระบบ ดังนั้นจะเกิดรอบของการประมวลผลทั้งสิ้น 1,987 ครั้ง ดังรูปที่ 4.9

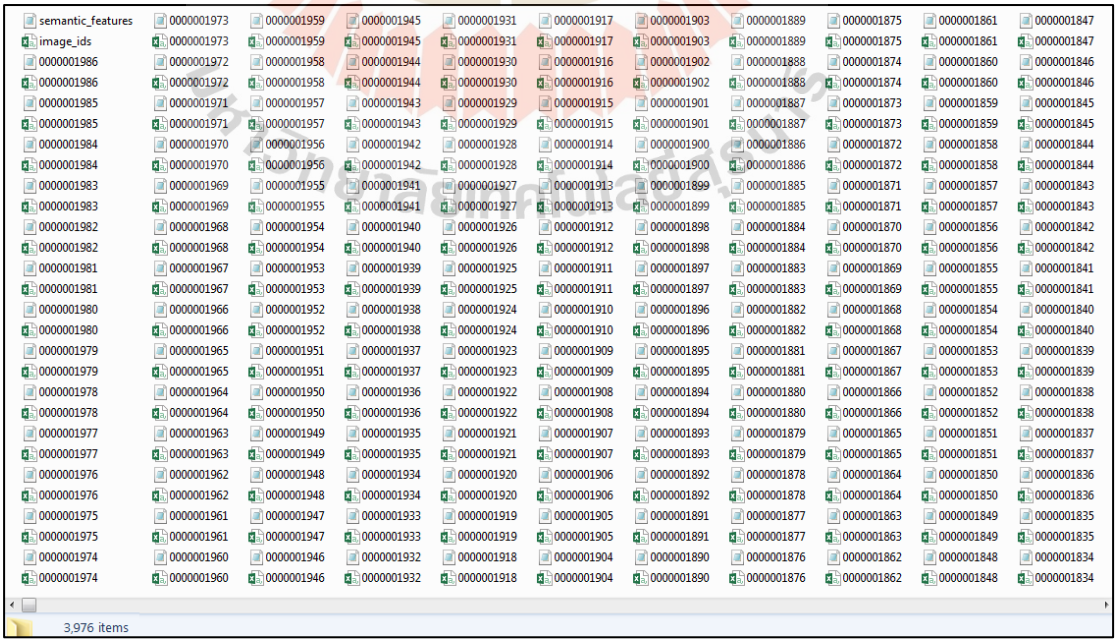
```
batch_size = 16

features_path = Path("semantic_features")
batches = math.ceil(len(image_files) / batch_size)

for i in range(batches):
    print(f"Processing batch {i+1}/{batches}")

    batch_ids_path = features_path / f"{i:010d}.csv"
    batch_features_path = features_path / f"{i:010d}.npy"

    if not batch_features_path.exists():
        try:
            batch_files = image_files[i*batch_size : (i+1)*batch_size]

            batch_features = compute_semantic_features(batch_files)
            np.save(batch_features_path, batch_features)

            image_ids = [photo_file.name.split(".")[0] for photo_file in batch_files]
            image_ids_data = pd.DataFrame(image_ids, columns=['image_id'])
            image_ids_data.to_csv(batch_ids_path, index=False)
        except:
            # check error
            print(f'Problem with batch {i}')
```

```
Processing batch 1/1987
Processing batch 2/1987
Processing batch 3/1987
...
Processing batch 1986/1987
Processing batch 1987/1987
```

รูปที่ 4.9 การเรียนรู้ซ้ำบนชุดข้อมูลที่กำหนดเอง

3.4) ผลการเรียนรู้ความคล้ายคลึงเชิงความหมายด้วยกระบวนการเรียนรู้ด้วยตนเอง เมื่อตัวแบบที่ผ่านการเรียนรู้เพื่อจดจำคุณลักษณะเชิงความหมายของรูปภาพแล้วจะสร้างป้ายกำกับเชิงความหมาย (Semantic labeled) สำหรับการค้นหา และใช้ประโยชน์จากความสัมพันธ์หรือความเกี่ยวข้องเชิงความหมายที่แตกต่างกัน จากนั้นจะสร้างเป็นเวกเตอร์คุณลักษณะรูปภาพขนาด ขนาด 2,048 และจัดเก็บไว้ในชุดคำอธิบายรูปภาพ (Image description set) สำหรับใช้ในกระบวนการค้นคืนรูปภาพต่อไป

หลังจากประมวลผลแล้ว ผลลัพธ์จะถูกจัดเก็บไว้ในรูปแบบไฟล์ .npy ซึ่งเหมาะสำหรับเก็บข้อมูลในรูปแบบอาร์เรย์ และในขณะเดียวกัน รหัสรูปภาพ (image ID) ของแต่ละภาพในแต่ละแบตช์นั้นจะถูกจัดเก็บไว้ในไฟล์ .csv เพื่อใช้เป็นข้อมูลในการจับคู่เวกเตอร์เพื่อค้นคืนรูปภาพเชิงความหมายต่อไป ดังรูปที่ 4.10 ผลลัพธ์จากกระบวนการสกัดคุณลักษณะเชิงความหมายจะประกอบด้วยไฟล์ semantic_features.npy ที่เก็บข้อมูลเวกเตอร์คุณลักษณะเชิงความหมายจากการประมวลผลรูปภาพในแบตช์ ซึ่งถ้ามีรูปภาพ 16 รูปในหนึ่งแบตช์ และแต่ละรูปมีเวกเตอร์ขนาด 768 มิติ ไฟล์จะมี shape (16, 768) ซึ่งจะเชื่อมโยงกับไฟล์ image_ids.csv ที่เก็บรหัสของภาพ (image IDs) ซึ่งตรงกับลำดับภาพที่ใช้สร้างฟีเจอร์ในไฟล์ .npy ที่มีชื่อเดียวกัน ดังนั้นผลลัพธ์ที่ได้จากแต่ละแบตช์ ประกอบด้วยไฟล์เวกเตอร์คุณลักษณะของแบตช์ที่ 1 ชื่อ 0000000001.npy จนถึง 0000001986.npy และไฟล์รายชื่อ image ID ของแบตช์ที่ 1 ชื่อ 0000000001.csv จนถึง 0000001986.csv ตามลำดับ



รูปที่ 4.10 รายละเอียดชุดคำอธิบายรูปภาพ

4.1.2 ผลการพัฒนามอดูลการประมวลผลข้อความค้นหา

งานวิจัยนี้ใช้ตัวแบบภาษา DistilBERT และทำการปรับแต่ง (Fine-tuning) ข้อมูลในส่วน Pre-trained โดยทำการเปลี่ยนแปลงเลเยอร์สุดท้ายของตัวแบบให้เหมาะสำหรับการระบุนิพจน์เชิงความหมาย ขั้นตอนการทำงานของตัวแบบจะนำข้อความทั้งหมดมาแบ่งออกเป็นโทเค็น (Tokenized) แล้วจึงแปลงแต่ละโทเค็นให้อยู่ในรูปแบบเวกเตอร์ ซึ่งผลลัพธ์ของ Token embedding จะเป็นข้อมูลที่แสดงความหมายของแต่ละโทเค็น Segment embedding เป็นส่วนช่วยให้ตัวแบบแยกส่วนต่าง ๆ ในประโยค และ Position embedding จะเป็นการให้ข้อมูลเกี่ยวกับตำแหน่งของโทเค็นในประโยครวมกันเป็นข้อมูลไปใช้ในขั้นตอนถัดไป เพื่อเรียนรู้ความหมายโดยรวมของข้อความค้นหาแต่ละประโยค ผลลัพธ์คือเวกเตอร์ตัวแทนคุณลักษณะข้อความค้นหาขนาด 768 ดังรูปที่ 4.11

```
search_query = "move forward"

with torch.no_grad():
    text_encoded = model.encode_text(clip.tokenize(search_query).to(device))
    text_encoded /= text_encoded.norm(dim=-1, keepdim=True)

print(text_encoded)
```

```
tensor([[ 5.3549e-04, -4.9347e-02, -3.7720e-02,  4.6425e-03,  4.5685e-02,
          5.0903e-02,  1.5251e-02, -6.3171e-02, -4.7333e-02,  4.4373e-02,
          3.0441e-02, -2.6199e-02, -3.3691e-02,  5.6725e-03,  3.3295e-02,
          2.8671e-02,  1.1292e-03, -7.4577e-03, -1.0689e-02,  4.7684e-03,
          6.7635e-03,  1.2177e-02,  3.6224e-02, -4.4594e-03,  9.4604e-03,
         -2.1484e-02,  7.2823e-03,  3.2745e-02, -3.7594e-03, -3.9337e-02,
          2.6474e-02,  1.5869e-02, -9.8228e-04, -4.5868e-02,  3.7994e-03,
         -1.1375e-02,  2.0386e-02,  2.8595e-02, -8.5678e-03,  2.5040e-02,
         -4.3762e-02,  1.1208e-02,  2.3865e-02, -1.6006e-02,  4.6234e-02,
          1.7487e-02,  3.4351e-03, -3.4637e-02,  8.9493e-03, -1.1032e-02,
         -2.3727e-03, -3.1815e-03,  2.1255e-02,  1.9272e-02,  2.7542e-02,
         -3.5278e-02,  3.9429e-02, -1.2894e-03, -2.9968e-02, -6.2065e-03,
          2.7618e-02, -1.1574e-02, -3.3539e-02,  4.2297e-02,  2.7832e-02,
         -2.7252e-02, -1.0292e-02,  1.6968e-02, -2.1027e-02, -1.2581e-02,
          4.5815e-03, -1.0307e-02,  3.2440e-02,  4.3427e-02, -3.3691e-02,
          ...
         -7.2266e-02, -1.5930e-02,  3.4271e-02,  1.4807e-01, -3.0899e-02,
          1.0063e-02, -2.0554e-02,  4.2358e-02, -5.6671e-02,  6.7711e-03,
         -1.1192e-02, -9.5367e-03,  2.5162e-02,  2.4052e-03,  5.3673e-03,
          2.5070e-02, -3.7781e-02,  6.2927e-02,  9.3536e-03,  1.3237e-02,
          2.2476e-02,  5.8289e-03]], device='cuda:0', dtype=torch.float16)
```

รูปที่ 4.11 ผลการแปลงข้อความค้นหาเป็นเวกเตอร์

จากรูปที่ 4.11 ผลการเข้ารหัสข้อความค้นหา “move forward” ผ่านกระบวนการแปลงเป็นเวกเตอร์เชิงความหมายด้วยฟังก์ชัน encode_text ซึ่งเป็นส่วนหนึ่งของโมเดล CLIP โดยผ่านขั้นตอนการ Tokenize ข้อความก่อน และดำเนินการเข้ารหัสภายใต้บริบทของ torch.no_grad() เพื่อป้องกันการคำนวณ Gradient ซึ่งไม่จำเป็นในขั้นตอนการใช้งานโมเดล (Inference) จากนั้นเวกเตอร์ที่ได้จะถูกปรับขนาดให้เป็นเวกเตอร์หน่วย (Unit vector) ด้วยการ Normalize ด้วยค่าความยาวเท่ากับ L2 Norm เพื่อความสอดคล้องในการเปรียบเทียบค่าความคล้ายเชิงมุม (Cosine similarity) ระหว่างเวกเตอร์ข้อความและเวกเตอร์ภาพที่อยู่ใน Latent space ร่วมกัน ซึ่งมีมิติเท่ากับ 512 มิติ ผลลัพธ์ที่ได้คือเทนเซอร์ที่ประกอบด้วยตัวเลขทศนิยมขนาดเล็กในช่วงประมาณ -0.1 ถึง 0.1

จำนวน 512 ค่า โดยมีการจัดเก็บอยู่ในรูปแบบ torch.float16 และทำงานบนอุปกรณ์ cuda ซึ่งเป็นการประมวลผลผ่าน GPU เพื่อเพิ่มประสิทธิภาพด้านความเร็ว

ดังนั้นเวกเตอร์ดังกล่าวถือเป็นตัวแทนของความหมายเชิงนามธรรมของข้อความค้นหา “move forward” ซึ่งสามารถนำไปใช้ในการคำนวณความคล้ายคลึงกับเวกเตอร์ของภาพที่ผ่านการเข้ารหัสด้วยโมเดลเดียวกัน เพื่อให้ระบบสามารถค้นคืนภาพที่มีความหมายหรือเนื้อหาสอดคล้องกับแนวคิด “การเคลื่อนไปข้างหน้า” ได้อย่างมีประสิทธิภาพ อาทิ ภาพของบุคคลที่กำลังก้าวเดิน ยานพาหนะที่กำลังเคลื่อนที่ หรือวัตถุที่แสดงทิศทางของการเคลื่อนไหว เป็นต้น

4.1.3 ผลการพัฒนามอดูลการจับคู่เวกเตอร์

การคำนวณค่าความคล้ายคลึงระหว่างเวกเตอร์รูปภาพและเวกเตอร์ข้อความค้นหาด้วยวิธีวัดความคลึงด้วยระยะทางแบบยุคลิด (Euclidean distance) มีรายละเอียดดังนี้

1) นำเข้า (Insert) ข้อมูลรูปภาพที่อยู่ในรูปแบบเวกเตอร์ผ่านตัวเข้ารหัสรูปภาพ โดยเวกเตอร์เหล่านี้ถูกจัดเก็บไว้ในชุดคำอธิบายรูปภาพ ซึ่งมีลักษณะเป็นฐานข้อมูลเวกเตอร์พร้อมกับการอ้างอิงถึงรูปภาพต้นฉบับทั้งหมดในชุดข้อมูล

2) แปลงข้อความค้นหาจากผู้ใช้ผ่านโมเดลการฝังเพื่อแปลงเป็นเวกเตอร์ผ่านตัวเข้ารหัสข้อความ เพื่อสร้างเวกเตอร์ข้อความคำค้นหาที่ผ่านการฝังสำหรับค้นหารูปภาพที่มีความคล้ายคลึงกันเชิงความหมาย

3) จากนั้นจะใช้เวกเตอร์ข้อความค้นหาเพื่อค้นหาเวกเตอร์การฝังที่คล้ายคลึงกันในฐานข้อมูลเวกเตอร์บนพื้นที่การฝังหลายรูปแบบ สำหรับแปลงเวกเตอร์ที่มีขนาดแตกต่างกันให้อยู่ในมิติเท่ากับ 256 เพื่อให้สามารถเปรียบเทียบความคล้ายกันได้ในปริภูมิเวกเตอร์มิติสูง (High-dimensional vector space) ด้วยวิธีวัดระยะทางระหว่างเวกเตอร์รูปภาพและเวกเตอร์ข้อความค้นหาแบบระยะทางแบบยุคลิด

4) ผลลัพธ์จากการค้นหาความคล้ายคลึงกันจะแสดงรายการเวกเตอร์รูปภาพที่คล้ายกับเวกเตอร์ข้อความค้นหามากที่สุด โดยเรียงลำดับจากมากที่สุดไปยังน้อยที่สุด จากนั้นจะเข้าถึงรูปภาพต้นฉบับเพื่อนำมาแสดงผลตามความเกี่ยวข้องแก่ผู้ใช้ต่อไป

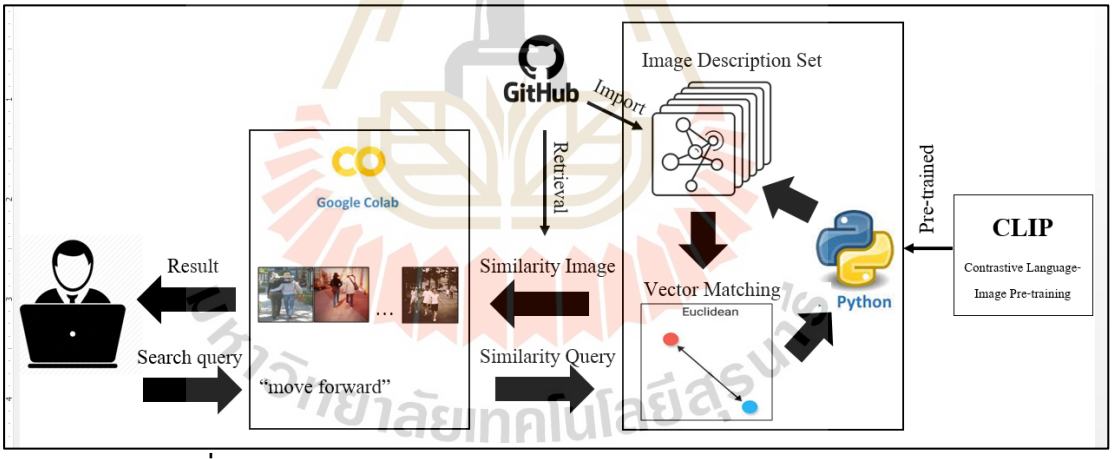
จากรูปที่ 4.12 ผลการค้นคืนรูปภาพจากชุดข้อมูลโดยใช้ข้อความค้นหา “move forward” ซึ่งสะท้อนแนวคิดด้านการเคลื่อนที่ไปข้างหน้า แล้วดำเนินการแปลงข้อความดังกล่าวให้เป็นเวกเตอร์เชิงความหมายผ่านฟังก์ชัน encode_search_query() โดยอาศัยตัวแบบ CLIP ที่ถูกฝึกฝนล่วงหน้า เพื่อสร้างตัวแทนเวกเตอร์ของข้อความในลักษณะที่สามารถนำไปเปรียบเทียบกับเวกเตอร์ของภาพได้โดยตรง ภายหลังจากได้เวกเตอร์ของข้อความแล้ว ฟังก์ชัน find_vector_matching() จะดำเนินการเปรียบเทียบเวกเตอร์ดังกล่าวกับเวกเตอร์ของภาพทั้งหมดในฐานข้อมูล (Image features) เพื่อจัดลำดับภาพตามความคล้ายคลึงสูงสุดเชิงความหมาย

โดยใช้วิธีการวัดความคล้ายคลึงโคไซน์ ซึ่งเป็นวิธีการที่นิยมใช้ในการประเมินความสัมพันธ์ของข้อมูลในเวกเตอร์สเปซ ผลลัพธ์คือรูปภาพจากการค้นคืนจำนวน 20 ภาพที่มีความคล้ายคลึงกับข้อความมากที่สุดตามลำดับ โดยแต่ละภาพแสดงถึงกิจกรรมที่มีลักษณะสอดคล้องกับแนวคิด “การเคลื่อนไปข้างหน้า” เช่น ภาพที่ 1, 6, 8, 11 และ 12 เป็นภาพของบุคคลที่กำลังเดิน เช่นเดียวกับภาพที่ 2, 7 และ 13 เป็นภาพของการข้ามถนน ตลอดจนภาพที่ 9, 10, และ 14 เป็นภาพของถนนหรือทางเดินที่นำสายตาไปข้างหน้า อย่างไรก็ดี ภาพที่ 16, 17 และ 18 แม้จะเป็นป้ายบอกทาง แต่มีบริบทเชิงความหมายที่สื่อถึงสัญลักษณ์ที่เกี่ยวข้องกับการเคลื่อนไหวไปข้างหน้า

จะเห็นได้ว่าแบบจำลองสามารถเข้าใจบริบทเชิงความหมายของข้อความ และจับคู่ (Mapping) เข้ากับภาพที่มีเนื้อหาสอดคล้องในระดับความหมาย ไม่จำกัดอยู่เพียงวัตถุในภาพแต่รวมถึงแนวคิดและบริบทที่เกี่ยวข้องด้วย เช่น การเคลื่อนที่ การเดินหน้า และการขับเคลื่อน ซึ่งเป็นลักษณะของความรู้เชิงมนุษย์ (Human-level semantic understanding) ดังนั้น จึงสามารถนำไปประยุกต์ใช้ในงานค้นคืนสารสนเทศจากภาพในหลายบริบท เช่น งานห้องสมุดดิจิทัล งานด้านหุ่นยนต์ที่ต้องเข้าใจคำสั่งภาษา และระบบแนะนำเนื้อหาภาพที่ตรงกับความตั้งใจของผู้ใช้งาน เป็นต้น

รูปที่ 4.12 ผลลัพธ์การค้นคืนรูปภาพที่เกี่ยวข้องกับข้อความค้นหา “move forward” มากที่สุด

ภาพรวมแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมาย โดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้าที่สร้างขึ้นจากภาษาไพทอนดัง รูปที่ 4.13 การทำงานผ่าน Google Colaboratory แบ่งเป็น 2 ส่วน ประกอบด้วย ส่วนหลัง (Backend) ในการเรียกใช้ตัวแบบ CLIP ที่ถูกฝึกฝนล่วงหน้า เพื่อเรียนรู้แบบคอนทราสต์ระหว่างรูปภาพที่สุ่มมาจากชุดข้อมูล Flickr30k จำนวน 100 รูปภาพและชุดข้อมูลที่ผู้วิจัยรวบรวมเองจำนวน 300 รูปภาพ รวมทั้งสิ้น 400 รูปภาพ ร่วมกับคำบรรยายรูปภาพ 5 รายการต่อ 1 รูปภาพที่กำหนดโดยผู้วิจัย สำหรับพยากรณ์ป้ายกำกับของรูปภาพก่อนจะเปรียบเทียบกับผลการประเมินความหมายของรูปภาพโดยผู้เชี่ยวชาญ เพื่อนำมาคำนวณหาค่าการสูญเสียก่อนจะนำไปปรับปรุงพารามิเตอร์ของตัวแบบตามแนวคิดการแพร่กระจายย้อนกลับ ก่อนจะถ่ายโอนความรู้ไปใช้เรียนรู้ซ้ำอีกครั้งบนชุดข้อมูล Flickr30k จำนวน 31,783 รูปภาพ และชุดข้อมูลที่ผู้วิจัยรวบรวมเองจำนวน 92,168 รูปภาพ รวมทั้งสิ้น 123,951 รูปภาพ เพื่อสร้างชุดคำอธิบายรูปภาพเพื่อใช้ค้นคืนรูปภาพจากการจับคู่ระหว่างเวกเตอร์รูปภาพและเวกเตอร์ข้อความค้นหาจากผู้ใช้ บนพื้นที่การฝังหลายรูปแบบด้วย จากนั้นจะปฏิสัมพันธ์กับผู้ใช้ผ่านส่วนหน้า (Frontend) เพื่อค้นคืนรูปภาพที่เกี่ยวข้องกับข้อความค้นหามากที่สุด

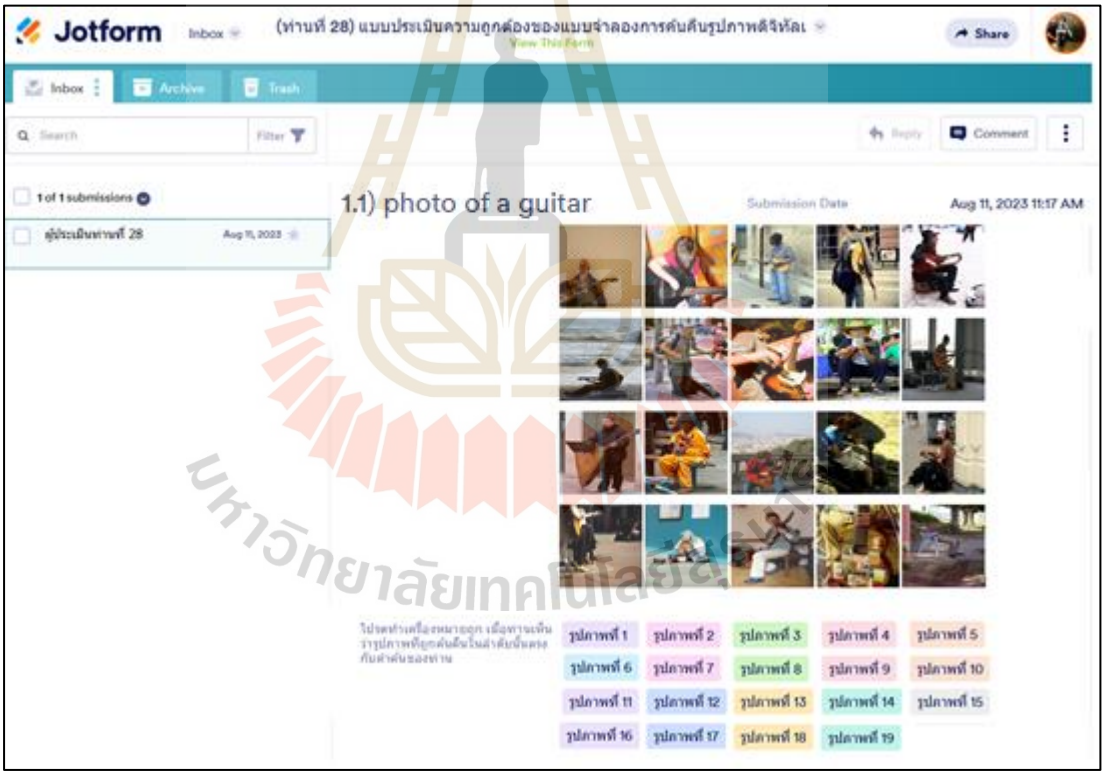


รูปที่ 4.13 ภาพรวมแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมาย โดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า

4.2 ผลการประเมินประสิทธิภาพการค้นคืนรูปภาพดิจิทัลเชิงความหมาย

การประเมินผลความหมายของรูปภาพ ดำเนินการทดสอบผ่าน Google Colaboratory โดยผู้วิจัยนำเข้าข้อความค้นหาไปยังตัวแบบที่พัฒนาขึ้นทีละข้อความด้วยมาตรวัดแบบไม่สนใจลำดับของผลลัพธ์ ประกอบด้วยค่าความแม่นยำที่ k อันดับ ค่าความครบถ้วนที่ k อันดับ และค่าประสิทธิภาพโดยรวมที่ k อันดับ บนชุดข้อมูล Flickr30k จำนวน 31,783 รูปภาพ และชุดข้อมูลที่ผู้วิจัยรวบรวมเองจำนวน 92,168 รูปภาพ รวมทั้งสิ้น 123,951 รูปภาพ โดยกำหนดจำนวน

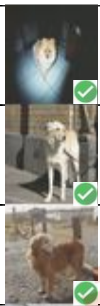
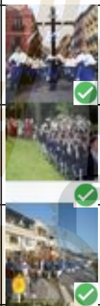
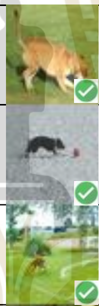
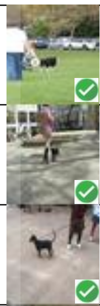
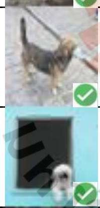
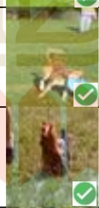
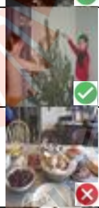
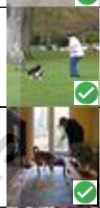
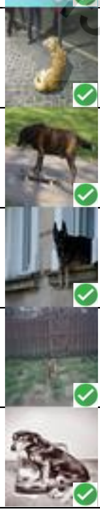
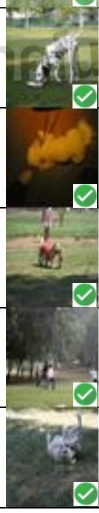
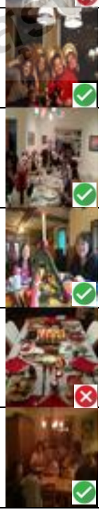
รูปภาพในการแสดงผลแต่ละครั้งเท่ากับ 20 รูปภาพ ภายใต้ข้อความค้นหาภาษาอังกฤษ 3 เงื่อนไข ได้แก่ 1) ข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ ในลักษณะป้ายกำกับของรูปภาพ 2) ข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพ และ 3) ข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ โดยกำหนดข้อความค้นหาจากผู้เชี่ยวชาญเงื่อนไขละ 10 รายการ รวมทั้งสิ้น 30 รายการ เมื่อกระบวนการเสร็จสิ้นจะปรากฏผลลัพธ์เป็นรูปภาพที่มีความเกี่ยวข้องกันมากที่สุดมาแสดงเป็นผลลัพธ์การค้นคืนรูปภาพแก่ผู้ใช้แต่ละครั้งเท่ากับ 20 รูปภาพ จากนั้นให้ผู้เชี่ยวชาญประเมินความถูกต้องของรูปภาพทีละรูปภาพผ่านแบบประเมินความถูกต้องทีละรายการ จนครบทั้ง 30 รายการต่อผู้เชี่ยวชาญ 1 ท่าน เมื่อครบทั้ง 30 ท่าน จะมีข้อความค้นหารวมทั้งสิ้น 900 รายการ แบ่งเป็นเงื่อนไขละ 300 รายการ ตัวอย่างการประเมินความถูกต้องของรูปภาพผ่านแบบประเมินออนไลน์ Jotform ดังรูปที่ 4.14



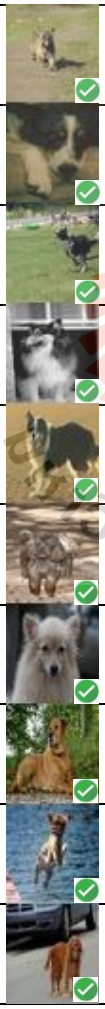
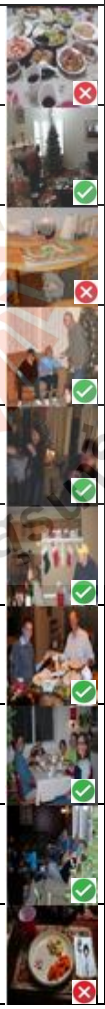
รูปที่ 4.14 ตัวอย่างการประเมินความถูกต้องของรูปภาพผ่านแบบประเมินออนไลน์ Jotform

จากรูปที่ 4.14 เมื่อรวบรวมข้อมูลเพื่อนำมาวิเคราะห์ทางสถิติ กำหนดให้ หากรูปภาพจากการค้นคืนถูกต้องตามข้อความค้นหา เท่ากับ 1 คะแนน ในทางกลับกันหากรูปภาพจากการค้นคืนไม่ถูกต้องหรือไม่ครบถ้วนตามข้อความค้นหา เท่ากับ 0 คะแนน ดังตารางที่ 4.2 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับที่ผ่านการประเมินความถูกต้องโดยผู้เชี่ยวชาญท่านที่ 1

ตารางที่ 4.2 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับ ที่ผ่านการประเมินความถูกต้องโดยผู้เชี่ยวชาญท่านที่ 1

ตัวอย่างข้อความค้นหา		เงื่อนไขที่ 1			เงื่อนไขที่ 2			เงื่อนไขที่ 3		
		1.1) photo of a dog	…	1.300) photo of tradition	2.1) the dog playing on ground	…	2.300) we enjoy our family traditions every year	3.1) the dog wagged its tail happily	…	3.300) family gathers to celebrate cultural traditions with love and joy
จำนวนรูปภาพจากการค้นคืน	3 ลำดับ									
	5 ลำดับ									
	10 ลำดับ									

ตารางที่ 4.2 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับ ที่ผ่านการประเมินความถูกต้องโดยผู้เชี่ยวชาญท่านที่ 1 (ต่อ)

ตัวอย่างข้อความค้นหา		เงื่อนไขที่ 1			เงื่อนไขที่ 2			เงื่อนไขที่ 3		
		1.1) photo of a dog	...	1.300) photo of tradition	2.1) the dog playing on ground	...	2.300) we enjoy our family traditions every year	3.1) the dog wagged its tail happily	...	3.300) family gathers to celebrate cultural traditions with love and joy
จำนวนรูปภาพจากการค้นคืน	15 ลำดับ									
	20 ลำดับ									

การพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายโดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า แบ่งผลประเมินประสิทธิภาพการค้นคืนรูปภาพดิจิทัลเชิงความหมายออกเป็น 2 ส่วน มีรายละเอียดดังนี้

4.2.1 ผลการประเมินประสิทธิภาพการค้นคืนรูปภาพดิจิทัลเชิงความหมายแบบไม่สนใจลำดับของผลลัพธ์ด้วยค่าความแม่นยำที่ k อันดับ ค่าความครบถ้วนที่ k อันดับ และค่าประสิทธิภาพโดยรวมที่ k อันดับ

การประเมินประสิทธิภาพการค้นคืนรูปภาพดิจิทัลเชิงความหมายภายใต้ข้อความค้นหา 3 เงื่อนไข ได้แก่ 1) ข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ ในลักษณะป้ายกำกับของรูปภาพ 2) ข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของภาพ และ 3) ข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ เงื่อนไขละ 100 รายการ รวมทั้งสิ้น 300 รายการดังตารางที่ 4.3

ตารางที่ 4.3 ผลการประเมินประสิทธิภาพการค้นคืนรูปภาพแบบไม่สนใจลำดับผลลัพธ์

เงื่อนไขการค้นหา		**1) ข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ ในลักษณะป้ายกำกับของรูปภาพ**			**2) ข้อความค้นหาด้วยสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพ**			**3) ข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ**		
การประเมินประสิทธิภาพการค้นคืนรูปภาพ		Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure
จำนวนรูปภาพจากการค้นคืน	**k@3**	0.939	0.603	0.734	0.864	0.590	0.701	0.830	0.578	0.682
	k@5	0.935	0.726	0.817	0.860	0.706	0.775	0.825	0.701	0.758
	k@10	0.907	0.818	0.860	0.859	0.813	0.835	0.819	0.802	0.810
	k@15	0.902	0.867	0.884	0.843	0.862	0.852	0.804	0.836	0.820
	k@20	0.894	1.000	0.994	0.833	1.000	0.909	0.771	1.00	0.870

จากตารางที่ 4.3 ผลการประเมินประสิทธิภาพการค้นคืนรูปภาพแบบไม่สนใจลำดับผลลัพธ์ด้วยค่าความแม่นยำ ค่าความครบถ้วน และค่าประสิทธิภาพโดยรวม มีรายละเอียดดังนี้

1) ผลการประเมินประสิทธิภาพการค้นคืนรูปภาพด้วยค่าความแม่นยำที่ k อันดับภายใต้เงื่อนไขข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ในลักษณะป้ายกำกับของรูปภาพ และข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพนั้นมีค่าความแม่นยำอยู่ในระดับดีมาก โดยเฉพาะผลลัพธ์การค้นคืนจำนวน 3 ลำดับแรก ค่าเฉลี่ย 0.939 และ 0.864 ตามลำดับ เช่นเดียวกับข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพที่อยู่ในระดับดี ค่าเฉลี่ยเท่ากับ 0.830 ซึ่งค่าความแม่นยำนั้นลดลงเรื่อย ๆ แปรผันตามจำนวนผลลัพธ์การค้นคืนรูปภาพที่เพิ่มมากขึ้น

2) ผลการประเมินประสิทธิภาพการค้นคืนรูปภาพด้วยค่าความครบถ้วนที่ k อันดับภายใต้เงื่อนไขข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ในลักษณะป้ายกำกับของรูปภาพและข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพนั้นมีค่าความครบถ้วนอยู่ในระดับดี เมื่อผลลัพธ์จากการค้นคืนเท่ากับ 3 ลำดับ ค่าเฉลี่ย 0.603 และ 0.590 ตามลำดับ ตรงกันข้ามกับข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพที่อยู่ในระดับปานกลาง ค่าเฉลี่ยเท่ากับ 0.578 อย่างไรก็ดี แม้ค่าความครบถ้วนจะค่อย ๆ เพิ่มขึ้นตามจำนวนผลลัพธ์ที่เพิ่มมากขึ้น และจะเข้าใกล้ 1 เมื่อจำนวนผลลัพธ์เท่ากับ 10 ที่ค่าเฉลี่ย 0.818, 0.813 และ 0.802 ตามลำดับ เมื่อเปรียบเทียบกับงานวิจัยของ Le et al. (2022) ที่ประยุกต์ใช้ตัวแบบ CLIP มาเรียนรู้คำอธิบายภาษาธรรมชาติเพื่อจำแนกรถยนต์ในมุมมองที่แตกต่างกัน เช่นเดียวกับ งานวิจัยของ Bianchi et al. (2021) ที่ประยุกต์ใช้ตัวแบบ CLIP มาใช้ค้นคืนรูปภาพจากการฝึกฝนตัวแบบให้เรียนรู้ข้อความบรรยายรูปภาพที่แปลงจากภาษาอังกฤษเป็นภาษาอิตาลีที่มีค่าความครบถ้วนที่ลำดับ จำนวน 1, 3 และ 5 รูปภาพบนชุดข้อมูล Flickr30k นั้นมีค่าความครบถ้วนเฉลี่ย 0.585, 0.664 และ 0.761 ตามลำดับ ถือว่าอยู่ในระดับที่น่าพึงพอใจ แต่ผลลัพธ์นั้นยังห่างไกลจากความคาดหวังของผู้ใช้ เนื่องจากพฤติกรรมของผู้ใช้ต้องการเพียงผลลัพธ์ที่ตนเองสนใจเท่านั้น (Jansen and Spink, 2006) ผลลัพธ์จำนวนมากจากค้นคืนทั้งที่ตรงและไม่ตรงกับความต้องการ อาจทำให้ผู้ใช้เสียเวลาคัดเลือกรูปภาพที่ตรงกับความต้องการอย่างแท้จริง โดย Jansen and Spink (2003) พบว่า ผู้ใช้จะเรียกดูเอกสารไม่เกิน 5 รายการที่ค้นคืน ทั้งนี้มีเพียงผู้ใช้ร้อยละ 28.3 เท่านั้นที่เรียกดูเพียง 2-3 รายการ อีกทั้งร้อยละ 55 จะเรียกดูผลลัพธ์เพียง 1 รายการต่อ 1 คำค้นหา และปัจจุบันมีแนวโน้มจะลดลงเรื่อย ๆ ซึ่งนับเป็นความสูญเปล่าทางสารสนเทศ เนื่องจากเป็นได้น้อยมากที่ผู้ใช้จะเลือกดูรูปภาพจากผลการค้นคืนทั้งหมด จึงเป็นการเสียโอกาสสำหรับผู้ใช้หากมีรูปภาพบางส่วนที่ไม่ถูกเรียกดู ดังนั้นการค้นคืนรูปภาพจึงต้องมุ่งเพิ่มค่าความครบถ้วนอันดับ 3, 5 และ 10 เป็นหลักเพื่อตอบสนองต่อความต้องการของผู้ใช้

3) ผลการประเมินประสิทธิภาพการค้นคืนรูปภาพด้วยค่าประสิทธิภาพโดยรวมที่ k อันดับภายใต้เงื่อนไขข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ในลักษณะป้ายกำกับของรูปภาพและข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพนั้นมีค่าความครบถ้วนอยู่ในระดับดีเช่นเดียวกับข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ อย่างไรก็ดี ค่าประสิทธิภาพโดยรวมจะค่อย ๆ เพิ่มขึ้นตามจำนวนผลลัพธ์ที่เพิ่มมากขึ้น และจะเข้าใกล้ 1 มากยิ่งขึ้นเมื่อจำนวนผลลัพธ์เท่ากับ 10 ค่าเฉลี่ย 0.860, 0.835 และ 0.810 ตามลำดับ

4) เมื่อพิจารณาจากค่าความแม่นยำที่ k อันดับ ค่าความครบถ้วนที่ k อันดับ และค่าประสิทธิภาพโดยรวมที่ k อันดับ พบว่าเป็นไปในทิศทางเดียวกัน โดยผลลัพธ์ดังกล่าวนั้นถือว่าอยู่ในระดับที่ยอมรับได้ และไม่เกิดปัญหาที่ตัวแบบจะมีประสิทธิภาพลดลง เมื่อทำงานกับข้อมูลที่ไม่เคยพบมาก่อน (Poor real-world performance)

5) การปริทัศน์งานวิจัยที่มุ่งค้นคืนรูปภาพเชิงความหมายในปัจจุบัน โดย Caicedo, Gonzalez, and Romero (2008) นำเสนอวิธีการค้นคืนรูปภาพเชิงเนื้อหาด้วยการแยกคุณสมบัติระดับสูงจากการวิเคราะห์ความหมายของรูปภาพค้นหา (Image query) พบว่า การค้นคืนด้วยคุณลักษณะระดับต่ำเพียงอย่างเดียว มีค่าความแม่นยำสูงสุดเมื่อจำนวนที่ 1 รูปภาพเท่ากับ 0.67 แต่เมื่อใช้ร่วมกับคุณลักษณะเชิงความหมาย ค่าความแม่นยำเพิ่มเป็น 0.80 แต่จะลดลงเรื่อย ๆ แปรปรวนตามจำนวนผลลัพธ์การค้นคืนที่เพิ่มมากขึ้น ในขณะที่ Khodaskar and Ladhake (2015) นำเสนอการวิเคราะห์เนื้อหารูปภาพค้นหาจากการแทนคุณลักษณะเนื้อหาเชิงความหมาย (Semantic content representation) ในฐานความรู้ ร่วมกับการดึงข้อมูลและการจัดทำดัชนีสำหรับการค้นคืนรูปภาพ พบว่า ค่าความแม่นยำอยู่ระหว่าง 0.79 ถึง 0.89 จากนั้น Thanh et al. (2020) นำเสนอระบบค้นคืนรูปภาพเชิงความหมายด้วยรูปภาพค้นหา พบว่า การใช้เทคนิคเคมีน (K-mean) เพียงอย่างเดียวมีค่าความแม่นยำเท่ากับ 0.65 แต่เมื่อผสมผสานเทคนิคเพื่อนบ้านใกล้ที่สุด (K-nearest neighbor) เข้าด้วยกัน ค่าความแม่นยำเพิ่มเป็น 0.70 และ Nhi et al. (2020) นำเสนอตัวแบบการค้นคืนรูปภาพเชิงเนื้อหาจากการวิเคราะห์รูปภาพค้นหาด้วยกราฟ C-tree ร่วมกับเทคนิคเพื่อนบ้านใกล้ที่สุด พบว่า ค่าความแม่นยำบนชุดข้อมูล COREL เท่ากับ 0.89 รายละเอียดดังตารางที่ 4.4

ตารางที่ 4.4 การเปรียบเทียบผลการวิจัยกับงานวิจัยอื่น ๆ ที่มุ่งค้นคืนรูปภาพเชิงความหมายโดยใช้รูปภาพค้นหา

งานวิจัย		ค่าความแม่นยำ
Caicedo, Gonzalez, and Romero (2008)	A semantic content-based retrieval method for histopathology images	0.67 - 0.80
Khodaskar and Ladhake (2015)	Content based image retrieval with semantic features using object ontology	0.79 - 0.89
Thanh et al. (2020)	A semantic-based image retrieval system using a hybrid method k-means and k-nearest-neighbor	0.65 - 0.70
Nhi et al. (2020)	A model of semantic-based image retrieval using c-tree and neighbor graph	0.89
งานวิจัยนี้	The development of a semantic-based image retrieval model using deep neural network by pre-training approach	k@3 = 0.830 k@5 = 0.825 k@10 = 0.819

จากตารางที่ 4.4 เมื่อพิจารณาในรายละเอียด พบว่า ส่วนใหญ่มุ่งใช้รูปภาพค้นหาในการค้นคืนรูปภาพเชิงความหมาย เนื่องจากระบบจะวิเคราะห์คุณลักษณะรูปภาพเพื่อค้นคืนรูปภาพตามความคล้ายคลึงกับรูปภาพค้นหา จึงสามารถลดปัญหาช่องว่างความหมายลงได้ แต่ในความเป็นจริงพฤติกรรมผู้ใช้โดยมากจะค้นคืนรูปภาพโดยใช้ความสัมพันธ์ระหว่างคุณลักษณะระดับต่ำผสมผสานแนวคิดระดับสูงที่เป็นรูปธรรม และแนวคิดระดับสูงที่เป็นนามธรรม รวมเข้าด้วยกันเป็นข้อความค้นหาภาษาธรรมชาติที่อยู่ในลักษณะแนวคิดระดับสูง ในทางกลับกันคุณลักษณะที่ถูกแยกอัตโนมัติโดยใช้คอมพิวเตอร์มักเป็นคุณลักษณะระดับต่ำเท่านั้น ซึ่งโดยทั่วไปไม่มีการเชื่อมโยงระหว่างคุณลักษณะระดับต่ำและแนวคิดระดับสูง ส่งผลให้ผลลัพธ์จากการค้นคืนมีประสิทธิภาพไม่สูงเท่าที่ควร อีกทั้งการใช้รูปภาพค้นหายังไม่สะดวก และเป็นการผลักภาระให้กับผู้ใช้ในการหารูปภาพต้นฉบับอีกด้วย ดังนั้นเมื่อเปรียบเทียบกับผลการค้นคืนรูปภาพด้วยข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพในงานวิจัยนี้มีค่าความแม่นยำที่ k อันดับเท่ากับ 0.830, 0.825 และ 0.819 ที่จำนวนรูปภาพเท่ากับ 3 รูปภาพ 5 รูปภาพ และ 10 รูปภาพตามลำดับ อยู่ในระดับดีและเป็นระดับที่ยอมรับได้ เนื่องจากความหมายของรูปภาพที่อยู่ในลักษณะนามธรรมนั้นยากที่จะ

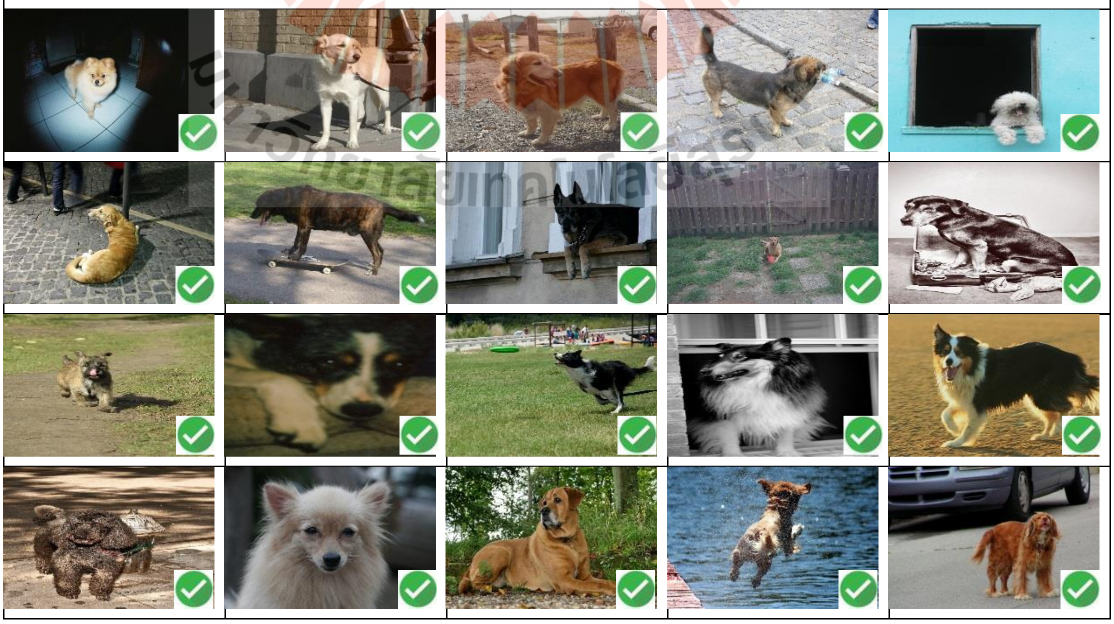
ประเมินผลว่าถูกต้องหรือไม่ถูกต้องเท่านั้น ดังนั้นการแปลความหมายตามหลักการรับรู้ของมนุษย์ (Human perception) (Dix et al., 2004) ด้วยประสบการณ์หรือความรู้เดิมจากการเรียนรู้ของแต่ละบุคคลจึงแตกต่างกัน เห็นได้จากผลลัพธ์การค้นคืนรูปภาพเชิงความหมายจากข้อความค้นหามีค่าความแม่นยำไม่แตกต่างจากการใช้รูปภาพค้นหามากนัก

เมื่อพิจารณาผลการประเมินประสิทธิภาพการค้นคืนรูปภาพดิจิทัลเชิงความหมายแบบไม่สนใจลำดับของผลลัพธ์ โดยจำแนกจากกลุ่มผู้เชี่ยวชาญ 3 กลุ่ม ได้แก่ 1) กลุ่มผู้เชี่ยวชาญด้านคอมพิวเตอร์และเทคโนโลยี 2) กลุ่มผู้เชี่ยวชาญด้านการค้นคืนสารสนเทศ และ 3) กลุ่มผู้ใช้ทั่วไป มีรายละเอียดดังนี้

4.2.1.1 ผลการประเมินประสิทธิภาพการค้นคืนรูปภาพด้วยค่าความแม่นยำที่ k อันดับ ค่าความครบถ้วนที่ k อันดับ และค่าประสิทธิภาพโดยรวมที่ k อันดับภายใต้เงื่อนไขข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ในลักษณะป้ายกำกับของรูปภาพ ที่เปรียบได้กับการค้นคืนรูปภาพด้วยข้อความด้วยข้อมูลที่ได้รับการติดแท็กด้วยป้ายกำกับตั้งแต่หนึ่งรายการขึ้นไป โดยทั่วไปจะใช้กับชุดข้อมูลที่ไม่มีป้ายกำกับและเพิ่มแต่ละส่วนด้วยแท็กข้อมูล เพื่อจำแนกคลาสของข้อมูลที่ได้จากการเรียนรู้ เช่น รูปภาพคน สัตว์ หรือสิ่งของ เป็นต้น ดังตารางที่ 4.5

ตารางที่ 4.5 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพด้วยชื่อรูปภาพและหมวดหมู่ ในลักษณะป้ายกำกับของรูปภาพที่ผ่านการประเมินความถูกต้องโดยผู้เชี่ยวชาญท่านที่ 1

1) photo of a dog



ตารางที่ 4.5 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพด้วยชื่อรูปภาพและหมวดหมู่ ในลักษณะป้ายกำกับของรูปภาพที่ผ่านการประเมินความถูกต้องโดยผู้เชี่ยวชาญท่านที่ 1 (ต่อ)

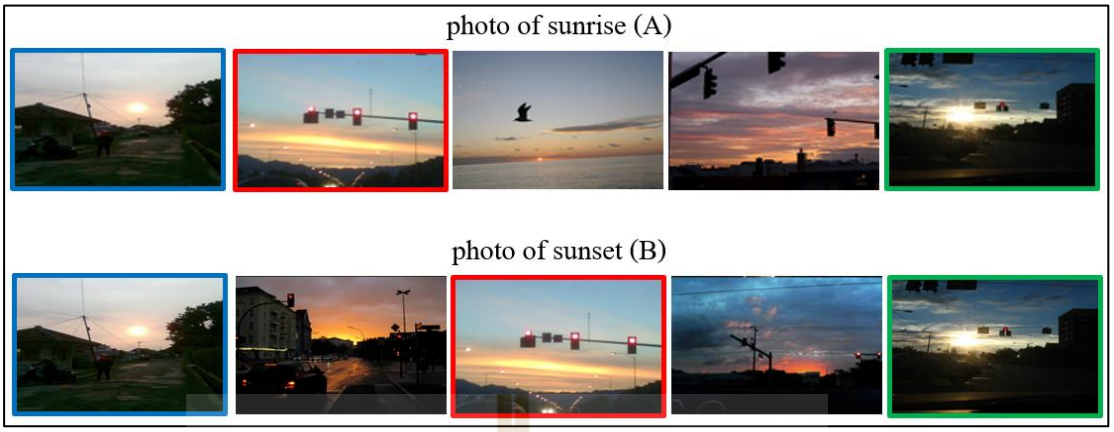
2) photo of a baby				
				
n) ...				
...	...	...	...	...
300) photo of a tradition				
				

จากตารางที่ 4.5 ผลลัพธ์จากการค้นคืนรูปภาพด้วยชื่อรูปภาพและหมวดหมู่ในลักษณะป้ายกำกับของรูปภาพที่ผ่านการประเมินความถูกต้องโดยผู้เชี่ยวชาญ พบว่าค่าความแม่นยำที่ 3 อันดับจากกลุ่มผู้เชี่ยวชาญด้านคอมพิวเตอร์และเทคโนโลยี กลุ่มผู้เชี่ยวชาญด้านการค้นคืนสารสนเทศ และกลุ่มผู้ใช้ทั่วไปอยู่ในระดับดีมาก ค่าเฉลี่ยเท่ากับ 0.943, 0.926 และ 0.853 ตามลำดับ ดังตารางที่ 4.6 อันแสดงถึงตัวแบบมีประสิทธิภาพการค้นคืนรูปภาพอยู่ในระดับสูงเทียบเท่ากับการค้นคืนรูปภาพด้วยข้อความ จากการเปรียบเทียบคำสำคัญ (Keyword) กับคุณลักษณะของรูปภาพที่ถูกเก็บไว้ในลักษณะของเมทาดาตา (Metadata) ของรูปภาพและใช้ดัชนีแบบหลายมิติในการค้นคืนรูปภาพ

ตารางที่ 4.6 ผลการประเมินประสิทธิภาพการค้นคืนรูปภาพด้วยชื่อรูปภาพและหมวดหมู่ในลักษณะป้ายกำกับของรูปภาพโดยจำแนกจากกลุ่มผู้เชี่ยวชาญ

ผู้เชี่ยวชาญ		**1) กลุ่มผู้เชี่ยวชาญด้านคอมพิวเตอร์และเทคโนโลยี**			**2) กลุ่มผู้เชี่ยวชาญด้านการค้นคืนสารสนเทศ**			**3) กลุ่มผู้ใช้ทั่วไป**		
การประเมินประสิทธิภาพการค้นคืนรูปภาพ		Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure
จำนวนรูปภาพจากการค้นคืน	**k@3**	0.943	0.609	0.740	0.926	0.595	0.854	0.853	0.604	0.694
	k@5	0.930	0.727	0.816	0.974	0.730	0.847	0.900	0.721	0.771
	k@10	0.903	0.825	0.868	0.943	0.836	0.848	0.875	0.839	0.833
	k@15	0.901	0.880	0.890	0.939	0.882	0.830	0.867	0.886	0.838
	k@20	0.892	1.00	0.943	0.930	1.00	0.822	0.862	1.00	0.865

นอกจากนี้ ผลการสังเกตและสัมภาษณ์ผู้เชี่ยวชาญในการประเมินความถูกต้อง พบว่า ยังมีข้อจำกัดในการค้นคืนรูปภาพที่มีความกำกวมเช่นเดียวกับมนุษย์ เช่น ข้อความค้นหา “photo of sunset” กับ “photo of sunrise” ดังรูปที่ 4.15 ที่มีผลลัพธ์ซ้ำกันถึง 3 จาก 5 รูปภาพ ซึ่งยากที่จะแยกความแตกต่างของรูปภาพได้ หากปราศจากองค์ประกอบอื่น ๆ มาเกี่ยวข้อง

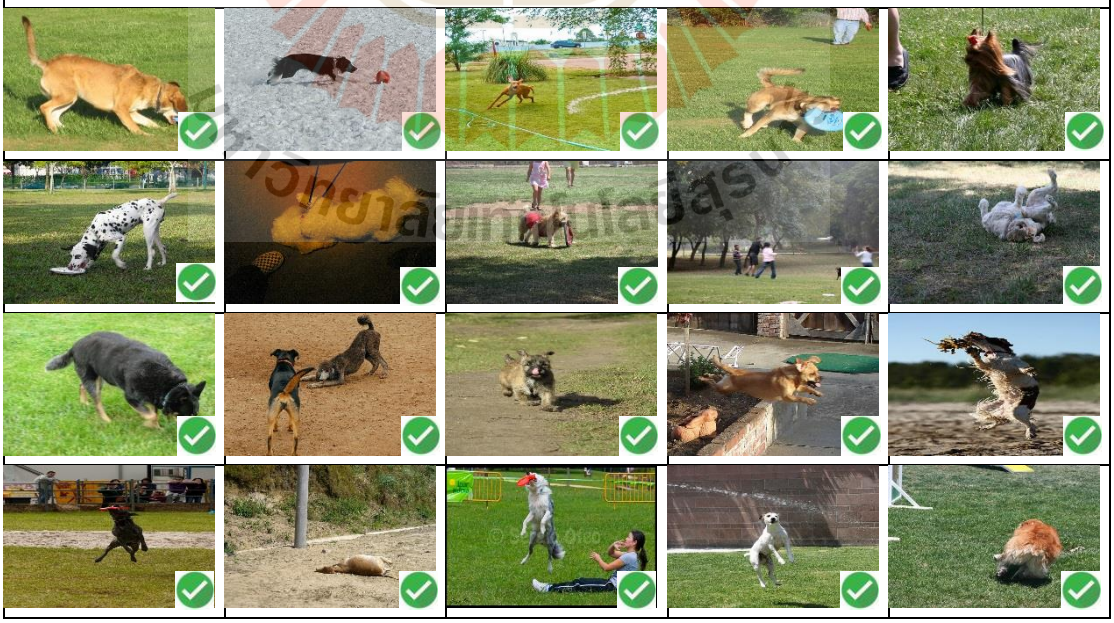


รูปที่ 4.15 ข้อความค้นหา ระหว่าง “photo of sunset” กับ “photo of sunrise”

4.2.1.2 ผลการประเมินประสิทธิภาพการค้นคืนรูปภาพด้วยค่าความแม่นยำที่ k อันดับ ค่าความครบถ้วนที่ k อันดับ และค่าประสิทธิภาพโดยรวมที่ k อันดับ ภายใต้เงื่อนไขข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของภาพที่เปรียบได้กับการค้นคืนรูปภาพเชิงเนื้อหา เช่น คน สัตว์ หรือสิ่งของที่เชื่อมโยงไปยังวัตถุ กิจกรรม หรือเหตุการณ์ที่อธิบายเนื้อหาของรูปภาพ

ตารางที่ 4.7 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพด้วยข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพ โดยผู้เชี่ยวชาญท่านที่ 1

1) the dog running on a ground



ตารางที่ 4.7 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพด้วยข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพ โดยผู้เชี่ยวชาญท่านที่ 1 (ต่อ)

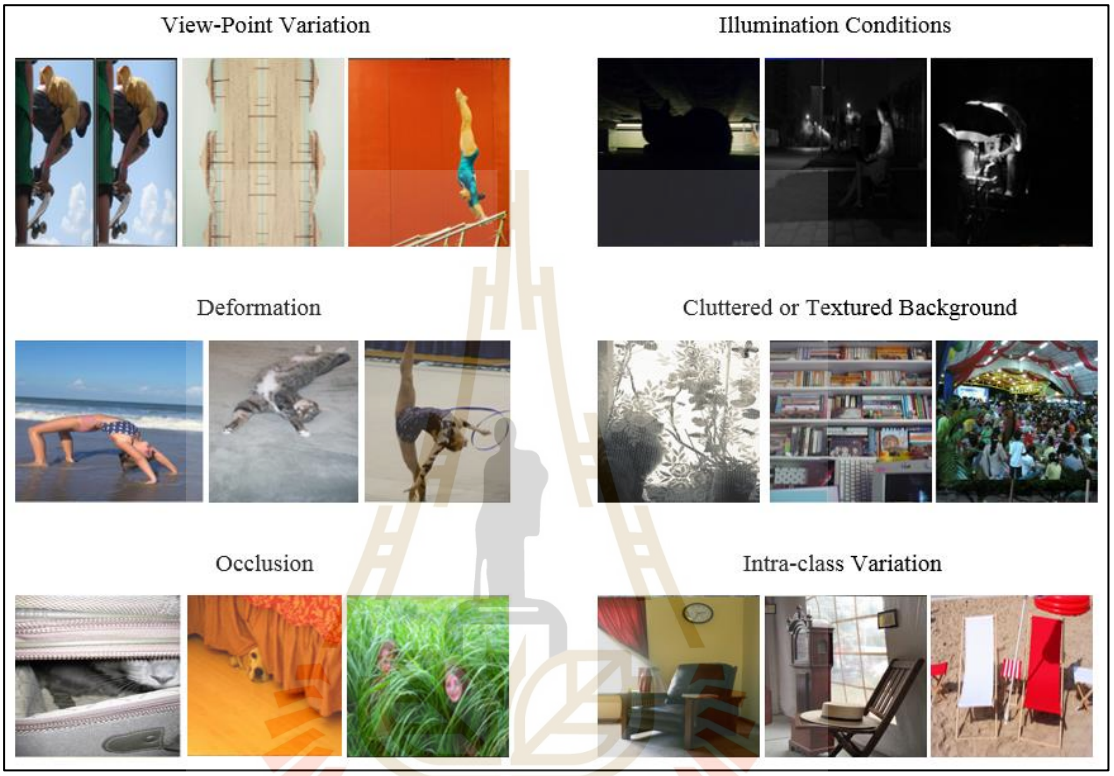
2) the baby sleeps peacefully in the bed				
				
n) ...				
...	...	...	...	...
300) we enjoy our family traditions every year				
				

จากตารางที่ 4.7 ผลลัพธ์จากการค้นคืนรูปภาพด้วยข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพที่ผ่านการประเมินความถูกต้องโดยผู้เชี่ยวชาญ พบว่าค่าความแม่นยำที่ 3 อันดับจากกลุ่มผู้เชี่ยวชาญด้านคอมพิวเตอร์และเทคโนโลยี กลุ่มผู้เชี่ยวชาญด้านการค้นคืนสารสนเทศ และกลุ่มผู้ใช้ทั่วไปอยู่ในระดับดีมาก ค่าเฉลี่ยเท่ากับ 0.827, 0.883 และ 0.883 ตามลำดับ ดังตารางที่ 4.8 อันแสดงถึงตัวแบบมีประสิทธิภาพการค้นคืนรูปภาพในระดับสูงเทียบเท่ากับการค้นคืนรูปภาพเชิงเนื้อหา โดยการเปรียบเทียบระหว่างคุณลักษณะข้อความค้นหากับคุณลักษณะรูปภาพ ซึ่งอยู่ในรูปแบบเวกเตอร์จากการใช้เทคนิคและวิธีการต่าง ๆ ในการประมวลผลเพื่อดึงเอาคุณลักษณะของรูปภาพ (Visual features extraction) ออกมาสร้างดัชนีจากคุณลักษณะของรูปภาพ (Multidimensional indexing) และค้นคืนรูปภาพโดยใช้คอนเทนท์หรือเนื้อหาหลักของรูปภาพ เช่น วัตถุ กิจกรรม ฉาก หรือเหตุการณ์

ตารางที่ 4.8 ผลการประเมินประสิทธิภาพการค้นคืนรูปภาพด้วยข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพโดยจำแนกจากกลุ่มผู้เชี่ยวชาญ

ผู้เชี่ยวชาญ		**1) กลุ่มผู้เชี่ยวชาญด้านคอมพิวเตอร์และเทคโนโลยี**			**2) กลุ่มผู้เชี่ยวชาญด้านการค้นคืนสารสนเทศ**			**3) กลุ่มผู้ใช้ทั่วไป**		
การประเมินประสิทธิภาพการค้นคืนรูปภาพ		Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure
จำนวนรูปภาพจากการค้นคืน	**k@3**	0.827	0.602	0.725	0.883	0.604	0.717	0.883	0.600	0.724
	k@5	0.824	0.705	0.835	0.872	0.699	0.776	0.884	0.714	0.780
	k@10	0.824	0.837	0.886	0.874	0.832	0.852	0.878	0.835	0.849
	k@15	0.803	0.884	0.910	0.859	0.888	0.873	0.865	0.890	0.851
	k@20	0.802	1.000	0.964	0.844	1.000	0.915	0.855	1.000	0.881

เมื่อพิจารณาในรายละเอียดของผลลัพธ์ที่ไม่ถูกต้อง พบว่า ความแปรปรวนของรูปภาพเป็นอุปสรรคสำคัญที่ทำให้ตัวแบบจำแนกข้อมูลภาพได้ไม่ดีเท่าที่ควร (Felzenszwalb, Girshick, Ramanan and McAllester, 2010) ดังรูปที่ 4.16



รูปที่ 4.16 ตัวอย่างรูปภาพที่มีความแปรปรวนที่พบในการจำแนกข้อมูล

จากรูปที่ 4.6 ความแปรปรวนที่พบในการจำแนกข้อมูล ประกอบด้วย 1) ความแปรปรวนของมุมมอง (View-point variation) ที่วัตถุเดียวกันแต่สามารถวางแนว หรือหมุนได้หลายมิติตามวิธีการถ่ายภาพและการจับภาพวัตถุ 2) การเปลี่ยนรูป (Deformation) การกระทำต่อวัตถุให้เปลี่ยนรูปร่างหรือบิดเบี้ยวส่งผลให้วัตถุเปลี่ยนลักษณะจากเดิมอย่างที่ไม่ควรจะเป็น 3) วัตถุบางส่วนถูกบดบัง (Occlusion) หรือซ่อนอยู่หลังวัตถุอื่น ๆ ส่งผลให้ประสิทธิภาพการจำแนกจะลดลงเรื่อย ๆ ตามขนาดของพื้นที่ที่บดบังวัตถุ 4) ความสว่าง (Illumination conditions) เนื่องจากรูปภาพนั้นจะมีค่าของแสงแตกต่างกัน ดังนั้นการจำแนกรูปภาพที่มีแสงสว่างน้อย หรือวัตถุในที่มืดนั้นจะมีประสิทธิภาพน้อยกว่าวัตถุในแสงสว่างปกติ 5) พื้นหลังยุ่งเหยิง (Cluttered or textured background) จากการที่วัตถุมีสีหรือลวดลายกลมกลืนไปกับพื้นหลัง หรือมีวัตถุจำนวนมากในภาพ ทำให้ยากต่อการจำแนกวัตถุ เนื่องจากสิ่งรบกวนมากเกินไป และ 6) ความแปรปรวนภายในคลาส (Intra-class variation) ที่มีวัตถุหลายประเภทที่ถูกจัดอยู่ในคลาสเดียวกัน

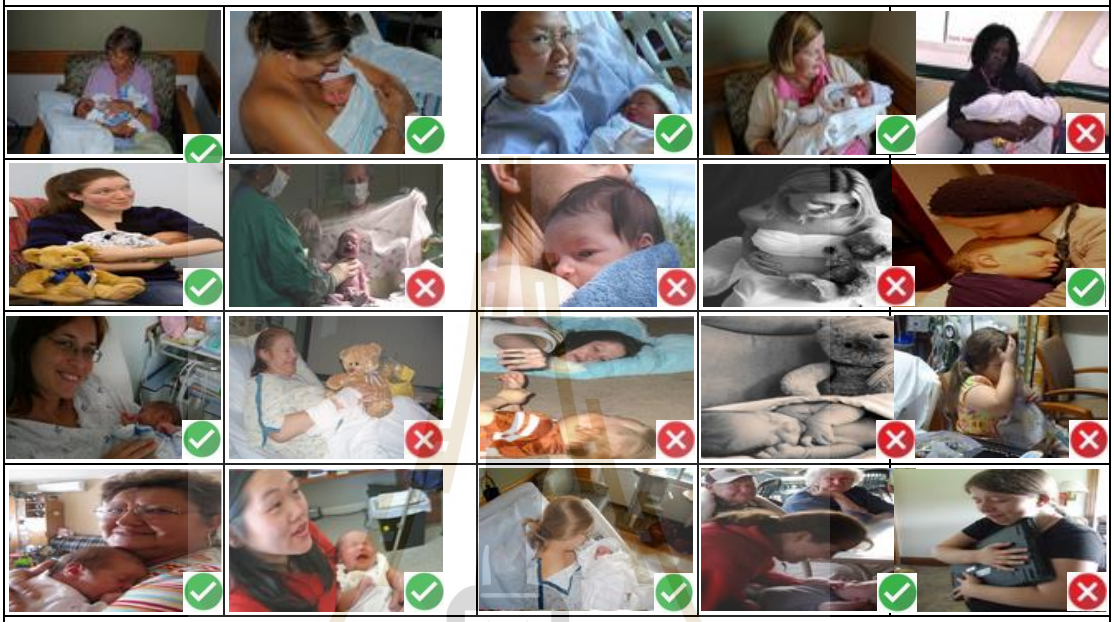
จึงจำเป็นต้องลดความแตกต่างภายในคลาส ขณะเดียวกันก็ขยายความแตกต่างระหว่างคลาส (Inter-class variation) เพื่อเพิ่มประสิทธิภาพการจำแนก

4.2.1.3 ผลการประเมินประสิทธิภาพการค้นคืนรูปภาพด้วยค่าความแม่นยำที่ k อันดับ ค่าความครบถ้วนที่ k อันดับ และค่าประสิทธิภาพโดยรวมที่ k อันดับภายใต้เงื่อนไขข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพที่เปรียบได้กับการค้นคืนรูปภาพเชิงความหมายด้วยความหมายของคุณลักษณะของรูปภาพที่เป็นนามธรรมที่เป็นได้ทั้งความคิด ความเห็น หรือข้อความที่อ้างถึงปรากฏการณ์ที่ก่อให้เกิดเหตุการณ์ที่เป็นรูปธรรม โดยนามธรรมคือกระบวนการคิดที่ไม่ได้ก่อเกิดจากตัวตนของวัตถุนั้นจริงและไม่มีรูปร่าง แต่นามธรรมเกิดมาจากความคิดและอารมณ์เข้าสัมผัสปรุงแต่งด้วยจิตก่อเกิดเป็นความหมายอย่างอารมณ์และความรู้สึกให้คอมพิวเตอร์สามารถเข้าใจความหมายของรูปภาพได้เช่นเดียวกับการรับรู้ของมนุษย์ ดังตารางที่ 4.9

ตารางที่ 4.9 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพด้วยข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพที่ผ่านการประเมินความถูกต้องโดยผู้เชี่ยวชาญท่านที่ 1

1) the dog wagged its tail happily

ตารางที่ 4.9 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพด้วยข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพที่ผ่านการประเมินความถูกต้องโดยผู้เชี่ยวชาญ (ต่อ)

2) She experienced a profound sense of joy as she held her newborn baby in her arms

n) ...
... ...
300) family gathers to celebrate cultural traditions with love and joy


จากตารางที่ 4.9 ผลลัพธ์จากการค้นคืนรูปภาพด้วยข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพที่ผ่านการประเมินความถูกต้องโดยผู้เชี่ยวชาญ พบว่า ค่าความแม่นยำที่ 3 อันดับจากกลุ่มผู้เชี่ยวชาญด้านคอมพิวเตอร์และเทคโนโลยี กลุ่มผู้เชี่ยวชาญด้านการค้นคืนสารสนเทศ และกลุ่มผู้ใช้ทั่วไปอยู่ในระดับดีมาก ค่าเฉลี่ยเท่ากับ 0.833, 0.883 และ 0.773 ตามลำดับ ดังตารางที่ 4.10 อันแสดงถึงตัวแบบมีประสิทธิภาพการค้นคืนรูปภาพอยู่ในเกณฑ์ดี และอยู่ในระดับที่ยอมรับในการค้นคืนรูปภาพเชิงความหมายจากการฝึกฝนตัวแบบโครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า เพื่อจำแนกประเภทข้อมูลรูปภาพเชิงความหมายจากการเรียนรู้ความสัมพันธ์ระหว่างรูปภาพกับข้อความที่อยู่ในลักษณะข้อความภาษาธรรมชาติ แทนที่จะเรียนรู้ภายใต้คลาสที่จำกัดอย่างสุนัขหรือแมวเช่นเดิม ดังนั้นเมื่อตัวแบบได้รับการฝึกฝนจากข้อความทั้งประโยค จะส่งผลให้ตัวแบบสามารถเรียนรู้สิ่งต่าง ๆ ได้มากขึ้น และสามารถค้นหารูปแบบความสัมพันธ์บางอย่างระหว่างรูปภาพกับข้อความด้วยการเรียนรู้ด้วยตนเอง ภายใต้ข้อความค้นหาในรูปแบบภาษาธรรมชาติที่ยึดโยงกับความหมายของรูปภาพแทนที่จะยึดตามหลักไวยากรณ์ของภาษา

ตารางที่ 4.10 ผลการประเมินประสิทธิภาพการค้นคืนรูปภาพด้วยข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพโดยจำแนกจากกลุ่มผู้เชี่ยวชาญ

ผู้เชี่ยวชาญ		**1) กลุ่มผู้เชี่ยวชาญด้านคอมพิวเตอร์และเทคโนโลยี**			**2) กลุ่มผู้เชี่ยวชาญด้านการค้นคืนสารสนเทศ**			**3) กลุ่มผู้ใช้ทั่วไป**		
การประเมินประสิทธิภาพการค้นคืนรูปภาพ		Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure
จำนวนรูปภาพจากการค้นคืน	**k@3**	0.833	0.595	0.707	0.883	0.613	0.714	0.773	0.601	0.676
	k@5	0.840	0.712	0.801	0.864	0.711	0.790	0.772	0.712	0.741
	k@10	0.818	0.848	0.857	0.850	0.847	0.856	0.788	0.794	0.791
	k@15	0.788	0.895	0.876	0.813	0.892	0.877	0.811	0.907	0.856
	k@20	0.762	1.000	0.926	0.787	1.000	0.922	0.764	1.000	0.866

เมื่อพิจารณาในรายละเอียด พบว่า การค้นคืนรูปภาพเชิงความหมายแม้จะมีค่าความถูกต้องอยู่ในระดับดี แต่แนวคิดระดับสูงที่เป็นนามธรรมนั้นยากที่จะประเมินผลเพียงถูกต้องหรือไม่ถูกต้องเท่านั้น ดังนั้นการแปลความหมายจึงเป็นไปตามหลักการรับรู้ของมนุษย์ที่เกิดจากประสบการณ์หรือความรู้เดิมของแต่ละบุคคล ดังตารางที่ 4.11 ตัวอย่างผลลัพธ์การค้นคืนรูปภาพด้วยข้อความค้นหา “she experienced a profound sense of joy as she held her newborn baby in her arms” จำนวน 5 รูปภาพ

ตารางที่ 4.11 ตัวอย่างผลการประเมินความถูกต้องของผลลัพธ์การค้นคืนรูปภาพด้วยแนวคิดระดับสูงโดยผู้เชี่ยวชาญ

ลำดับการค้นคืนรูปภาพ			1	2	3	4	5
Concrete Concept	Object	she, her newborn baby					
	Activities	held					
	Event	in her arms					
Abstract Concept	Feelings	she experienced a profound					
	Emotion	sense of joy					

จากตารางที่ 4.11 จะเห็นได้ว่า แนวคิดระดับสูงที่เป็นนามธรรมนั้นการประเมินความถูกต้องนั้นขึ้นอยู่กับการตีความหมายตามหลักการรับรู้ของมนุษย์ อันจะเห็นได้จากรูปภาพที่ 1 ถึง 5 นั้นเป็นภาพหญิงสาวอุ้มทารกไว้ในอ้อมแขนทั้งสิ้น แต่เมื่อประเมินคุณลักษณะเชิงความหมายนั้น พบว่า รูปภาพที่ 5 นั้นไม่ถูกต้อง เนื่องจากเป็นภาพที่หญิงสาวอุ้มทารกไว้ในอ้อมแขนนั้นกำลังหลับ แตกต่างจากอีก 4 รูปภาพที่เหลือที่ปรากฏภาพหญิงสาวยิ้มขณะอุ้มทารกไว้ในอ้อมแขน จะเห็นได้ว่าการสกัดคุณลักษณะรูปภาพตามแนวคิดของสถาปัตยกรรม ViT จากการ

เปรียบเทียบความสัมพันธ์ระหว่างแพตช์เพื่อคำนวณหาผังคุณลักษณะในการหาความสัมพันธ์กับบริเวณทั้งภาพ จึงใช้ทรัพยากรในการคำนวณมากกว่าการสกัดคุณลักษณะของรูปภาพทีละแพตช์เพียงอย่างเดียว อย่างไรก็ดี การสกัดคุณลักษณะของรูปภาพทีละแพตช์นั้นก็มีประสิทธิภาพสูงกว่าหากทำงานบนชุดข้อมูลที่มีจำนวนไม่มากนัก ในทางกลับกันสถาปัตยกรรม ViT จะให้ความแม่นยำสูงกว่าหากชุดข้อมูลมีจำนวนมากพอ (Liu, Sun and Zhou, 2022)

4.2.2 ผลการประเมินประสิทธิภาพการค้นคืนรูปภาพดิจิทัลเชิงความหมายด้วยมาตรวัดความเกี่ยวข้องแบบให้คะแนนด้วยค่า NDCG ที่ k อันดับ

การเปรียบเทียบผลค้นคืนรูปภาพระหว่างตัวแบบ CLIP ดั้งเดิมกับการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนัก กำหนดสภาพแวดล้อมการทำงานดังตารางที่ 4.12

ตารางที่ 4.12 การเปรียบเทียบสภาพแวดล้อมระหว่างการค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิมกับการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนัก

ค่าที่กำหนด	ตัวแบบ CLIP ดั้งเดิม	ตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนัก
image_path	image_path	image_path
captions_path	captions_path	captions_path
device	torch.device(“cuda” if torch.cuda.is_available() else “cpu”)	torch.device(“cuda” if torch.cuda.is_available() else “cpu”)
model_name	resnet50	vit-b/32
image_embedding	2048	2048
text_encoder_model	GPT-2	distilbert-base-uncased
text_embedding	768	768
text_tokenizer	GPT-2	distilbert-base-uncased
max_length	200	200
temperature	1	1
Image_size	224	384
num_projection_layers	1	1
projection_dim	256	512

การศึกษานี้ได้มีการเปรียบเทียบระหว่างตัวแบบ CLIP ดั้งเดิมและตัวแบบที่ผ่านการปรับปรุง โดยมีความแตกต่างด้านพารามิเตอร์หลายประการที่ส่งผลต่อประสิทธิภาพของการฝังข้อมูล และการจับคู่ภาพกับข้อความอย่างมีนัยสำคัญ หนึ่งในความเปลี่ยนแปลงหลักคือโครงสร้างของตัวเข้ารหัสภาพและข้อความ โดยตัวแบบ CLIP ดั้งเดิมใช้ ResNet-50 สำหรับการเข้ารหัสภาพและ GPT-2 สำหรับการเข้ารหัสข้อความ ในขณะที่ตัวแบบที่ได้รับการปรับปรุงเปลี่ยนไปใช้ ViT-B/32 และ DistilBERT ตามลำดับ พร้อมทั้งใช้ Tokenizer ที่สอดคล้องกัน นอกจากนี้ ขนาดของภาพที่ได้รับการปรับจาก 224 * 224 พิกเซล เป็น 384 * 384 พิกเซล เพื่อเพิ่มความสามารถในการจับรายละเอียดเชิงพื้นที่ โดยเฉพาะเมื่อใช้ร่วมกับโครงข่าย ViT ซึ่งแบ่งภาพเป็นแพตช์เพื่อประมวลผลขณะเดียวกัน ขนาดของ Projection dimension ได้เพิ่มขึ้นจาก 256 เป็น 512 เพื่อขยายความสามารถในการแยกแยะและจับความสัมพันธ์เชิงลึกของภาพและข้อความได้ดียิ่งขึ้น และกำหนดค่า Temperature ในการคำนวณความคล้ายคลึงของเวกเตอร์เท่ากันในทั้งสองตัวแบบ เพื่อให้การเปรียบเทียบผลลัพธ์มีความเป็นธรรมและสามารถวัดประสิทธิภาพได้อย่างแม่นยำ

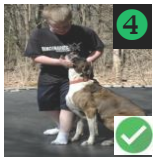
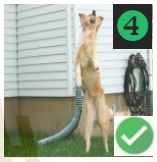
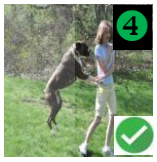
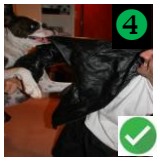
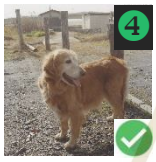
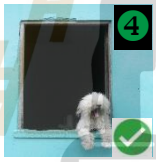
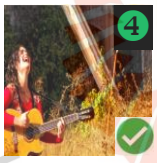
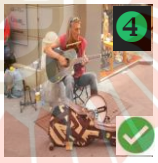
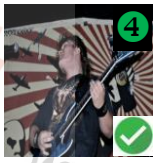
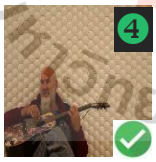
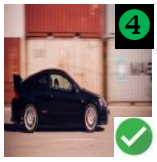
แม้โมเดลทั้งสองจะใช้โครงสร้างพื้นฐานเดียวกัน แต่การปรับค่าน้ำหนักทำให้โมเดลมีความสามารถในการเรียนรู้รายละเอียดเชิงโดเมนที่จำเพาะเจาะจงมากขึ้น เช่น คุณลักษณะบางประการในภาพที่สัมพันธ์กับชุดข้อมูลฝึกฝน หรือการลดความคลาดเคลื่อนในการจับคู่เวกเตอร์คุณลักษณะรูปภาพและคุณลักษณะข้อความค้นหาเชิงความหมายได้ดีขึ้น (Radford et al., 2021) ส่งผลให้คุณลักษณะที่ได้มีความละเอียดและแยกแยะความต่างได้ดีขึ้นในการค้นคืนรูปภาพ อย่างไรก็ตาม โมเดลผ่านการปรับแต่งที่ถึงแม้ผลการวัดประสิทธิภาพอาจแสดงให้เห็นถึงความเปลี่ยนแปลงเพียงเล็กน้อย แต่ก็สามารถตีความได้ว่าตัวแบบที่ผ่านการปรับค่าน้ำหนักอาจมีศักยภาพสูงกว่าในแง่ของ การจับความสัมพันธ์เฉพาะด้าน โดยเฉพาะเมื่อนำไปประยุกต์ใช้ในงานที่แตกต่างจากโดเมนการฝึกดั้งเดิมของตัวแบบ CLIP ที่ผ่านการฝึกฝนล่วงหน้าในการค้นคืนรูปภาพ เชิงความหมาย ซึ่งตัวแบบดั้งเดิมอาจไม่มีข้อมูลพื้นฐานเพียงพอ

ผู้วิจัยคัดเลือกข้อความค้นหาที่มีค่าความแม่นยำที่ 5 ลำดับสูงที่สุด เพื่อประเมินผลความหมายของรูปภาพในแต่ละรายการบนชุดข้อมูล Flickr30k ด้วยค่า NDCG@k โดยกำหนดคะแนนความเกี่ยวข้องเป็นเกณฑ์ 5 ระดับ ประกอบด้วยตัวเลขจำนวนเต็ม ตั้งแต่ 0 ถึง 4 ได้แก่

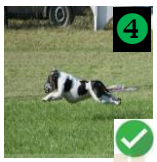
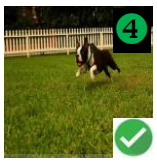
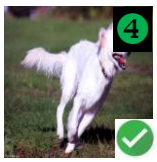
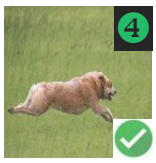
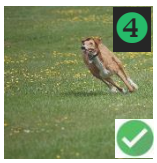
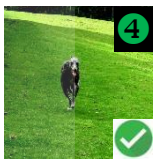
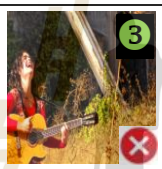
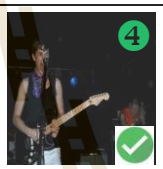
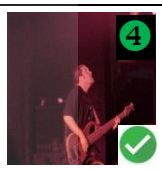
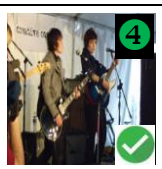
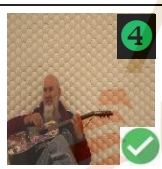
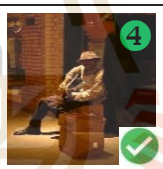
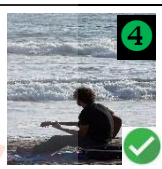
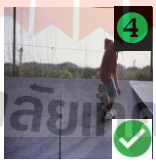
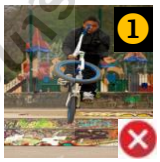
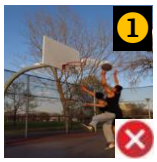
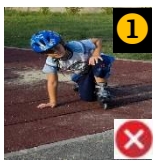
- 0 หมายถึง ไม่มีความเกี่ยวข้องเลย (Non-relevant item)
- 1 หมายถึง มีความเกี่ยวข้องน้อย (Relevant item)
- 2 หมายถึง มีความเกี่ยวข้องปานกลาง (Medium relevant item)
- 3 หมายถึง มีความเกี่ยวข้องมาก (High relevant item)
- 4 หมายถึง มีความเกี่ยวข้องมากที่สุด (Very high relevant item)

ผลการค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิมกับการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนัก จากข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ ในลักษณะป้ายกำกับของรูปภาพ ข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพ และข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพเงื่อนไขละ 100 รายการ รวมทั้งสิ้น 300 รายการ ดังตารางที่ 4.13

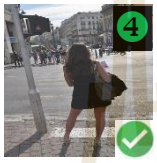
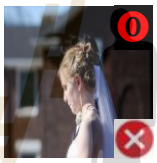
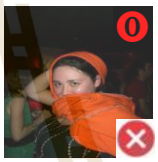
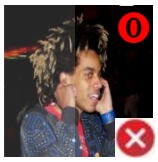
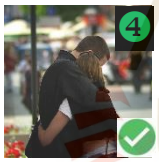
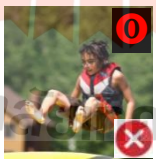
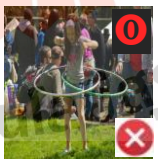
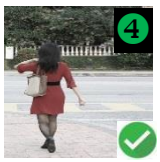
ตารางที่ 4.13 ตัวอย่างผลลัพธ์การค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิมกับการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนัก บนชุดข้อมูล Flickr30k

1) ข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ ในลักษณะป้ายกำกับของรูปภาพ					
1.1) photo of a dog					
ตัวแบบ CLIP ดั้งเดิม					
ตัวแบบ CLIP ที่ปรับค่าน้ำหนัก					
1.2) photo of a guitar					
ตัวแบบ CLIP ดั้งเดิม					
ตัวแบบ CLIP ที่ปรับค่าน้ำหนัก					
...					
1.100) photo of a car					
ตัวแบบ CLIP ดั้งเดิม					
ตัวแบบ CLIP ที่ปรับค่าน้ำหนัก					

ตารางที่ 4.13 ตัวอย่างผลลัพธ์การค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิมกับการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนัก บนชุดข้อมูล Flickr30k (ต่อ)

2) ข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพ					
2.1) the dog running on a grass					
ตัวแบบ CLIP ดั้งเดิม					
ตัวแบบ CLIP ที่ปรับค่าน้ำหนัก					
2.2) the man playing the guitar					
ตัวแบบ CLIP ดั้งเดิม					
ตัวแบบ CLIP ที่ปรับค่าน้ำหนัก					
...					
2.100) a boy jumping in the skateboard					
ตัวแบบ CLIP ดั้งเดิม					
ตัวแบบ CLIP ที่ปรับค่าน้ำหนัก					

ตารางที่ 4.13 ตัวอย่างผลลัพธ์การค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิมกับการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนัก บนชุดข้อมูล Flickr30k (ต่อ)

3) ข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ					
3.1) feel alone in the city					
ตัวแบบ CLIP ดั้งเดิม					
ตัวแบบ CLIP ที่ปรับค่าน้ำหนัก					
3.2) the feeling when you love someone					
ตัวแบบ CLIP ดั้งเดิม					
ตัวแบบ CLIP ที่ปรับค่าน้ำหนัก					
...					
3.100) move forward					
ตัวแบบ CLIP ดั้งเดิม					
ตัวแบบ CLIP ที่ปรับค่าน้ำหนัก					

จากตารางที่ 4.13 ผลการประเมินค่า NDCG@k ของผลลัพธ์การค้นคืนรูปภาพที่ลำดับ 1, 3 และ 5 ตามลำดับ เปรียบเทียบระหว่างการค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิมกับการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนัก เมื่อดำเนินการเสร็จสิ้นจะทำการประเมินอีกครั้งจนครบ 3 รอบ เพื่อให้ความคลาดเคลื่อนลดลงจนอยู่ในระดับที่ยอมรับได้ เมื่อนำผลการประเมินมาหาค่าเฉลี่ย ดังตารางที่ 4.14

ตารางที่ 4.14 ค่า NDCG ของการค้นคืนลำดับที่ 1, 3 และ 5 ระหว่างการค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิมกับการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนัก

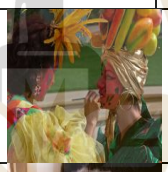
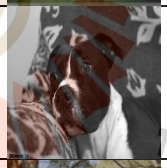
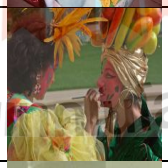
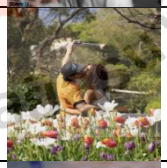
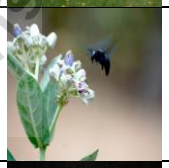
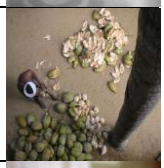
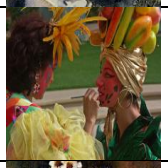
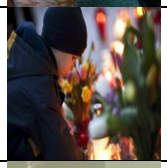
เงื่อนไขการค้นหา	**การค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิม**			**การค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนัก**		
	NDCG@1	NDCG@3	NDCG@5	NDCG@1	NDCG@3	NDCG@5
1) ข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ ในลักษณะป้ายกำกับของรูปภาพ	0.863	0.899	0.902	0.900	0.902	0.928
2) ข้อความค้นหาด้วยสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของ	0.808	0.885	0.890	0.885	0.887	0.903
3) ข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ	0.545	0.562	0.563	0.760	0.761	0.791

จากตารางที่ 4.14 ผลการประเมินผลลัพธ์การค้นคืนรูปภาพ เปรียบเทียบระหว่างการค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิมกับการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนัก ภายใต้ 3 เงื่อนไข คือ 1) ข้อความบรรยายรูปภาพในลักษณะป้ายกำกับของรูปภาพ

2) ข้อความบรรยายรูปภาพเกี่ยวกับแนวคิดระดับสูงของรูปภาพ และ 3) ข้อความบรรยายรูปภาพที่อธิบายความหมายเชิงคุณภาพของรูปภาพ เมื่อพิจารณาในรายละเอียด พบว่า

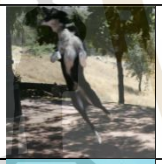
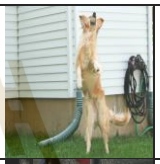
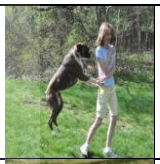
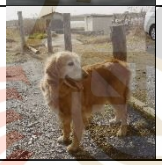
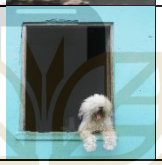
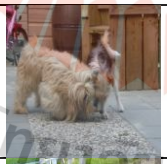
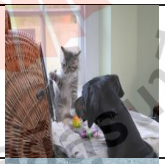
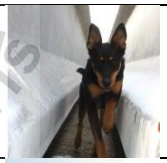
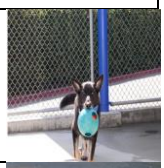
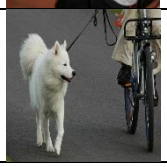
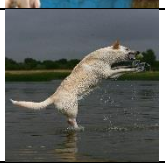
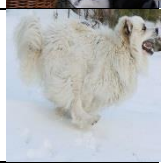
1) ผลการค้นคืนรูปภาพด้วยข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ในลักษณะป้ายกำกับของรูปภาพ พบว่า ค่า NDCG ที่ลำดับ 1, 3 และ 5 จากการค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิม เท่ากับ 0.863, 0.899 และ 0.902 ตามลำดับ ไม่แตกต่างกันมากนักกับค่า NDCG ที่ลำดับ 1, 3 และ 5 จากการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนักเท่ากับ 0.900, 0.902 และ 0.928 ตามลำดับ คิดเป็นเพิ่มขึ้นร้อยละ 4.29, 0.33 และ 2.88 ตามลำดับ แสดงถึงการกำหนดข้อความค้นหา ส่งผลอย่างมากต่อผลลัพธ์จากการค้นคืนรูปภาพ เนื่องจากหากกำหนดเพียงชื่อคลาส เช่น “daisy” ผลลัพธ์จากการค้นคืนจะไม่สูงมากนัก แต่หากเพิ่มบริบทของคำจำกัดความ เช่น “daisy flower” ผลลัพธ์จากการค้นคืนจะดีขึ้นเรื่อย ๆ และหากกำหนดเป็น “photo of a daisy flower” จะให้ผลลัพธ์จากการค้นคืนสูงที่สุด รายละเอียดดังตารางที่ 4.15

ตารางที่ 4.15 ตัวอย่างผลการเปรียบเทียบผลลัพธ์จากข้อความค้นหาจำนวน 5 รูปภาพ

ข้อความค้นหา		@1	@2	@3	@4	@5
daisy	ตัวแบบ CLIP แบบดั้งเดิม					
	ตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนัก					
daisy flower	ตัวแบบ CLIP แบบดั้งเดิม					
	ตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนัก					
photo of a daisy flower	ตัวแบบ CLIP แบบดั้งเดิม					
	ตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนัก					

2) ผลการค้นคืนรูปภาพด้วยข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพ พบว่า ค่า NDCG ที่ลำดับ 1, 3 และ 5 จากการค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิม เท่ากับ 0.808, 0.885 และ 0.890 ตามลำดับ ในขณะที่ค่า NDCG ที่ลำดับ 1, 3 และ 5 จากการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนัก นั้นเพิ่มขึ้นเป็น 0.885, 0.887 และ 0.903 ตามลำดับ คิดเป็นเพิ่มขึ้นร้อยละ 9.52, 0.23 และ 1.46 ตามลำดับ แสดงถึงประสิทธิภาพในการจำแนกวัตถุ (Recognizing common objects) และยังคงประสิทธิภาพแม้จะอยู่ในสถานการณ์ที่มีการนับจำนวนวัตถุในรูปภาพ ตลอดจนการจำแนกรายละเอียดเชิงลึก (Fine-grained classification) ดังตารางที่ 4.16 เช่น สายพันธุ์สุนัข สายพันธุ์ดอกไม้ หรือยี่ห้อรถยนต์ เป็นต้น

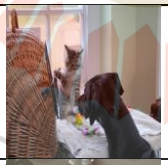
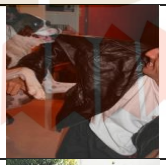
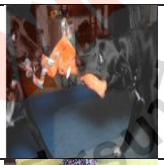
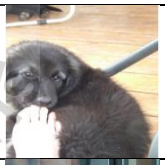
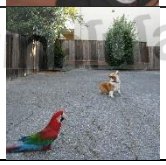
ตารางที่ 4.16 ตัวอย่างผลการเปรียบเทียบผลลัพธ์จากข้อความค้นหาแบบจำแนกรายละเอียดเชิงลึก

Search Query	@1	@2	@3	@4	@5
photo of a dog					
ตัวแบบ CLIP แบบดั้งเดิม					
ตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนัก					
photo of two dog					
ตัวแบบ CLIP แบบดั้งเดิม					
ตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนัก					
photo of a Siberian Husky					
ตัวแบบ CLIP แบบดั้งเดิม					
ตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนัก					

3) ผลการค้นคืนรูปภาพด้วยข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ พบว่า ค่า NDCG ที่ลำดับ 1, 3 และ 5 จากการค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิม เท่ากับ 0.545, 0.562 และ 0.563 ตามลำดับ ในขณะที่ค่า NDCG ที่ลำดับ 1, 3 และ 5 จากการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนัก เพิ่มขึ้นเป็น 0.760, 0.761 และ 0.791 ตามลำดับ โดยคิดเป็นเพิ่มขึ้นร้อยละ 39.45, 35.41 และ 40.50 ตามลำดับแม้ผลการประเมินนั้นจะเกินมาตรฐาน และอยู่ในระดับที่ยอมรับได้ แต่ผลลัพธ์จากการค้นคืนรูปภาพเชิงความหมายนั้นจึงมีความซับซ้อน (Complex) และมีความเป็นนามธรรม (Abstract) เข้ามาเกี่ยวข้อง ดังนั้นการแปลความหมายตามหลักการรับรู้ของมนุษย์ (Human perception) ด้วยประสบการณ์หรือความรู้เดิมจากการเรียนรู้ของแต่ละบุคคลจึงแตกต่างกัน

4) การค้นคืนรูปภาพนั้นยังไม่ทำงานไม่ดีนักหากใช้งานกับข้อความค้นหาในรูปประโยคปฏิเสธ เช่น หากข้อความบรรยายรูปภาพคือ “photo without a cat” หรือ “a picture containing no cat” ผลลัพธ์การค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิม รวมถึงการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนักจะไม่สามารถค้นคืนรูปภาพที่ถูกต้องได้ โดยตัวอย่างผลการเปรียบเทียบผลลัพธ์จากข้อความค้นหาแบบปฏิเสธ ดังตารางที่ 4.17

ตารางที่ 4.17 ตัวอย่างผลการเปรียบเทียบผลลัพธ์จากข้อความค้นหาแบบปฏิเสธ

Search Query	@1	@2	@3	@4	@5
photo without a cat					
ตัวแบบ CLIP แบบดั้งเดิม					
ตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนัก					

5) ผลการค้นคืนกับประเภทรูปภาพที่ไม่มีอยู่ในชุดข้อมูลเรียนรู้ล่วงหน้า (Pre-training dataset) เช่น ภาพขาวดำ ภาพวาด หรือภาพการ์ตูนนั้น การค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิมจะให้ผลลัพธ์จากการค้นคืนรูปภาพจะค่อนข้างต่ำ (Poor generalization) และเป็นไปในทิศทางเดียวกับการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนัก

บทที่ 5
สรุปและข้อเสนอแนะ

การพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมาย โดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายโดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า และเพื่อประเมินประสิทธิภาพการค้นคืนรูปภาพดิจิทัลเชิงความหมาย เนื้อหาในบทนี้ประกอบด้วยสรุปผลการวิจัย การประยุกต์ผลการวิจัย และข้อเสนอแนะในการวิจัยครั้งต่อไป มีรายละเอียดดังนี้

5.1 สรุปผลการวิจัย

งานวิจัยนี้แบ่งสรุปผลการวิจัยออกเป็น 2 ส่วน ตามวัตถุประสงค์ ได้แก่ 1) เพื่อออกแบบและพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายโดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า และ 2) เพื่อประเมินประสิทธิภาพการค้นคืนรูปภาพดิจิทัลเชิงความหมาย มีรายละเอียดดังนี้

5.1.1 สรุปผลการพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายโดยประยุกต์ใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า

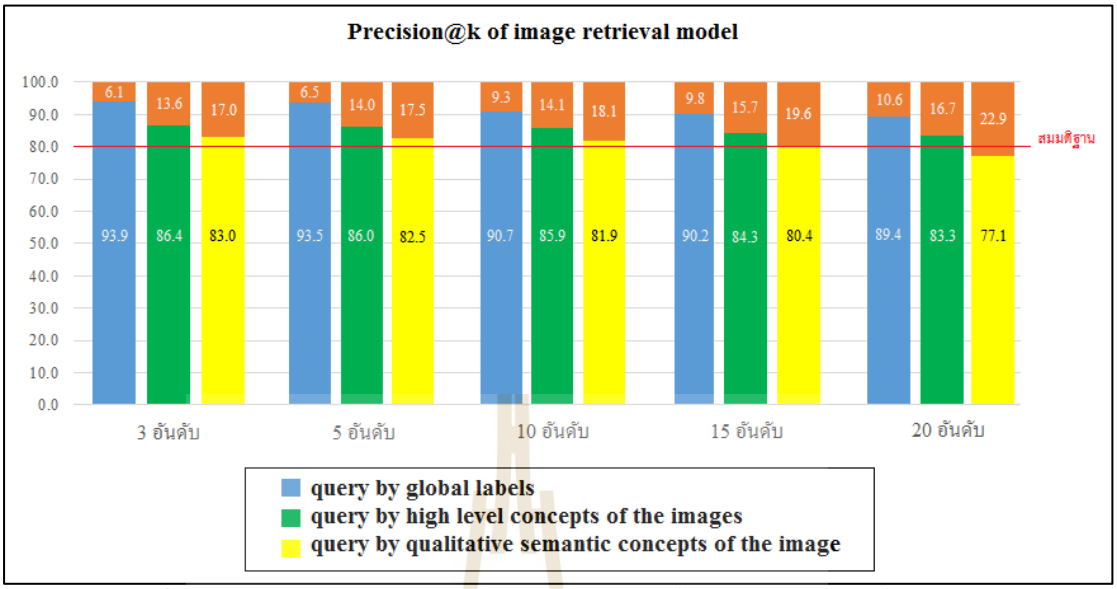
การพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายด้วยการเขียนโปรแกรมภาษาไพทอนบน Google Colaboratory ประกอบด้วย 3 มอดูล ได้แก่ 1) มอดูลการสร้างชุดคำอธิบายรูปภาพ โดยประยุกต์ใช้ตัวแบบ CLIP ที่ถูกฝึกฝนล่วงหน้ารุ่น ViT-B/32 สำหรับฝึกฝนตัวเข้ารหัสรูปภาพ และตัวเข้ารหัสข้อความบนพื้นที่การฝังหลายรูปแบบ จากการคำนวณหาค่าความผิดพลาดเปรียบเทียบระหว่างผลลัพธ์การพยากรณ์ป้ายกำกับจากตัวแบบกับผลลัพธ์ที่ถูกประเมินความหมายเชิงคุณภาพของรูปภาพโดยผู้เชี่ยวชาญจำนวน 10 ท่านกับรูปภาพจำนวน 400 รูปภาพที่สุ่มจากชุดข้อมูล จากนั้นนำค่าการสูญเสียมาปรับปรุงพารามิเตอร์ที่ใช้ในการเรียนรู้ของโมเดลด้วยการปรับแต่งด้วยตนเอง เพื่อแยกความแตกต่างของแต่ละคลาสด้วยค่าความคล้ายคลึงโคไซน์ของเวกเตอร์รูปภาพและข้อความบรรยายรูปภาพตามหลักการเรียนรู้แบบคอนทราสต์ จากการเปรียบเทียบความคล้ายคลึงเชิงความหมายให้ใกล้เคียงกับมนุษย์มากที่สุดก่อนจะเรียนรู้ซ้ำบนชุดข้อมูลที่กำหนดเอง เพื่อสกัดคุณลักษณะเชิงความหมายของรูปภาพบนชุดข้อมูล Flickr30k จำนวน 31,783 รูปภาพ และชุดข้อมูลที่ผู้วิจัยรวบรวมเองจำนวน 92,168 รูปภาพ รวมทั้งสิ้น 123,951 รูปภาพ เพื่อสร้างเป็นเวกเตอร์คุณลักษณะรูปภาพขนาด 2,048 แล้วจัดเก็บไว้ในชุดคำอธิบายรูปภาพ

2) มอดูลการประมวลผลข้อความค้นหาจากผู้ใช้ในรูปแบบภาษาธรรมชาติด้วยโมเดลภาษา DistilBERT ที่ถูกฝึกฝนล่วงหน้ามาปรับแต่งให้เหมาะสมสำหรับเข้ารหัสข้อความค้นหาภาษาธรรมชาติในรูปแบบภาษาอังกฤษจากผู้ใช้ เพื่อสร้างเป็นเวกเตอร์คุณลักษณะข้อความค้นหา ขนาด 768 ใช้ในการค้นคืนรูปภาพต่อไป 3) มอดูลการจับคู่เวกเตอร์คุณลักษณะรูปภาพและเวกเตอร์คุณลักษณะข้อความค้นหาบนพื้นที่การฝังหลายรูปแบบเพื่อแปลงเป็นเวกเตอร์ขนาด 256 และนำมาเปรียบเทียบค่าความคล้ายคลึงของเวกเตอร์ ก่อนจะเรียงลำดับตามความเกี่ยวข้องจากมากไปน้อย และแสดงเป็นผลลัพธ์การค้นคืนรูปภาพทั้งในบริบทของเนื้อหาและบริบทเชิงความหมายแก่ผู้ใช้

5.1.2 สรุปผลการประเมินประสิทธิภาพการค้นคืนรูปภาพดิจิทัลเชิงความหมายด้วยค่าความแม่นยำที่ k อันดับ ค่าความครบถ้วนที่ k อันดับ และค่าประสิทธิภาพโดยรวมที่ k อันดับ

การประเมินประสิทธิภาพการค้นคืนรูปภาพดิจิทัลเชิงความหมายภายใต้ข้อความค้นหา 3 เงื่อนไข ได้แก่ 1) ข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ในลักษณะป้ายกำกับของรูปภาพ 2) ข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพ และ 3) ข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ ค้นหาเงื่อนไขละ 100 รายการ รวมทั้งสิ้น 300 รายการ โดยผู้วิจัยระบุข้อความค้นหาผ่าน Google Colaboratory เมื่อการค้นคืนรูปภาพเสร็จสิ้นจะแสดงผลรูปภาพแต่ละครั้งเท่ากับ 20 รูปภาพ เมื่อสิ้นสุดการค้นคืนผลลัพธ์ใน แต่ละรายการผู้วิจัยประเมินความถูกต้องทีละรูปภาพ กำหนดให้ หากรูปภาพจากการค้นคืนถูกต้องตามข้อความค้นหา เท่ากับ 1 คะแนน ในทางกลับกันหากรูปภาพจากการค้นคืนไม่ถูกต้องหรือไม่ครบถ้วนตามข้อความค้นหาเท่ากับ 0 คะแนน มีรายละเอียดดังนี้

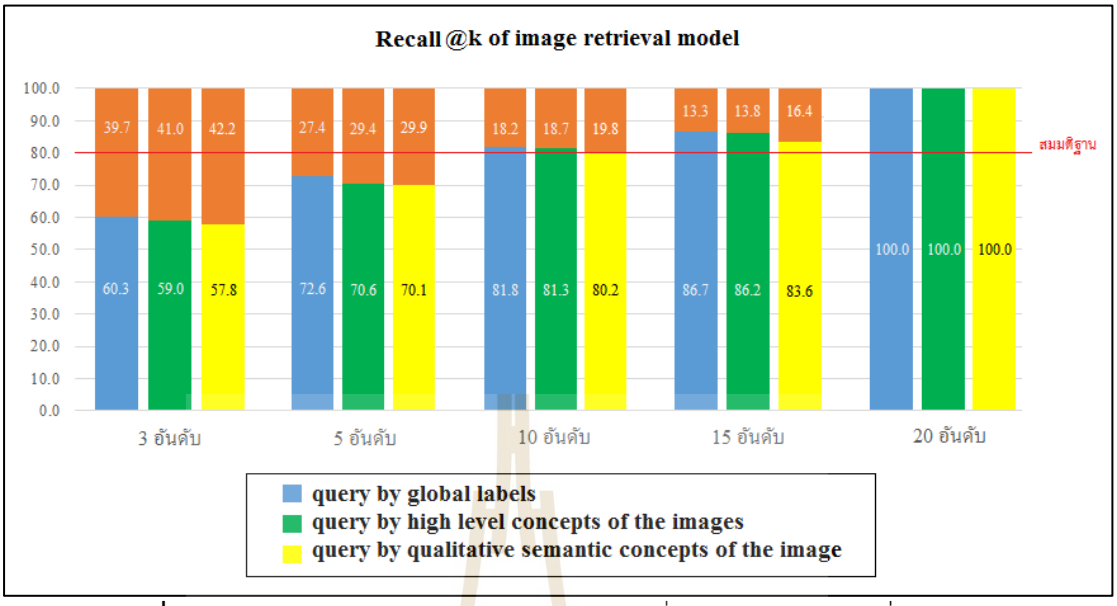
5.1.2.1 สรุปผลการประเมินประสิทธิภาพการค้นคืนรูปภาพดิจิทัลเชิงความหมายเมื่อพิจารณาค่าความแม่นยำที่ k อันดับ อันแสดงถึงประสิทธิภาพของระบบในการค้นคืนรูปภาพโดยดูจากอัตราส่วนของจำนวนรูปภาพที่ถูกต้องจากรูปภาพที่ถูกเลือกมาทั้งหมด แสดงถึงแบบจำลองมีความสามารถในการขจัดรูปภาพที่ไม่เกี่ยวข้อง พบว่า ผลลัพธ์การค้นคืนจำนวน 3, 5 และ 10 ลำดับแรกภายใต้เงื่อนไขข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ในลักษณะป้ายกำกับของรูปภาพคิดเป็นร้อยละ 93.9, 93.5 และ 90.7 ตามลำดับ อยู่ในระดับดีมาก โดยลักษณะข้อความบรรยายรูปภาพส่งผลอย่างมากต่อผลลัพธ์จากการค้นคืน เช่นเดียวกับข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพนั้นมีค่าความแม่นยำคิดเป็นร้อยละ 86.4, 86.0 และ 85.9 ตามลำดับ อยู่ในระดับดีมากแสดงถึงตัวแบบมีประสิทธิภาพสูงการจำแนกวัตถุภายในรูปภาพ แต่ยังมีประสิทธิภาพลดลงเมื่อมีการนับจำนวนวัตถุในภาพ ตลอดจนการจำแนกที่มีรายละเอียดเชิงลึก และข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ คิดเป็นร้อยละ 83.0, 82.5 และ 81.9 ตามลำดับ อยู่ในระดับดี



รูปที่ 5.1 ประสิทธิภาพการค้นคืนรูปภาพดิจิทัล เมื่อค่าความแม่นยำที่ k อันดับ เท่ากับ 3, 5, 10, 15 และ 20 ตามลำดับ

จากรูปที่ 5.1 ค่าความแม่นยำนั้นลดลงเรื่อย ๆ แปรผันตามจำนวนผลลัพธ์การค้นคืนรูปภาพที่เพิ่มมากขึ้น ดังรูปที่ 5.1 แสดงให้เห็นถึงข้อความที่อยู่ในรูปแบบของป้ายกำกับภาษาธรรมชาตินั้นถือว่าเป็นแนวคิดระดับสูง ดังนั้นประสบการณ์ของแต่ละบุคคลจึงส่งผลให้การประเมินนั้นแตกต่างกันตามหลักการรับรู้ของมนุษย์ อย่างไรก็ดี ผลการประเมินประสิทธิภาพการค้นคืนรูปภาพด้วยค่าความแม่นยำสูงกว่าสมมติฐานการวิจัยที่กำหนดไว้ที่ร้อยละ 80

5.1.2.2 สรุปผลการประเมินประสิทธิภาพการค้นคืนรูปภาพด้วยค่าความครบถ้วนที่ k อันดับ อันแสดงถึงประสิทธิภาพของการค้นคืนรูปภาพโดยดูจากอัตราส่วนจำนวนรูปภาพที่ถูกต้องที่เลือกมาต่อจำนวนรูปภาพที่ถูกต้องทั้งหมดที่อยู่ในชุดข้อมูล แสดงถึงแบบจำลองมีความสามารถในการดึงรูปภาพที่เกี่ยวข้อง พบว่า ผลการประเมินประสิทธิภาพการค้นคืนรูปภาพด้วยค่าความครบถ้วนจะค่อย ๆ เพิ่มขึ้นตามจำนวนผลลัพธ์ที่เพิ่มมากขึ้น และจะเข้าใกล้ 1 มากขึ้นเมื่อจำนวนผลลัพธ์เท่ากับ 10 และ 15 ลำดับ ภายใต้เงื่อนไขข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ในลักษณะป้ายกำกับของรูปภาพอยู่ในระดับดี คิดเป็นร้อยละ 81.8 และ 86.7 ตามลำดับ และภายใต้เงื่อนไขข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพนั้นมีค่าความครบถ้วนอยู่ในระดับดีเช่นกัน คิดเป็นร้อยละ 81.3 และ 86.2 ตามลำดับ เป็นไปในทิศทางเดียวกันกับข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพที่อยู่ในระดับดี คิดเป็นร้อยละ 80.2 และ 83.6 ตามลำดับ

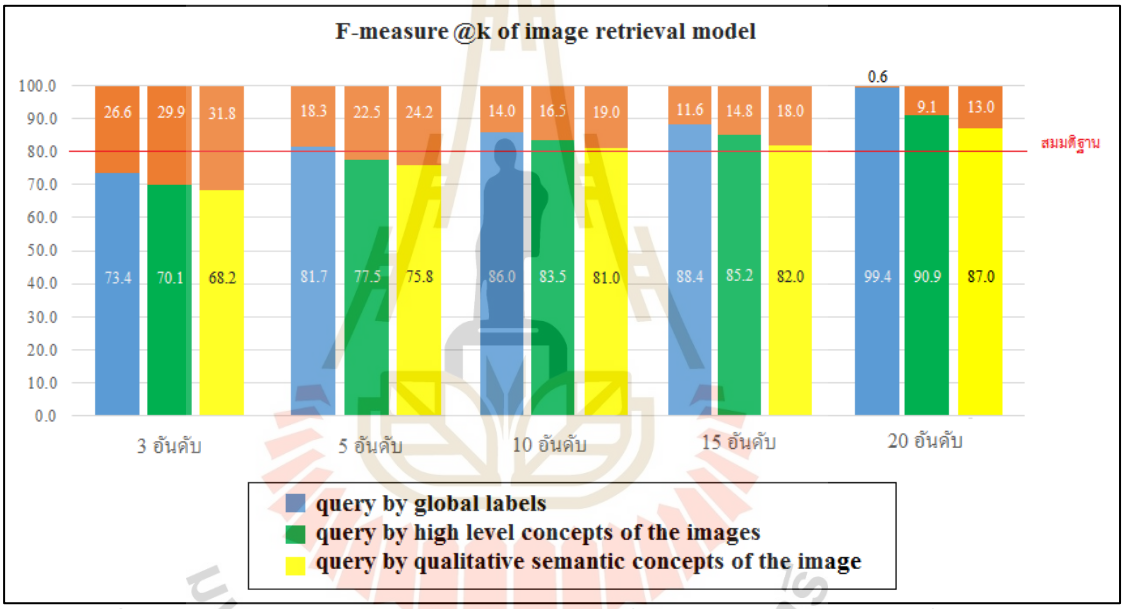


รูปที่ 5.2 ประสิทธิภาพการค้นคืนรูปภาพดิจิทัล เมื่อค่าความครบถ้วนที่ k อันดับ เท่ากับ 3, 5, 10, 15 และ 20 ตามลำดับ

จากรูปที่ 5.2 อย่างไรก็ดี ยังมีปัจจัยอื่นที่อาจส่งผลต่อค่าความถูกต้องและค่าความแม่นยำของแบบจำลองที่พัฒนาขึ้น ได้แก่ ปริมาณรูปภาพในฐานข้อมูล ซึ่งมีความเป็นได้น้อยมากที่จำนวนรูปภาพจะครอบคลุมความต้องการจากผู้ใช้ทั้งหมด อีกทั้งพฤติกรรมของผู้ใช้มักต้องการเพียงผลลัพธ์ที่ตนเองสนใจเท่านั้น ผู้ใช้จำนวนไม่น้อยที่ขาดความชำนาญในการใช้คำค้นหาหรือไม่สามารถกำหนดคำหลักหรือคำที่เฉพาะเจาะจงสำหรับการค้นหาที่เหมาะสมได้ ส่งผลให้เกิดเป็นผลลัพธ์รูปภาพจำนวนมากจากค้นคืนทั้งที่ตรงกับความต้องการและไม่ตรงกับความต้องการ อีกทั้งปริมาณผลลัพธ์ที่มากเกินไปอาจทำให้ผู้ใช้เสียเวลาในการคัดเลือกรูปภาพที่ตรงกับความต้องการอย่างแท้จริง อีกทั้งความต้องการของผู้ใช้ที่ไม่มีขีดจำกัดยังส่งผลต่อความคาดหวังต่อผลลัพธ์จากการค้นคืนอีกด้วย

5.1.2.3 สรุปผลการประเมินประสิทธิภาพการค้นคืนรูปภาพด้วยค่าประสิทธิภาพโดยรวมที่ k อันดับ เมื่อพิจารณาค่าความแม่นยำและค่าความครบถ้วนจะมีค่าอยู่ระหว่าง 0 กับ 1 หากการค้นคืนเอกสารนั้นได้ผลลัพธ์เป็นรูปภาพที่ตรงกับความต้องการทั้งหมดและไม่มีรูปภาพที่ไม่เกี่ยวข้องออกมาด้วยค่าความแม่นยำและค่าความครบถ้วนจะมีค่าเป็น 1 โดยทั่วไปค่าความแม่นยำและค่าความครบถ้วนจะมีความสัมพันธ์เป็นปฏิภาคผกผัน คือหากค่าความแม่นยำสูงขึ้นค่าความครบถ้วนก็จะมักลดลง และในทางตรงกันข้าม หากค่าความครบถ้วนสูงขึ้น ค่าความแม่นยำก็มักจะลดลง แต่ในบางกรณีค่าความแม่นยำและค่าความครบถ้วนอาจจะเพิ่มขึ้นหรือลดลงด้วยกันทั้งสองค่าก็ได้ ดังนั้นค่าประสิทธิภาพโดยรวมที่ k อันดับที่เกิดจากการเปรียบเทียบกันระหว่าง

ค่าความแม่นยำและค่าความครบถ้วน เพื่อเป็นมาตรวัดเดี่ยว ซึ่งเป็นสิ่งที่แสดงถึงประสิทธิภาพโดยรวมของแบบจำลองที่พัฒนาขึ้นที่สูงกว่าสมมติฐานการวิจัยที่กำหนดไว้ที่ร้อยละ 80 เมื่อจำนวนผลลัพธ์เท่ากับ 10 และ 15 ลำดับและจะสูงที่สุดเมื่อจำนวนผลลัพธ์เท่ากับ 20 อันดับ โดยเงื่อนไขข้อความค้นหาด้วย ชื่อรูปภาพและหมวดหมู่ในลักษณะป้ายกำกับของรูปภาพอยู่ในระดับดีมาก คิดเป็นร้อยละ 86.0, 88.4 และ 99.4 ตามลำดับ และภายใต้เงื่อนไขข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพนั้นมีค่าความครบถ้วนอยู่ในระดับดีเช่นกัน คิดเป็นร้อยละ 83.5, 85.2 และ 90.9 ตามลำดับ เป็นไปในทิศทางเดียวกันกับข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพที่อยู่ในระดับดี คิดเป็นร้อยละ 81.0, 82.0 และ 87.0 ตามลำดับ



รูปที่ 5.3 ประสิทธิภาพการค้นคืนรูปภาพดิจิทัล เมื่อค่าประสิทธิภาพโดยรวมที่ k อันดับเท่ากับ 3, 5, 10, 15 และ 20 ตามลำดับ

5.1.3 สรุปผลการประเมินประสิทธิภาพการค้นคืนรูปภาพดิจิทัลเชิงความหมายโดยจำแนกจากกลุ่มผู้เชี่ยวชาญ

งานวิจัยนี้แบ่งผู้เชี่ยวชาญออกเป็น 3 กลุ่ม ได้แก่ 1) กลุ่มผู้เชี่ยวชาญด้านคอมพิวเตอร์และเทคโนโลยี 2) กลุ่มผู้เชี่ยวชาญด้านการค้นคืนสารสนเทศ และ 3) กลุ่มผู้ใช้ทั่วไป เพื่อให้ผลการประเมินครอบคลุมทุกมิติ มีรายละเอียดดังนี้

5.1.3.1 สรุปผลการประเมินประสิทธิภาพการค้นคืนรูปภาพด้วยค่าความแม่นยำที่ k อันดับ ค่าความครบถ้วนที่ k อันดับ และค่าประสิทธิภาพโดยรวมที่ k อันดับภายใต้เงื่อนไขข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ในลักษณะป้ายกำกับของรูปภาพ พบว่า ค่าความแม่นยำที่ 3, 5

และ 10 อันดับจากกลุ่มผู้เชี่ยวชาญด้านคอมพิวเตอร์และเทคโนโลยี อยู่ในระดับดีมาก คิดเป็นร้อยละ 94.3, 93.0 และ 90.3 ตามลำดับ เช่นเดียวกับกลุ่มผู้เชี่ยวชาญด้านการค้นคืนสารสนเทศ คิดเป็นร้อยละ 92.6, 97.4 และ 94.3 ตามลำดับ และเป็นไปในทิศทางเดียวกันกับกลุ่มผู้ใช้ทั่วไปอยู่ในระดับดีมาก คิดเป็นร้อยละ 85.3, 90.0 และ 87.5 ตามลำดับ ซึ่งสูงกว่าสมมติฐานการวิจัยที่กำหนดไว้ที่ร้อยละ 80 แสดงถึงตัวแบบมีประสิทธิภาพการค้นคืนรูปภาพสูงเทียบเท่ากับการค้นคืนรูปภาพด้วยข้อความ จากการเปรียบเทียบคำสำคัญ (Keyword) กับคุณลักษณะของรูปภาพที่ถูกเก็บไว้ในลักษณะของเมทาดาตาของรูปภาพและใช้ดัชนีแบบหลายมิติในการค้นคืนรูปภาพ อย่างไรก็ดี การกำหนดข้อความค้นหาที่ส่งผลอย่างมากต่อความถูกต้องของผลการค้นคืนรูปภาพ นอกจากนี้ แบบจำลองยังมีข้อจำกัดในการค้นคืนรูปภาพที่มีความกำกวมเช่นเดียวกับมนุษย์ เช่น ข้อความค้นหาระหว่าง “photo of sunset” กับ “photo of sunrise” ซึ่งยากที่จะแยกความแตกต่างของรูปภาพได้ หากปราศจากองค์ประกอบอื่น ๆ มาเกี่ยวข้อง

5.1.3.2 สรุปผลการประเมินประสิทธิภาพการค้นคืนรูปภาพด้วยค่าความแม่นยำที่ k อันดับ ค่าความครบถ้วนที่ k อันดับ และค่าประสิทธิภาพโดยรวมที่ k อันดับภายใต้เงื่อนไขข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพ พบว่า ค่าความแม่นยำที่ 3, 5 และ 10 อันดับจากกลุ่มผู้เชี่ยวชาญด้านคอมพิวเตอร์และเทคโนโลยี อยู่ในระดับดีมาก คิดเป็นร้อยละ 82.7, 82.4 และ 82.4 ตามลำดับ เช่นเดียวกับกลุ่มผู้เชี่ยวชาญด้านการค้นคืนสารสนเทศ อยู่ในระดับดีมาก คิดเป็นร้อยละ 88.3, 87.2 และ 87.4 ตามลำดับ เป็นไปในทิศทางเดียวกับกลุ่มผู้ใช้ทั่วไปอยู่ในระดับดีมาก คิดเป็นร้อยละ 88.3, 88.4 และ 87.8 ตามลำดับ ซึ่งสูงกว่าสมมติฐานการวิจัยที่กำหนดไว้ที่ร้อยละ 80 อย่างไรก็ดี ความแปรปรวนของรูปภาพเป็นอุปสรรคสำคัญที่ทำให้ตัวแบบจำแนกข้อมูลภาพได้ไม่ดีเท่าที่ควร โดยเฉพาะอย่างยิ่ง ความคล้ายคลึงกันระหว่างคลาส (Inter-class similarity) และความต่างภายในคลาสเดียวกัน (Intra-class variance) ซึ่งตัวแบบที่มีอยู่ในปัจจุบัน เช่น ResNet50 มีแนวโน้มที่จะเลือกรู้จำวัตถุภายในภาพจากคุณลักษณะสีมากกว่าคุณลักษณะอื่น ๆ

5.1.3.3 สรุปผลการประเมินประสิทธิภาพการค้นคืนรูปภาพด้วยค่าความแม่นยำที่ k อันดับ ค่าความครบถ้วนที่ k อันดับ และค่าประสิทธิภาพโดยรวมที่ k อันดับ ภายใต้เงื่อนไขข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ พบว่า ค่าความแม่นยำที่ 3, 5 และ 10 อันดับจากกลุ่มผู้เชี่ยวชาญด้านคอมพิวเตอร์และเทคโนโลยี อยู่ในระดับดีมาก คิดเป็นร้อยละ 83.3, 84.8 และ 81.8 ตามลำดับ เช่นเดียวกับกลุ่มผู้เชี่ยวชาญด้านการค้นคืนสารสนเทศ อยู่ในระดับดีมาก คิดเป็นร้อยละ 88.3, 86.4 และ 85.0 ตามลำดับ เป็นไปในทิศทางเดียวกับกลุ่มผู้ใช้ทั่วไปอยู่ในระดับดี คิดเป็นร้อยละ 77.3, 77.2 และ 78.8 ตามลำดับ ซึ่งสูงกว่าสมมติฐานการวิจัยที่กำหนดไว้ที่ร้อยละ 80 อย่างไรก็ดี ความหมายของรูปภาพที่อยู่ในลักษณะนามธรรมนั้นยากที่จะประเมินผลว่าถูกต้องหรือไม่ถูกต้องเท่านั้น ดังนั้นการแปลความหมายจึงเป็นไปตามหลักการรับรู้ของมนุษย์ (Human

perception) ด้วยประสบการณ์หรือความรู้เดิมจากการเรียนรู้ของแต่ละบุคคลจึงแตกต่างกัน จะเห็นได้ว่า ปัญหาช่องว่างความหมายของรูปภาพนั้นยากที่จะทำให้หมดไป ทำได้เพียงลดช่องว่างความหมายระหว่างคอมพิวเตอร์ ผู้ใช้ และรูปภาพลงด้วยเทคนิคและวิธีการต่าง ๆ เท่านั้น จึงควรให้ความสำคัญกับการลดช่องว่างความหมายของข้อความค้นหาจากผู้ใช้ควบคู่ไปกับการวิเคราะห์ความหมายของรูปภาพ เพื่อเพิ่มประสิทธิภาพในการค้นคืน

5.1.4 สรุปผลการประเมินประสิทธิภาพการค้นคืนรูปภาพดิจิทัลเชิงความหมายด้วยค่า NDCG เปรียบเทียบระหว่างการค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิมกับการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนัก

โดยผู้วิจัยคัดเลือกข้อความค้นหาภาษาอังกฤษที่มีค่าความแม่นยำที่ 5 ลำดับแรกสูงที่สุด ภายใต้ 3 เงื่อนไข ได้แก่ 1) ข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ ในลักษณะป้ายกำกับของรูปภาพ 2) ข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพ และ 3) ข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ เงื่อนไขละ 100 รายการ รวมทั้งสิ้น 300 รายการ บนชุดข้อมูล Flickr30k เมื่อพิจารณาค่า NDCG ที่ลำดับ 1, 3 และ 5 พบว่า ผลลัพธ์จากการค้นคืนรูปภาพด้วยข้อความบรรยายรูปภาพในลักษณะป้ายกำกับของรูปภาพ โดยการค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิมเปรียบเทียบกับการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนักนั้นมีค่า NDCG ที่ลำดับ 1, 3 และ 5 ไม่แตกต่างกันมากนัก เช่นเดียวกับผลลัพธ์จากการค้นคืนรูปภาพด้วยข้อความบรรยายรูปภาพเกี่ยวกับแนวคิดระดับสูงของรูปภาพ โดยการค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิม เปรียบเทียบกับการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนักนั้นมีค่า NDCG ที่ลำดับ 1, 3 และ 5 เพิ่มขึ้นจากเดิมเล็กน้อย ตรงกันข้ามกับผลลัพธ์จากการค้นคืนรูปภาพด้วยข้อความบรรยายรูปภาพที่อธิบายความหมายเชิงคุณภาพของรูปภาพ โดยการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนักนั้นมี ค่า NDCG ที่ลำดับ 1, 3 และ 5 เพิ่มขึ้นจากการค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิมถึงร้อยละ 25.77, 18.23 และ 31.33 ตามลำดับ

อย่างไรก็ดี งานวิจัยนี้ไม่ได้เพียงเสนอวิธีการปรับแต่งโมเดลเชิงเทคนิคเท่านั้น แต่ยังเปิดมุมมองใหม่ในการพัฒนาโมเดลที่เข้าใจภาษาธรรมชาติและรูปภาพเชิงความหมายในแบบที่ใกล้เคียงกับการรับรู้ของมนุษย์ ทั้งในด้านอารมณ์ ความรู้สึก และบริบทอย่างลึกซึ้ง ซึ่งสามารถจำแนกผลจากการวิจัยนี้ว่าก่อให้เกิดประโยชน์ (Contribution) สำคัญใน 4 มิติที่ครอบคลุมทั้งด้านวิชาการ ทฤษฎี ปฏิบัติ และเทคโนโลยี รายละเอียดดังตารางที่ 5.1

ตารางที่ 5.1 ประโยชน์จากการพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายโดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า

Contribution (ประโยชน์จากการวิจัย)	What (สิ่งที่ดำเนินการ)	Why (เหตุผลในการดำเนินการ)	How (แนวทางการดำเนินการ)	Result (ผลที่ได้รับจากการดำเนินการ)
1) ประโยชน์เชิงวิชาการ (Academic contribution)	การปรับแต่งน้ำหนักของตัวแบบ CLIP ที่ถูกฝึกล่วงหน้าด้วยการกำหนดค่าน้ำหนักโครงข่ายด้วยตนเอง (Manually set weights) โดยอิงจากผลการประเมินของผู้เชี่ยวชาญ	เพื่อเพิ่มความสามารถของโมเดลในการจับคู่เวกเตอร์ข้อความกับรูปภาพอย่างมีประสิทธิภาพ โดยเฉพาะในด้านความหมายเชิงบริบท หรือรูปแบบที่เป็นนามธรรม ซึ่งสอดคล้องกับการรับรู้ของมนุษย์ มากกว่าการจับคู่อย่างผิวเผินจากลักษณะทางกายภาพ	งานวิจัยนี้นำผลการประเมินจากผู้เชี่ยวชาญจำนวน 400 ค่า มาสร้างเป็นฐานน้ำหนัก (base_weights) จากนั้นแปลงเป็น list และเติมค่าที่เหลือจนครบ 393,216 ค่าด้วยค่าที่สุ่มจากการแจกแจงแบบปกติ จากฐานน้ำหนัก แล้วแปลงเป็น PyTorch tensor และปรับรูปทรงให้ตรงกับเลเยอร์ visual_projection ก่อนอัปเดตเข้าไปในโมเดล CLIP เพื่อให้ได้ชุดน้ำหนักใหม่ที่ผสานระหว่างข้อมูลจริงและการสุ่มแบบควบคุม	แบบจำลองสามารถเรียนรู้แนวคิดนามธรรม เช่น ความสงบ โดดเดี่ยว หรือความอบอุ่น ที่ไม่ปรากฏชัดในรูปภาพได้ดีขึ้น เพิ่มความแม่นยำของระบบค้นคืนภาพในแง่ความหมายเชิงลึก (Semantic relevance) และสามารถประเมินภาพในลักษณะเดียวกับมนุษย์ที่ตีความความหมายผ่านบริบทและบรรยากาศโดยรวม ซึ่งมิใช่แค่ลักษณะวัตถุ กิจกรรม ฉาก หรือเหตุการณ์เท่านั้น

ตารางที่ 5.1 ประโยชน์จากการพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายโดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า (ต่อ)

Contribution (ประโยชน์จากการวิจัย)	What (สิ่งที่ดำเนินการ)	Why (เหตุผลในการดำเนินการ)	How (แนวทางการดำเนินการ)	Result (ผลที่ได้รับจากการดำเนินการ)
2) ประโยชน์เชิงทฤษฎี (Theoretical contribution)	การนำเสนอแนวทางใหม่ในการผสานข้อมูลเชิงมนุษย์ (Human-labeled semantics) เข้ากับการเรียนรู้ของแบบจำลองเชิงคณิตศาสตร์ โดยมุ่งให้แบบจำลองสามารถแปลความหมายรูปภาพได้ใกล้เคียงกับการรับรู้ของมนุษย์	เพื่อยกระดับการเรียนรู้ของโมเดลจากระดับสถิติ (Statistical learning) ไปสู่การตีความที่มีความหมาย (Semantic-level understanding) ซึ่งเป็นการลดช่องว่างความหมายของการค้นคืนรูปภาพจากเนื้อหาทั่วไปที่ยังมีข้อจำกัดในการประมวลผลภาษาธรรมชาติหรือการวิเคราะห์ความหมายเชิงนามธรรมจากรูปภาพ	งานวิจัยนี้ผสมผสานผลการประเมินจากผู้เชี่ยวชาญมาสร้างเป็นชุดฐานน้ำหนัก และออกแบบวิธีการสุ่มเติมน้ำหนักให้สอดคล้องทางสถิติกับค่าจริง โดยใช้ค่าที่สุ่มตามการแจกแจงแบบปกติและ tensor transformation พร้อมการปรับแต่งเลเยอร์เฉพาะใน CLIP เพื่อไม่กระทบโครงสร้างโดยรวม	สามารถเพิ่มขีดความสามารถของแบบจำลองในการประมวลผลรูปภาพเชิงความหมายที่มีบริบทหลากหลาย และสามารถใช้แนวคิดนี้ต่อยอดไปยังโมเดลอื่นด้านมัลติโมดัล (Multimodal) หรือ การรู้จำอารมณ์จากรูปภาพ (Emotional recognition) ได้

ตารางที่ 5.1 ประโยชน์จากการพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายโดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า (ต่อ)

Contribution (ประโยชน์จากการวิจัย)	What (สิ่งที่ดำเนินการ)	Why (เหตุผลในการดำเนินการ)	How (แนวทางการดำเนินการ)	Result (ผลที่ได้รับจากการดำเนินการ)
3) ประโยชน์เชิงปฏิบัติ/สังคม (Practical/ Social contribution)	การสนับสนุนผู้ใช้ด้วยข้อความค้นหาภาษาธรรมชาติที่ยึดโยงกับความหมายของรูปภาพ แม้จะอยู่ในรูปแบบที่ไม่เป็นทางการ เช่น “ภาพที่รู้สึกเหงา” หรือ “บรรยากาศอบอุ่น”	เนื่องจากผู้ใช้จำนวนไม่น้อยที่ขาดความชำนาญในการใช้คำค้นหาหรือไม่สามารถกำหนดคำหลักหรือคำที่เฉพาะเจาะจงสำหรับการค้นหาที่เหมาะสมได้ ส่งผลให้เกิดเป็นผลลัพธ์รูปภาพจำนวนมากจากค้นคืนทั้งที่ตรงกับความต้องการและไม่ตรงกับความต้องการ จึงต้องมีระบบที่สามารถวิเคราะห์ข้อความค้นหาภาษาธรรมชาติยึดโยงกับความหมายของรูปภาพแทนที่จะยึดตามหลักไวยากรณ์ของภาษาเพื่อง่ายต่อการใช้งาน	งานวิจัยนี้ปรับค่าน้ำหนักโครงข่ายด้วยตนเอง เพื่อให้แบบจำลองสามารถเรียนรู้คุณลักษณะเชิงความหมายของรูปภาพ และสามารถค้นคืนผลลัพธ์ได้ใกล้เคียงกับการรับรู้ของมนุษย์	แบบจำลองสามารถค้นคืนรูปภาพทั้งเชิงเนื้อหาและเชิงความหมายที่ตรงกับความต้องการของผู้ใช้ได้ แม้ไม่ใช้คำศัพท์ทางเทคนิค ทำให้ผู้ใช้ทั่วไปสามารถใช้งานได้โดยไม่จำเป็นต้องมีความรู้เฉพาะทาง ส่งเสริมการเข้าถึงเทคโนโลยีอย่างเสมอภาค และสามารถนำไปใช้ในสื่อดิจิทัล การศึกษา การดูแลสุขภาพจิต หรือแอปพลิเคชันแนะนำเนื้อหาเชิงความหมายได้ในอนาคต

ตารางที่ 5.1 ประโยชน์จากการพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายโดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า (ต่อ)

Contribution (ประโยชน์จากการวิจัย)	What (สิ่งที่ดำเนินการ)	Why (เหตุผลในการดำเนินการ)	How (แนวทางการดำเนินการ)	Result (ผลที่ได้รับจากการดำเนินการ)
4) ประโยชน์เชิงเทคโนโลยี (Technological contribution)	การผสมผสานข้อมูลแบบสุ่มและข้อมูลจริงในการปรับค่าน้ำหนักโครงข่ายของโมเดล ที่ปรับเปลี่ยนแบบเจาะจงเฉพาะบางชั้น (Layer-specific) ของโครงข่าย แทนที่จะทำการปรับเปลี่ยนทั้งโมเดล	เพื่อให้เกิดการควบคุมความเปลี่ยนแปลงของพารามิเตอร์ในลักษณะที่ไม่ส่งผลเสียต่อคุณลักษณะเชิงโครงสร้างของโมเดล ตลอดจนควบคุมความแม่นยำของการปรับค่าพารามิเตอร์ต่อการเรียนรู้ความหมายเชิงนามธรรมของโมเดล ซึ่งยากที่จะทำได้ด้วยกระบวนการเรียนรู้อัตโนมัติทั่วไป	งานวิจัยนี้ใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของฐานน้ำหนัก (base_weights) มาเป็นแนวทางกำหนดขอบเขตของค่าที่สุ่มลงในพารามิเตอร์ของเลเยอร์ visual_projection แบบควบคุม จากนั้นปรับรูปทรงและอัปเดตน้ำหนักโครงข่ายโดยตรงในโมเดล CLIP	ได้ชุดน้ำหนักใหม่ที่กลมกลืนกับโมเดลเดิม และสามารถเก็บรักษาโครงสร้างความรู้จากการฝึกฝนเดิมไว้ในขณะเดียวกันก็เพิ่มขีดความสามารถในการประมวลผลข้อมูลเชิงความหมายที่ใกล้เคียงกับการรับรู้ของมนุษย์

จากตารางที่ 5.1 งานวิจัยนี้นำเสนอกรอบแนวคิดใหม่เชิงบูรณาการที่รวมความรู้ทางเทคนิคเชิงลึกเข้ากับองค์ความรู้จากมนุษย์ โดยนำเสนอแนวทางใหม่ในการผสานข้อมูลเชิงมนุษย์เข้าสู่กระบวนการเรียนรู้ของแบบจำลองเชิงคณิตศาสตร์ที่ถูกฝึกไว้ล่วงหน้า เพื่อยกระดับความสามารถของการค้นคืนรูปภาพให้เข้าใจความหมายเชิงนามธรรมในลักษณะเดียวกับการรับรู้ของมนุษย์ ด้วยการปรับแต่งน้ำหนักของโครงข่ายเฉพาะบางชั้น โดยอิงข้อมูลจากผลการประเมินของผู้เชี่ยวชาญ ซึ่งถูกนำมาใช้เป็นฐานน้ำหนักสำหรับสุ่มค่าน้ำหนักในลักษณะที่ควบคุมทางสถิติ และไม่กระทบโครงสร้างโดยรวมของโมเดลเดิม ดังนั้นแบบจำลองจึงสามารถวิเคราะห์ข้อความค้นหาในรูปแบบภาษาธรรมชาติที่ไม่เป็นทางการ เช่น “อบอุ่น”, “เหงา” หรือ “สันโดษ” ที่สะท้อนคุณลักษณะเชิงนามธรรมได้อย่างมีประสิทธิภาพ ส่งผลให้ผู้ใช้ทั่วไปสามารถค้นคืนภาพผ่านภาษาธรรมชาติได้อย่างสะดวก โดยไม่ต้องพึ่งพาคำค้นหาเชิงเทคนิคหรือเฉพาะทาง ซึ่งถือเป็นการลดข้อจำกัดในการเข้าถึงเทคโนโลยีด้านปัญญาประดิษฐ์ และสามารถต่อยอดไปยังการประยุกต์ด้านมัลติโมดัลการวิเคราะห์และรู้จำอารมณ์จากภาพ และการแนะนำเนื้อหาที่อิงความหมายที่ไม่เพียงแม่นยำเชิงสถิติ แต่ยังสามารถเข้าใจความหมายเชิงบริบทได้อย่างลึกซึ้ง และสอดคล้องกับการใช้งานในสถานการณ์จริงได้อย่างครอบคลุมและเป็นรูปธรรม

นอกจากนี้แนวทางดังกล่าว ยังเป็นการขยายองค์ความรู้ในระดับวิชาการเกี่ยวกับการเรียนรู้ของแบบจำลองเชิงลึก ตลอดจนเสนอกรอบแนวคิดเชิงทฤษฎีที่สะท้อนความเป็นไปได้ในการเชื่อมโยงระหว่างการเรียนรู้เชิงลึกกับข้อมูลที่สะท้อนคุณลักษณะเชิงความหมายของมนุษย์ (Human-centered semantics) อย่างมีนัยสำคัญ ซึ่งเป็นประเด็นสำคัญในการพัฒนาแบบจำลองที่สามารถตีความภาพในระดับความหมาย มากกว่าการรับรู้เพียงลักษณะทางกายภาพ

5.2 การประยุกต์ผลการวิจัย

5.2.1 การประยุกต์ใช้ผลการวิจัยทางตรง คือการนำตัวแบบโครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้าจากการเรียนรู้ร่วมกันระหว่างรูปภาพและข้อความบรรยายรูปภาพที่ผ่านการปรับค่าน้ำหนักให้ใกล้เคียงกับมนุษย์ไปใช้เป็นกรอบการพัฒนาระบบเพื่อค้นคืนรูปภาพดิจิทัลจากชุดข้อมูลคุณลักษณะเชิงความหมายของรูปภาพที่รวบรวมไว้ ซึ่งไม่จำกัดเพียงการค้นคืนรูปภาพด้วยเนื้อหาจากการระบุแนวคิดระดับสูงที่เป็นรูปธรรมอย่างวัตถุ กิจกรรม ฉาก หรือเหตุการณ์เท่านั้น แต่ยังรวมถึงแนวคิดระดับสูงที่เป็นนามธรรมอย่างอารมณ์ และความรู้สึก อีกทั้งช่วยสนับสนุนผู้ใช้ที่ไม่สามารถกำหนดคำหลักหรือคำที่เฉพาะเจาะจงสำหรับการค้นหาที่เหมาะสมได้ด้วยการใช้ข้อความค้นหาภายใต้รูปแบบของภาษาธรรมชาติที่ยึดโยงกับความหมายของรูปภาพแทนที่จะยึดตามหลักไวยากรณ์ของภาษา เนื่องจากพฤติกรรมของผู้ใช้นั้นต้องการเพียงผลลัพธ์ที่ตนเองสนใจเท่านั้น จึงมีผู้ใช้จำนวนไม่น้อยที่ขาดความชำนาญในการใช้คำค้นหา ไม่สามารถกำหนดคำหลักหรือคำที่

เฉพาะเจาะจงสำหรับการค้นหาที่เหมาะสมได้ ส่งผลให้เกิดเป็นผลลัพธ์รูปภาพจำนวนมากจากค้นคืน ทั้งที่ตรงกับความต้องการและไม่ตรงกับความต้องการ จึงอาจเป็นการเสียโอกาสสำหรับผู้ใช้หากมีรูปภาพบางส่วนที่ไม่ถูกเรียกดู ตลอดจนอำนวยความสะดวกให้กับผู้ใช้ในการใช้ข้อความค้นหาในรูปแบบภาษาธรรมชาติแทนที่จะใช้ภาพค้นหาที่ยังไม่สะดวก และเป็นการผลักภาระให้กับผู้ใช้ในการหารูปภาพต้นฉบับอีกด้วย

5.2.2 การประยุกต์ใช้ผลการวิจัยทางอ้อม คือการนำตัวแบบโครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้าที่ผ่านการปรับค่าน้ำหนักให้ใกล้เคียงกับมนุษย์ไปใช้เรียนรู้แบบคอนทราสต์ (Contrastive learning) เพิ่มเติมกับสารสนเทศในรูปแบบวีดีโอ เพื่อเพิ่มประสิทธิภาพการค้นคืนสารสนเทศเชิงความหมายของมัลติมีเดียให้ครอบคลุมทุกมิติ และลดช่องว่างความหมายระหว่างคอมพิวเตอร์ ผู้ใช้ และสารสนเทศ และสามารถต่อยอดไปยังการประยุกต์ใช้ด้านมัลติโมดัลการวิเคราะห์และรู้จำอารมณ์จากภาพ และการแนะนำเนื้อหาที่อิงความหมายที่ไม่เพียงแม่นยำเชิงสถิติ อันเป็นแนวทางในการพัฒนาการค้นคืนสารสนเทศในรูปแบบที่เกี่ยวข้องต่อไปในอนาคต

5.3 ข้อเสนอแนะในการวิจัยครั้งต่อไป

5.3.1 ปัญหาช่องว่างความหมายของรูปภาพนั้นยากที่จะทำให้หมดไป ทำได้เพียงลดช่องว่างความหมายระหว่างคอมพิวเตอร์ ผู้ใช้ และรูปภาพลงด้วยเทคนิคและวิธีการต่าง ๆ เท่านั้น งานวิจัยในครั้งต่อไปจึงควรให้ความสำคัญกับการลดช่องว่างความหมายของข้อความค้นหาจากผู้ใช้ควบคู่ไปกับการวิเคราะห์ความหมายของรูปภาพ เพื่อเพิ่มประสิทธิภาพในการค้นคืน

5.3.2 การพัฒนาแบบจำลองการค้นคืนรูปภาพนั้นควรให้ความสำคัญที่การปรับค่าไฮเปอร์พารามิเตอร์ (Hyperparameter) ก่อนที่ตัวแบบจะทำการเรียนรู้ เนื่องจากค่าไฮเปอร์พารามิเตอร์ที่เหมาะสมจะส่งผลให้ตัวแบบมีความแม่นยำที่สูงขึ้น หรือลดค่าการสูญเสียให้ต่ำลงได้

5.3.3 การเพิ่มประสิทธิภาพจากการค้นคืนควรมุ่งที่ผลลัพธ์ไม่เกิน 5 ลำดับ เป็นหลักเพื่อตอบสนองต่อความต้องการของผู้ใช้ เนื่องจากพฤติกรรมของผู้ใช้ต้องการเพียงผลลัพธ์ที่ตนเองสนใจเท่านั้น ผลลัพธ์จำนวนมากจากค้นคืนอาจทำให้ผู้ใช้เสียเวลาคัดเลือกรูปภาพที่ตรงกับความต้องการอย่างแท้จริง ซึ่งนับเป็นความสูญเปล่าทางสารสนเทศ เนื่องจากเป็นได้น้อยมากที่ผู้ใช้จะเลือกดูรูปภาพจากผลการค้นคืนทั้งหมด

5.3.4 ความหมายเชิงนามธรรมนั้นหลากหลาย และมีลักษณะที่ขึ้นอยู่กับบริบททางวัฒนธรรม อารมณ์ หรือประสบการณ์ส่วนบุคคล จึงควรมีการพัฒนาแบบจำลองที่เปิดรับข้อมูลจากผู้ใช้ เช่น การให้คะแนนหรือข้อเสนอแนะเกี่ยวกับผลลัพธ์ของการค้นคืน เพื่อให้โมเดลสามารถเรียนรู้และปรับปรุงตนเองได้อย่างต่อเนื่อง (Aaptive learning) ซึ่งจะช่วยเพิ่มศักยภาพในการเรียนรู้ความหมายของรูปภาพได้อย่างครอบคลุมและสอดคล้องกับการรับรู้ของมนุษย์ในสถานการณ์จริง

รายการอ้างอิง

รายการอ้างอิง

ณัฏฐ์ธนัน ทองดี. (2553). สถิติเพื่องานวิจัย (พิมพ์ครั้งที่ 2). กรุงเทพฯ: สำนักพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

เรวัต แสงสุริยงค์. (2565). ความเสี่ยงของการเกิดความคลาดเคลื่อนในการวิจัยเชิงปริมาณด้านสังคมวิทยา. **วารสารวิชาการมนุษยศาสตร์และสังคมศาสตร์ มหาวิทยาลัยบูรพา**, 30(1), 158-185.

อรนุช ศรีสะอาด. (2561). การตรวจสอบความเที่ยงตรงของเครื่องมือวัดผลโดยผู้เชี่ยวชาญ . **วารสารการวัดผลการศึกษา มหาวิทยาลัยมหาสารคาม.** 1(1): 45-49.

อารัมภ์ กาวีวงศ์. (2555). บทบาทของเทคโนโลยีการประมวลผลภาพดิจิทัลต่องานทางด้านสัตวแพทย์. **เชียงใหม่สัตวแพทยสาร.** 10(3): 203-217.

AbdElrazek, E. E. (2017). A Comparative Study of Image Retrieval Algorithms for Enhancing a Content-based Image Retrieval System. **Global Journal of Computer Science and Technology: (F) Graphics & Vision**. 17(3): 1-9.

Aggarwal, C. C. (2015). **Data Classification Algorithms and Applications**. Boca Raton, FL: CRC Press.

Aggarwal, C. C. (2018). **Neural Networks and Deep Learning A Textbook**. Cham: Springer.

Anandh, A., Mala, K., and Suresh Babu, R. (2019). Combined global and local semantic feature–based image retrieval analysis with interactive feedback. **Measurement and Control**. 53(1-2): 3–17.

Baltrusaitis, T., Ahuja, C., and Morency, L. (2019). Multimodal Machine Learning: A Survey and Taxonomy. **IEEE Transactions on Pattern Analysis and Machine Intelligence**. 41(2): 423-443.

Barz B., and Denzler J. (2020). **Content-based Image Retrieval and the Semantic Gap in the Deep Learning Era. Pattern Recognition**. ICPR International Workshops and Challenges. Cham: Springer.

Barz, B. (2020). **Semantic and Interactive Content-based Image Retrieval**. Ph.D. Dissertation, University of Jena, Germany.

Bianchi, F., Attanasio, G., Pisoni, R., Terragni, S., Sarti, G., and Lakshmi, S. (2021). **Contrastive Language-Image Pre-training for the Italian Language**. [On-line]. Available: https://arxiv.org/pdf/2108.08688.pdf

Brase, C.H. and Brase, C.P. (2018). **Understanding Basic Statistics**. Boston: Cengage Learning.

Broz, M. (2024). **How Many Pictures Are There (2024): Statistics, Trends, and Forecasts**. [On-line]. Available: https://photutorial.com/photos-statistics/.

Buda, M., Maki, A., and Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. **Neural Networks**, 106, 249–259.

Caicedo, J.C., Gonzalez, F.A., and Romero, E. (2008). A Semantic Content-Based Retrieval Method for Histopathology Images. In **Proceedings of the 4th Asia Information Retrieval Symposium (AIRS 2008)**. (pp. 51-60). Harbin: Springer-Verlag.Canty, M. J. (2014). **Image Analysis, Classification and Change Detection in Remote Sensing - With Algorithms for ENVI/IDL and Python**, 3rd ed. Boca Raton, FL: CRC Press.

Carnevali, L. (2023). **Evaluation Measures in Information Retrieval**. [On-line]. Available: https://www.pinecone.io/learn/offline-evaluation/

Chaudhary, A. (2020). **Evaluation Metrics For Information Retrieval**. [On-line]. Available: https://amitness.com/2020/08/information-retrieval-evaluation/

Chen, H., Trouve, A., Murakami, K. J., and Fukuda, A. (2017). Semantic image retrieval for complex queries using a knowledge parser. **Multimedia Tools and Applications**. 77(9): 10733–10751.

Chen, Y., Lu, E.J. and Lin, S. (2020). Ontology-based Dynamic Semantic Annotation for Social Image Retrieval. In **Proceeding of the 21st IEEE International Conference on Mobile Data Management (MDM)**. (pp.337-341). Versailles: IEEE.

Davies, J., Studer, R., and Warren, P. (2006). **Semantic Web Technologies: Trends and Research in Ontology-based Systems**. West Sussex: John Wiley & Sons Ltd.

Dix, A., Finlay, J., Abowd, G. D., and Beale, R. (2004). **Human–computer Interaction**, 3rd ed. Haddington, Scotland: Pearson Education Limited.

Dong, H., Wang, Z., Qiu, Q., and Sapiro, G. (2020). **Using Text to Teach Image Retrieval**. [On-line]. Available: https://arxiv.org/abs/2011.09928.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. **In Proceeding of the 9th International Conference on Learning Representations 2021 (ICLR 2021)**. (pp. 1-21.) Virtual Event, Austria: ICLR.

Felzenszwalb, P.F., Girshick, R.B., Ramanan, D. and McAllester, D. (2010) Object Detection with Discriminatively Trained Part-Based Models. **IEEE Transactions on Pattern Analysis and Machine Intelligence**. 32: 1627-1645.

Ghosh, A., Sufian, A., Sultana, F., Chakrabarti, A., and De, D. (2019). Fundamental concepts of convolutional neural network. In V. E. Balas, R. Kumar, and R. Srivastava (eds), **Recent Trends and Advances in Artificial Intelligence and Internet of Things**. (pp. 519-567). Cham: Springer.

Gkeliosa, S., Sophokleousb, A., Plakiasa, S., Boutalisa, Y., and Chatzichristofisb, S.A. (2021). Deep Convolutional Features for Image Retrieval. **Expert Systems with Applications**. 177(2021): 1-17.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). **Deep Learning**. MA: MIT press.

Guilford, J.P. (1967). **The nature of human intelligence**. New York: McGraw-Hill.

Gupta, R., and Singh, V. (2018). A Framework for Semantic based Image Retrieval from Cyberspace by mapping low level features with high level semantics. In **Proceeding of the 2018 3rd International Conference On Internet of Things: Smart Innovation and Usages (IoT-SIU)** (pp. 1-6). Bhimtal: IEEE.

Jansen, B.J. and Spink A. (2003). An Analysis of Web Documents Retrieved and Viewed. In **Proceedings of the 4th International Conference on Internet Computing**. (pp. 65-69). Chennai: H-Index.

Jansen, B.J., and Spink, A. (2006). How are we searching the world wide web? A comparison of nine search engine transaction logs. **Inform Process Manag**. 42(1): 248-63.

Kalimuthu, M., and Krishnamurthi, I. (2015). Semantic-Based Facial Image-Retrieval System with Aid of Adaptive Particle Swarm Optimization and Squared Euclidian Distance. **Journal of Applied Mathematics**. 2015: 1–12.

Khodaskar, A. and Ladke, S. A. (2012). Content Based Image Retrieval with Semantic Features using Object Ontology. **International Journal of Engineering Research & Technology (IJERT)**. 1(4): 1-6.

Le, H.D., Nguyen, Q.Q., Nguyen, V.A., Nguyen, T.D., Chung, N.M., Thái, T., and Ha, S.V. (2022). Tracked-Vehicle Retrieval by Natural Language Descriptions With Domain Adaptive Knowledge. In **Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops**. (pp. 3300-3309). New Orleans, LA: IEEE.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. **In European conference on computer vision** (pp. 740–755). Springer.

Liu, C. and Song, G. (2011). A Method of Measuring the Semantic Gap in Image Retrieval: Using the Information Theory. In **Proceeding of the 2011 International Conference on Image Analysis and Signal Processing (IASP 2011)**. (pp. 1-5). Hubei, China: IEEE.

Liu, H., and Motoda, H. (Eds.). (2012). **Feature extraction, construction and selection: A data mining perspective** (2nd ed.). Springer US.

Liu, J., Sun, J., and Zhou, X. (2022). Comparison of ResNet-50 and vision transformer models for trash classification. In **Proceedings of the 3rd International Conference on Artificial Intelligence and Computer Engineering (ICAICE 2022)**. Wuhan: AEIC.

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhao, L., Zhu, D., Li, X., Qiang, N., Shen, D., Liu, T., and Ge, B. (2023). **Summary of ChatGPT-Related Research and Perspective Towards the Future of Large Language Models**. [On-line]. Available: https://arxiv.org/pdf/2304.01852.pdf

Ma, Z., Liu, F., Dong, J., Qu, X., He, Y., and Ji, S. (2021). Hierarchical Similarity Learning for Language-Based Product Image Retrieval. In **Proceeding of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (**pp. 4335-4339**)**. Toronto, Ontario: IEEE.

Mai, L., Jin, H., Lin, Z., Fang, C., Brandt, J. and Liu, F. (2017). Spatial-Semantic Image Search by Visual Feature Synthesis. In **Proceeding of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)** (pp. 4718-4727). Honolulu, HI: IEEE.

Mane, A. L. (2016). Semantic based image retrieval. **Indian Journal of Computer Science and Engineering**. 7(4): 110-117.

Manning, C.D., Raghavan, P. and Schütze, H., (2008). **Introduction to Information Retrieval**. Cambridge: Cambridge University Press.

Manzoor, U., Balubaid, M. A., Zafar, B., Umar, H., and Khan, M. S. (2015). Semantic Image Retrieval: An Ontology Based Approach. **International Journal of Advanced Research in Artificial Intelligence (IJARAI)**. 4(4): 1-8.

Mezaris, V., Kompatsiaris, I. and Strintzis, M. G. (2003). An ontology approach to object-based image retrieval. In **Proceeding of the 2003 International Conference on Image Processing** (pp. 511-514). Barcelona: IEEE.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to WordNet: An On-line Lexical Database*. **International Journal of Lexicography**, 3(4), 235–244.

Mitchell R. (2018). **Web Scraping with Python: Collecting Data from the Modern Web**. 2nd ed. Sebastopol, CA: O'Reilly Media Inc.

Momin, R. (2021). **Vision Transformer (ViT) Fine-tuning**. [On-line]. Available: https://www.kaggle.com/code/raufmomin/vision-transformer-vit-fine-tuning

Mumuni, A., and Mumuni, F. (2022). Data augmentation: A comprehensive survey of modern approaches. **Array**. 16(2022): 1-27.

Nhi, N.T.U., Le, T.M. and Van, T.T. (2022). A Model of Semantic-Based Image Retrieval Using C-Tree and Neighbor Graph. **International journal on Semantic Web and information systems**. 18(1): 1-23.

Nwankpa, C., Ijomah, W., Gachagan, A., and Marshall, S. (2018). **Activation Functions: Comparison of trends in Practice and Research for Deep Learning**. [On-line]. Available: https://arxiv.org/pdf/1811.03378.pdf

Office of General Counsel the California State University. (2021). **Conflict of Interest Handbook**. CA: California State University.

Patel, S. B., and Sampat, A. (2018). **Semantic image search using queries**. [On-line]. Available: https://cs224d.stanford.edu/reports/PatelShabaz.pdf

Pan, S. J., and Yang, Q. (2010). A survey on transfer learning. **IEEE Transactions on Knowledge and Data Engineering**, *22*(10), 1345–1359.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In R. Plutchik, and H. Kellerman (eds), **Emotion: Theory, research, and experience: Vol. 1. Theories of emotion**. (pp. 3-33). New York: Academic.

Potapov, A., Zhdanov, I., Scherbakov, O., Skorobogatko, N., Latapie, H., and Fenoglio, E. (2018). Semantic Image Retrieval by Uniting Deep Neural Networks and Cognitive Architectures. In **Proceeding of the 11th conference on Artificial General Intelligence, AGI'18** (pp. 1-10). Prague: Springer.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskeve, I. (2019). **Language models are unsupervised multitask learners**. OpenAI blog. 1(8). 9.

Royce, W.W. (1 9 7 0). Managing the Development of Large Software Systems. In **Proceeding of the IEEE WESCON** (pp. 328-338). LA: TRW.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). **DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter**. [On-line]. Available: https://arxiv.org/pdf/1910.01108.pdf

Sawarkar, K. (2022). **Deep Learning with PyTorch Lightning: Build and train high-performance artificial intelligence and self-supervised models using Python**. Birmingham: Packt Publishing.

Talebi, H., and Milanfar, P. (2017). **NIMA: Neural Image Assessment**. [On-line]. Available: https://arxiv.org/pdf/1709.05424.pdf

Thanh, L.M., Nhi, N.T.U., Thi, N.T.U., Han, P.T.A. and Thanh, V. T. (2020). A Semantic-Based Image Retrieval System Using A Hybrid Method K-Means And K-Nearest-Neighbor. **Annales Univ. Sci. Budapest**., Sec. Comp. 51: 253-274.

Torrey, L., and Shavlik, J. (2010). Transfer learning. In T. Ganchev, M. Sokolova, R. Rada, P.J. García-Laencina and V. Ravi (eds), **Handbook of research on machine learning applications and trends: algorithms, methods, and techniques**. (pp. 242-264). Hershey, PA: IGI Global.

Turing, A. M. (1950). Computing machinery and intelligence. **Mind**, 59(236), 433–460.

Tyagi, V. (2017). **Content-Based Image Retrieval Ideas, Influences, and Current Trends**. Gateway East: Springer Nature.

Vasilev, I., Slater, D., Spacagna, G., Roelants, P., and Zocca, V. (2019). **Python Deep Learning Exploring deep learning techniques and neural network architectures with PyTorch, Keras, and TensorFlow**. 2nd ed. Birmingham: Packt Publishing.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I. (2017). Attention is All you Need. In **Proceedings of the 31st International Conference on Neural Information Processing Systems**. (pp: 6000–6010). Long beach, CA: Curran Associates Inc.

Weers, F., Shankar, V., Katharopoulos, A., Yang, Y., and Gunte, T. (2023). Masked Autoencoding Does Not Help Natural Language Supervision at Scale. In: **Proceedings of the 2023 Conference on Computer Vision and Pattern Recognition (CVPR)**. (pp. 1-19). Vancouver: IEEE.

Wozniak, M. (2014). **Hybrid classifiers methods of data, knowledge, and classifier combination**. Heidelberg: Springer-Verlag.

Xie, C., Sun, S., Xiong, X., Zheng, Y., Zhao, D., Zhou, J. (2023) .RA-CLIP: Retrieval Augmented Contrastive Language-Image Pre-training. In **Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).** (pp. 19265-74). Vancouver: IEEE.

Xu, M., Yoon, S, Fuentes, A., and Park, D.S. (2023). A Comprehensive Survey of Image Augmentation Techniques for Deep Learning. **Pattern Recognition**. 127(2023): 1-12.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. **Transactions of the Association for Computational Linguistics**, *2,* 67–78.

Zhang, D. (2019). **Fundamentals of Image Data Mining Analysis, Features, classification and Retrieval**. Cham: Springer.

Zhang, H., Koh, J. Y., Baldridge, J., Lee, H., and Yang, Y. (2020). Cross-Modal Contrastive Learning for Text-to-Image Generation. [On-line]. Available: https://arxiv.org/abs/2101.04702.

Zhang, P., Lan, C., Zeng, W., Xing, J., Xue, J., and Zheng, N. (2020). Semantics-Guided Neural Networks for Efficient Skeleton-Based Human Action Recognition. In **Proceeding of the 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).** (pp: 1112-1121). Virtual: IEEE.

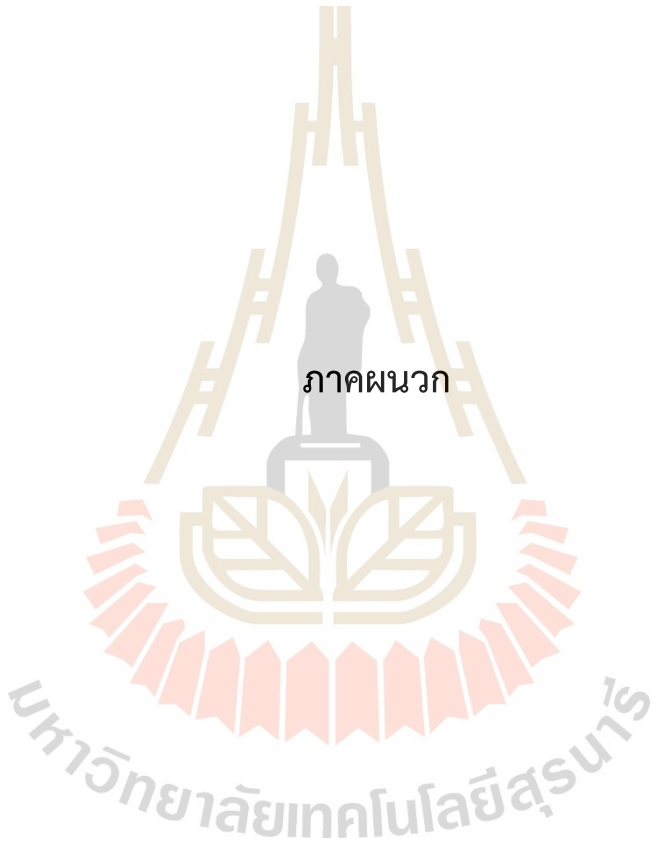
Zhang, Y. (2007). **Semantic-Based Visual Information Retrieval**. Hershey: IRM Press.

Zhao, X., Pei, L., Zhang, Z., and Li, T. (2020). Online multi-core ranking model for image retrieval. **Journal of Physics: Conference Series**. 1682(2020): 1-7.

Zhao, Z. Q., Zheng, P., Xu, S. T., and Wu, X. (2019). Object detection with deep learning: A review. **IEEE Transactions on Neural Networks and Learning Systems**, *30*(11), 3212–3232.

Zizka, J., Darena, F., and Svoboda, A. (2020). **Text Mining with Machine Learning Principles and Techniques**. Boca Raton: CRC Press.

ภาคผนวก



ภาคผนวก ก
ตัวอย่างรูปภาพต่อข้อความบรรยายความหมายเชิงคุณภาพ

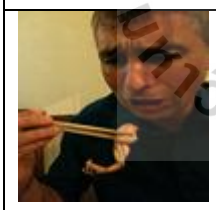
ตารางที่ ก.1 ตัวอย่างรูปภาพที่สุ่มจากชุดข้อมูล Flickr30k และชุดข้อมูลที่ผู้วิจัยรวบรวมเองต่อข้อความบรรยายรูปภาพที่อธิบายความหมายเชิงคุณภาพ

โดเมน		ข้อความบรรยายรูปภาพที่อธิบายความหมายเชิงคุณภาพ
ความรื่นเริง (Joy)		(A) two women walking down a street, hugging each other, enjoying the moment together.
		(B) two women walking arm in arm, enjoying each other's company.
		(C) two women walking down a street, hugging each other, enjoying the moment.
		(D) two women walking down a street, holding hands and smiling, enjoying each other's company.
		(E) two women walking down the street, holding hands and sharing a love that transcends time.
		(A) a woman sits on the ground playing with a cat.
		(B) a woman sitting on a brick walkway playing with a cat.
		(C) a woman sitting on a brick street, petting a cat.
		(D) a woman sitting on the ground petting a cat, surrounded by potted plants and a handbag.
		(E) a woman sitting on the ground playing with a cat.
ความวางใจ (Trust)		(A) the little girl is so adorable as she prays in front of the lit candle.
		(B) a young girl praying in front of a lit candle.
		(C) the little girl's eyes are closed as she blows out the candles on her birthday cake.
		(D) a beautiful birthday cake with lit candles, ready to be blown out by a little girl.
		(E) a young girl praying in front of a lit candle, and the girl is making a wish.
		(A) the man is in a meditative state on the wooden platform.
		(B) the man is in a peaceful and contemplative state of mind as he sits on the wooden floor, kneeling on a green rug.
		(C) a man praying on a wooden deck with a green rug.
		(D) the man is in a peaceful and contemplative state of mind while kneeling on a rug.
		(E) a man in a white shirt and black pants is kneeling on a green rug, praying on a deck.

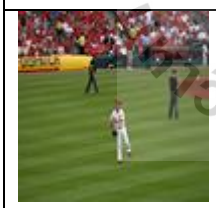
ตารางที่ ก.1 ตัวอย่างรูปภาพที่สุ่มจากชุดข้อมูล Flickr30k และชุดข้อมูลที่ผู้วิจัยรวบรวมเองต่อข้อความบรรยายรูปภาพที่อธิบายความหมายเชิงคุณภาพ (ต่อ)

โดเมน		ข้อความบรรยายรูปภาพที่อธิบายความหมายเชิงคุณภาพ
ความกลัว (Fear)		(A) the boy's face is covered in makeup, giving him a scary appearance.
		(B) the boy's face is smeared with fake blood, making him look scared.
		(C) he is sitting on the ground, holding a hot dog, which adds to the eerie atmosphere of the scene.
		(D) he appears to be enjoying the thrill of being scary and possibly participating in a halloween event or a costume party.
		(E) the boy is feeling a sense of excitement or playfulness as he is dressed up as a zombie and eating a hot dog.
		(A) the two apples are screaming.
		(B) a pair of apples with faces drawn on them, one with a screaming face and the other with a smiling face.
		(C) the apple with the knife stuck in it has a scared face.
		(D) the apple's face is being sliced, and it's not happy about it.
		(E) the image of the two apples with faces drawn on them and a knife sticking out of one of them evokes a sense of shock or surprise.
ความประหลาดใจ (Surprise)		(A) the woman is amazed by the taste of her food.
		(B) the woman is surprised or shocked by something she has just seen or heard.
		(C) the woman is laughing at something.
		(D) the woman is shocked by something she sees or hears.
		(E) the woman is covering her mouth with her hand, possibly expressing disapproval or shock at something she has seen or heard.
		(A) the two women are amazed at the large amount of food in front of them.
		(B) the two women are surprised to see a man in the background.
		(C) the two women are enjoying a meal together, with one of them smiling as she eats. they are surrounded by various bowls of food, including a large bowl of meat.
		(D) the two women are eating meat, but they should be eating more vegetables for a healthier diet.
		(E) two women sitting at a table, enjoying a meal together and laughing.

ตารางที่ ก.1 ตัวอย่างรูปภาพที่สุ่มจากชุดข้อมูล Flickr30k และชุดข้อมูลที่ผู้วิจัยรวบรวมเองต่อข้อความบรรยายรูปภาพที่อธิบายความหมายเชิงคุณภาพ (ต่อ)

โดเมน		ข้อความบรรยายรูปภาพที่อธิบายความหมายเชิงคุณภาพ
ความเศร้า (Sadness)		(A) the woman is crying.
		(B) a woman lying on a couch, feeling sad and lonely.
		(C) a girl with a broken heart laying on a couch.
		(D) the woman is feeling sad or lonely while laying on the couch.
		(E) the woman is feeling a sense of frustration or disappointment as she lays on the couch with her mouth open.
		(A) a small wooden figure sits next to a remote control, looking sad.
		(B) the two wooden figures sitting on the couch, one of them holding a remote, appear to be sad.
		(C) a small wooden figure sits next to a remote control, seemingly lonely and longing for companionship.
		(D) a small wooden figure sitting next to a remote control.
		(E) (a small wooden figure is sitting on a couch next to a remote control. the figure appears to be looking at the remote with a sense of curiosity or longing, as if it wishes to control the television.
ความรังเกียจ (Disgust)		(A) a woman in a red sweater looks at a plate of food with disgust.
		(B) a woman is disgusted by the dessert on the table.
		(C) a woman in a red sweater is bored at the table with a plate of food.
		(D) a woman sitting at a table with a plate of food in front of her, looking at the food with a frown on her face.
		(E) a woman in a red sweater is about to eat a dessert.
		(A) a man is eating a piece of food that looks like a muscle man. he is making a face and appears to be loathing the food.
		(B) a man is making a disgusted face while holding a piece of food on a toothpick.
		(C) a man is bored eating chopsticks.
		(D) the man looks sad as he eats a piece of food shaped like a man.
		(E) the man is eating a piece of food that looks like a man.

ตารางที่ ก.1 ตัวอย่างรูปภาพที่สุ่มจากชุดข้อมูล Flickr30k และชุดข้อมูลที่ผู้วิจัยรวบรวมเองต่อข้อความบรรยายรูปภาพที่อธิบายความหมายเชิงคุณภาพ (ต่อ)

โดเมน		ข้อความบรรยายรูปภาพที่อธิบายความหมายเชิงคุณภาพ
ความโกรธ (Anger)		(A) a dog is growling at another dog while laying on the floor.
		(B) the angry dog is about to bite the smaller dog.
		(C) the dog is yawning and the other dog is trying to bite him.
		(D) a white dog with its mouth open and a brown dog underneath it.
		(E) a large white dog is growling at a small brown dog, showing its teeth and aggressive behavior.
		(A) a man is yelling and jumping in the air while surrounded by people.
		(B) the man is yelling with his arms raised, showing his anger.
		(C) a man yelling with his arms up in the air, annoying everyone around him.
		(D) a man is being lifted up by his friends while he is yelling.
		(E) a man is being lifted off the ground by a group of people, and he is yelling with his mouth open.
ความคาดหวัง (Anticipation)		(A) a group of friends are sitting around a table with a dog and a bottle of beer.
		(B) the group of friends is eagerly waiting for the man to finish his beer so they can continue playing the wii.
		(C) a group of friends enjoying a night in, playing video games and drinking beer.
		(D) the group of friends is enjoying a fun and lighthearted moment together.
		(E) a group of friends sitting around a table, sharing laughter and camaraderie while playing video games and drinking beer.
		(A) a baseball player in a red hat and uniform is walking across the field, holding a baseball glove.
		(B) a baseball player is walking across the field, holding a baseball glove, as he anticipates the next play in the game.
		(C) a baseball player in a red and white uniform walks across the field.
		(D) the baseball player is walking on the field with a smile on his face, showing his optimism and determination to perform well in the game.
		(E) a baseball player in a red cap and white uniform is walking across the field.

ภาคผนวก ข

ผลการพยากรณ์คำบรรยายรูปภาพจากตัวแบบเปรียบเทียบกับผลการประเมินจากผู้เชี่ยวชาญ

ตารางที่ ข.1 ผลการพยากรณ์คำบรรยายรูปภาพจากตัวแบบเปรียบเทียบกับผลการประเมินจากผู้เชี่ยวชาญ จำนวน 10 ท่าน

รูปภาพ	คำบรรยายรูปภาพ	ผลการพยากรณ์	ผลการประเมินจากผู้เชี่ยวชาญ									
			1	2	3	4	5	6	7	8	9	10
 Domain: Joy	(A) two women walking down a street, hugging each other, enjoying the moment together.	0.2467	☑	☑		☑			☑			☑
	(B) two women walking arm in arm, enjoying each other's company.	0.1262			☑							
	(C) two women walking down a street, hugging each other, enjoying the moment.	0.2418					☑	☑				
	(D) two women walking down a street, holding hands and smiling, enjoying each other's company.	0.2166								☑	☑	
	(E) two women walking down the street, holding hands and sharing a love that transcends time.	0.1687										
 Domain: Joy	(A) a woman sits on the ground playing with a cat.	0.5106		☑	☑	☑						☑
	(B) a woman sitting on a brick walkway playing with a cat.	0.0219	☑				☑				☑	
	(C) a woman sitting on a brick street, petting a cat.	0.0512						☑				
	(D) a woman sitting on the ground petting a cat, surrounded by potted plants and a handbag.	0.3103							☑			
	(E) a woman sitting on the ground playing with a cat.	0.1059								☑		

ตารางที่ ข.1 ผลการพยากรณ์คำบรรยายรูปภาพจากตัวแบบเปรียบเทียบกับผลการประเมินจากผู้เชี่ยวชาญ จำนวน 10 ท่าน (ต่อ)

รูปภาพ	คำบรรยายรูปภาพ	ผลการพยากรณ์	ผลการประเมินจากผู้เชี่ยวชาญ									
			1	2	3	4	5	6	7	8	9	10
	(A) the little girl is so adorable as she prays in front of the lit candle.	0.2175	☑	☑		☑						
	(B) a young girl praying in front of a lit candle.	0.2506							☑		☑	☑
	(C) the little girl's eyes are closed as she blows out the candles on her birthday cake.	0.1860			☑			☑				
Domain: Trust	(D) a beautiful birthday cake with lit candles, ready to be blown out by a little girl.	0.1946					☑					
	(E) a young girl praying in front of a lit candle, and the girl is making a wish.	0.1513								☑		
	(A) the man is in a meditative state on the wooden platform.	0.2603	☑				☑					
	(B) the man is in a peaceful and contemplative state of mind as he sits on the wooden floor, kneeling on a green rug.	0.0028		☑	☑						☑	
	(C) a man praying on a wooden deck with a green rug.	0.1688						☑	☑			☑
Domain: Trust	(D) the man is in a peaceful and contemplative state of mind while kneeling on a rug.	0.3699				☑						
	(E) a man in a white shirt and black pants is kneeling on a green rug, praying on a deck.	0.1982								☑		

ตารางที่ ข.1 ผลการพยากรณ์คำบรรยายรูปภาพจากตัวแบบเปรียบเทียบกับผลการประเมินจากผู้เชี่ยวชาญ จำนวน 10 ท่าน (ต่อ)

รูปภาพ	คำบรรยายรูปภาพ	ผลการพยากรณ์	ผลการประเมินจากผู้เชี่ยวชาญ									
			1	2	3	4	5	6	7	8	9	10
	(A) the boy's face is covered in makeup, giving him a scary appearance.	0.0508					☑	☑				
	(B) the boy's face is smeared with fake blood, making him look scared.	0.8423	☑	☑	☑						☑	☑
	(C) he is sitting on the ground, holding a hot dog, which adds to the eerie atmosphere of the scene.	0.0046				☑			☑			
Domain: Fear	(D) he appears to be enjoying the thrill of being scary and possibly participating in a halloween event or a costume party.	0.0051								☑		
	(E) the boy is feeling a sense of excitement or playfulness as he is dressed up as a zombie and eating a hot dog.	0.0972										
	(A) the two apples are screaming.	0.0016	☑	☑								☑
	(B) a pair of apples with faces drawn on them, one with a screaming face and the other with a smiling face.	0.2540			☑	☑	☑				☑	
	(C) the apple with the knife stuck in it has a scared face.	0.0020							☑	☑		
Domain: Fear	(D) the apple's face is being sliced, and it's not happy about it.	0.0055						☑				
	(E) the image of the two apples with faces drawn on them and a knife sticking out of one of them evokes a sense of shock or surprise.	0.7369										

ตารางที่ ข.1 ผลการพยากรณ์คำบรรยายรูปภาพจากตัวแบบเปรียบเทียบกับผลการประเมินจากผู้เชี่ยวชาญ จำนวน 10 ท่าน (ต่อ)

รูปภาพ	คำบรรยายรูปภาพ	ผลการพยากรณ์	ผลการประเมินจากผู้เชี่ยวชาญ									
			1	2	3	4	5	6	7	8	9	10
 **Domain:** Surprise	(A) the woman is amazed by the taste of her food.	0.0234								☑		
	(B) the woman is surprised or shocked by something she has just seen or heard.	0.3369	☑	☑	☑				☑		☑	☑
	(C) the woman is laughing at something.	0.0233					☑					
	(D) the woman is shocked by something she sees or hears.	0.0621				☑		☑				
	(E) the woman is covering her mouth with her hand, possibly expressing disapproval or shock at something she has seen or heard.	0.5542										
 **Domain:** Surprise	(A) the two women are amazed at the large amount of food in front of them.	0.0338	☑			☑	☑		☑			☑
	(B) the two women are surprised to see a man in the background.	0.0011		☑				☑		☑		
	(C) the two women are enjoying a meal together, with one of them smiling as she eats. they are surrounded by various bowls of food, including a large bowl of meat.	0.4959									☑	
	(D) the two women are eating meat, but they should be eating more vegetables for a healthier diet.	0.0007			☑							
	(E) two women sitting at a table, enjoying a meal together and laughing.	0.4685										

ตารางที่ ข.1 ผลการพยากรณ์คำบรรยายรูปภาพจากตัวแบบเปรียบเทียบกับผลการประเมินจากผู้เชี่ยวชาญ จำนวน 10 ท่าน (ต่อ)

รูปภาพ	คำบรรยายรูปภาพ	ผลการพยากรณ์	ผลการประเมินจากผู้เชี่ยวชาญ									
			1	2	3	4	5	6	7	8	9	10
 **Domain:** Sadness	(A) the woman is crying.	0.0124			☑							
	(B) a woman lying on a couch, feeling sad and lonely.	0.4753	☑	☑						☑	☑	☑
	(C) a girl with a broken heart laying on a couch.	0.0078										
	(D) the woman is feeling sad or lonely while laying on the couch.	0.0650				☑	☑					
	(E) the woman is feeling a sense of frustration or disappointment as she lays on the couch with her mouth open.	0.4394						☑	☑			
 **Domain:** Sadness	(A) a small wooden figure sits next to a remote control, looking sad.	0.7742	☑				☑	☑				
	(B) the two wooden figures sitting on the couch, one of them holding a remote, appear to be sad.	0.0238			☑	☑			☑		☑	☑
	(C) a small wooden figure sits next to a remote control, seemingly lonely and longing for companionship.	0.0896		☑								
	(D) a small wooden figure sitting next to a remote control.	0.0372										
	(E) a small wooden figure is sitting on a couch next to a remote control. the figure appears to be looking at the remote with a sense of curiosity or longing, as if it wishes to control the television.	0.0752								☑		

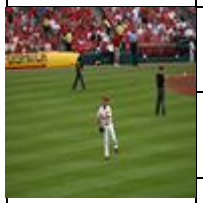
ตารางที่ ข.1 ผลการพยากรณ์คำบรรยายรูปภาพจากตัวแบบเปรียบเทียบกับผลการประเมินจากผู้เชี่ยวชาญ จำนวน 10 ท่าน (ต่อ)

รูปภาพ	คำบรรยายรูปภาพ	ผลการพยากรณ์	ผลการประเมินจากผู้เชี่ยวชาญ									
			1	2	3	4	5	6	7	8	9	10
 **Domain:** disgust	(A) a woman in a red sweater looks at a plate of food with disgust.	0.7537	☑	☑	☑			☑			☑	☑
	(B) a woman is disgusted by the dessert on the table.	0.1362										
	(C) a woman in a red sweater is bored at the table with a plate of food.	0.0682								☑		
	(D) a woman sitting at a table with a plate of food in front of her, looking at the food with a frown on her face.	0.0291				☑	☑					
	(E) a woman in a red sweater is about to eat a dessert.	0.0129							☑			
 **Domain:** disgust	(A) a man is eating a piece of food that looks like a muscle man. he is making a face and appears to be loathing the food.	0.0887			☑							☑
	(B) a man is making a disgusted face while holding a piece of food on a toothpick.	0.3700	☑	☑		☑	☑	☑			☑	
	(C) a man is bored eating chopsticks.	0.4718										
	(D) the man looks sad as he eats a piece of food shaped like a man.	0.0658								☑		
	(E) the man is eating a piece of food that looks like a man.	0.0036							☑			

ตารางที่ ข.1 ผลการพยากรณ์คำบรรยายรูปภาพจากตัวแบบเปรียบเทียบกับผลการประเมินจากผู้เชี่ยวชาญ จำนวน 10 ท่าน (ต่อ)

รูปภาพ	คำบรรยายรูปภาพ	ผลการพยากรณ์	ผลการประเมินจากผู้เชี่ยวชาญ									
			1	2	3	4	5	6	7	8	9	10
	(A) a dog is growling at another dog while laying on the floor.	0.0018					☑					
	(B) the angry dog is about to bite the smaller dog.	0.0295		☑		☑		☑				☑
	(C) the dog is yawning and the other dog is trying to bite him.	0.0219								☑		
	(D) a white dog with its mouth open and a brown dog underneath it.	0.1144			☑				☑			
Domain: Anger	(E) a large white dog is growling at a small brown dog, showing its teeth and aggressive behavior.	0.8323	☑								☑	
	(A) a man is yelling and jumping in the air while surrounded by people.	0.0403						☑	☑	☑		
	(B) the man is yelling with his arms raised, showing his anger.	0.5550										
	(C) a man yelling with his arms up in the air, annoying everyone around him.	0.2096		☑	☑							☑
Domain: Anger	(D) a man is being lifted up by his friends while he is yelling.	0.1794	☑				☑					☑
	(E) a man is being lifted off the ground by a group of people, and he is yelling with his mouth open.	0.0157				☑						

ตารางที่ ข.1 ผลการพยากรณ์คำบรรยายรูปภาพจากตัวแบบเปรียบเทียบกับผลการประเมินจากผู้เชี่ยวชาญ จำนวน 10 ท่าน (ต่อ)

รูปภาพ	คำบรรยายรูปภาพ	ผลการพยากรณ์	ผลการประเมินจากผู้เชี่ยวชาญ									
			1	2	3	4	5	6	7	8	9	10
 **Domain:** Anticipation	(A) a group of friends are sitting around a table with a dog and a bottle of beer.	0.0022	☑			☑						
	(B) the group of friends is eagerly waiting for the man to finish his beer so they can continue playing the wii.	0.2843		☑	☑						☑	
	(C) a group of friends enjoying a night in, playing video games and drinking beer.	0.0990					☑			☑		
	(D) the group of friends is enjoying a fun and lighthearted moment together.	0.0130										☑
	(E) a group of friends sitting around a table, sharing laughter and camaraderie while playing video games and drinking beer.	0.6015						☑	☑			
 **Domain:** Anticipation	(A) a baseball player in a red hat and uniform is walking across the field, holding a baseball glove.	0.3038			☑						☑	
	(B) a baseball player is walking across the field, holding a baseball glove, as he anticipates the next play in the game.	0.0718							☑	☑		
	(C) a baseball player in a red and white uniform walks across the field.	0.3487					☑					
	(D) the baseball player is walking on the field with a smile on his face, showing his optimism and determination to perform well in the game.	0.0007	☑	☑		☑		☑				☑
	(E) a baseball player in a red cap and white uniform is walking across the field.	0.2750										

ภาคผนวก ค

ตัวอย่างการคำนวณน้ำหนักโครงข่ายประสาทเทียม เมื่อกำหนดค่าน้ำหนักเริ่มต้นเท่ากับ 2

ตารางที่ ค.1 ตัวอย่างการคำนวณน้ำหนักโครงข่ายประสาทเทียม เมื่อกำหนดค่าน้ำหนักเริ่มต้นเท่ากับ 2

รูปภาพ	คำบรรยายรูปภาพ	ผู้เชี่ยวชาญ	ตัวเลือก	$\hat{y}$	Error	$\frac{\partial L}{\partial \hat{y}}$	$\frac{\partial z}{\partial W_1}$	Error at W_1	Update W_1
 Domain: Joy	(A) two women walking down a street, hugging each other, enjoying the moment together. (B) two women walking arm in arm, enjoying each other's company. (C) two women walking down a street, hugging each other, enjoying the moment. (D) two women walking down a street, holding hands and smiling, enjoying each other's company. (E) two women walking down the street, holding hands and sharing a love that transcends time.	1	A	0.2467	2.0192	-1.5066	0.1858	-0.0140	0.2070
		2	A	0.2467	2.0192	-1.5066	0.1858	-0.0140	0.2070
		3	B	0.1262	-2.9862	-1.7476	0.2156	-0.0188	0.2094
		4	A	0.2467	2.0192	-1.5066	0.1858	-0.0140	0.2070
		5	C	0.2418	-2.0481	-1.5164	0.1870	-0.0142	0.2071
		6	C	0.2418	-2.0481	-1.5164	0.1870	-0.0142	0.2071
		7	A	0.2467	2.0192	-1.5066	0.1858	-0.0140	0.2070
		8	D	0.2166	-2.2069	-1.5668	0.1933	-0.0151	0.2076
		9	D	0.2166	-2.2069	-1.5668	0.1933	-0.0151	0.2076
		10	A	0.2467	2.0192	-1.5066	0.1858	-0.0140	0.2070
		ค่าถ่วงน้ำหนักเฉลี่ย							**0.2074**
 Domain: Joy	(A) a woman sits on the ground playing with a cat. (B) a woman sitting on a brick walkway playing with a cat. (C) a woman sitting on a brick street, petting a cat. (D) a woman sitting on the ground petting a cat, surrounded by potted plants and a handbag. (E) a woman sitting on the ground playing with a cat.	1	B	0.0219	-5.5116	-1.9562	0.0214	-0.0021	0.2070
		2	A	0.5106	0.9697	-0.9788	0.2499	-0.0122	0.2061
		3	A	0.5106	0.9697	-0.9788	0.2499	-0.0122	0.2061
		4	A	0.5106	0.9697	-0.9788	0.2499	-0.0122	0.2061
		5	B	0.0219	-5.5116	-1.9562	0.0214	-0.0021	0.2010
		6	C	0.0512	-4.2867	-1.8975	0.0486	-0.0046	0.2023
		7	D	0.3103	-1.6884	-1.3794	0.2140	-0.0148	0.2074
		8	E	0.1059	-3.2386	-1.7881	0.0947	-0.0085	0.2042
		9	B	0.0219	-5.5116	-1.9562	0.0214	-0.0021	0.2010
		10	A	0.5106	0.9697	-0.9788	0.2499	-0.0122	0.2061
		ค่าถ่วงน้ำหนักเฉลี่ย							**0.2047**

ตารางที่ ค.1 ตัวอย่างการคำนวณน้ำหนักโครงข่ายประสาทเทียม เมื่อกำหนดค่าน้ำหนักเริ่มต้นเท่ากับ 2 (ต่อ)

รูปภาพ	คำบรรยายรูปภาพ	ผู้เชี่ยวชาญ	ตัวเลือก	$\hat{y}$	Error	$\frac{\partial L}{\partial \hat{y}}$	$\frac{\partial z}{\partial W_1}$	Error at W_1	Update W_1
 **Domain:** Trust	(A) the little girl is so adorable as she prays in front of the lit candle. (B) a young girl praying in front of a lit candle. (C) the little girl's eyes are closed as she blows out the candles on her birthday cake. (D) a beautiful birthday cake with lit candles, ready to be blown out by a little girl. (E) a young girl praying in front of a lit candle, and the girl is making a wish.	1	A	0.2175	2.2008	-1.5650	0.1702	-0.0133	0.2070
		2	A	0.2175	2.2008	-1.5650	0.1702	-0.0133	0.2067
		3	C	0.1860	-2.4269	-1.6281	0.1514	-0.0123	0.2062
		4	A	0.2175	2.2008	-1.5650	0.1702	-0.0133	0.2067
		5	E	0.1513	-2.7245	-1.6974	0.1284	-0.0109	0.2054
		6	C	0.1860	-2.4269	-1.6281	0.1514	-0.0123	0.2062
		7	B	0.2506	-1.9967	-1.4988	0.1878	-0.0141	0.2070
		8	E	0.1513	-2.7245	-1.6974	0.1284	-0.0109	0.2054
		9	B	0.2506	-1.9967	-1.4988	0.1878	-0.0141	0.2070
		10	B	0.2506	-1.9967	-1.4988	0.1878	-0.0141	0.2070
		ค่าถ่วงน้ำหนักเฉลี่ย							**0.2065**
 **Domain:** Trust	(A) the man is in a meditative state on the wooden platform. (B) the man is in a peaceful and contemplative state of mind as he sits on the wooden floor, kneeling on a green rug. (C) a man praying on a wooden deck with a green rug. (D) the man is in a peaceful and contemplative state of mind while kneeling on a rug. (E) a man in a white shirt and black pants is kneeling on a green rug, praying on a deck.	1	A	0.2603	1.9419	-1.4794	0.1925	-0.0142	0.2070
		2	B	0.0028	-8.4631	-1.9943	0.0028	-0.0003	0.2001
		3	B	0.0028	-8.4631	-1.9943	0.0028	-0.0003	0.2001
		4	D	0.3699	-1.4350	-1.2603	0.2331	-0.0147	0.2073
		5	A	0.2603	1.9419	-1.4794	0.1925	-0.0142	0.2071
		6	C	0.1688	-2.5664	-1.6623	0.1403	-0.0117	0.2058
		7	C	0.1688	-2.5664	-1.6623	0.1403	-0.0117	0.2058
		8	E	0.1982	-2.3350	-1.6036	0.1589	-0.0127	0.2064
		9	A	0.2603	1.9419	-1.4794	0.1925	-0.0142	0.2071
		10	B	0.0028	-8.4631	-1.9943	0.0028	-0.0003	0.2001
		ค่าถ่วงน้ำหนักเฉลี่ย							**0.2047**

ตารางที่ ค.1 ตัวอย่างการคำนวณน้ำหนักโครงข่ายประสาทเทียม เมื่อกำหนดค่าน้ำหนักเริ่มต้นเท่ากับ 2 (ต่อ)

รูปภาพ	คำบรรยายรูปภาพ	ผู้เชี่ยวชาญ	ตัวเลือก	$\hat{y}$	Error	$\frac{\partial L}{\partial \hat{y}}$	$\frac{\partial z}{\partial W_1}$	Error at W_1	Update W_1
 **Domain:** Fear	(A) the boy's face is covered in makeup, giving him a scary appearance. (B) the boy's face is smeared with fake blood, making him look scared. (C) he is sitting on the ground, holding a hot dog, which adds to the eerie atmosphere of the scene. (D) he appears to be enjoying the thrill of being scary and possibly participating in a halloween event or a costume party. (E) the boy is feeling a sense of excitement or playfulness as he is dressed up as a zombie and eating a hot dog.	1	B	0.8423	-0.2476	-0.3154	0.1328	-0.0021	0.2070
		2	B	0.8423	-0.2476	-0.3154	0.1328	-0.0021	0.2010
		3	B	0.8423	-0.2476	-0.3154	0.1328	-0.0021	0.2010
		4	C	0.0046	-7.7642	-1.9908	0.0046	-0.0005	0.2002
		5	A	0.0508	4.2990	-1.8984	0.0482	-0.0046	0.2023
		6	A	0.0508	4.2990	-1.8984	0.0482	-0.0046	0.2023
		7	C	0.0046	-7.7642	-1.9908	0.0046	-0.0005	0.2002
		8	D	0.0051	-7.6153	-1.9898	0.0051	-0.0005	0.2003
		9	B	0.8423	-0.2476	-0.3154	0.1328	-0.0021	0.2010
		10	B	0.8423	-0.2476	-0.3154	0.1328	-0.0021	0.2010
		ค่าถ่วงน้ำหนักเฉลี่ย							**0.2016**
 **Domain:** Fear	(A) the two apples are screaming. (B) a pair of apples with faces drawn on them, one with a screaming face and the other (C) the apple with the knife stuck in it has a scared face. (D) the apple's face is being sliced, and it's not happy about it. (E) the image of the two apples with faces drawn on them and a knife sticking out of one of them evokes a sense of shock or surprise.	1	A	0.0016	9.2877	-1.9968	0.0016	-0.0002	0.2070
		2	A	0.0016	9.2877	-1.9968	0.0016	-0.0002	0.2001
		3	B	0.2540	-1.9771	-1.4920	0.1895	-0.0141	0.2071
		4	B	0.2540	-1.9771	-1.4920	0.1895	-0.0141	0.2071
		5	B	0.2540	-1.9771	-1.4920	0.1895	-0.0141	0.2071
		6	D	0.0055	-7.5064	-1.9890	0.0055	-0.0005	0.2003
		7	C	0.0020	-8.9658	-1.9960	0.0020	-0.0002	0.2001
		8	C	0.0020	-8.9658	-1.9960	0.0020	-0.0002	0.2001
		9	B	0.2540	-1.9771	-1.4920	0.1895	-0.0141	0.2071
		10	A	0.0016	9.2877	-1.9968	0.0016	-0.0002	0.2001
		ค่าถ่วงน้ำหนักเฉลี่ย							**0.2036**

ตารางที่ ค.1 ตัวอย่างการคำนวณน้ำหนักโครงข่ายประสาทเทียม เมื่อกำหนดค่าน้ำหนักเริ่มต้นเท่ากับ 2 (ต่อ)

รูปภาพ	คำบรรยายรูปภาพ	ผู้เชี่ยวชาญ	ตัวเลือก	$\hat{y}$	Error	$\frac{\partial L}{\partial \hat{y}}$	$\frac{\partial z}{\partial W_1}$	Error at W_1	Update W_1
 **Domain:** Surprise	(A) the woman is amazed by the taste of her food. (B) the woman is surprised or shocked by something she has just seen or heard. (C) the woman is laughing at something. (D) the woman is shocked by something she sees or hears. (E) the woman is covering her mouth with her hand, possibly expressing disapproval or shock at something she has seen or heard.	1	B	0.3369	-1.5696	-1.3262	0.2234	-0.0148	0.2070
		2	B	0.3369	-1.5696	-1.3262	0.2234	-0.0148	0.2074
		3	B	0.3369	-1.5696	-1.3262	0.2234	-0.0148	0.2074
		4	D	0.0621	-4.0093	-1.8758	0.0582	-0.0055	0.2027
		5	C	0.0233	-5.4235	-1.9534	0.0228	-0.0022	0.2011
		6	D	0.0621	-4.0093	-1.8758	0.0582	-0.0055	0.2027
		7	B	0.3369	-1.5696	-1.3262	0.2234	-0.0148	0.2074
		8	A	0.0234	5.4173	-1.9532	0.0229	-0.0022	0.2011
		9	B	0.3369	-1.5696	-1.3262	0.2234	-0.0148	0.2074
		10	B	0.3369	-1.5696	-1.3262	0.2234	-0.0148	0.2074
		ค่าถ่วงน้ำหนักเฉลี่ย							**0.2052**
 **Domain:** Surprise	(A) the two women are amazed at the large amount of food in front of them. (B) the two women are surprised to see a man in the background. (C) the two women are enjoying a meal together, with one of them smiling as she eats. they are surrounded by various bowls of food, including a large bowl of meat. (D) the two women are eating meat, but they should be eating more vegetables for a healthier diet. (E) two women sitting at a table, enjoying a meal together and laughing.	1	A	0.0338	4.8855	-1.9323	0.0327	-0.0032	0.2070
		2	B	0.0011	-9.8319	-1.9978	0.0011	-0.0001	0.2001
		3	D	0.0007	-10.5646	-1.9987	0.0007	-0.0001	0.2000
		4	A	0.0338	4.8855	-1.9323	0.0327	-0.0032	0.2016
		5	A	0.0338	4.8855	-1.9323	0.0327	-0.0032	0.2016
		6	B	0.0011	-9.8319	-1.9978	0.0011	-0.0001	0.2001
		7	A	0.0338	4.8855	-1.9323	0.0327	-0.0032	0.2016
		8	B	0.0011	-9.8319	-1.9978	0.0011	-0.0001	0.2001
		9	C	0.4959	-1.0118	-1.0082	0.2500	-0.0126	0.2063
		10	A	0.0338	4.8855	-1.9323	0.0327	-0.0032	0.2016
		ค่าถ่วงน้ำหนักเฉลี่ย							**0.2020**

ตารางที่ ค.1 ตัวอย่างการคำนวณน้ำหนักโครงข่ายประสาทเทียม เมื่อกำหนดค่าน้ำหนักเริ่มต้นเท่ากับ 2 (ต่อ)

รูปภาพ	คำบรรยายรูปภาพ	ผู้เชี่ยวชาญ	ตัวเลือก	$\hat{y}$	Error	$\frac{\partial L}{\partial \hat{y}}$	$\frac{\partial z}{\partial W_1}$	Error at W_1	Update W_1
 **Domain:** Sadness	(A) the woman is crying. (B) a woman lying on a couch, feeling sad and lonely. (C) a girl with a broken heart laying on a couch. (D) the woman is feeling sad or lonely while laying on the couch. (E) the woman is feeling a sense of frustration or disappointment as she lays on the couch with her mouth open.	1	B	0.4753	-1.0731	-1.0494	0.2494	-0.0131	0.2070
		2	B	0.4753	-1.0731	-1.0494	0.2494	-0.0131	0.2065
		3	A	0.0124	6.3335	-1.9752	0.0122	-0.0012	0.2006
		4	C	0.0078	-7.0023	-1.9844	0.0077	-0.0008	0.2004
		5	C	0.0078	-7.0023	-1.9844	0.0077	-0.0008	0.2004
		6	D	0.0650	-3.9434	-1.8700	0.0608	-0.0057	0.2028
		7	D	0.0650	-3.9434	-1.8700	0.0608	-0.0057	0.2028
		8	B	0.4753	-1.0731	-1.0494	0.2494	-0.0131	0.2065
		9	B	0.4753	-1.0731	-1.0494	0.2494	-0.0131	0.2065
		10	B	0.4753	-1.0731	-1.0494	0.2494	-0.0131	0.2065
		ค่าถ่วงน้ำหนักเฉลี่ย							**0.2040**
 **Domain:** Sadness	(A) a small wooden figure sits next to a remote control, looking sad. (B) the two wooden figures sitting on the couch, one of them holding a remote, appear to be sad. (C) a small wooden figure sits next to a remote control, seemingly lonely and longing for companionship. (D) a small wooden figure sitting next to a remote control. (E) a small wooden figure is sitting on a couch next to a remote control. the figure appears to be looking at the remote with a sense of curiosity or longing, as if it wishes to control the television.	1	A	0.7742	0.3692	-0.4516	0.1748	-0.0039	0.2070
		2	C	0.0896	-3.4804	-1.8208	0.0816	-0.0074	0.2037
		3	B	0.0238	-5.3929	-1.9524	0.0232	-0.0023	0.2011
		4	B	0.0238	-5.3929	-1.9524	0.0232	-0.0023	0.2011
		5	A	0.7742	0.3692	-0.4516	0.1748	-0.0039	0.2020
		6	A	0.7742	0.3692	-0.4516	0.1748	-0.0039	0.2020
		7	B	0.0238	-5.3929	-1.9524	0.0232	-0.0023	0.2011
		8	E	0.0752	-3.7331	-1.8496	0.0695	-0.0064	0.2032
		9	B	0.0238	-5.3929	-1.9524	0.0232	-0.0023	0.2011
		10	B	0.0238	-5.3929	-1.9524	0.0232	-0.0023	0.2011
		ค่าถ่วงน้ำหนักเฉลี่ย							**0.2024**

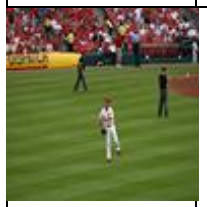
ตารางที่ ค.1 ตัวอย่างการคำนวณน้ำหนักโครงข่ายประสาทเทียม เมื่อกำหนดค่าน้ำหนักเริ่มต้นเท่ากับ 2 (ต่อ)

รูปภาพ	คำบรรยายรูปภาพ	ผู้เชี่ยวชาญ	ตัวเลือก	$\hat{y}$	Error	$\frac{\partial L}{\partial \hat{y}}$	$\frac{\partial z}{\partial W_1}$	Error at W_1	Update W_1
 **Domain:** Disgust	(A) a woman in a red sweater looks at a plate of food with disgust. (B) a woman is disgusted by the dessert on the table. (C) a woman in a red sweater is bored at the table with a plate of food. (D) a woman sitting at a table with a plate of food in front of her, looking at the food with a frown on her face. (E) a woman in a red sweater is about to eat a dessert.	1	A	0.7537	0.4079	-0.4926	0.1856	-0.0046	0.2070
		2	A	0.7537	0.4079	-0.4926	0.1856	-0.0046	0.2023
		3	A	0.7537	0.4079	-0.4926	0.1856	-0.0046	0.2023
		4	D	0.0291	-5.1028	-1.9418	0.0283	-0.0027	0.2014
		5	D	0.0291	-5.1028	-1.9418	0.0283	-0.0027	0.2014
		6	A	0.7537	0.4079	-0.4926	0.1856	-0.0046	0.2023
		7	E	0.0129	-6.2765	-1.9742	0.0127	-0.0013	0.2006
		8	C	0.0682	-3.8741	-1.8636	0.0635	-0.0059	0.2030
		9	A	0.7537	0.4079	-0.4926	0.1856	-0.0046	0.2023
		10	A	0.7537	0.4079	-0.4926	0.1856	-0.0046	0.2023
		ค่าถ่วงน้ำหนักเฉลี่ย							**0.2025**
 **Domain:** Disgust	(A) a man is eating a piece of food that looks like a muscle man. he is making a face and appears to be loathing the food. (B) a man is making a disgusted face while holding a piece of food on a toothpick. (C) a man is bored eating chopsticks. (D) the man looks sad as he eats a piece of food shaped like a man. (E) the man is eating a piece of food that looks like a man.	1	B	0.3700	-1.4344	-1.2600	0.2331	-0.0147	0.2070
		2	B	0.3700	-1.4344	-1.2600	0.2331	-0.0147	0.2073
		3	A	0.0887	3.4949	-1.8226	0.0808	-0.0074	0.2037
		4	B	0.3700	-1.4344	-1.2600	0.2331	-0.0147	0.2073
		5	B	0.3700	-1.4344	-1.2600	0.2331	-0.0147	0.2073
		6	B	0.3700	-1.4344	-1.2600	0.2331	-0.0147	0.2073
		7	E	0.0036	-8.1178	-1.9928	0.0036	-0.0004	0.2002
		8	D	0.0658	-3.9258	-1.8684	0.0615	-0.0057	0.2029
		9	B	0.3700	-1.4344	-1.2600	0.2331	-0.0147	0.2073
		10	A	0.0887	3.4949	-1.8226	0.0808	-0.0074	0.2037
		ค่าถ่วงน้ำหนักเฉลี่ย							**0.2054**

ตารางที่ ค.1 ตัวอย่างการคำนวณน้ำหนักโครงข่ายประสาทเทียม เมื่อกำหนดค่าน้ำหนักเริ่มต้นเท่ากับ 2 (ต่อ)

รูปภาพ	คำบรรยายรูปภาพ	ผู้เชี่ยวชาญ	ตัวเลือก	$\hat{y}$	Error	$\frac{\partial L}{\partial \hat{y}}$	$\frac{\partial z}{\partial W_1}$	Error at W_1	Update W_1
 **Domain:** Anger	(A) a dog is growling at another dog while laying on the floor. (B) the angry dog is about to bite the smaller dog. (C) the dog is yawning and the other dog is trying to bite him. (D) a white dog with its mouth open and a brown dog underneath it. (E) a large white dog is growling at a small brown dog, showing its teeth and aggressive behavior.	1	E	0.8323	-0.2648	-0.3354	0.1396	-0.0023	0.2070
		2	B	0.0295	-5.0831	-1.9410	0.0286	-0.0028	0.2014
		3	D	0.1144	-3.1278	-1.7712	0.1013	-0.0090	0.2045
		4	B	0.0295	-5.0831	-1.9410	0.0286	-0.0028	0.2014
		5	A	0.0018	9.1178	-1.9964	0.0018	-0.0002	0.2001
		6	B	0.0295	-5.0831	-1.9410	0.0286	-0.0028	0.2014
		7	D	0.1144	-3.1278	-1.7712	0.1013	-0.0090	0.2045
		8	C	0.0219	-5.5129	-1.9562	0.0214	-0.0021	0.2010
		9	E	0.8323	-0.2648	-0.3354	0.1396	-0.0023	0.2012
		10	B	0.0295	-5.0831	-1.9410	0.0286	-0.0028	0.2014
		ค่าถ่วงน้ำหนักเฉลี่ย							**0.2024**
 **Domain:** Anger	(A) a man is yelling and jumping in the air while surrounded by people. (B) (B the man is yelling with his arms raised, showing his anger. (C) a man yelling with his arms up in the air, annoying everyone around him. (D) a man is being lifted up by his friends while he is yelling. (E) a man is being lifted off the ground by a group of people, and he is yelling with his mouth open.	1	D	0.1794	-2.4787	-1.6412	0.1472	-0.0121	0.2070
		2	C	0.2096	-2.2543	-1.5808	0.1657	-0.0131	0.2065
		3	C	0.2096	-2.2543	-1.5808	0.1657	-0.0131	0.2065
		4	E	0.0157	-5.9931	-1.9686	0.0155	-0.0015	0.2008
		5	D	0.1794	-2.4787	-1.6412	0.1472	-0.0121	0.2060
		6	A	0.0403	4.6331	-1.9194	0.0387	-0.0037	0.2019
		7	A	0.0403	4.6331	-1.9194	0.0387	-0.0037	0.2019
		8	A	0.0403	4.6331	-1.9194	0.0387	-0.0037	0.2019
		9	C	0.2096	-2.2543	-1.5808	0.1657	-0.0131	0.2065
		10	C	0.2096	-2.2543	-1.5808	0.1657	-0.0131	0.2065
		ค่าถ่วงน้ำหนักเฉลี่ย							**0.2046**

ตารางที่ ค.1 ตัวอย่างการคำนวณน้ำหนักโครงข่ายประสาทเทียม เมื่อกำหนดค่าน้ำหนักเริ่มต้นเท่ากับ 2 (ต่อ)

รูปภาพ	คำบรรยายรูปภาพ	ผู้เชี่ยวชาญ	ตัวเลือก	$\hat{y}$	Error	$\frac{\partial L}{\partial \hat{y}}$	$\frac{\partial z}{\partial W_1}$	Error at W_1	Update W_1
 **Domain:** Anticipation	(A) a group of friends are sitting around a table with a dog and a bottle of beer. (B) the group of friends is eagerly waiting for the man to finish his beer so they can continue playing the wii. (C) a group of friends enjoying a night in, playing video games and drinking beer. (D) the group of friends is enjoying a fun and lighthearted moment together. (E) a group of friends sitting around a table, sharing laughter and camaraderie while playing video games and drinking beer.	1	A	0.0022	8.8550	-1.9957	0.0022	-0.0002	0.2070
		2	B	0.2843	-1.8143	-1.4313	0.2035	-0.0146	0.2073
		3	B	0.2843	-1.8143	-1.4313	0.2035	-0.0146	0.2073
		4	A	0.0022	8.8550	-1.9957	0.0022	-0.0002	0.2001
		5	C	0.0990	-3.3360	-1.8019	0.0892	-0.0080	0.2040
		6	E	0.6015	-0.7334	-0.7970	0.2397	-0.0096	0.2048
		7	E	0.6015	-0.7334	-0.7970	0.2397	-0.0096	0.2048
		8	C	0.0990	-3.3360	-1.8019	0.0892	-0.0080	0.2040
		9	B	0.2843	-1.8143	-1.4313	0.2035	-0.0146	0.2073
		10	D	0.0130	-6.2662	-1.9740	0.0128	-0.0013	0.2006
		ค่าถ่วงน้ำหนักเฉลี่ย							**0.2047**
 **Domain:** Anticipation	(A) a baseball player in a red hat and uniform is walking across the field, holding a baseball glove. (B) a baseball player is walking across the field, holding a baseball glove, as he anticipates the next play in the game. (C) a baseball player in a red and white uniform walks across the field. (D) the baseball player is walking on the field with a smile on his face, showing his optimism and determination to perform well in the game. (E) a baseball player in a red cap and white uniform is walking across the field.	1	D	0.0007	-10.3928	-1.9985	0.0007	-0.0001	0.2070
		2	D	0.0007	-10.3928	-1.9985	0.0007	-0.0001	0.2000
		3	A	0.3038	1.7188	-1.3924	0.2115	-0.0147	0.2074
		4	D	0.0007	-10.3928	-1.9985	0.0007	-0.0001	0.2000
		5	C	0.3487	-1.5198	-1.3025	0.2271	-0.0148	0.2074
		6	D	0.0007	-10.3928	-1.9985	0.0007	-0.0001	0.2000
		7	B	0.0718	-3.8006	-1.8565	0.0666	-0.0062	0.2031
		8	B	0.0718	-3.8006	-1.8565	0.0666	-0.0062	0.2031
		9	A	0.3038	1.7188	-1.3924	0.2115	-0.0147	0.2074
		10	D	0.0007	-10.3928	-1.9985	0.0007	-0.0001	0.2000
		ค่าถ่วงน้ำหนักเฉลี่ย							**0.2038**

ภาคผนวก ง
การสุ่มข้อความค้นหาที่สร้างจาก ChatGPT

ตารางที่ ง.1 คำศัพท์จากการสุ่มและตัวอย่างข้อความค้นหาที่สร้างจากคำศัพท์ด้วย ChatGPT

คำศัพท์		เงื่อนไขที่ 1	เงื่อนไขที่ 2	เงื่อนไขที่ 3
หมวดคน สัตว์ สิ่งของ	1. dog	photo of a dog	the dog playing on ground	the dog wagged its tail happily
	2. baby	photo of a baby	the baby sleeps peacefully in bed	the baby sleeps peacefully, cuddled in her mother's arms
	3. man	photo of a man	the man wears a blue suit	the man smiles warmly as he plays catch with his dog
	4. woman	photo of a woman	the woman smiles brightly	the woman laughs joyfully while chatting with her friends
	5. guitar	photo of a guitar	the man playing the guitar	the guitarist strums passionately, lost in the melody
	6. cat	photo of a cat	a black cat lounges on the windowsill	the cat purrs contentedly as it curls up in her lap
	7. mirror	photo of a mirror	the mirror reflects the room	she gazes into the mirror, admiring her reflection with confidence
	8. chair	photo of a chair	a wooden chair sits by the table	he sinks into the chair, feeling relieved after a long day
	9. glasses	photo of a glasses	she wears round glasses	she adjusts her glasses, feeling focused and ready to work
	10. hat	photo of a hat	the hat sits on the coat rack	he tips his hat with a grin, greeting his neighbors on the street
หมวดหมวดอาหารและเครื่องดื่ม	11. pizza	photo of a pizza	we enjoyed a delicious pizza for dinner	she felt excited to order her favorite pizza for dinner
	12. salad	photo of a salad	she tossed a fresh salad with greens and tomatoes	he felt refreshed after eating a crisp, green salad
	13. beer	photo of a beer	he drank a cold beer on a hot summer day	they felt satisfied as they sipped on their favorite beer
	14. water	photo of a water	she quenched her thirst with a glass of cold water	she felt rejuvenated after drinking a glass of cold water
	15. coffee	photo of a coffee	he savored a cup of hot coffee in the morning	he felt energized after his morning cup of coffee
	16. bread	photo of a bread	the bakery smelled of freshly baked bread	the warm aroma of freshly baked bread made her feel cozy
	17. milk	photo of a milk	she poured a glass of cold milk for her cereal	they felt comforted by the creamy texture of cold milk
	18. tea	photo of a tea	he brewed a soothing cup of tea before bed	she felt relaxed as she sipped on a soothing cup of tea
	19. cheese	photo of a cheese	she sliced some cheese to go with crackers	he felt indulgent as he nibbled on a piece of rich cheese
	20. juice	photo of a juice	they enjoyed refreshing fruit juice on a afternoon	they felt refreshed by the sweet and tangy taste of fresh fruit juice

ตารางที่ ง.1 คำศัพท์จากการสุ่มและตัวอย่างข้อความค้นหาที่สร้างจากคำศัพท์ด้วย ChatGPT (ต่อ)

คำศัพท์		เงื่อนไขที่ 1	เงื่อนไขที่ 2	เงื่อนไขที่ 3
หมวดการท่องเที่ยว	21. tourist	photo of a tourist	the tourist snapped photos of the famous landmark	the tourist felt excited to explore the new city
	22. hotel	photo of a hotel	they checked into the cozy hotel for the night	they felt relieved to finally check into the comfortable hotel
	23. park	photo of a park	families enjoyed picnics in the sunny park	families felt happy as they played together in the sunny park
	24. postcard	photo of a postcard	she sent a postcard with a picture of the beach	she felt nostalgic as she wrote a postcard to her family back home
	25. beach	photo of a beach	children built sandcastles on the sandy beach	visitors felt relaxed as they lounged on the sandy beach
	26. car	photo of a car	they drove along the scenic route in their car	they felt adventurous as they embarked on a road trip in their car
	27. souvenir	photo of a souvenir	he bought a keychain as a souvenir of his trip	he felt delighted to find the perfect souvenir to remember his trip
	28. museum	photo of a museum	visitors explored ancient artifacts in the museum	visitors felt fascinated as they explored the exhibits in the museum
	29. train	photo of a train	they boarded the train for an exciting adventure	they felt excited as they boarded the train for their journey
	30. zoo	photo of a zoo	families admired the animals at the local zoo	families felt amazed as they observed the animals at the zoo
หมวดการศึกษา	31. test	photo of a test	she studied hard for the upcoming test	she felt anxious about the upcoming test
	32. homework	photo of a homework	he finished his homework before dinner	he felt relieved after completing his homework
	33. subject	photo of a subject	math is her favorite subject at school	math is his least favorite subject in school
	34. lesson	photo of a lesson	the teacher taught a valuable lesson about kindness	the teacher taught a valuable lesson about perseverance
	35. classroom	photo of a classroom	students sat quietly in the classroom during the lesson	students felt excited as they entered the classroom
	36. teacher	photo of a teacher	the teacher explained the lesson clearly to the students	the teacher felt proud of her students' progress
	37. exam	photo of an exam	they felt nervous before the final exam	they felt nervous before the final exam
	38. school	photo of a school	children walked to school together in the morning	children felt excited to start the new school year
	39. student	photo of a student	he is a dedicated student who always completes his assignments	he felt determined to excel as a student
	40. library	photo of a library	she borrowed books from the library for her research project	she felt peaceful while studying in the quiet library

ตารางที่ ง.1 คำศัพท์จากการสุ่มและตัวอย่างข้อความค้นหาที่สร้างจากคำศัพท์ด้วย ChatGPT (ต่อ)

คำศัพท์		เงื่อนไขที่ 1	เงื่อนไขที่ 2	เงื่อนไขที่ 3
หมวดสุขภาพ	41. doctor	photo of a doctor	the doctor smiled kindly at the patient	the doctor's reassuring smile comforted the patient
	42. healthy	photo of a healthy	eating fruits and vegetables keeps you healthy	eating well and staying active keeps you healthy and happy
	43. soccer	photo of a soccer	they had fun playing a soccer game at the park	the soccer game stirred up feelings of teamwork, competitiveness and excitement
	44. therapy	photo of a therapy	she attends therapy sessions to manage her anxiety	therapy sessions help people manage their emotions and stress
	45. nurse	photo of a nurse	the nurse checked the patient's vitals	the nurse's caring touch provided comfort during the hospital stay
	46. exercise	photo of an exercise	regular exercise is good for your body	regular exercise boosts mood and reduces stress levels
	47. stress	photo of a stress	deep breathing helps to reduce stress	deep breathing exercises can ease stress and tension
	48. hospital	photo of a hospital	he was admitted to the hospital for surgery	being in the hospital can be stressful, but the staff works to make patients comfortable
	49. sleep	photo of a sleep	good sleep is essential for overall well-being	good sleep is vital for mental and physical well-being
	50. medicine	photo of a medicine	she took her medicine after breakfast	taking medicine as prescribed helps maintain good health
หมวดธรรมชาติ	51. lake	photo of a lake	the lake shimmered under the afternoon sun	sitting by the tranquil lake filled her with peace
	52. daisy flower	photo of a daisy flower	colorful daisy flowers bloomed in the garden	admiring the colorful daisy flowers in bloom filled her with happiness
	53. moon	photo of a moon	the moon cast a soft glow over the nighttime landscape	gazing at the full moon stirred up feelings of nostalgia
	54. star	photo of a star	twinkling stars filled the dark sky	counting the stars in the clear sky sparked joy in her heart
	55. mountain	photo of a mountain	the mountain peak stood tall against the horizon	climbing the mountain filled him with a sense of accomplishment
	56. tree	photo of a tree	a solitary tree stood at the edge of the field	finding shade under the tall tree brought relief from the heat
	57. weather	photo of a weather	the weather was warm and sunny, perfect for a picnic	the sunny weather lifted everyone's spirits
	58. beach	photo of a beach	children played in the sand on the beach	walking along the sandy beach made her feel free and alive
	59. sunset	Photo of a sunset	she admired the beautiful sunset over the ocean	she felt peaceful watching the vibrant colors of the sunset
	60. sunrise	photo of a sunrise	the sunrise set behind the hills, painting the sky with hues of orange and pink	sunrise brought a feeling of hope and renewal

ตารางที่ ง.1 คำศัพท์จากการสุ่มและตัวอย่างข้อความค้นหาที่สร้างจากคำศัพท์ด้วย ChatGPT (ต่อ)

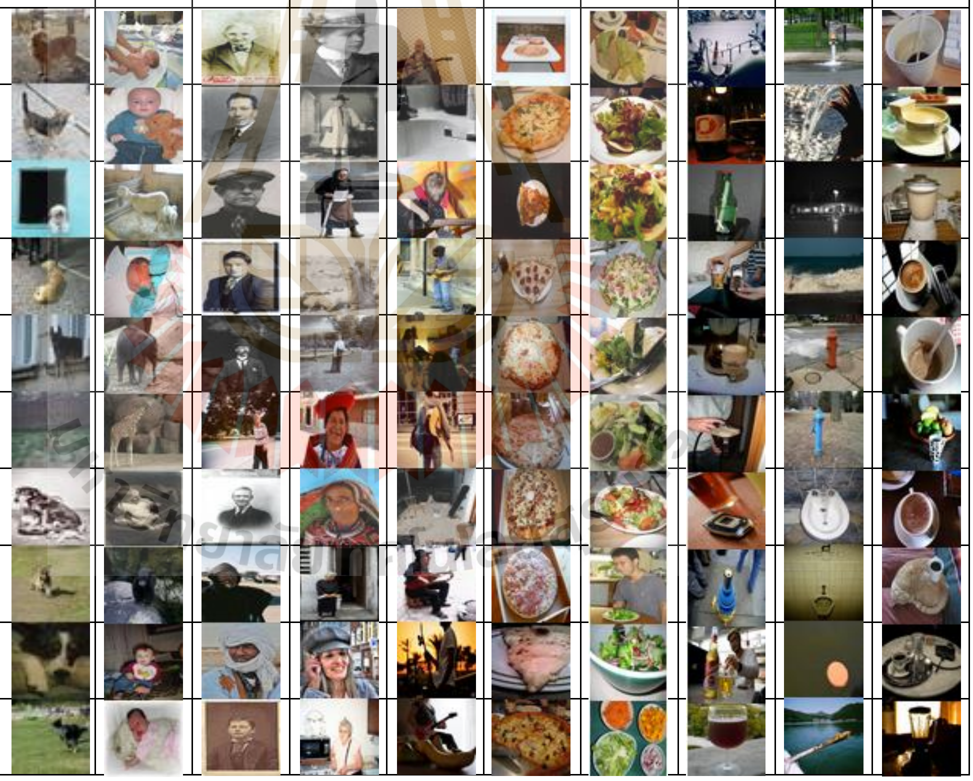
คำศัพท์		เงื่อนไขที่ 1	เงื่อนไขที่ 2	เงื่อนไขที่ 3
หมวดเทคโนโลยี	61. email	photo of an email	she sent an email to her coworker	she felt excited to receive an email from her friend
	62. internet	photo of an internet	he browsed the internet for news updates	he was relieved to find the information he needed on the internet
	63. video	photo of a video	they watched a funny video on their tv	they laughed together while watching a funny video
	64. game	photo of a game	he played a game on his console	he felt accomplished after winning the game
	65. social media	photo of a social media	she scrolled through social media on her tablet	she felt connected to her friends while scrolling through social media
	66. computer	photo of a computer	he worked on his computer to finish his project	he felt productive after finishing his work on the computer
	67. website	photo of a website	she visited her favorite website to read articles	she felt entertained while browsing her favorite website
	68. application	photo of an application	he downloaded a new application for his smartphone	he felt satisfied after downloading a useful application
	69. phone	photo of a phone	she answered her phone when it rang	she felt happy to hear her friend's voice on the phone
	70. camera	photo of a camera	he took a photo with his digital camera	he felt nostalgic as he looked through old photos on his camera
หมวดสื่อและความบันเทิง	71. book	photo of a book	she enjoyed reading a book in the cozy corner	she felt immersed in the story while reading a book
	72. movie	photo of a movie	they watched a movie together on friday night	they felt entertained watching a movie together
	73. concert	photo of a concert	he attended a concert with his friends last weekend	he felt exhilarated after attending the concert
	74. drama	photo of a drama	she loves watching drama series on tv	she felt captivated by the drama unfolding on stage
	75. theater	photo of a theater	they went to the theater to see a live performance	they felt enchanted by the magic of the theater
	76. music	photo of a music	he listened to music while studying	he felt relaxed listening to soothing music
	77. comedy	photo of a comedy	they laughed throughout the comedy show	they laughed heartily at the comedy show
	78. art	photo of an art	she admired the artwork at the gallery	she felt inspired by the beauty of the art gallery
	79. dance	photo of a dance	he danced with his partner at the party	he felt energized dancing with his friends
	80. museum	photo of a museum	she explored the exhibits at the museum	she felt curious exploring the museum's exhibits

ตารางที่ ง.1 คำศัพท์จากการสุ่มและตัวอย่างข้อความค้นหาที่สร้างจากคำศัพท์ด้วย ChatGPT (ต่อ)

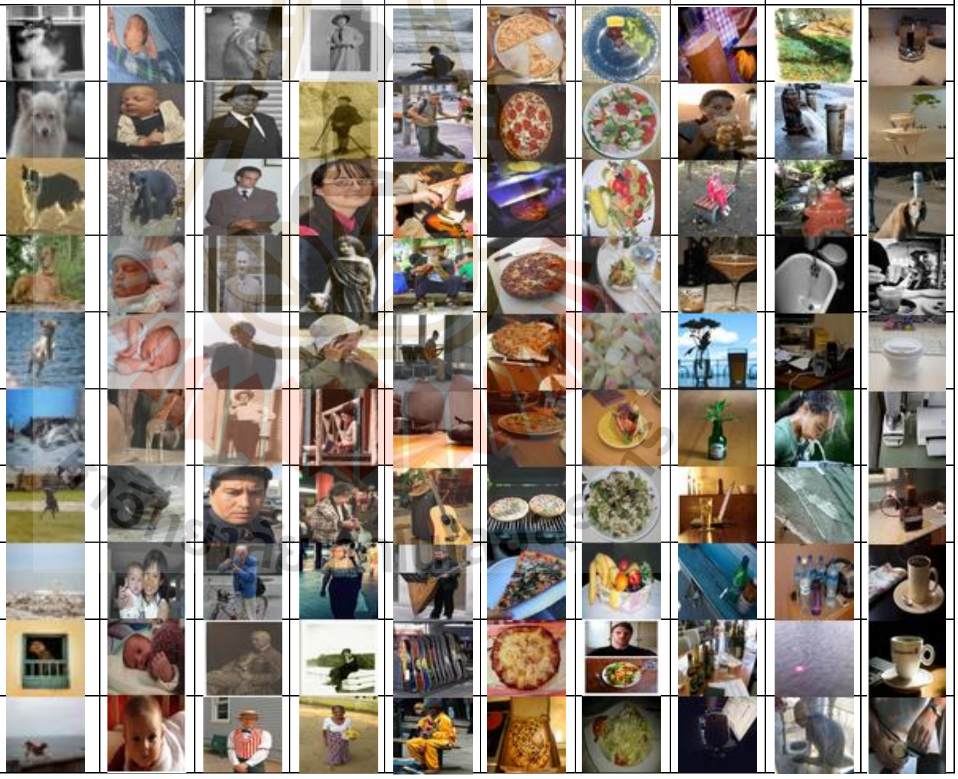
คำศัพท์		เงื่อนไขที่ 1	เงื่อนไขที่ 2	เงื่อนไขที่ 3
หมวดธุรกิจและเศรษฐกิจ	81. business	photo of a business	he started his own business last year	she felt excited to start her own business
	82. salary	photo of a salary	she was happy with her new salary increase	he was relieved when he received his salary
	83. marketing	photo of a marketing	they discussed marketing strategies for their product	they were proud of their successful marketing campaign
	84. budget	photo of a budget	he created a budget to manage his expenses	she felt organized after creating a budget
	85. customer	photo of a customer	she listened carefully to the customer's feedback	he smiled as he greeted each customer
	86. money	photo of a money	they saved money for their upcoming vacation	they felt secure knowing they had enough money saved
	87. profit	photo of a profit	the company made a significant profit this quarter	she was thrilled when the company made a profit
	88. product	photo of a product	he launched a new product on the market	he was confident in the quality of his product
	89. job	photo of a job	she found a job at a local cafe	she felt grateful for her new job opportunity
	90. service	photo of a service	they provided excellent service to their customers	they were happy to provide excellent service to their clients
หมวดวัฒนธรรมและสังคม	91. culture	photo of a culture	she learned about different cultures during her travels	she felt curious to explore the local culture
	92. community	photo of a community	he felt a sense of belonging in his community	he felt supported by his tight-knit community
	93. vacation	photo of a vacation	they enjoyed a relaxing vacation at the beach	they were excited to go on a vacation to the beach
	94. celebration	photo of a celebration	she attended a celebration for her friend's birthday	she felt joyful during the celebration of her graduation
	95. language	photo of a language	he practiced speaking a new language every day	he struggled to learn a new language but felt determined
	96. family	photo of a family	they gathered with their family for a holiday dinner	they cherished spending time together as a family
	97. country	photo of a country	she felt proud of her country's accomplishments	she felt patriotic pride for her country
	98. religion	photo of a religion	many people find comfort in their religion	her faith in her religion gives her strength during difficult times
	99. service	photo of a service	a true friend is always there for you	a true friend is always there to lend a helping hand when needed
	100. tradition	photo of a tradition	we enjoy our family traditions every year	family gathers to celebrate cultural traditions with love and joy

ภาคผนวก จ
ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับ

ตารางที่ จ.1 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับ ภายใต้เงื่อนไขข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ ในลักษณะป้ายกำกับของรูปภาพ

ข้อความค้นหา		หมวดคน สัตว์ สิ่งของ					หมวดอาหารและเครื่องดื่ม				
		1) photo of a dog	2) photo of a baby	3) photo of a man	4) photo of a woman	5) photo of a guitar	1) photo of a pizza	2) photo of a salad	3) photo of a beer	4) photo of a water	5) photo of a coffee
จำนวนรูปภาพจากการค้นคืน	3 ลำดับ										
	5 ลำดับ										
	10 ลำดับ										

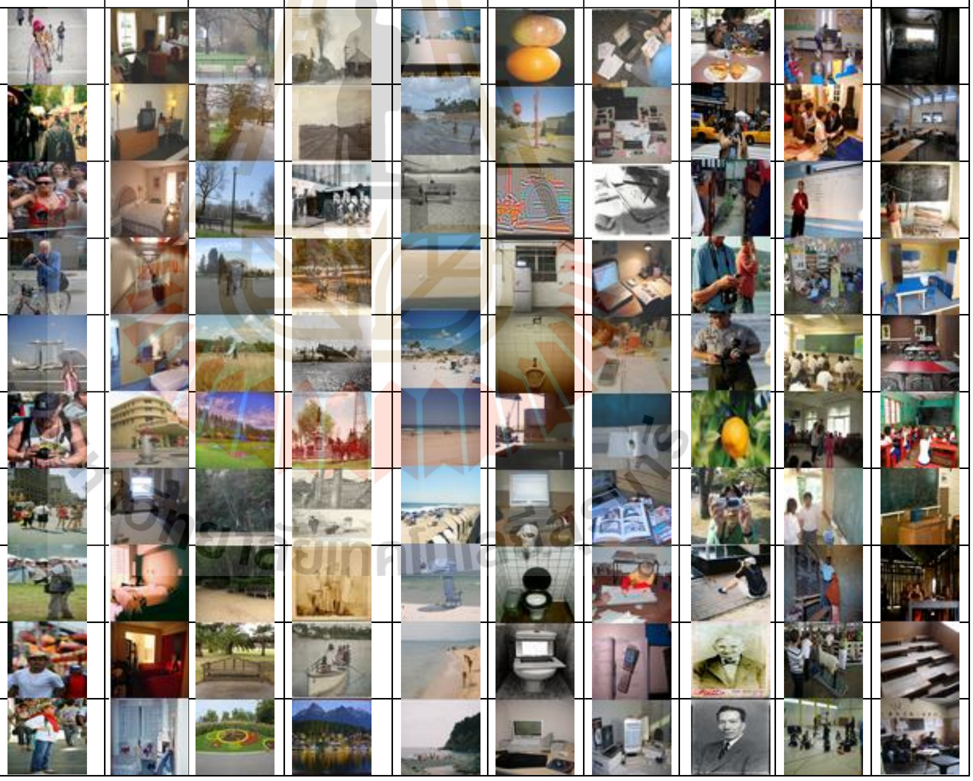
ตารางที่ จ.1 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับ ภายใต้เงื่อนไขข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ ในลักษณะป้ายกำกับของรูปภาพ (ต่อ)

ข้อความค้นหา		หมวดคน สัตว์ สิ่งของ					หมวดอาหารและเครื่องดื่ม				
		1) photo of a dog	2) photo of a baby	3) photo of a man	4) photo of a woman	5) photo of a guitar	1) photo of a pizza	2) photo of a salad	3) photo of a beer	4) photo of a water	5) photo of a coffee
จำนวนรูปภาพจากการค้นคืน	15 ลำดับ										
	20 ลำดับ										

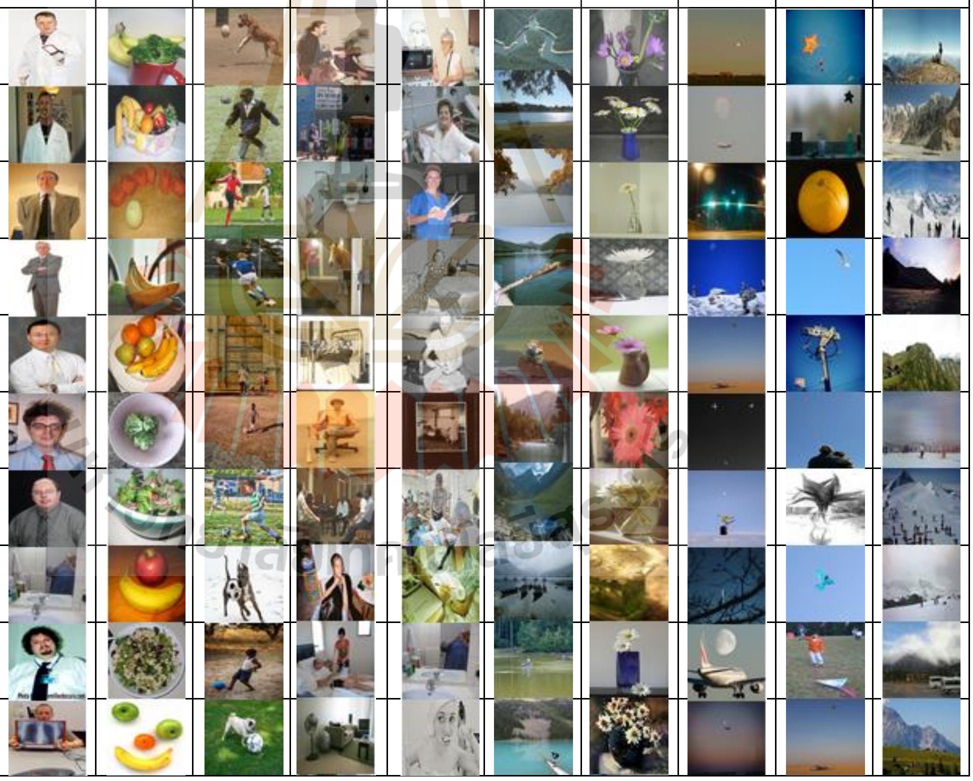
ตารางที่ จ.1 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับ ภายใต้เงื่อนไขข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ ในลักษณะป้ายกำกับของรูปภาพ (ต่อ)

ข้อความค้นหา		หมวดการท่องเที่ยว					หมวดการศึกษา				
		1) photo of a tourist	2) photo of a hotel	3) photo of a park	4) photo of a postcard	5) photo of a beach	1) photo of a test	2) photo of a homework	3) photo of a subject	4) photo of a lesson	5) photo of a classroom
จำนวนรูปภาพจากการค้นคืน	3 ลำดับ										
	5 ลำดับ										
	10 ลำดับ										

ตารางที่ จ.1 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับ ภายใต้เงื่อนไขข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ ในลักษณะป้ายกำกับของรูปภาพ (ต่อ)

ข้อความค้นหา		**หมวดการท่องเที่ยว**					**หมวดการศึกษา**				
		1) photo of a tourist	2) photo of a hotel	3) photo of a park	4) photo of a postcard	5) photo of a beach	1) photo of a test	2) photo of a homework	3) photo of a subject	4) photo of a lesson	5) photo of a classroom
จำนวนรูปภาพจากการค้นคืน	15 ลำดับ										
	20 ลำดับ										

ตารางที่ จ.1 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับ ภายใต้เงื่อนไขข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ ในลักษณะป้ายกำกับของรูปภาพ (ต่อ)

ข้อความค้นหา		หมวดสุขภาพ					หมวดธรรมชาติ				
		1) photo of a doctor	2) photo of a healthy	3) photo of a soccer	4) photo of a therapy	5) photo of a nurse	1) photo of a lake	2) photo of a daisy flower	3) photo of a moon	4) photo of a star	5) photo of a mountain
จำนวนรูปภาพจากการค้นคืน	3 ลำดับ										
	5 ลำดับ										
	10 ลำดับ										

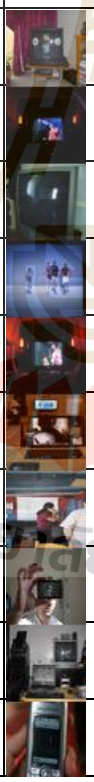
ตารางที่ จ.1 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับ ภายใต้เงื่อนไขข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ ในลักษณะป้ายกำกับของรูปภาพ (ต่อ)

ข้อความค้นหา		หมวดสุขภาพ					หมวดธรรมชาติ				
		1) photo of a doctor	2) photo of a healthy	3) photo of a soccer	4) photo of a therapy	5) photo of a nurse	1) photo of a lake	2) photo of a daisy flower	3) photo of a moon	4) photo of a star	5) photo of a mountain
จำนวนรูปภาพจากการค้นคืน	15 ลำดับ										
	20 ลำดับ										

ตารางที่ จ.1 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับ ภายใต้เงื่อนไขข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ ในลักษณะป้ายกำกับของรูปภาพ (ต่อ)

ข้อความค้นหา		หมวดเทคโนโลยี					หมวดศิลปะและการบันเทิง				
		1) photo of an email	2) photo of an internet	3) photo of a video	4) photo of a game	5) photo of a social media	1) photo of a book	2) photo of a movie	3) photo of a concert	4) photo of a drama	5) photo of a theater
จำนวนรูปภาพจากการค้นคืน	3 ลำดับ										
	5 ลำดับ										
	10 ลำดับ										

ตารางที่ จ.1 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับ ภายใต้เงื่อนไขข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ ในลักษณะป้ายกำกับของรูปภาพ (ต่อ)

ข้อความค้นหา		หมวดเทคโนโลยี					หมวดศิลปะและการบันเทิง				
		1) photo of a email	2) photo of a internet	3) photo of a video	4) photo of a game	5) photo of a social media	1) photo of a book	2) photo of a movie	3) photo of a concert	4) photo of a drama	5) photo of a theater
จำนวนรูปภาพจากการค้นคืน	15 ลำดับ										
	20 ลำดับ										

ตารางที่ จ.1 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับ ภายใต้เงื่อนไขข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ ในลักษณะป้ายกำกับของรูปภาพ (ต่อ)

ข้อความค้นหา		หมวดธุรกิจและการเงิน					หมวดสังคมและวัฒนธรรม				
		1) photo of a business	2) photo of a salary	3) photo of a marketing	4) photo of a budget	5) photo of a customer	1) photo of a culture	2) photo of a community	3) photo of a vocation	4) photo of a celebration	5) photo of a language
จำนวนรูปภาพจากการค้นคืน	3 ลำดับ										
	5 ลำดับ										
	10 ลำดับ										

ตารางที่ จ.1 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับ ภายใต้เงื่อนไขข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ ในลักษณะป้ายกำกับของรูปภาพ (ต่อ)

ข้อความค้นหา		หมวดธุรกิจและการเงิน					หมวดสังคมและวัฒนธรรม				
		1) photo of a business	2) photo of a salary	3) photo of a marketing	4) photo of a budget	5) photo of a customer	1) photo of a culture	2) photo of a community	3) photo of a vocation	4) photo of a celebration	5) photo of a language
จำนวนรูปภาพจากการค้นคืน	15 ลำดับ										
	20 ลำดับ										

ตารางที่ จ.2 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับ ภายใต้เงื่อนไขข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพ

ข้อความค้นหา		หมวดคน สัตว์ ลิ่งของ					หมวดอาหารและเครื่องดื่ม				
		1) the dog playing on ground	2) the baby sleeps peacefully in bed	3) the man wears a blue suit	4) the woman smiles brightly	5) the man playing the guitar	1) we enjoyed a delicious pizza for dinner	2) she tossed a fresh salad with greens and tomatoes	3) he drank a cold beer on a hot summer day	4) she quenched her thirst with a glass of cold water	5) he savored a cup of hot coffee in the morning
จำนวนรูปภาพจากการค้นคืน	3 ลำดับ										
	5 ลำดับ										
	10 ลำดับ										

ตารางที่ จ.2 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับ ภายใต้เงื่อนไขข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพ (ต่อ)

ข้อความค้นหา		หมวดคน สัตว์ สิ่งของ					หมวดอาหารและเครื่องดื่ม				
		1) the dog playing on ground	2) the baby sleeps peacefully in bed	3) the man wears a blue suit	4) the woman smiles brightly	5) the man playing the guitar	1) we enjoyed a delicious pizza for dinner	2) she tossed a fresh salad with greens and tomatoes	3) he drank a cold beer on a hot summer day	4) she quenched her thirst with a glass of cold water	5) he savored a cup of hot coffee in the morning
จำนวนรูปภาพจากการค้นคืน	15 ลำดับ										
	20 ลำดับ										

ตารางที่ จ.2 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับ ภายใต้เงื่อนไขข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพ (ต่อ)

ข้อความค้นหา		หมวดการท่องเที่ยว					หมวดการศึกษา				
		1) the tourist snapped photos of the famous landmark	2) they checked into the cozy hotel for the night	3) families enjoyed picnics in the sunny park	4) she sent a postcard with a picture of the beach	5) children built sandcastles on the sandy beach	1) she studied hard for the upcoming test	2) he finished his homework before dinner	3) math is her favorite subject at school	4) the teacher taught a valuable lesson about kindness	5) students sat quietly in the classroom during the lesson
จำนวนรูปภาพจากการค้นคืน	3 ลำดับ										
	5 ลำดับ										
	10 ลำดับ										

ตารางที่ จ.2 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับ ภายใต้เงื่อนไขข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพ (ต่อ)

ข้อความค้นหา		หมวดการท่องเที่ยว					หมวดการศึกษา				
		1) the tourist snapped photos of the famous landmark	2) they checked into the cozy hotel for the night	3) families enjoyed picnics in the sunny park	4) she sent a postcard with a picture of the beach	5) children built sandcastles on the sandy beach	1) she studied hard for the upcoming test	2) he finished his homework before dinner	3) math is her favorite subject at school	4) the teacher taught a valuable lesson about kindness	5) students sat quietly in the classroom during the lesson
จำนวนรูปภาพจากการค้นคืน	15 ลำดับ										
	20 ลำดับ										

ตารางที่ จ.2 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับ ภายใต้เงื่อนไขข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพ (ต่อ)

ข้อความค้นหา		หมวดสุขภาพ					หมวดธรรมชาติ				
		1) the doctor smiled kindly at the patient	2) eating fruits and vegetables keeps you healthy	3) they had fun playing a soccer game at the park	4) she attends therapy sessions to manage her anxiety	5) the nurse checked the patient's vitals	1) the lake shimmered under the afternoon sun	2) colorful daisy flowers bloomed in the garden	3) the moon cast a soft glow over the nighttime landscape	4) twinkling stars filled the dark sky	5) the mountain peak stood tall against the horizon
จำนวนรูปภาพจากการค้นคืน	3 ลำดับ										
	5 ลำดับ										
	10 ลำดับ										

ตารางที่ จ.2 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับ ภายใต้เงื่อนไขข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพ (ต่อ)

ข้อความค้นหา	หมวดสุขภาพ					หมวดธรรมชาติ				
	1) the doctor smiled kindly at the patient	2) eating fruits and vegetables keeps you healthy	3) they had fun playing a soccer game at the park	4) she attends therapy sessions to manage her anxiety	5) the nurse checked the patient's vitals	1) the lake shimmered under the afternoon sun	2) colorful daisy flowers bloomed in the garden	3) the moon cast a soft glow over the nighttime landscape	4) twinkling stars filled the dark sky	5) the mountain peak stood tall against the horizon
จำนวนรูปภาพจากการค้นคืน 15 ลำดับ										
20 ลำดับ										

ตารางที่ จ.2 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับ ภายใต้เงื่อนไขข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพ (ต่อ)

ข้อความค้นหา		หมวดเทคโนโลยี					หมวดศิลปะและการบันเทิง				
		1) she sent an email to her coworker	2) he browsed the internet for news updates	3) they watched a funny video on their tv	4) he played a game on his console	5) she scrolled through social media on her tablet	1) she enjoyed reading a book in the cozy corner	2) they watched a movie together on friday night	3) he attended a concert with his friends last weekend	4) she loves watching drama series on tv	5) they went to the theater to see a live performance
จำนวนรูปภาพจากการค้นคืน	3 ลำดับ										
	5 ลำดับ										
	10 ลำดับ										

ตารางที่ จ.2 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับ ภายใต้เงื่อนไขข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพ (ต่อ)

ข้อความค้นหา		หมวดเทคโนโลยี					หมวดศิลปะและการบันเทิง				
		1) she sent an email to her coworker	2) he browsed the internet for news updates	3) they watched a funny video on their tv	4) he played a game on his console	5) she scrolled through social media on her tablet	1) she enjoyed reading a book in the cozy corner	2) they watched a movie together on friday night	3) he attended a concert with his friends last weekend	4) she loves watching drama series on tv	5) they went to the theater to see a live performance
จำนวนรูปภาพจากการค้นคืน	15 ลำดับ										
	20 ลำดับ										

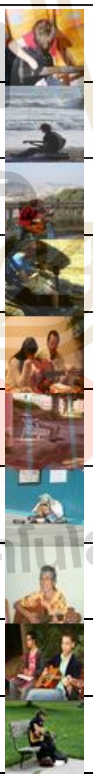
ตารางที่ จ.2 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับ ภายใต้เงื่อนไขข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพ (ต่อ)

ข้อความค้นหา		หมวดธุรกิจและการเงิน					หมวดสังคมและวัฒนธรรม				
		1) he started his own business last year	2) she was happy with her new salary increase	3) they discussed marketing strategies for their product	4) he created a budget to manage his expenses	5) she listened carefully to the customer's feedback	1) she learned about different cultures during her travels	2) he felt a sense of belonging in his community	3) they enjoyed a relaxing vacation at the beach	4) she attended a celebration for her friend's birthday	5) he practiced speaking a new language every day
จำนวนรูปภาพจากการค้นคืน	3 ลำดับ										
	5 ลำดับ										
	10 ลำดับ										

ตารางที่ จ.2 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับ ภายใต้เงื่อนไขข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพ (ต่อ)

ข้อความค้นหา		หมวดธุรกิจและการเงิน					หมวดสังคมและวัฒนธรรม				
		1) he started his own business last year	2) she was happy with her new salary increase	3) they discussed marketing strategies for their product	4) he created a budget to manage his expenses	5) she listened carefully to the customer's feedback	1) she learned about different cultures during her travels	2) he felt a sense of belonging in his community	3) they enjoyed a relaxing vacation at the beach	4) she attended a celebration for her friend's birthday	5) he practiced speaking a new language every day
จำนวนรูปภาพจากการค้นคืน	15 ลำดับ										
	20 ลำดับ										

ตารางที่ จ.3 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับ ภายใต้เงื่อนไข ข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ

ข้อความค้นหา	หมวดคน สัตว์ สิ่งของ					หมวดอาหารและเครื่องดื่ม				
	1) the dog wagged its tail happily	2) the baby sleeps peacefully, cuddled in her mother's arms	3) the man smiles warmly as he plays catch with his dog	4) the woman laughs joyfully while chatting with her friends	5) the guitarist strums passionately, lost in the melody	1) she felt excited to order her favorite pizza for dinner	2) he felt refreshed after eating a crisp, green salad	3) they felt satisfied as they sipped on their favorite beer	4) she felt rejuvenated after drinking a glass of cold water	5) he felt energized after his morning cup of coffee
จำนวนรูปภาพจากการค้นคืน										
3 ลำดับ										
5 ลำดับ										
10 ลำดับ										

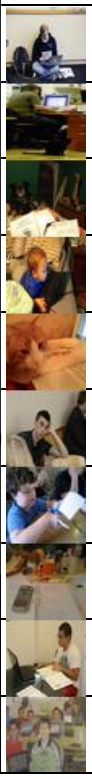
ตารางที่ จ.3 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับ ภายใต้เงื่อนไขข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ (ต่อ)

ข้อความค้นหา		หมวดคน สัตว์ สิ่งของ					หมวดอาหารและเครื่องดื่ม				
		1) the dog wagged its tail happily	2) the baby sleeps peacefully, cuddled in her mother's arms	3) the man smiles warmly as he plays catch with his dog	4) the woman laughs joyfully while chatting with her friends	5) the guitarist strums passionately, lost in the melody	1) she felt excited to order her favorite pizza for dinner	2) he felt refreshed after eating a crisp, green salad	3) they felt satisfied as they sipped on their favorite beer	4) she felt rejuvenated after drinking a glass of cold water	5) he felt energized after his morning cup of coffee
จำนวนรูปภาพจากการค้นคืน	15 ลำดับ										
	20 ลำดับ										

ตารางที่ จ.3 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับ ภายใต้เงื่อนไข ข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ (ต่อ)

ข้อความค้นหา	หมวดการท่องเที่ยว					หมวดการศึกษา				
	1) the tourist felt excited to explore the new city	2) they felt relieved to finally check into the comfortable hotel	3) families felt happy as they played together in the sunny park	4) she felt nostalgic as she wrote a postcard to her family back home	5) visitors felt relaxed as they lounged on the sandy beach	1) she felt anxious about the upcoming test	2) he felt relieved after completing his homework	3) math is his least favorite subject in school	4) the teacher taught a valuable lesson about perseverance	5) students felt excited as they entered the classroom
จำนวนรูปภาพจากการค้นคืน 3 ลำดับ										
5 ลำดับ										
10 ลำดับ										

ตารางที่ จ.3 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับ ภายใต้เงื่อนไขข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ (ต่อ)

ข้อความค้นหา		หมวดการท่องเที่ยว					หมวดการศึกษา				
		1) the tourist felt excited to explore the new city	2) they felt relieved to finally check into the comfortable hotel	3) families felt happy as they played together in the sunny park	4) she felt nostalgic as she wrote a postcard to her family back home	5) visitors felt relaxed as they lounged on the sandy beach	1) she felt anxious about the upcoming test	2) he felt relieved after completing his homework	3) math is his least favorite subject in school	4) the teacher taught a valuable lesson about perseverance	5) students felt excited as they entered the classroom
จำนวนรูปภาพจากการค้นคืน	15 ลำดับ										
	20 ลำดับ										

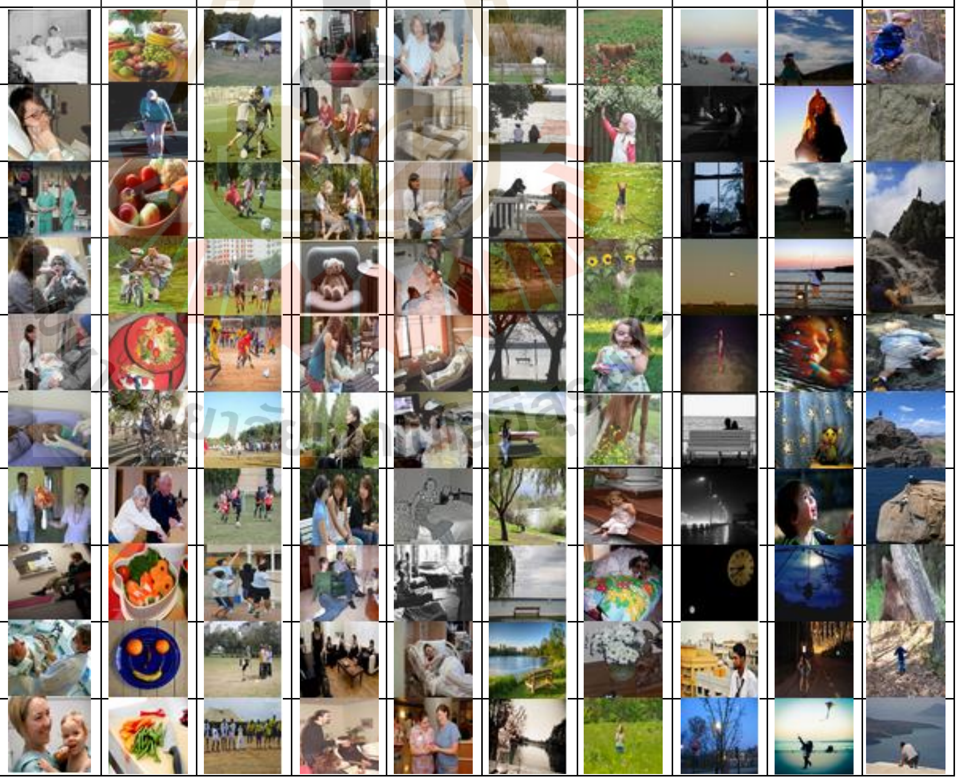
ตารางที่ จ.3 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับ ภายใต้เงื่อนไข ข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ (ต่อ)

ข้อความค้นหา	หมวดสุขภาพ					หมวดธรรมชาติ				
	1) the doctor's reassuring smile comforted the patient	2) eating well and staying active keeps you healthy and happy	3) the soccer game stirred up feelings of teamwork, competitiveness and excitement	4) therapy sessions help people manage their emotions and stress	5) the nurse's caring touch provided comfort during the hospital stay	1) sitting by the tranquil lake filled her with peace	2) admiring the colorful daisy flowers in bloom filled her with happiness	3) gazing at the full moon stirred up feelings of nostalgia	4) counting the stars in the clear sky sparked joy in her heart	5) climbing the mountain filled him with a sense of accomplishment
จำนวนรูปภาพจากการค้นคืน										
3 ลำดับ										
5 ลำดับ										
10 ลำดับ										

ตารางที่ จ.3 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับ ภายใต้เงื่อนไขข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ (ต่อ)

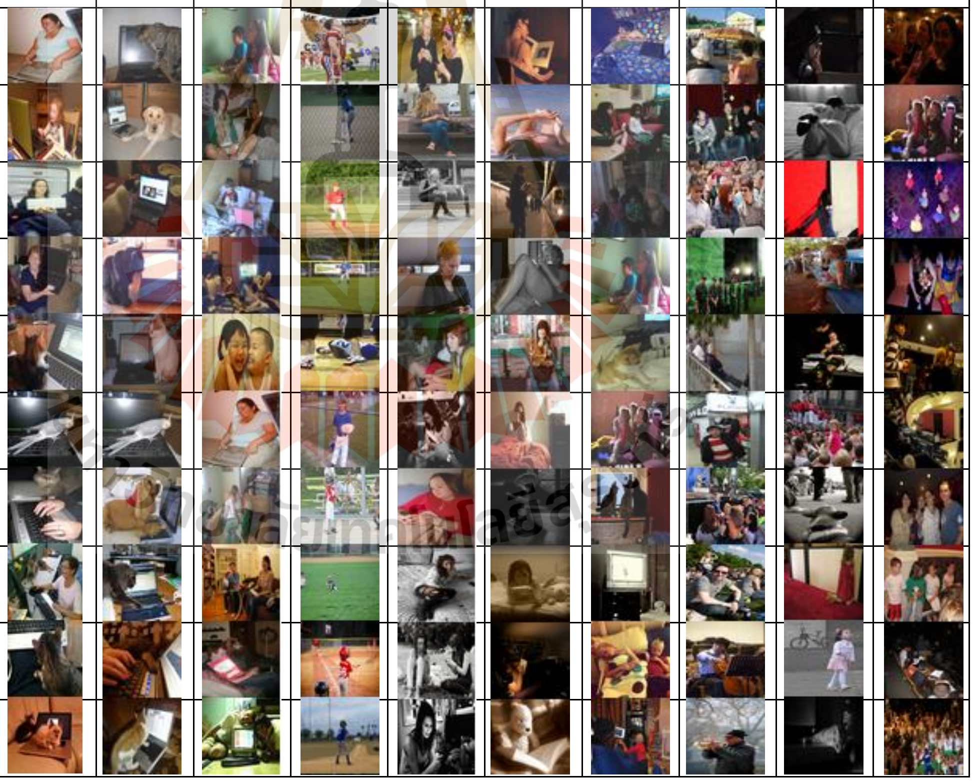
หมวด	ข้อความค้นหา
หมวดสุขภาพ	1) the doctor's reassuring smile comforted the patient
หมวดสุขภาพ	2) eating well and staying active keeps you healthy and happy
หมวดสุขภาพ	3) the soccer game stirred up feelings of teamwork, competitiveness and excitement
หมวดสุขภาพ	4) therapy sessions help people manage their emotions and stress
หมวดสุขภาพ	5) the nurse's caring touch provided comfort during the hospital stay
หมวดธรรมชาติ	1) sitting by the tranquil lake filled her with peace
หมวดธรรมชาติ	2) admiring the colorful daisy flowers in bloom filled her with happiness
หมวดธรรมชาติ	3) gazing at the full moon stirred up feelings of nostalgia
หมวดธรรมชาติ	4) counting the stars in the clear sky sparked joy in her heart
หมวดธรรมชาติ	5) climbing the mountain filled him with a sense of accomplishment

จำนวนรูปภาพจากการค้นคืน: 15 ลำดับ / 20 ลำดับ

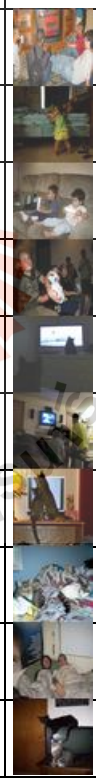


ตารางที่ จ.3 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับ ภายใต้เงื่อนไข ข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ (ต่อ)

ข้อความค้นหา		หมวดเทคโนโลยี					หมวดศิลปะและการบันเทิง				
		1) she felt excited to receive an email from her friend	2) he was relieved to find the information he needed on the internet	3) they laughed together while watching a funny video	4) he felt accomplished after winning the game	5) she felt connected to her friends while scrolling through social media	1) she felt immersed in the story while reading a book	2) they felt entertained watching a movie together	3) he felt exhilarated after attending the concert	4) she felt captivated by the drama unfolding on stage	5) they felt enchanted by the magic of the theater
จำนวนรูปภาพจากการค้นคืน	3 ลำดับ										
	5 ลำดับ										
	10 ลำดับ										



ตารางที่ จ.3 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับ ภายใต้เงื่อนไขข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ (ต่อ)

ข้อความค้นหา	หมวดเทคโนโลยี					หมวดศิลปะและการบันเทิง				
	1) she felt excited to receive an email from her friend	2) he was relieved to find the information he needed on the internet	3) they laughed together while watching a funny video	4) he felt accomplished after winning the game	5) she felt connected to her friends while scrolling through social media	1) she felt immersed in the story while reading a book	2) they felt entertained watching a movie together	3) he felt exhilarated after attending the concert	4) she felt captivated by the drama unfolding on stage	5) they felt enchanted by the magic of the theater
จำนวนรูปภาพจากการค้นคืน										
15 ลำดับ										
20 ลำดับ										

ตารางที่ จ.3 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับ ภายใต้เงื่อนไข ข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ (ต่อ)

ข้อความค้นหา	หมวดธุรกิจและการเงิน					หมวดสังคมและวัฒนธรรม				
	1) she felt excited to start her own business	2) he was relieved when he received his salary	3) they were proud of their successful marketing campaign	4) she felt organized after creating a budget	5) he smiled as he greeted each customer	1) she felt curious to explore the local culture	2) he felt supported by his tight-knit community	3) they were excited to go on a vacation to the beach	4) she felt joyful during the celebration of her graduation	5) he struggled to learn a new language but felt determined
จำนวนรูปภาพจากการค้นคืน: 3 ลำดับ										
5 ลำดับ										
10 ลำดับ										

ตารางที่ จ.3 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพ 20 อันดับ ภายใต้เงื่อนไขข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ (ต่อ)

ข้อความค้นหา		หมวดธุรกิจและการเงิน					หมวดสังคมและวัฒนธรรม				
		1) she felt excited to start her own business	2) he was relieved when he received his salary	3) they were proud of their successful marketing campaign	4) she felt organized after creating a budget	5) he smiled as he greeted each customer	1) she felt curious to explore the local culture	2) he felt supported by his tight-knit community	3) they were excited to go on a vacation to the beach	4) she felt joyful during the celebration of her graduation	5) he struggled to learn a new language but felt determined
จำนวนรูปภาพจากการค้นคืน	15 ลำดับ										
	20 ลำดับ										

ภาคผนวก ฉ
ตัวอย่างแบบประเมินความถูกต้องการค้นคืนรูปภาพโดยผู้เชี่ยวชาญ

งานวิจัยนี้มีวัตถุประสงค์เพื่อประเมินประสิทธิภาพแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายโดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า ผู้วิจัยเล็งเห็นว่าท่านมีคุณสมบัติครบถ้วนในการเป็นผู้เชี่ยวชาญ จึงขอความอนุเคราะห์เก็บข้อมูลผลการประเมินความถูกต้องการค้นคืนรูปภาพภายใต้ข้อความค้นหา 3 เงื่อนไข ประกอบด้วย

1) ข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ ในลักษณะป้ายกำกับของรูปภาพ (Query by global labels) เช่น dog, cat หรือ animal เป็นต้น

2) ข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพ (Query by high level concepts of the images) เช่น the dog running on a grass เป็นต้น

3) ข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ (Query by qualitative semantic concepts of the image) เช่น how do you feel when you finish work เป็นต้น

เงื่อนไขละ 10 ข้อความ รวมทั้งสิ้น 30 ข้อความ กำหนดเกณฑ์การประเมินในแต่ละข้อดังนี้

☑ ถูกต้อง หมายถึง หากความหมายระดับสูงของรูปภาพเชิงรูปธรรม[1] และเชิงนามธรรม[2] นั้นถูกต้องตามข้อความค้นหา

☐ ไม่ถูกต้อง หมายถึง หากความหมายระดับสูงของรูปภาพนั้น ไม่ถูกต้องตามข้อความค้นหา หรือถูกต้องเพียงบางส่วนแต่ไม่ครอบคลุมประเด็นการค้นหา

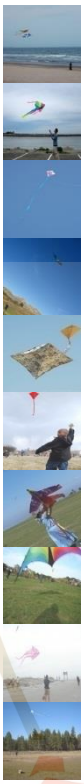
[1]ความหมายระดับสูงของรูปภาพเชิงรูปธรรม (Concrete Concept) ได้แก่ วัตถุ กิจกรรม หรือเหตุการณ์ภายในรูปภาพ เป็นต้น

[2]ความหมายระดับสูงของรูปภาพเชิงนามธรรม (Abstract Concept) ได้แก่ อารมณ์ และความรู้สึกภายในรูปภาพ เป็นต้น

1) ข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ ในลักษณะป้ายกำกับของรูปภาพ

1.1) photo of a kite

☐ รูปภาพที่ 1
☑ รูปภาพที่ 2
☑ รูปภาพที่ 3
☐ รูปภาพที่ 4
☑ รูปภาพที่ 5
☑ รูปภาพที่ 6
☑ รูปภาพที่ 7
☑ รูปภาพที่ 8
☑ รูปภาพที่ 9
☑ รูปภาพที่ 10



☑ รูปภาพที่ 11
☑ รูปภาพที่ 12
☑ รูปภาพที่ 13
☐ รูปภาพที่ 14
☑ รูปภาพที่ 15
☑ รูปภาพที่ 16
☑ รูปภาพที่ 17
☑ รูปภาพที่ 18
☑ รูปภาพที่ 19
☑ รูปภาพที่ 20

1.2) photo of a knife

☐ รูปภาพที่ 1
☑ รูปภาพที่ 2
☑ รูปภาพที่ 3
☑ รูปภาพที่ 4
☐ รูปภาพที่ 5
☑ รูปภาพที่ 6
☑ รูปภาพที่ 7
☑ รูปภาพที่ 8
☑ รูปภาพที่ 9
☑ รูปภาพที่ 10

☑ รูปภาพที่ 11
☑ รูปภาพที่ 12
☐ รูปภาพที่ 13
☑ รูปภาพที่ 14
☐ รูปภาพที่ 15
☐ รูปภาพที่ 16
☐ รูปภาพที่ 17
☑ รูปภาพที่ 18
☐ รูปภาพที่ 19
☑ รูปภาพที่ 20

1.3) photo of a library

☐ รูปภาพที่ 1
☑ รูปภาพที่ 2
☑ รูปภาพที่ 3
☑ รูปภาพที่ 4
☐ รูปภาพที่ 5
☐ รูปภาพที่ 6
☑ รูปภาพที่ 7
☐ รูปภาพที่ 8
☑ รูปภาพที่ 9
☐ รูปภาพที่ 10

☑ รูปภาพที่ 11
☑ รูปภาพที่ 12
☐ รูปภาพที่ 13
☑ รูปภาพที่ 14
☑ รูปภาพที่ 15
☐ รูปภาพที่ 16
☑ รูปภาพที่ 17
☑ รูปภาพที่ 18
☐ รูปภาพที่ 19
☑ รูปภาพที่ 20

1.4) photo of a light

☑ รูปภาพที่ 1
☑ รูปภาพที่ 2
☐ รูปภาพที่ 3
☐ รูปภาพที่ 4
☑ รูปภาพที่ 5
☑ รูปภาพที่ 6
☑ รูปภาพที่ 7
☐ รูปภาพที่ 8
☑ รูปภาพที่ 9
☐ รูปภาพที่ 10

☐ รูปภาพที่ 11
☐ รูปภาพที่ 12
☐ รูปภาพที่ 13
☑ รูปภาพที่ 14
☐ รูปภาพที่ 15
☐ รูปภาพที่ 16
☐ รูปภาพที่ 17
☑ รูปภาพที่ 18
☑ รูปภาพที่ 19
☐ รูปภาพที่ 20

1.5) photo of a mirror


☑ รูปภาพที่ 1


☑ รูปภาพที่ 2


☑ รูปภาพที่ 3


☑ รูปภาพที่ 4


☑ รูปภาพที่ 5


☑ รูปภาพที่ 6


☑ รูปภาพที่ 7


☑ รูปภาพที่ 8


☑ รูปภาพที่ 9


☑ รูปภาพที่ 10


☑ รูปภาพที่ 11


☑ รูปภาพที่ 12


☑ รูปภาพที่ 13


☑ รูปภาพที่ 14


☑ รูปภาพที่ 15


☑ รูปภาพที่ 16


☑ รูปภาพที่ 17


☑ รูปภาพที่ 18


☑ รูปภาพที่ 19


☑ รูปภาพที่ 20

1.6) photo of a money


☑ รูปภาพที่ 1


☐ รูปภาพที่ 2


☑ รูปภาพที่ 3


☑ รูปภาพที่ 4


☑ รูปภาพที่ 5


☑ รูปภาพที่ 6


☑ รูปภาพที่ 7


☐ รูปภาพที่ 8


☐ รูปภาพที่ 9


☐ รูปภาพที่ 10


☐ รูปภาพที่ 11


☑ รูปภาพที่ 12


☐ รูปภาพที่ 13


☐ รูปภาพที่ 14


☐ รูปภาพที่ 15


☐ รูปภาพที่ 16


☐ รูปภาพที่ 17


☐ รูปภาพที่ 18


☑ รูปภาพที่ 19


☐ รูปภาพที่ 20

1.7) photo of a moon

☑ รูปภาพที่ 1
☑ รูปภาพที่ 2
☐ รูปภาพที่ 3
☑ รูปภาพที่ 4
☐ รูปภาพที่ 5
☑ รูปภาพที่ 6
☑ รูปภาพที่ 7
☐ รูปภาพที่ 8
☑ รูปภาพที่ 9
☑ รูปภาพที่ 10

☐ รูปภาพที่ 11
☑ รูปภาพที่ 12
☐ รูปภาพที่ 13
☑ รูปภาพที่ 14
☑ รูปภาพที่ 15
☐ รูปภาพที่ 16
☑ รูปภาพที่ 17
☐ รูปภาพที่ 18
☐ รูปภาพที่ 19
☑ รูปภาพที่ 20

1.8) photo of a mountain

☑ รูปภาพที่ 1
☐ รูปภาพที่ 2
☑ รูปภาพที่ 3
☑ รูปภาพที่ 4
☑ รูปภาพที่ 5
☑ รูปภาพที่ 6
☑ รูปภาพที่ 7
☑ รูปภาพที่ 8
☑ รูปภาพที่ 9
☑ รูปภาพที่ 10

☑ รูปภาพที่ 11
☑ รูปภาพที่ 12
☑ รูปภาพที่ 13
☑ รูปภาพที่ 14
☑ รูปภาพที่ 15
☑ รูปภาพที่ 16
☑ รูปภาพที่ 17
☑ รูปภาพที่ 18
☑ รูปภาพที่ 19
☑ รูปภาพที่ 20

1.9) photo of a museum

☐ รูปภาพที่ 1
☑ รูปภาพที่ 2
☐ รูปภาพที่ 3
☐ รูปภาพที่ 4
☐ รูปภาพที่ 5
☑ รูปภาพที่ 6
☑ รูปภาพที่ 7
☑ รูปภาพที่ 8
☑ รูปภาพที่ 9
☐ รูปภาพที่ 10

☐ รูปภาพที่ 11
☐ รูปภาพที่ 12
☑ รูปภาพที่ 13
☐ รูปภาพที่ 14
☐ รูปภาพที่ 15
☐ รูปภาพที่ 16
☐ รูปภาพที่ 17
☑ รูปภาพที่ 18
☐ รูปภาพที่ 19
☐ รูปภาพที่ 20

1.10) photo of a music

☑ รูปภาพที่ 1
☑ รูปภาพที่ 2
☑ รูปภาพที่ 3
☑ รูปภาพที่ 4
☑ รูปภาพที่ 5
☑ รูปภาพที่ 6
☑ รูปภาพที่ 7
☑ รูปภาพที่ 8
☑ รูปภาพที่ 9
☑ รูปภาพที่ 10

☑ รูปภาพที่ 11
☑ รูปภาพที่ 12
☑ รูปภาพที่ 13
☑ รูปภาพที่ 14
☑ รูปภาพที่ 15
☑ รูปภาพที่ 16
☐ รูปภาพที่ 17
☑ รูปภาพที่ 18
☑ รูปภาพที่ 19
☑ รูปภาพที่ 20

2) ข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพ

2.1) children flew a kite in a wide-open field

☑ รูปภาพที่ 1
☑ รูปภาพที่ 2
☑ รูปภาพที่ 3
☑ รูปภาพที่ 4
☑ รูปภาพที่ 5
☑ รูปภาพที่ 6
☑ รูปภาพที่ 7
☑ รูปภาพที่ 8
☑ รูปภาพที่ 9
☑ รูปภาพที่ 10

☐ รูปภาพที่ 11
☑ รูปภาพที่ 12
☑ รูปภาพที่ 13
☑ รูปภาพที่ 14
☑ รูปภาพที่ 15
☐ รูปภาพที่ 16
☐ รูปภาพที่ 17
☑ รูปภาพที่ 18
☑ รูปภาพที่ 19
☑ รูปภาพที่ 20

2.2) a chef skillfully chopped ingredients with a knife

☑ รูปภาพที่ 1
☑ รูปภาพที่ 2
☑ รูปภาพที่ 3
☑ รูปภาพที่ 4
☑ รูปภาพที่ 5
☐ รูปภาพที่ 6
☑ รูปภาพที่ 7
☐ รูปภาพที่ 8
☑ รูปภาพที่ 9
☑ รูปภาพที่ 10

☑ รูปภาพที่ 11
☑ รูปภาพที่ 12
☐ รูปภาพที่ 13
☑ รูปภาพที่ 14
☐ รูปภาพที่ 15
☑ รูปภาพที่ 16
☐ รูปภาพที่ 17
☐ รูปภาพที่ 18
☐ รูปภาพที่ 19
☐ รูปภาพที่ 20

2.3) students studied in a quiet library

☑ รูปภาพที่ 1
☑ รูปภาพที่ 2
☑ รูปภาพที่ 3
☑ รูปภาพที่ 4
☑ รูปภาพที่ 5
☑ รูปภาพที่ 6
☑ รูปภาพที่ 7
☑ รูปภาพที่ 8
☑ รูปภาพที่ 9
☑ รูปภาพที่ 10

☐ รูปภาพที่ 11
☑ รูปภาพที่ 12
☐ รูปภาพที่ 13
☑ รูปภาพที่ 14
☐ รูปภาพที่ 15
☐ รูปภาพที่ 16
☑ รูปภาพที่ 17
☑ รูปภาพที่ 18
☑ รูปภาพที่ 19
☑ รูปภาพที่ 20

2.4) a light illuminated a dark room

☑ รูปภาพที่ 1
☑ รูปภาพที่ 2
☐ รูปภาพที่ 3
☐ รูปภาพที่ 4
☑ รูปภาพที่ 5
☑ รูปภาพที่ 6
☑ รูปภาพที่ 7
☑ รูปภาพที่ 8
☑ รูปภาพที่ 9
☑ รูปภาพที่ 10

☑ รูปภาพที่ 11
☑ รูปภาพที่ 12
☐ รูปภาพที่ 13
☑ รูปภาพที่ 14
☐ รูปภาพที่ 15
☑ รูปภาพที่ 16
☑ รูปภาพที่ 17
☑ รูปภาพที่ 18
☑ รูปภาพที่ 19
☑ รูปภาพที่ 20

2.5) a person counted their money carefully

☐ รูปภาพที่ 1
☑ รูปภาพที่ 2
☐ รูปภาพที่ 3
☐ รูปภาพที่ 4
☐ รูปภาพที่ 5
☐ รูปภาพที่ 6
☐ รูปภาพที่ 7
☐ รูปภาพที่ 8
☐ รูปภาพที่ 9
☐ รูปภาพที่ 10

☐ รูปภาพที่ 11
☐ รูปภาพที่ 12
☐ รูปภาพที่ 13
☐ รูปภาพที่ 14
☑ รูปภาพที่ 15
☐ รูปภาพที่ 16
☐ รูปภาพที่ 17
☐ รูปภาพที่ 18
☐ รูปภาพที่ 19
☐ รูปภาพที่ 20

2.6) the moon shone brightly in the night sky

☑ รูปภาพที่ 1
☐ รูปภาพที่ 2
☑ รูปภาพที่ 3
☑ รูปภาพที่ 4
☑ รูปภาพที่ 5
☑ รูปภาพที่ 6
☑ รูปภาพที่ 7
☑ รูปภาพที่ 8
☐ รูปภาพที่ 9
☑ รูปภาพที่ 10

☑ รูปภาพที่ 11
☑ รูปภาพที่ 12
☑ รูปภาพที่ 13
☑ รูปภาพที่ 14
☑ รูปภาพที่ 15
☐ รูปภาพที่ 16
☐ รูปภาพที่ 17
☑ รูปภาพที่ 18
☐ รูปภาพที่ 19
☑ รูปภาพที่ 20

2.7) hikers admired the breathtaking view from the mountaintop

 ☑ รูปภาพที่ 1
 ☑ รูปภาพที่ 2
 ☑ รูปภาพที่ 3
 ☑ รูปภาพที่ 4
 ☑ รูปภาพที่ 5
 ☑ รูปภาพที่ 6
 ☑ รูปภาพที่ 7
 ☑ รูปภาพที่ 8
 ☐ รูปภาพที่ 9
 ☐ รูปภาพที่ 10

 ☑ รูปภาพที่ 11
 ☑ รูปภาพที่ 12
 ☐ รูปภาพที่ 13
 ☐ รูปภาพที่ 14
 ☑ รูปภาพที่ 15
 ☑ รูปภาพที่ 16
 ☑ รูปภาพที่ 17
 ☑ รูปภาพที่ 18
 ☐ รูปภาพที่ 19
 ☐ รูปภาพที่ 20

2.8) visitors explored artifacts in a museum

 ☑ รูปภาพที่ 1
 ☑ รูปภาพที่ 2
 ☑ รูปภาพที่ 3
 ☐ รูปภาพที่ 4
 ☑ รูปภาพที่ 5
 ☑ รูปภาพที่ 6
 ☑ รูปภาพที่ 7
 ☐ รูปภาพที่ 8
 ☐ รูปภาพที่ 9
 ☑ รูปภาพที่ 10

 ☑ รูปภาพที่ 11
 ☐ รูปภาพที่ 12
 ☑ รูปภาพที่ 13
 ☑ รูปภาพที่ 14
 ☑ รูปภาพที่ 15
 ☑ รูปภาพที่ 16
 ☐ รูปภาพที่ 17
 ☑ รูปภาพที่ 18
 ☑ รูปภาพที่ 19
 ☐ รูปภาพที่ 20

2.9) musicians played instruments and created beautiful melodies

☑ รูปภาพที่ 1
☑ รูปภาพที่ 2
☑ รูปภาพที่ 3
☑ รูปภาพที่ 4
☐ รูปภาพที่ 5
☑ รูปภาพที่ 6
☑ รูปภาพที่ 7
☑ รูปภาพที่ 8
☑ รูปภาพที่ 9
☑ รูปภาพที่ 10

☑ รูปภาพที่ 11
☑ รูปภาพที่ 12
☑ รูปภาพที่ 13
☑ รูปภาพที่ 14
☑ รูปภาพที่ 15
☑ รูปภาพที่ 16
☑ รูปภาพที่ 17
☑ รูปภาพที่ 18
☑ รูปภาพที่ 19
☑ รูปภาพที่ 20

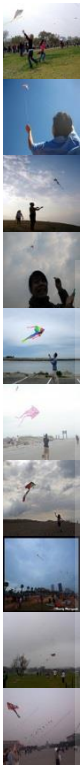
2.10) waves crashed against the shore in the ocean

☑ รูปภาพที่ 1
☑ รูปภาพที่ 2
☑ รูปภาพที่ 3
☑ รูปภาพที่ 4
☑ รูปภาพที่ 5
☑ รูปภาพที่ 6
☑ รูปภาพที่ 7
☑ รูปภาพที่ 8
☑ รูปภาพที่ 9
☐ รูปภาพที่ 10

☑ รูปภาพที่ 11
☑ รูปภาพที่ 12
☑ รูปภาพที่ 13
☑ รูปภาพที่ 14
☑ รูปภาพที่ 15
☑ รูปภาพที่ 16
☑ รูปภาพที่ 17
☑ รูปภาพที่ 18
☑ รูปภาพที่ 19
☑ รูปภาพที่ 20

3) ข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ

3.1) the kite soared high in the sky, creating a feeling of exhilaration

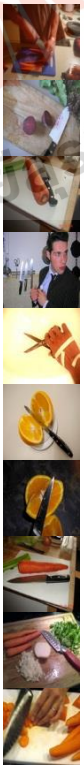


☑ รูปภาพที่ 1
☑ รูปภาพที่ 2
☑ รูปภาพที่ 3
☑ รูปภาพที่ 4
☑ รูปภาพที่ 5
☐ รูปภาพที่ 6
☐ รูปภาพที่ 7
☐ รูปภาพที่ 8
☐ รูปภาพที่ 9
☐ รูปภาพที่ 10

☐ รูปภาพที่ 11
☑ รูปภาพที่ 12
☑ รูปภาพที่ 13
☐ รูปภาพที่ 14
☑ รูปภาพที่ 15
☑ รูปภาพที่ 16
☐ รูปภาพที่ 17
☑ รูปภาพที่ 18
☑ รูปภาพที่ 19
☑ รูปภาพที่ 20

3.2) the knife felt sharp and precise

☐ รูปภาพที่ 1
☐ รูปภาพที่ 2
☐ รูปภาพที่ 3
☐ รูปภาพที่ 4
☐ รูปภาพที่ 5
☐ รูปภาพที่ 6
☐ รูปภาพที่ 7
☑ รูปภาพที่ 8
☐ รูปภาพที่ 9
☐ รูปภาพที่ 10



☐ รูปภาพที่ 11
☐ รูปภาพที่ 12
☐ รูปภาพที่ 13
☐ รูปภาพที่ 14
☐ รูปภาพที่ 15
☐ รูปภาพที่ 16
☑ รูปภาพที่ 17
☐ รูปภาพที่ 18
☐ รูปภาพที่ 19
☑ รูปภาพที่ 20

3.3) the library provided a sense of quiet and knowledge

☐ รูปภาพที่ 1
☑ รูปภาพที่ 2
☑ รูปภาพที่ 3
☑ รูปภาพที่ 4
☐ รูปภาพที่ 5
☐ รูปภาพที่ 6
☐ รูปภาพที่ 7
☐ รูปภาพที่ 8
☐ รูปภาพที่ 9
☐ รูปภาพที่ 10

☑ รูปภาพที่ 11
☑ รูปภาพที่ 12
☑ รูปภาพที่ 13
☑ รูปภาพที่ 14
☑ รูปภาพที่ 15
☑ รูปภาพที่ 16
☐ รูปภาพที่ 17
☐ รูปภาพที่ 18
☑ รูปภาพที่ 19
☐ รูปภาพที่ 20

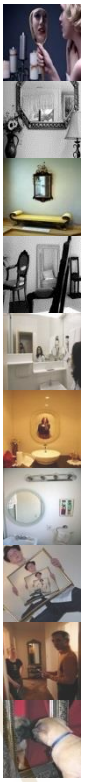
3.4) the light illuminated the space, creating a feeling of safety or visibility

☑ รูปภาพที่ 1
☑ รูปภาพที่ 2
☐ รูปภาพที่ 3
☑ รูปภาพที่ 4
☐ รูปภาพที่ 5
☑ รูปภาพที่ 6
☐ รูปภาพที่ 7
☐ รูปภาพที่ 8
☐ รูปภาพที่ 9
☐ รูปภาพที่ 10

☑ รูปภาพที่ 11
☐ รูปภาพที่ 12
☑ รูปภาพที่ 13
☐ รูปภาพที่ 14
☐ รูปภาพที่ 15
☑ รูปภาพที่ 16
☑ รูปภาพที่ 17
☑ รูปภาพที่ 18
☐ รูปภาพที่ 19
☑ รูปภาพที่ 20

3.5) the mirror reflected an image, evoking self-reflection or vanity

☑ รูปภาพที่ 1
☑ รูปภาพที่ 2
☑ รูปภาพที่ 3
☑ รูปภาพที่ 4
☐ รูปภาพที่ 5
☐ รูปภาพที่ 6
☑ รูปภาพที่ 7
☑ รูปภาพที่ 8
☑ รูปภาพที่ 9
☑ รูปภาพที่ 10



☐ รูปภาพที่ 11
☐ รูปภาพที่ 12
☐ รูปภาพที่ 13
☐ รูปภาพที่ 14
☑ รูปภาพที่ 15
☑ รูปภาพที่ 16
☑ รูปภาพที่ 17
☑ รูปภาพที่ 18
☐ รูปภาพที่ 19
☐ รูปภาพที่ 20

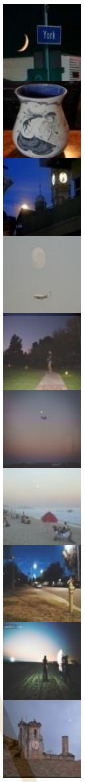
3.6) the money brought feelings of security, happiness, or stress

☐ รูปภาพที่ 1
☐ รูปภาพที่ 2
☑ รูปภาพที่ 3
☐ รูปภาพที่ 4
☐ รูปภาพที่ 5
☑ รูปภาพที่ 6
☐ รูปภาพที่ 7
☐ รูปภาพที่ 8
☐ รูปภาพที่ 9
☐ รูปภาพที่ 10

☐ รูปภาพที่ 11
☐ รูปภาพที่ 12
☐ รูปภาพที่ 13
☐ รูปภาพที่ 14
☐ รูปภาพที่ 15
☐ รูปภาพที่ 16
☐ รูปภาพที่ 17
☐ รูปภาพที่ 18
☑ รูปภาพที่ 19
☐ รูปภาพที่ 20

3.7) the moon cast a gentle glow, creating a sense of romance or mystery

☐ รูปภาพที่ 1
☐ รูปภาพที่ 2
☐ รูปภาพที่ 3
☑ รูปภาพที่ 4
☐ รูปภาพที่ 5
☐ รูปภาพที่ 6
☐ รูปภาพที่ 7
☐ รูปภาพที่ 8
☐ รูปภาพที่ 9
☑ รูปภาพที่ 10



☑ รูปภาพที่ 11
☐ รูปภาพที่ 12
☑ รูปภาพที่ 13
☐ รูปภาพที่ 14
☑ รูปภาพที่ 15
☐ รูปภาพที่ 16
☐ รูปภาพที่ 17
☑ รูปภาพที่ 18
☐ รูปภาพที่ 19
☐ รูปภาพที่ 20

3.8) the mountain stood as a symbol of strength and majesty

☑ รูปภาพที่ 1
☑ รูปภาพที่ 2
☑ รูปภาพที่ 3
☑ รูปภาพที่ 4
☑ รูปภาพที่ 5
☑ รูปภาพที่ 6
☐ รูปภาพที่ 7
☑ รูปภาพที่ 8
☑ รูปภาพที่ 9
☑ รูปภาพที่ 10

☑ รูปภาพที่ 11
☐ รูปภาพที่ 12
☑ รูปภาพที่ 13
☑ รูปภาพที่ 14
☑ รูปภาพที่ 15
☑ รูปภาพที่ 16
☑ รูปภาพที่ 17
☑ รูปภาพที่ 18
☐ รูปภาพที่ 19
☐ รูปภาพที่ 20

3.9) the museum exhibits evoked a range of emotions, from curiosity to awe

☑ รูปภาพที่ 1
☑ รูปภาพที่ 2
☑ รูปภาพที่ 3
☐ รูปภาพที่ 4
☐ รูปภาพที่ 5
☑ รูปภาพที่ 6
☑ รูปภาพที่ 7
☑ รูปภาพที่ 8
☑ รูปภาพที่ 9
☑ รูปภาพที่ 10

☑ รูปภาพที่ 11
☑ รูปภาพที่ 12
☑ รูปภาพที่ 13
☐ รูปภาพที่ 14
☑ รูปภาพที่ 15
☐ รูปภาพที่ 16
☑ รูปภาพที่ 17
☑ รูปภาพที่ 18
☐ รูปภาพที่ 19
☐ รูปภาพที่ 20

3.10) the music stirred various emotions, from joy to melancholy

☐ รูปภาพที่ 1
☐ รูปภาพที่ 2
☐ รูปภาพที่ 3
☑ รูปภาพที่ 4
☐ รูปภาพที่ 5
☑ รูปภาพที่ 6
☑ รูปภาพที่ 7
☐ รูปภาพที่ 8
☑ รูปภาพที่ 9
☐ รูปภาพที่ 10

☑ รูปภาพที่ 11
☐ รูปภาพที่ 12
☐ รูปภาพที่ 13
☐ รูปภาพที่ 14
☐ รูปภาพที่ 15
☐ รูปภาพที่ 16
☐ รูปภาพที่ 17
☐ รูปภาพที่ 18
☐ รูปภาพที่ 19
☐ รูปภาพที่ 20

ประวัติผู้เขียน

นายจักรินทร์ สันติรัตนภักดี เกิดเมื่อวันอังคารที่ 21 พฤษภาคม พ.ศ. 2528 ที่จังหวัดนครราชสีมา สำเร็จการศึกษาระดับปริญญาตรี สาขาวิชาคอมพิวเตอร์ธุรกิจ (เกียรตินิยมอันดับ 1) มหาวิทยาลัยวงษ์ชวลิตกุล ต่อมาในปี พ.ศ. 2553 เข้าศึกษาต่อปริญญาโท หลักสูตรวิทยาการสารสนเทศมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ สำนักเทคโนโลยีสังคม มหาวิทยาลัยเทคโนโลยีสุรนารี มีความสนใจในการออกแบบและพัฒนาระบบเทคโนโลยีสารสนเทศเพื่อธุรกิจ และการพัฒนาองค์กร และสำเร็จการศึกษาในปี พ.ศ. 2555 ปัจจุบันปฏิบัติหน้าที่อาจารย์ประจำหลักสูตรบริหารธุรกิจบัณฑิต สาขาวิชาเทคโนโลยีธุรกิจดิจิทัล มหาวิทยาลัยวงษ์ชวลิตกุล

จากนั้นในปี พ.ศ. 2563 ได้รับทุนสนับสนุนผู้มีผลการเรียนดีเด่นที่เข้าศึกษาต่อในระดับปริญญาเอก หลักสูตรปรัชญาดุษฎีบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ สำนักวิชาศาสตร์และศิลป์ดิจิทัล มหาวิทยาลัยเทคโนโลยีสุรนารี เพื่อพัฒนาความรู้ ความสามารถและความเชี่ยวชาญในการวิจัยเกี่ยวกับการพัฒนาแบบจำลองการค้นคืนรูปภาพเชิงความหมาย โดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า