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

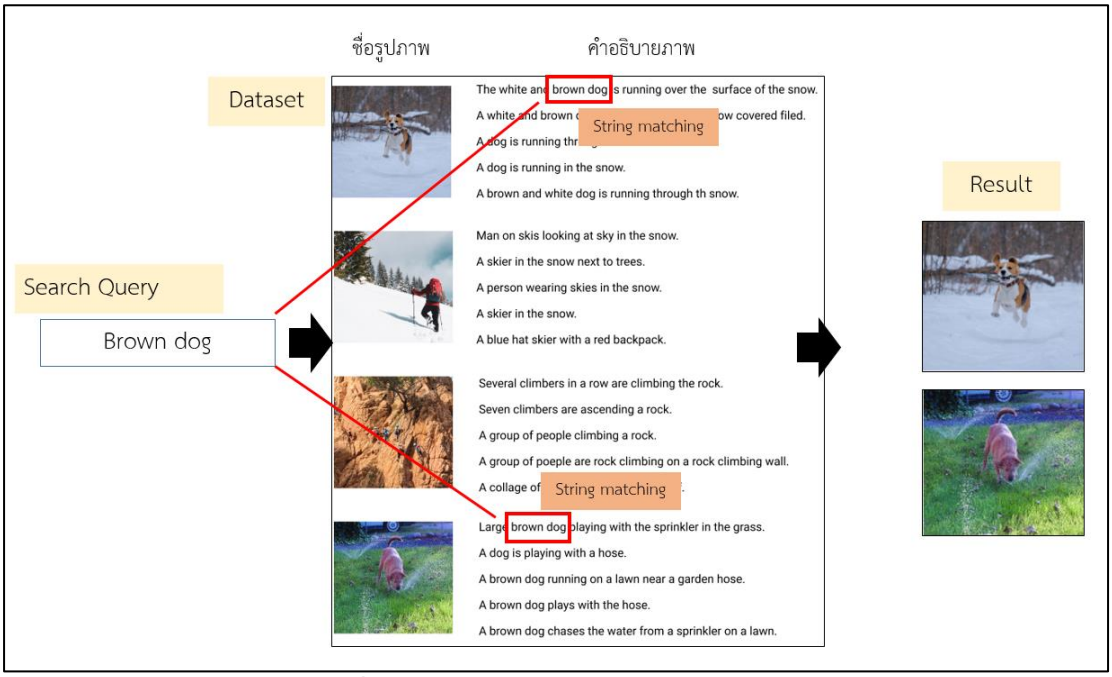
2.8 งานวิจัยที่เกี่ยวข้อง

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

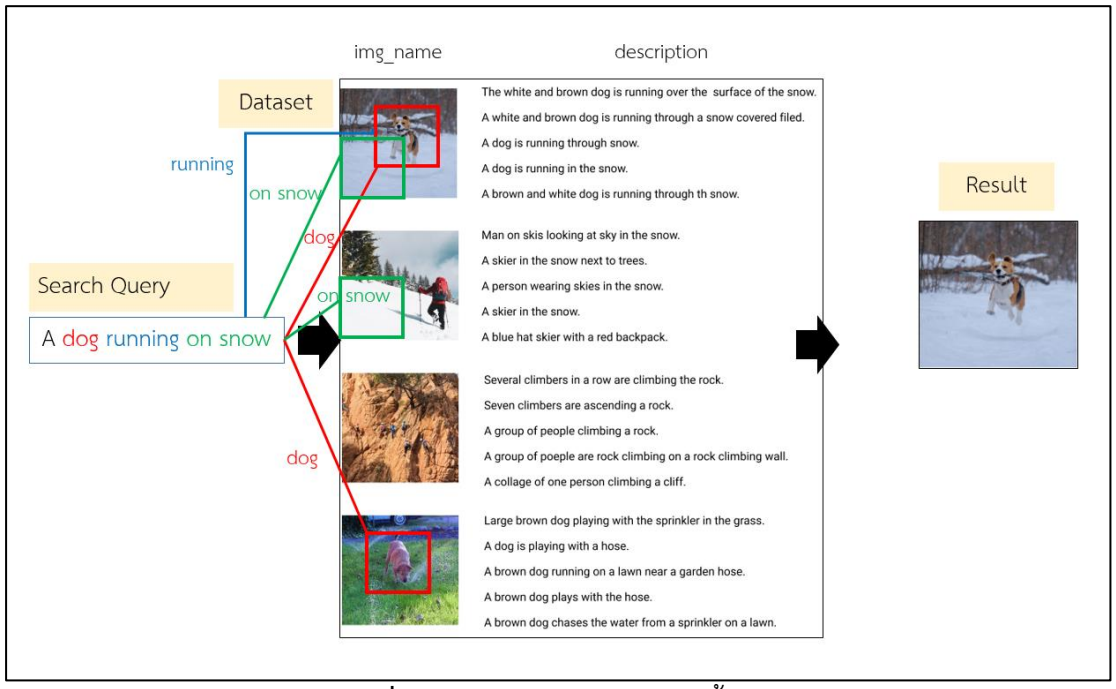


รูปที่ 2.1 การค้นคืนรูปภาพด้วยข้อความ

2.1.2 การค้นคืนรูปภาพจากเนื้อหา

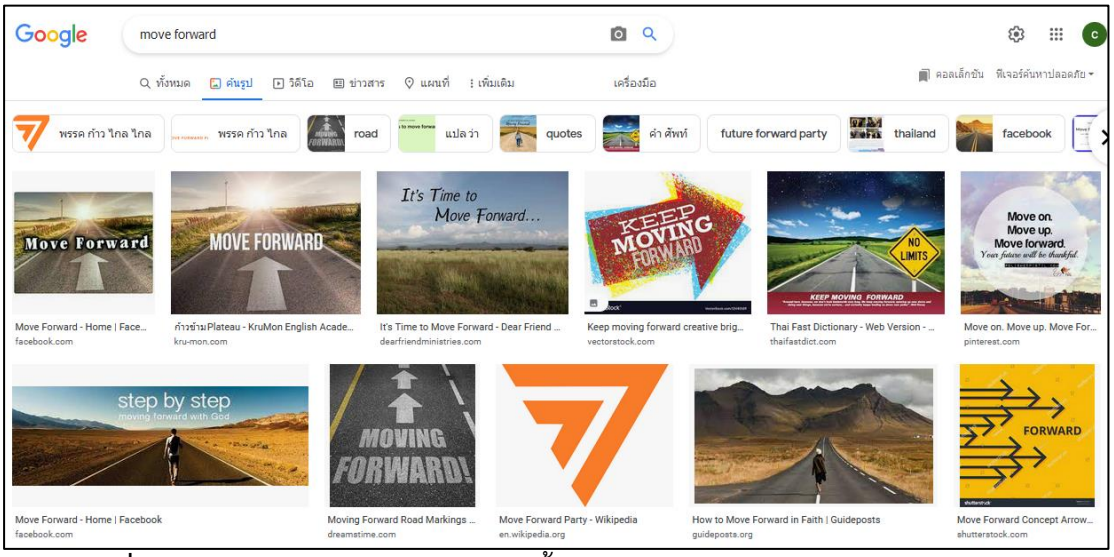
การค้นคืนรูปภาพจากเนื้อหา (Content-based Image Retrieval: CBIR) (Tyagi, 2017) คือการนำคุณลักษณะระดับต่ำ (Low-level features) ซึ่งเป็นข้อมูลที่ไม่มีความหมายในตัวเอง และไม่สื่อถึงความหมายใด ๆ ในรูปภาพมาเปรียบเทียบกับคำค้นหา ประกอบด้วยคุณลักษณะโกลบอล (Global features) ซึ่งเป็นคุณลักษณะแบบรวม ๆ ไม่เฉพาะเจาะจง 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

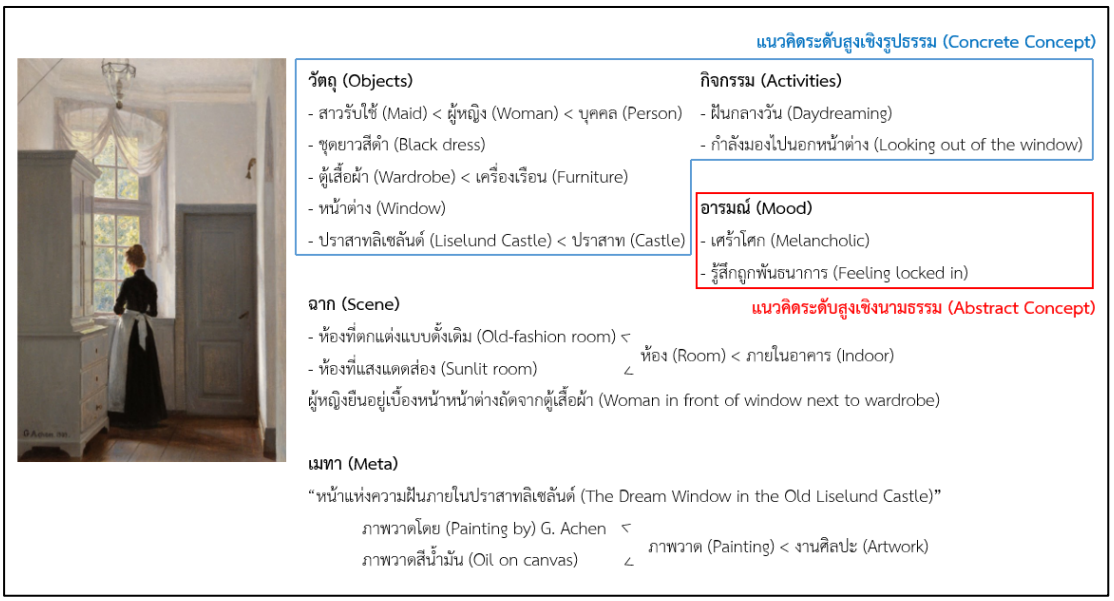
ปัจจุบันการค้นคืนรูปภาพจากเนื้อหาทำงานได้อย่างมีประสิทธิภาพ ดังรูปที่ 2.3 ผลลัพธ์การค้นคืนรูปภาพจากเนื้อหาบนเว็บไซต์ google.com ด้วยข้อความค้นหา “move forward” เป็นรูปภาพที่มีข้อความนั้นฝังอยู่ในภาพ หรือเป็นรูปภาพภายในเว็บเพจที่มีหัวข้อหรือเนื้อหาตรงกับข้อความค้นหาตามแนวคิดการใช้ข้อความที่อยู่รอบ ๆ ในการอธิบายความหมายของรูปภาพ 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



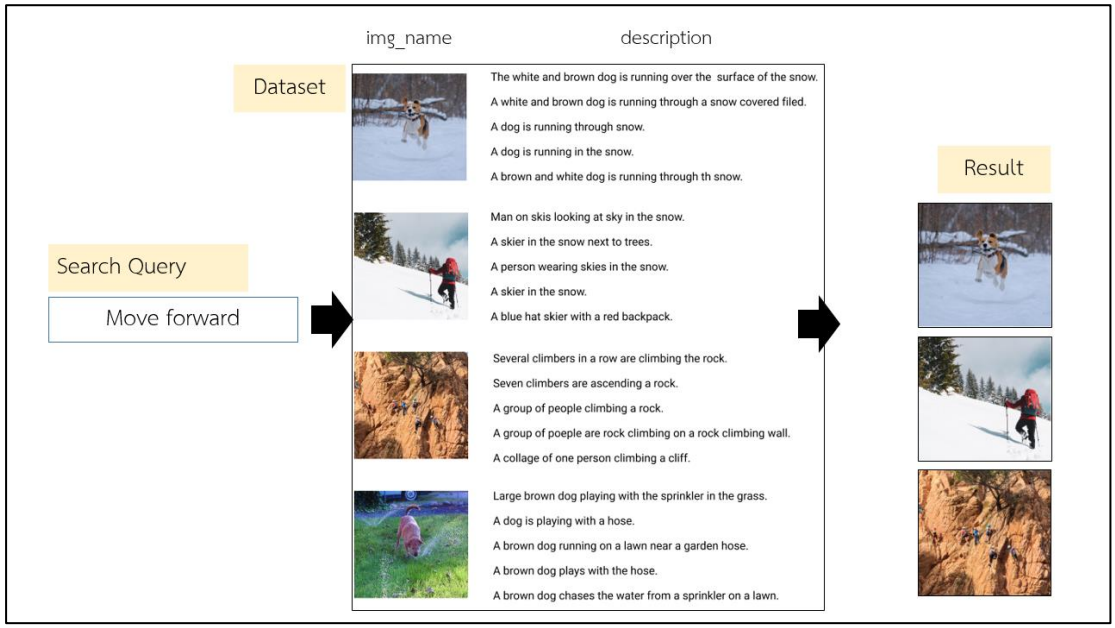
รูปที่ 2.4 ความสัมพันธ์ระหว่างแนวคิดระดับสูงที่เป็นรูปธรรมและที่เป็นนามธรรม
ที่มา: Barz and Denzler (2020)

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

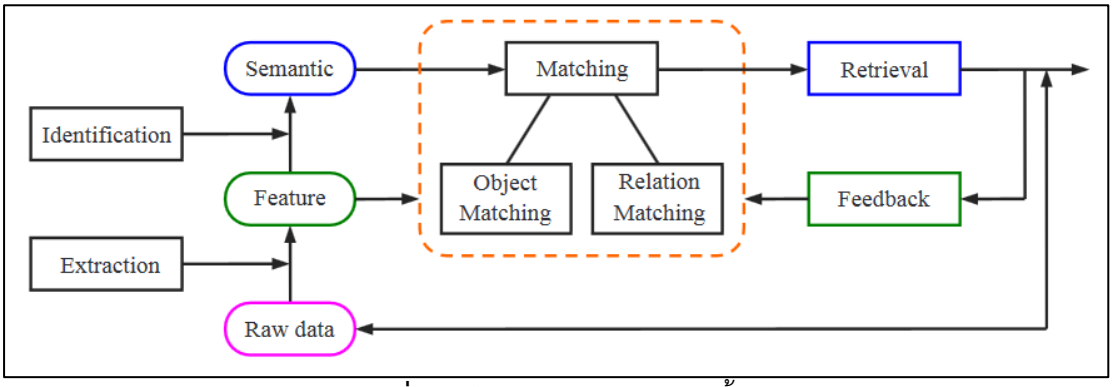
งานวิจัยนี้มุ่งพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมาย เพื่อเชื่อมโยงระหว่างแนวคิดระดับสูงที่เป็นรูปธรรมและที่เป็นนามธรรมให้เป็นแนวคิดเชิงความหมายในรูปแบบที่มนุษย์สามารถเข้าใจได้โดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า อันจะสามารถลดช่องว่างความหมายระหว่างความต้องการของผู้ใช้ที่อยู่ในลักษณะข้อความค้นหา 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



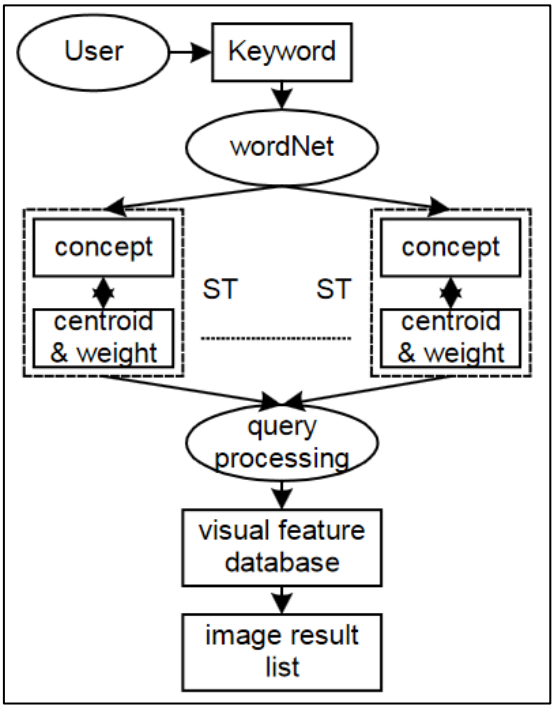
รูปที่ 2.6 ข้อมูลรูปภาพ 3 ระดับชั้น
ที่มา: Zhang (2007)

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

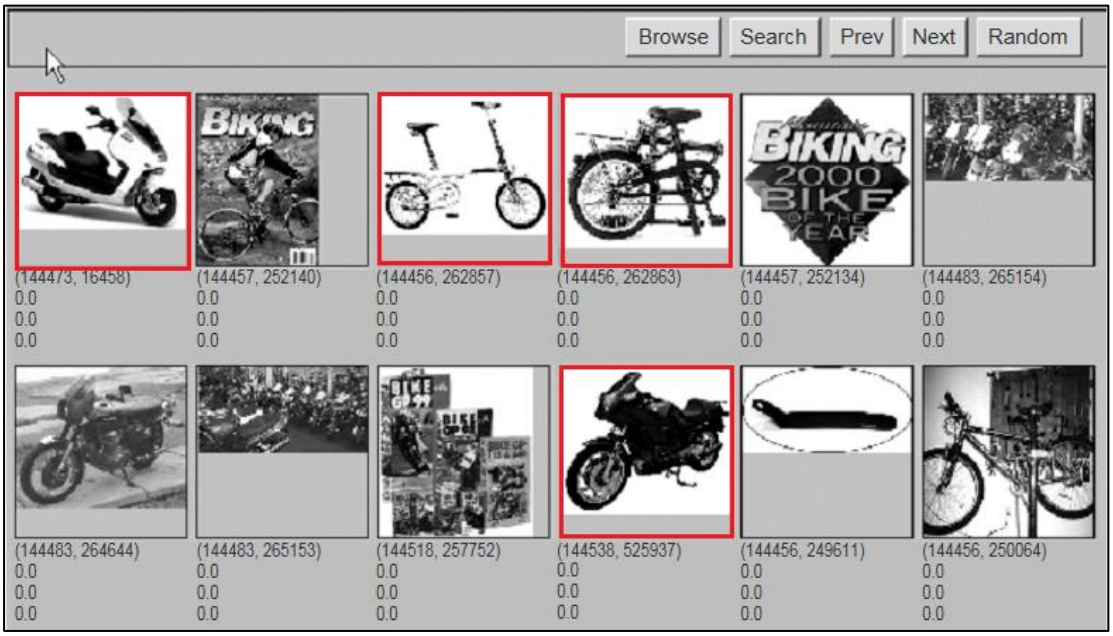
2.2.3 เทคนิคการใช้ข้อความเพื่อเรียนรู้ความหมายของรูปภาพ

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

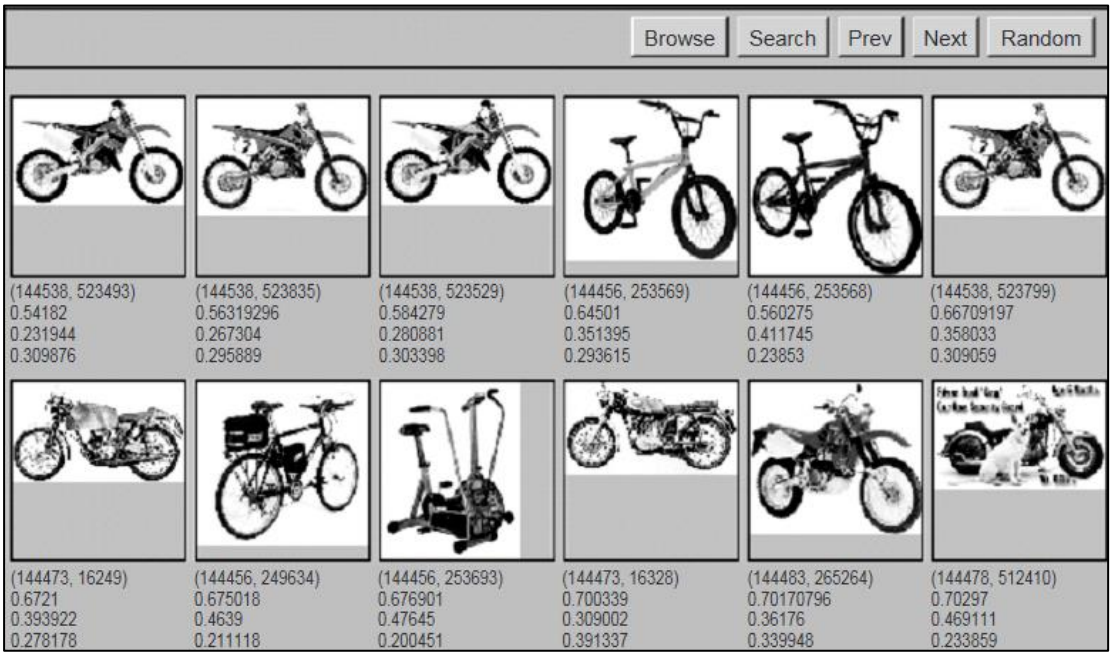
2.2.4 เทคนิคการตอบกลับของผู้ใช้

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)

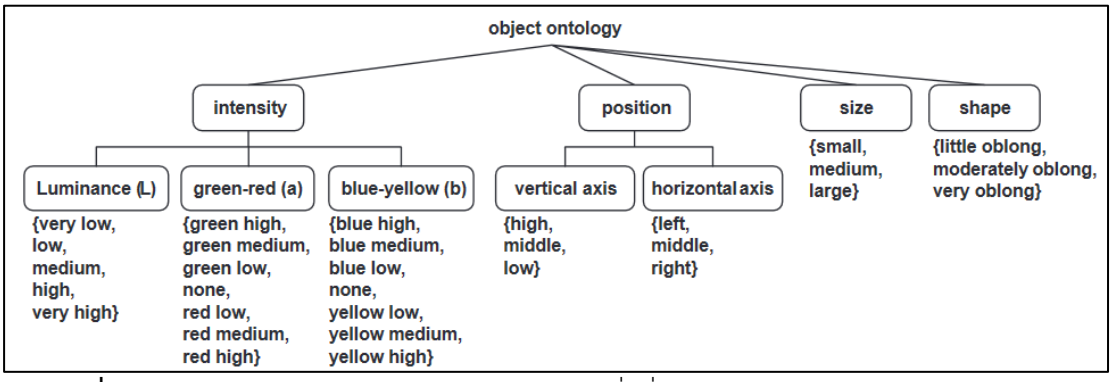
อย่างไรก็ตามเทคนิคการตอบกลับของผู้ใช้ อาจเป็นการเพิ่มภาระในการค้นหาให้กับผู้ใช้ เนื่องจากจำเป็นต้องให้ผู้ใช้ร่วมมือในการตอบกลับข้อคำถามให้กับระบบ เพื่อปรับปรุงผลลัพธ์ให้มีความถูกต้องมากขึ้น แต่วิธีการดังกล่าวอาจทำให้ผู้ใช้หมดความอดทน และเลิกใช้งานไปก่อน

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

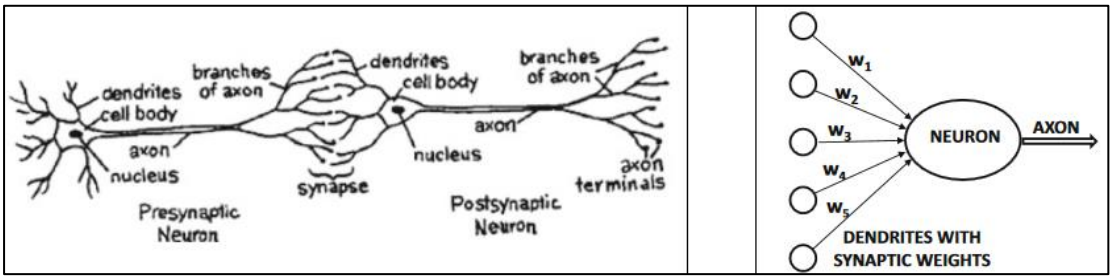
ปัจจุบันมีปหลายเทคนิคที่มักถูกนำมาใช้ในการสร้างตัวจำแนกข้อมูล (Wozniak, 2014) ประกอบด้วย 1) การจำแนกข้อมูลด้วยต้นไม้ตัดสินใจ (Decision tree classification) เป็นกระบวนการสร้างต้นไม้ขึ้นเพื่อใช้ในการตัดสินใจจากข้อมูลที่มีหมวดหมู่ข้อมูลระบุอยู่ 2) การจำแนกข้อมูลแบบเบย์เซียน (Bayesian classification) เป็นการสร้างตัวจำแนกข้อมูลโดยประยุกต์ใช้ทฤษฎีของเบย์ (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



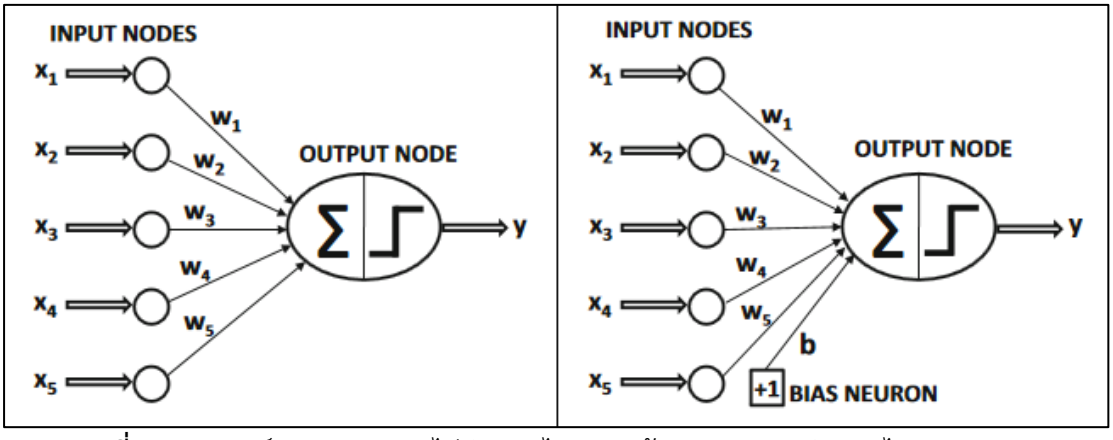
รูปที่ 2.11 แสดงระบบประสาทตามหลักชีววิทยา (ซ้าย) และโครงข่ายประสาทเทียม (ขวา)
ที่มา: Aggarwal (2018)

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

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

เมื่อแทนค่าการทำงานของเพอร์เซปตรอนจากรูปที่ 2.5 ในสมการที่ 2.1 แบบกำหนดค่าไบแอส จะมีรายละเอียดดังนี้

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

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

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

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

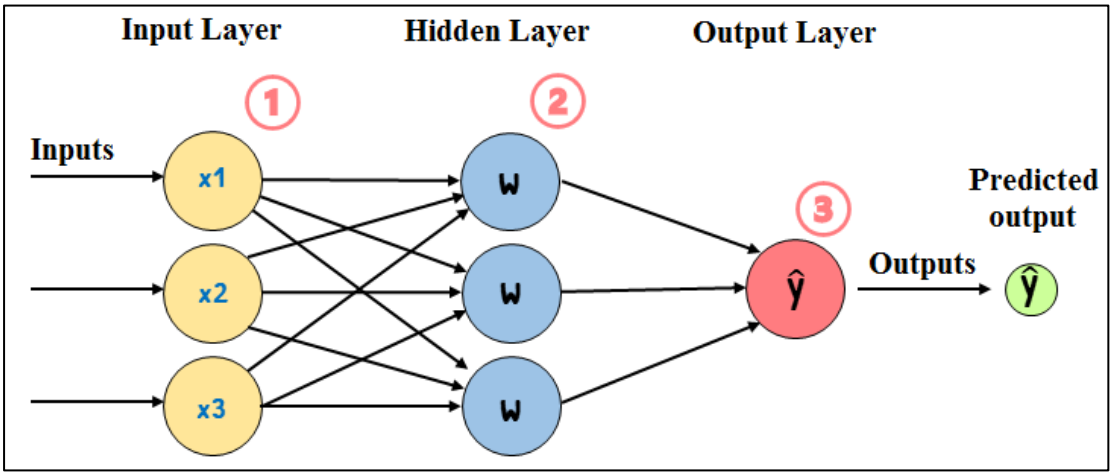
$$where, e = t - a$$

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

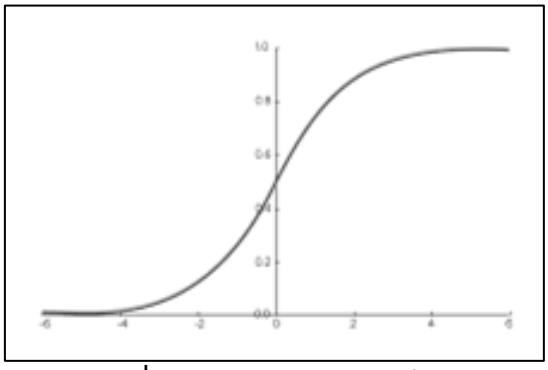
กำหนดสัญลักษณ์แทนการคำนวณไปข้างหน้า โดยให้ a_k^{l-1}แทนผลลัพธ์ของนิวรอนตัวที่ k ในลำดับที่ $l-1$ และ w_{jk}^l แทนน้ำหนักสำหรับนิวรอนตัวที่ j ในลำดับชั้น l ที่มีเส้นเชื่อมมาจากนิวรอนตัวที่ k ในระดับชั้นก่อนหน้า และb_j^l คือไบแอส นอกจากนี้ให้ g แทนฟังก์ชันกระตุ้นและให้ n แทนจำนวนนิวรอนในลำดับชั้นที่ $l-1$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

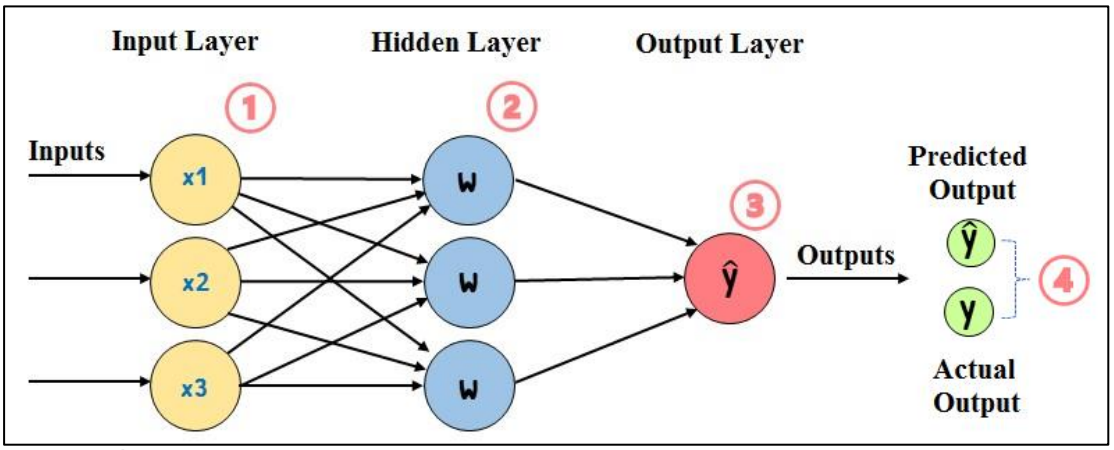
$exp(x_i)$ คือ ฟังก์ชันเอกซ์โพเนนเชียลมาตรฐานมีค่าคงที่เท่ากับ 2.71828 ยกกำลังด้วยเวกเตอร์อินพุต

$\sum_j exp(x_i)$ 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

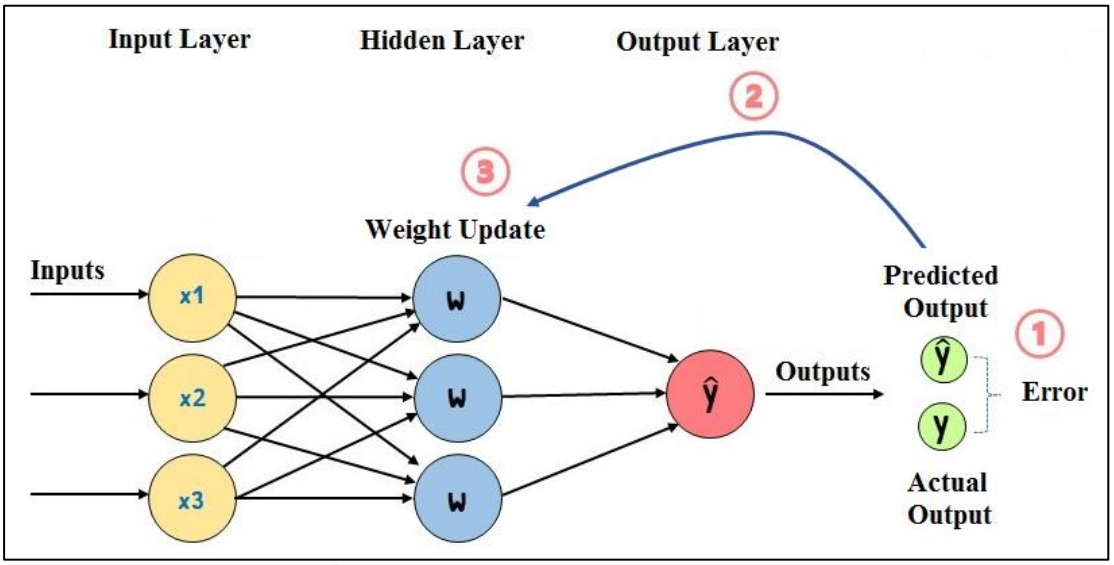
ค่าการสูญเสียครอสเอนโทรปี คือค่าที่เกิดจากการแจกแจงความน่าจะเป็นระหว่างผลลัพธ์จากการพยากรณ์กับผลลัพธ์จริง (True label) ในรูปแบบ One-hot encoding เพื่อจะให้ค่าความเชื่อมั่นที่ทุกคลาสรวมกันเท่ากับ 1 ต้องใช้งานร่วมกับฟังก์ชันซอฟต์แม็กซ์ในชั้นเอาต์พุตด้วยจากรูปที่ 2.5 การแปลงค่าให้เป็นมาตรฐานสำหรับการแจกแจงความน่าจะเป็นตามสัดส่วนของค่าเอาต์พุตจากฟังก์ชันซอฟต์แม็กซ์ด้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

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

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

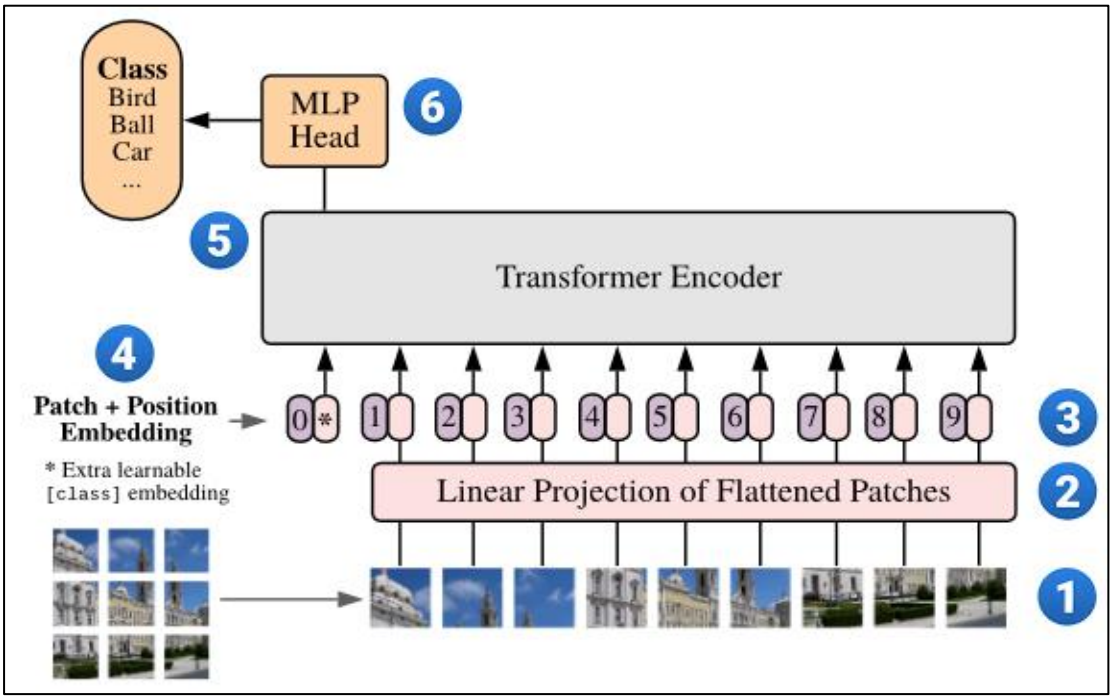
ดังนั้นการใช้สโตแคสติกเกรเดียนต์เดสเซนท์ (Stochastic Gradient Descent: SGD) เพื่อปรับปรุงค่าน้ำหนัก w ดังสมการที่ 2.12

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

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

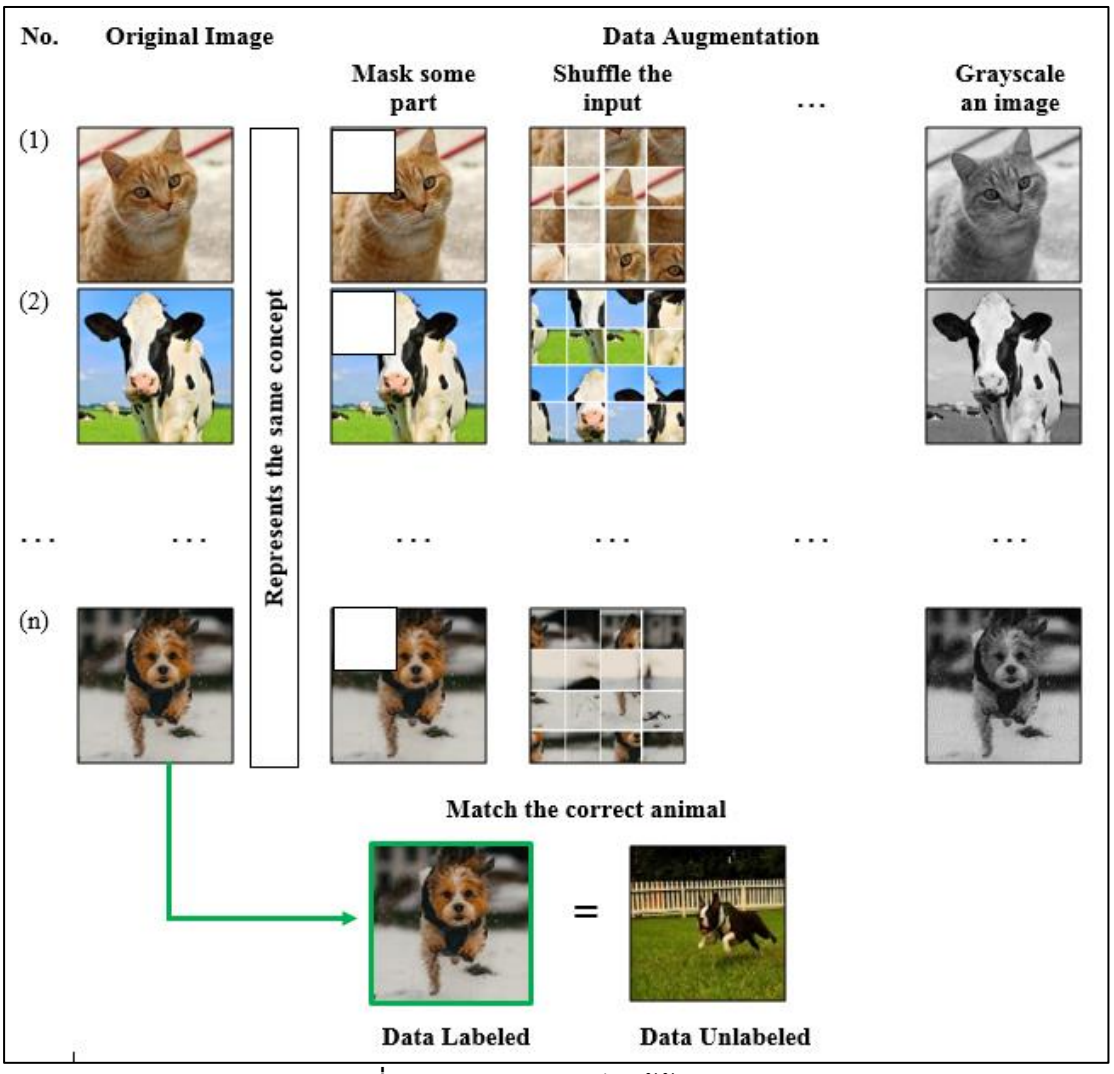
ที่มา: Dosovitskiy et al., (2021)

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

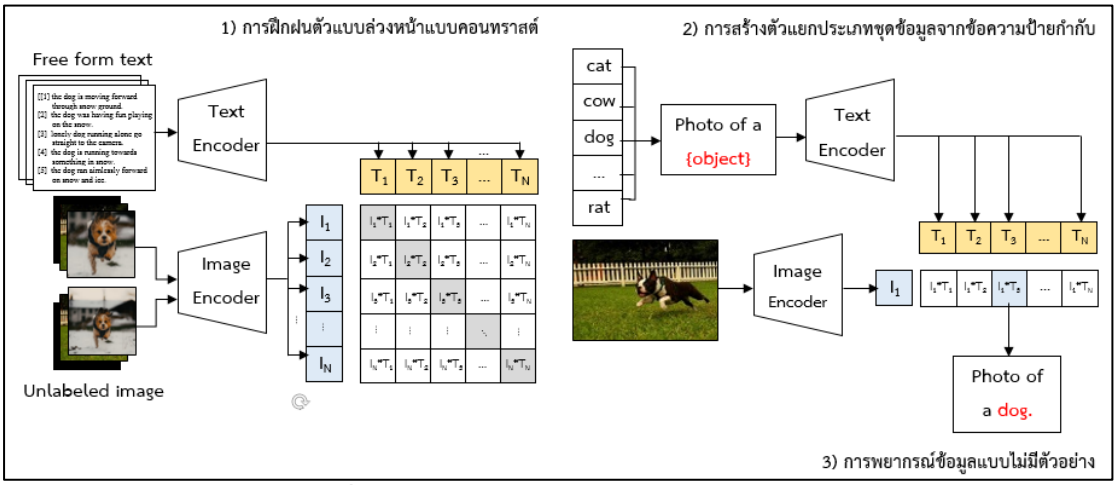
จากรูปที่ 2.18 แนวคิดการเรียนรู้ด้วยตนเองเพื่อทำความเข้าใจรูปภาพจากการเรียนรู้คุณลักษณะเฉพาะของแต่ละรูปภาพ (Represents the same concept) เช่น รูปภาพสุนัขต้นฉบับ (Original image) จากนั้นจะฝึกฝนตัวแบบให้เรียนรู้เพื่อจดจำรูปภาพสุนัขอื่น ๆ ที่มีรูปแบบหรือ

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) อย่างไรก็ดีในความเป็นจริงแล้วการมองภาพของมนุษย์มักจะพิจารณาจากความหมายของรูปภาพเป็นหลักโดยที่ไม่จำเป็นต้องมีคุณลักษณะสีหรือรูปทรงแบบเดียวกันก็สามารถตัดสินว่าเป็นภาพชนิดเดียวกันได้ ส่งผลให้การค้นคืนรูปภาพอาจไม่ได้ผลลัพธ์ที่ตรงกับความหมายในภาพอย่างแท้จริง

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