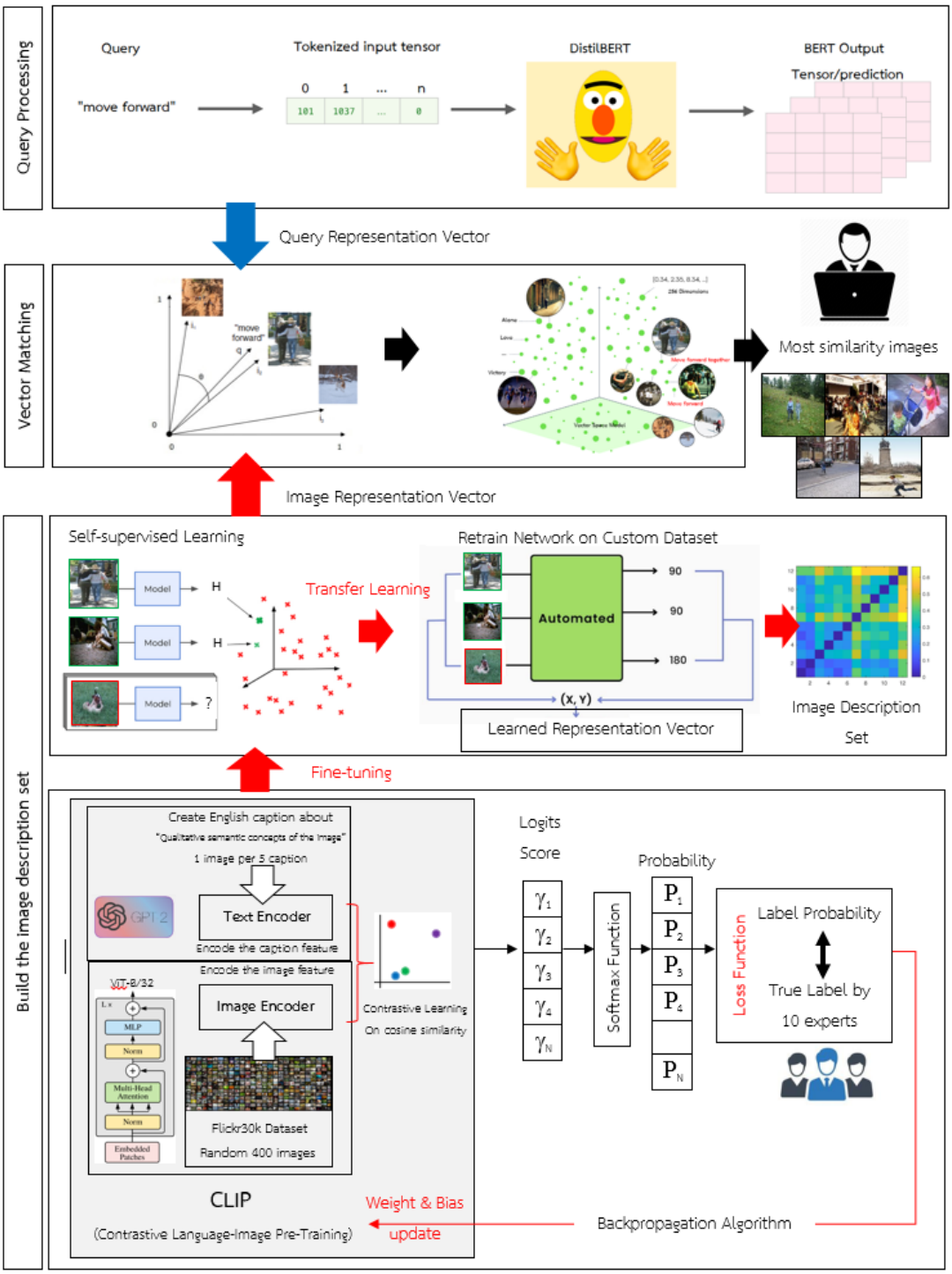
บทที่ 4
ผลการวิจัยและการอภิปรายผล

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

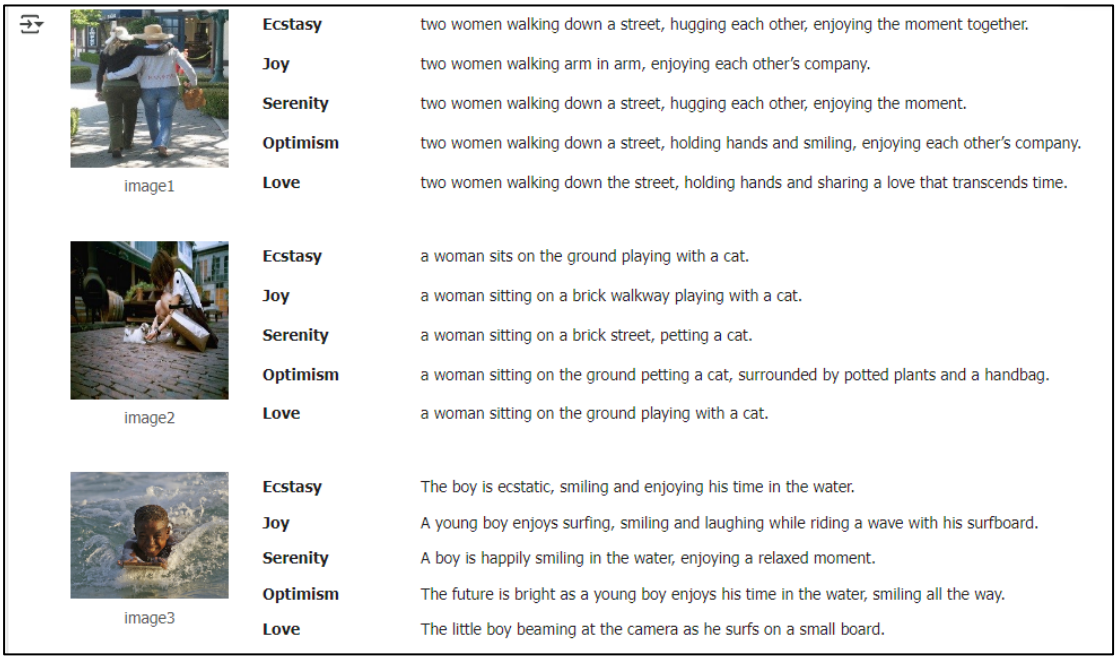


รูปที่ 4.1 การออกแบบและพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมาย โดยประยุกต์โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า

4.1.1 ผลการพัฒนามอดูลการสร้างชุดคำอธิบายรูปภาพ

การพัฒนามอดูลการสร้างชุดคำอธิบายรูปภาพสำหรับใช้ในการค้นคืนรูปภาพจากการเปรียบเทียบความคล้ายคลึงเชิงความหมาย ประกอบด้วย 3 มอดูล ได้แก่

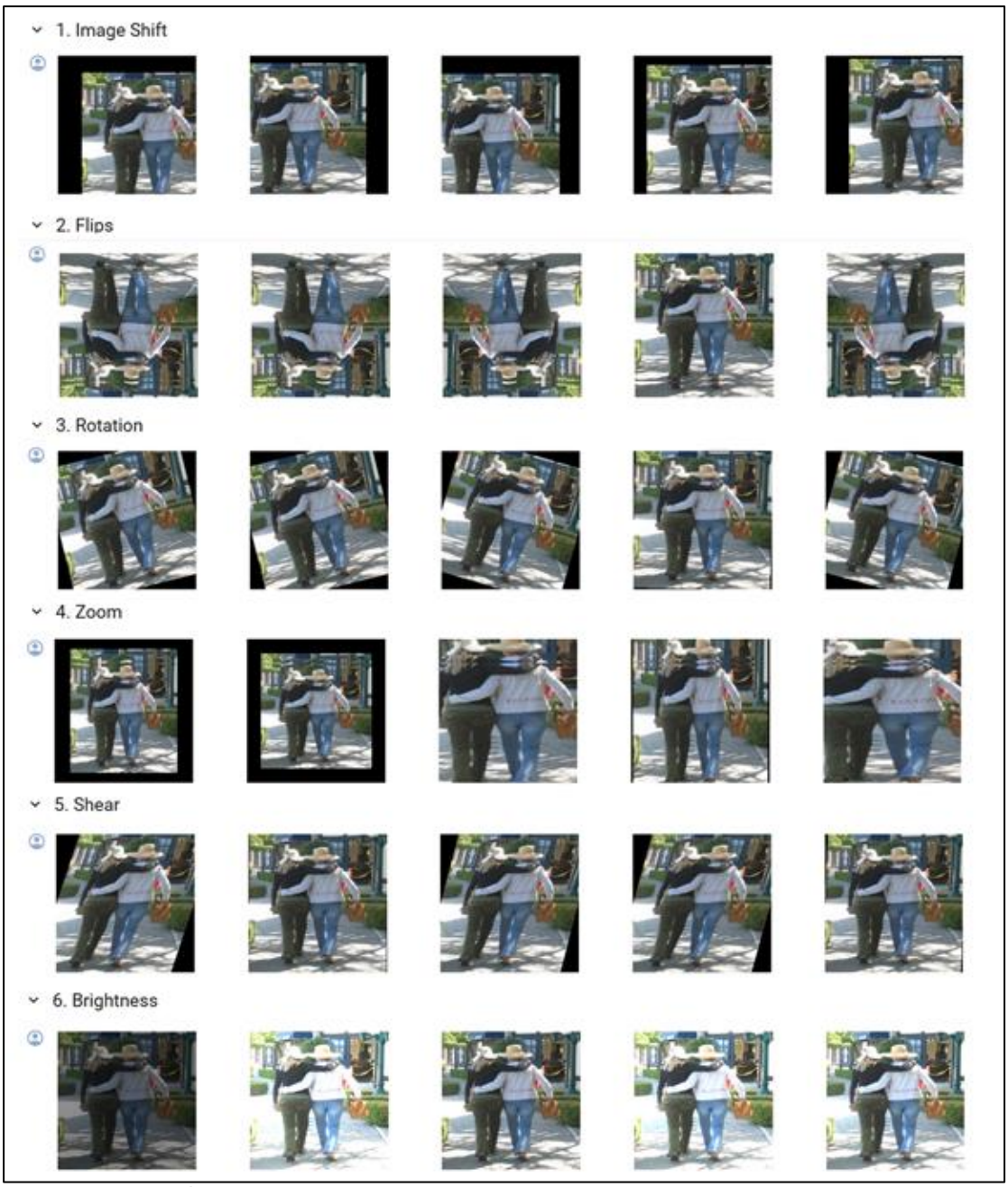
4.1.1.1 ผลการเตรียมข้อมูลก่อนการประมวลผล งานวิจัยนี้สุ่มรูปภาพจำนวน 400 รูปภาพ สำหรับฝึกฝนให้คอมพิวเตอร์เรียนรู้จากรูปภาพและข้อความภาษาธรรมชาติ แบ่งเป็นรูปภาพบนชุดข้อมูล Flickr30k จำนวน 100 รูปภาพ และอีก 300 รูปภาพจากชุดข้อมูลที่ผู้วิจัยรวบรวมเอง ตามสัดส่วนของจำนวนรูปภาพที่ใช้ในงานวิจัย มาสร้างข้อความบรรยายรูปภาพในรูปแบบภาษาอังกฤษ 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

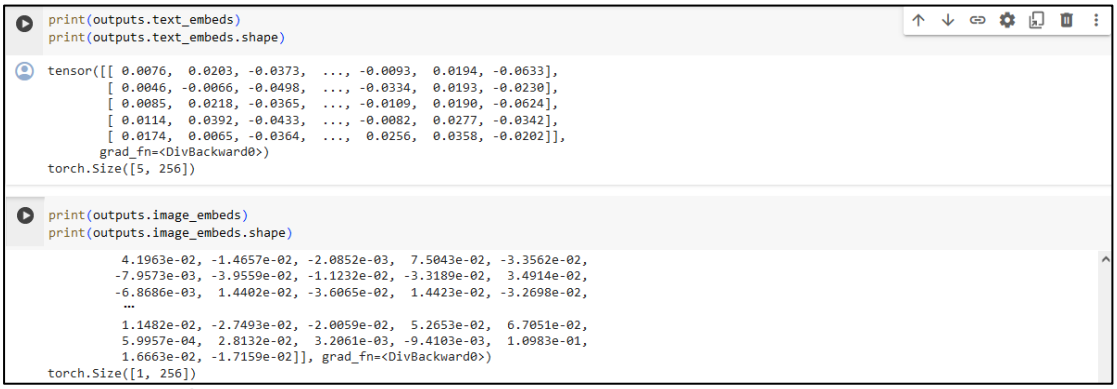
จากรูปที่ 4.5 ก่อนจะแยกข้อความออกเป็น 3 ส่วน ประกอบด้วยชื่อไฟล์รูปภาพในคอลัมน์ A ลำดับข้อความต่อรูปภาพในคอลัมน์ B และข้อความบรรยายรูปภาพในคอลัมน์ C ดังรูปที่ 4.5

2) ผลการฝึกฝนล่วงหน้าแบบคอนทราสต์ระหว่างรูปภาพและข้อความ โดยใช้ตัวแบบ CLIP ที่ถูกฝึกฝนล่วงหน้าบนชุดข้อมูลขนาดใหญ่จึงสามารถถ่ายโอนความรู้ไปใช้ได้เลยโดยไม่จำเป็นต้องฝึกฝนตัวแบบใหม่ตั้งแต่ต้น เพียงปรับแต่งตัวแบบเพิ่มเติมด้วยชุดข้อมูลที่เฉพาะเจาะจงในงานที่ต้องการ เนื่องจากตัวแบบได้เรียนรู้คุณลักษณะที่สำคัญของรูปภาพเอาไว้เป็นโมเดลฐาน เพื่อลดภาระการฝึกฝนตัวแบบจากเดิมที่ต้องการข้อมูลจำนวนมากในการเรียนรู้ อีกทั้งมีประสิทธิภาพดีกว่าตัวแบบที่ถูกสร้างขึ้นมาเองตั้งแต่ต้น มีรายละเอียดดังนี้

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

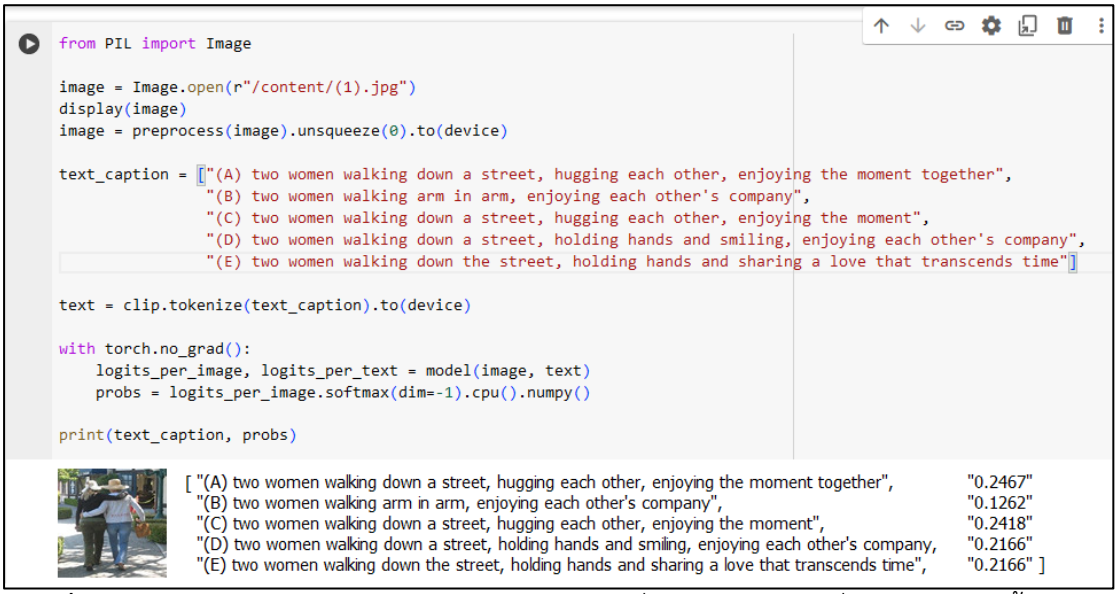
2.2) ผลการฝึกฝนตัวเข้ารหัสข้อความด้วยตัวแบบ GPT-2 บนพื้นฐานสถาปัตยกรรม Transformer โดยการเรียกใช้ CLIPTokenizer ที่ใช้เข้ารหัสข้อความ และ CLIPTextModel ที่ใช้ฝังข้อความบนพื้นที่ฝังหลายรูปแบบ เพื่อเรียนรู้ความหมายโดยรวมของข้อความบรรยายรูปภาพแต่ละประโยค ผลลัพธ์จากการฝังข้อความ (Text embedding) แปลงเป็นเวกเตอร์คุณลักษณะข้อความ (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

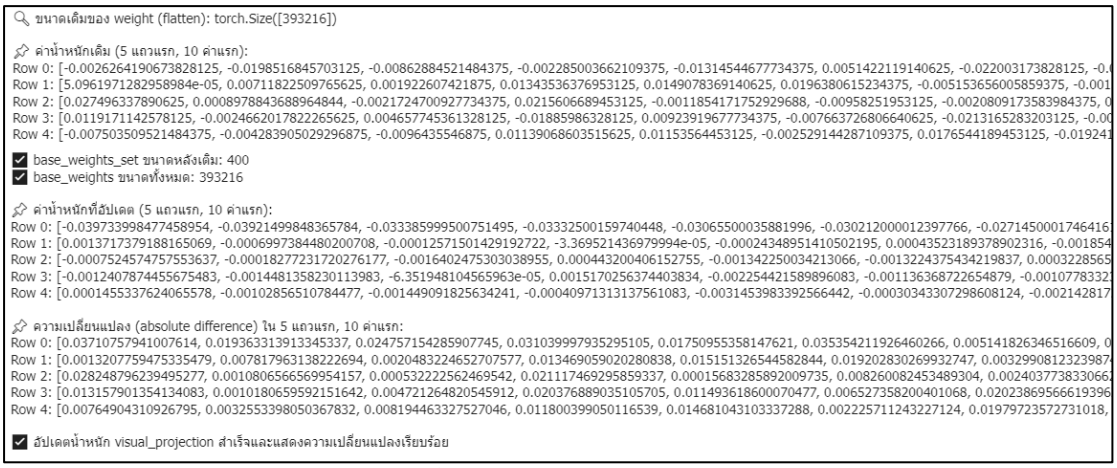
3.1) ผลการคำนวณค่าถ่วงน้ำหนักเฉลี่ย และสร้างค่าน้ำหนักใหม่โดยอิงผู้เชี่ยวชาญจากการเปรียบเทียบระหว่างผลลัพธ์จากตัวแบบกับผลลัพธ์ที่ถูกประเมินความหมายโดยผู้เชี่ยวชาญด้วยการหาค่าการสูญเสีย เพื่อย้อนกลับไปปรับแต่งค่าน้ำหนักด้วยตนเอง (Custom update) โดยตัวอย่างค่าถ่วงน้ำหนักเฉลี่ยจากผู้เชี่ยวชาญทั้ง 10 ท่านดังตารางที่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

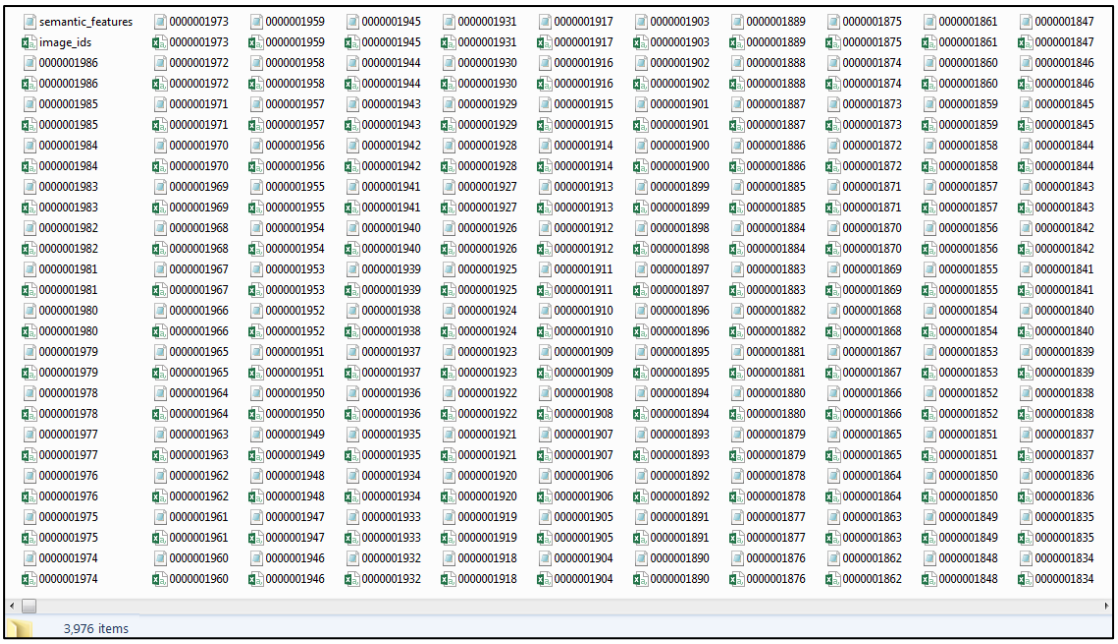
```
batch_size = 16

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

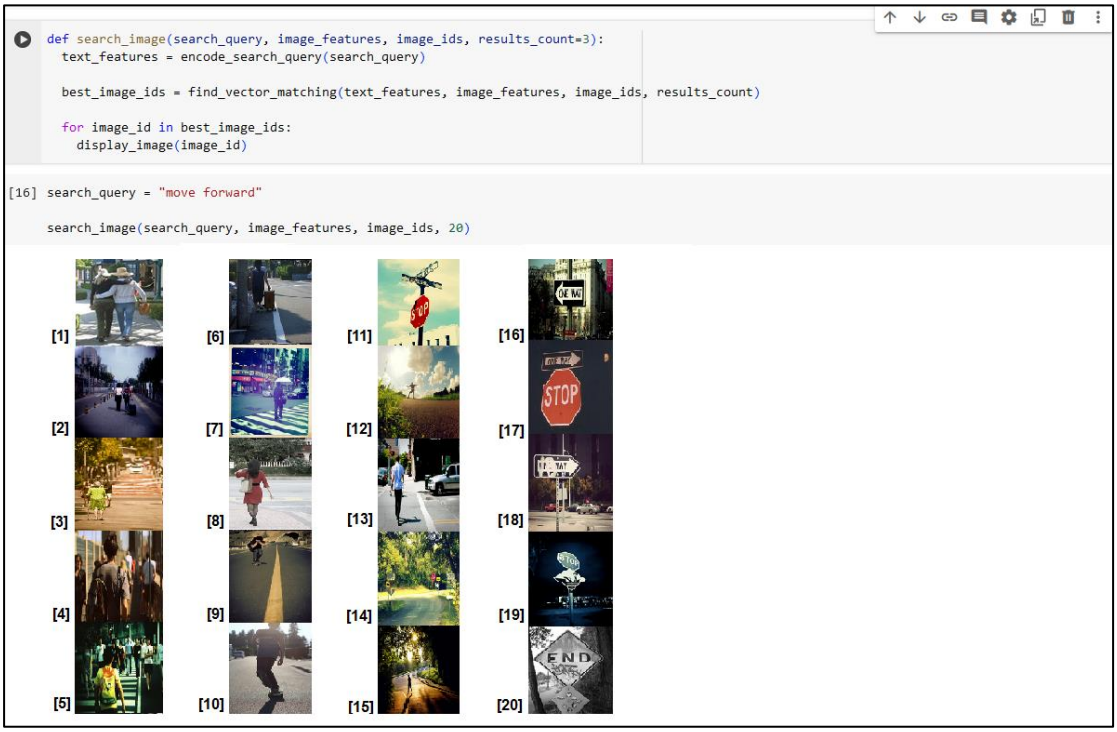
3) จากนั้นจะใช้เวกเตอร์ข้อความค้นหาเพื่อค้นหาเวกเตอร์การฝังที่คล้ายคลึงกันในฐานข้อมูลเวกเตอร์บนพื้นที่การฝังหลายรูปแบบ สำหรับแปลงเวกเตอร์ที่มีขนาดแตกต่างกันให้อยู่ในมิติเท่ากับ 256 เพื่อให้สามารถเปรียบเทียบความคล้ายกันได้ในปริภูมิเวกเตอร์มิติสูง (High-dimensional vector space) ด้วยวิธีวัดระยะทางระหว่างเวกเตอร์รูปภาพและเวกเตอร์ข้อความค้นหาแบบระยะทางแบบยุคลิด

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

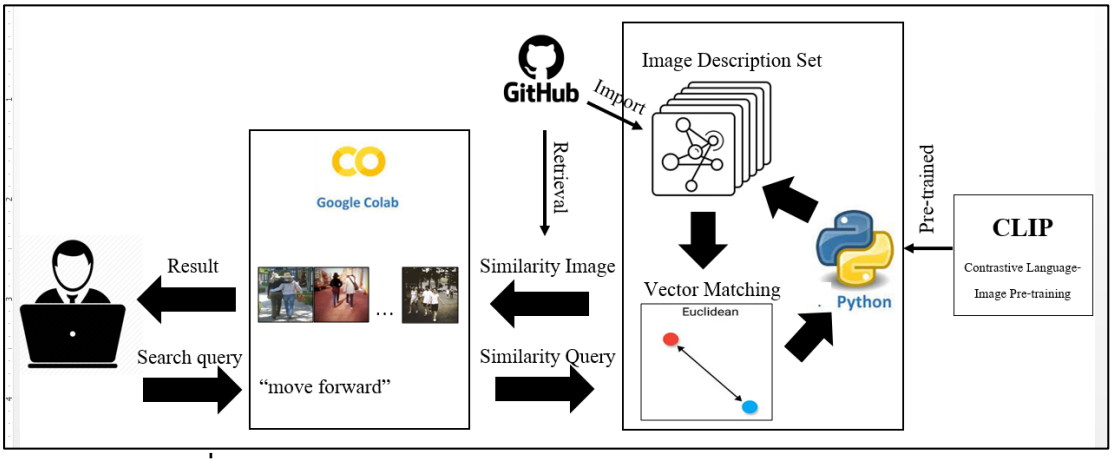
โดยใช้วิธีการวัดความคล้ายคลึงโคไซน์ ซึ่งเป็นวิธีการที่นิยมใช้ในการประเมินความสัมพันธ์ของข้อมูลในเวกเตอร์สเปซ ผลลัพธ์คือรูปภาพจากการค้นคืนจำนวน 20 ภาพที่มีความคล้ายคลึงกับข้อความมากที่สุดตามลำดับ โดยแต่ละภาพแสดงถึงกิจกรรมที่มีลักษณะสอดคล้องกับแนวคิด “การเคลื่อนไปข้างหน้า” เช่น ภาพที่ 1, 6, 8, 11 และ 12 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

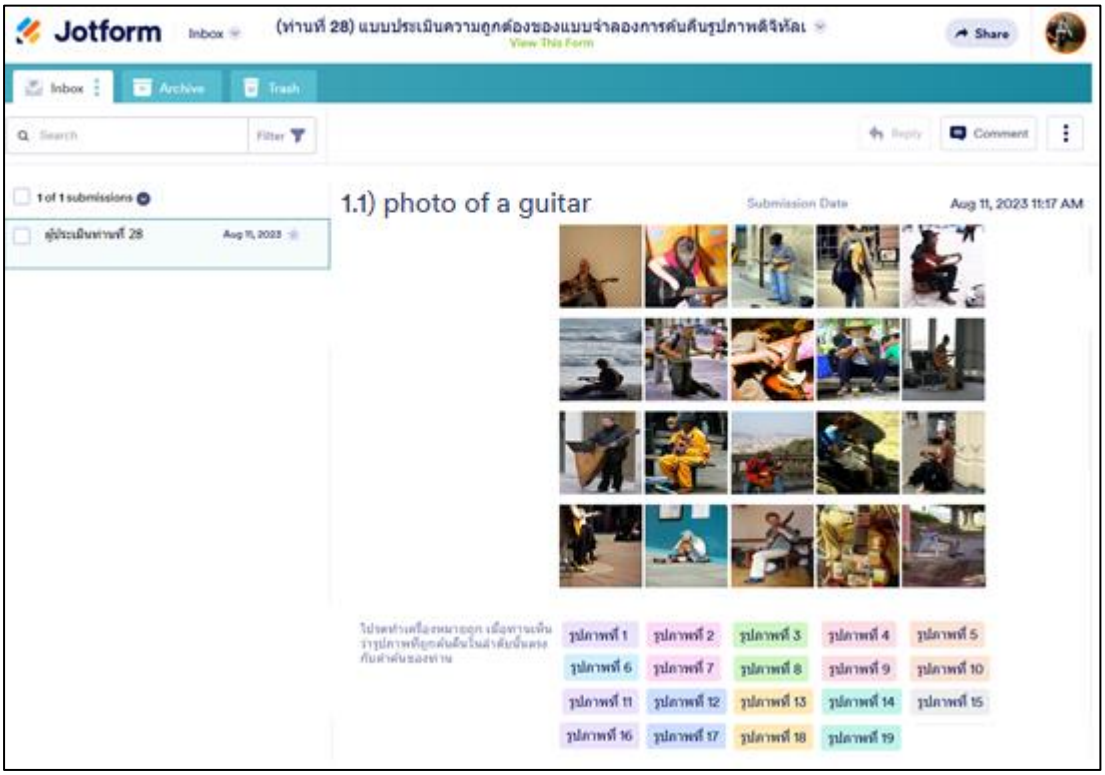


รูปที่ 4.13 ภาพรวมแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมาย โดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า

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

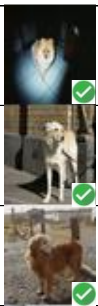
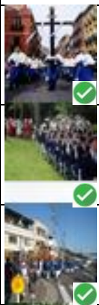
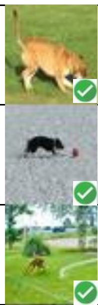
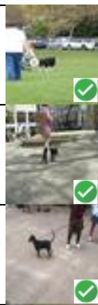
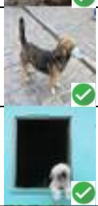
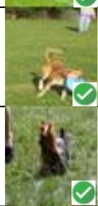
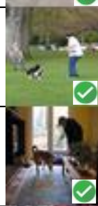
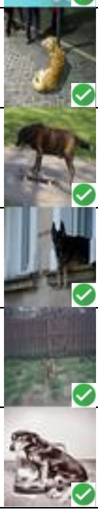
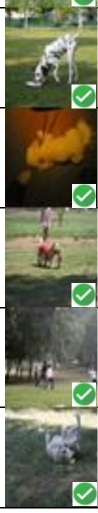
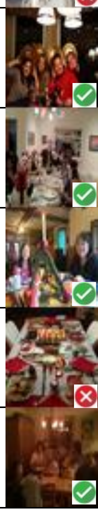
รูปภาพในการแสดงผลแต่ละครั้งเท่ากับ 20 รูปภาพ ภายใต้ข้อความค้นหาภาษาอังกฤษ 3 เงื่อนไข ได้แก่ 1) ข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ ในลักษณะป้ายกำกับของรูปภาพ 2) ข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพ และ 3) ข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ โดยกำหนดข้อความค้นหาจากผู้เชี่ยวชาญเงื่อนไขละ 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



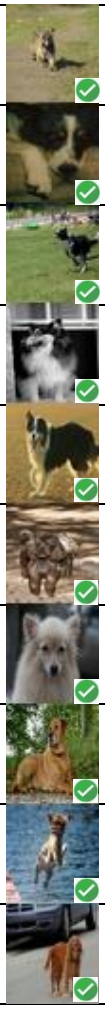
รูปที่ 4.14 ตัวอย่างการประเมินความถูกต้องของรูปภาพผ่านแบบประเมินออนไลน์ Jotform

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 ที่ผ่านการประเมินความถูกต้อง โดยผู้เชี่ยวชาญท่านที่ 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) ข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ในลักษณะป้ายกำกับของรูปภาพ**			**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 เกี่ยวกับแนวคิดระดับสูงของรูปภาพนั้นมีค่าความครบถ้วนอยู่ในระดับดีเมื่อผลลัพธ์จากการค้นคืนเท่ากับ 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

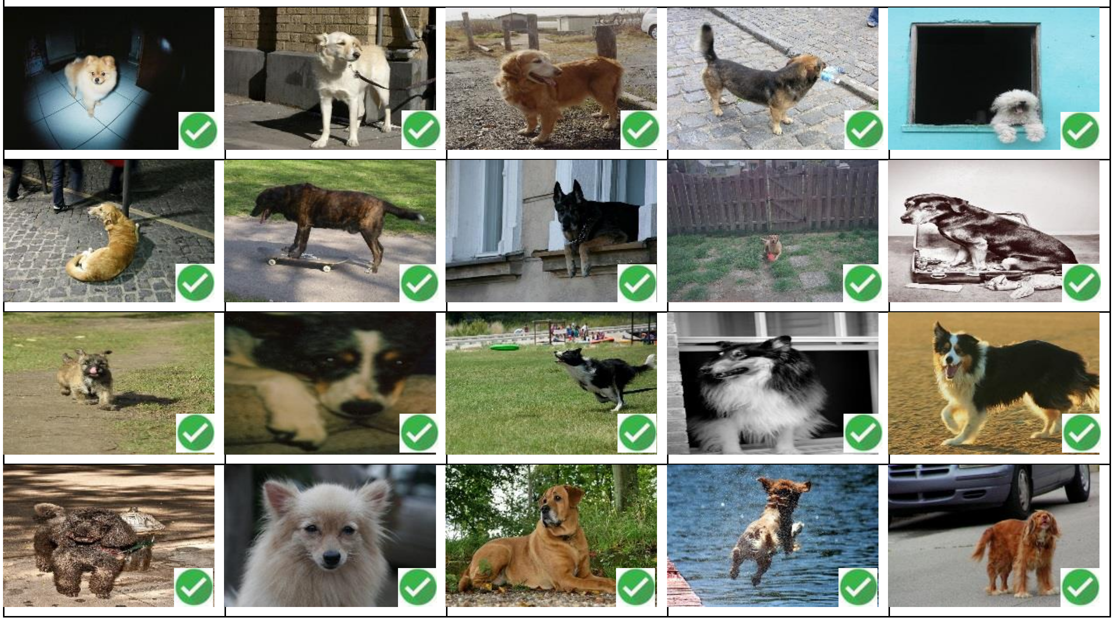
ประเมินผลว่าถูกต้องหรือไม่ถูกต้องเท่านั้น ดังนั้นการแปลความหมายตามหลักการรับรู้ของมนุษย์ (Human perception) (Dix et al., 2004) ด้วยประสบการณ์หรือความรู้เดิมจากการเรียนรู้ของแต่ละบุคคลจึงแตกต่างกัน เห็นได้จากผลลัพธ์การค้นคืนรูปภาพเชิงความหมายจากข้อความค้นหามีค่าความแม่นยำไม่แตกต่างจากการใช้รูปภาพค้นหามากนัก

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

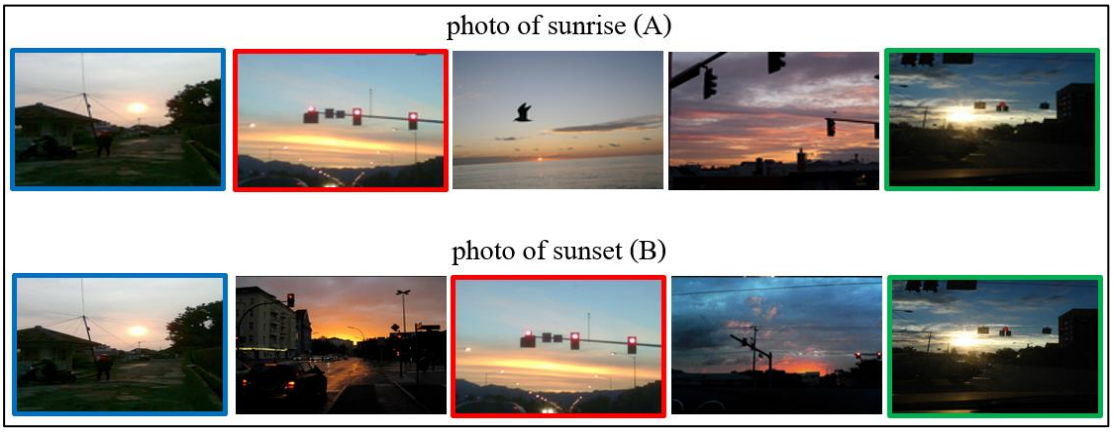
n) ...
... ...
300) photo of a tradition


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

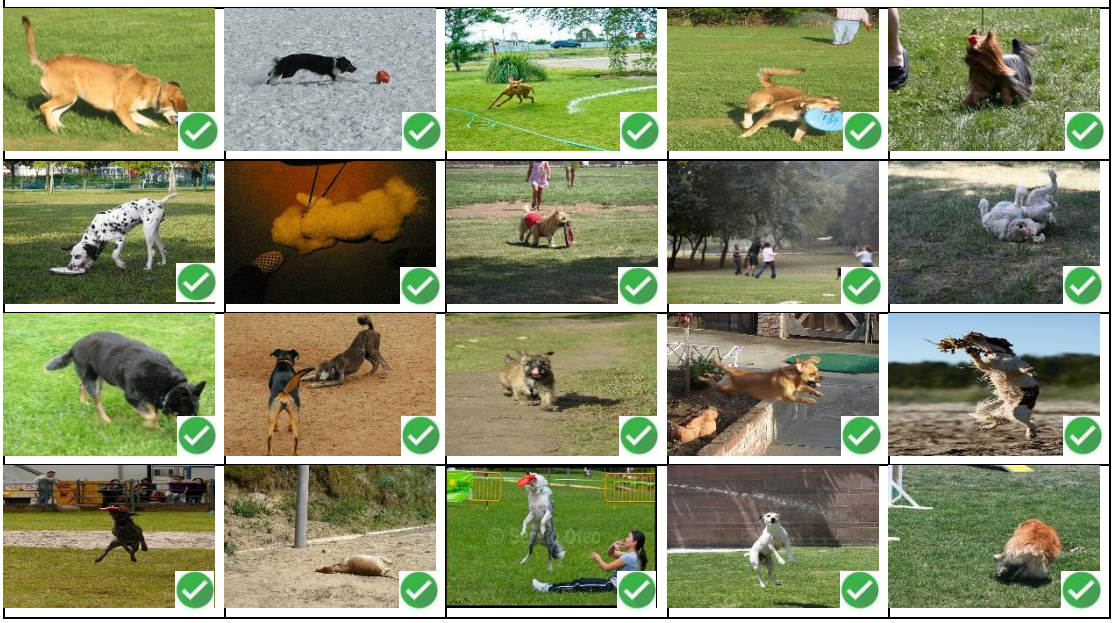
นอกจากนี้ ผลการสังเกตและสัมภาษณ์ผู้เชี่ยวชาญในการประเมินความถูกต้อง พบว่า ยังมีข้อจำกัดในการค้นคืนรูปภาพที่มีความกำกวมเช่นเดียวกับมนุษย์ เช่น ข้อความค้นหา “photo of sunset” กับ “photo of sunrise” ดังรูปที่ 4.15 ที่มีผลลัพธ์ซ้ำกันถึง 3 จาก 5 รูปภาพ ซึ่งยากที่จะแยกความแตกต่างของรูปภาพได้ หากปราศจากองค์ประกอบอื่น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

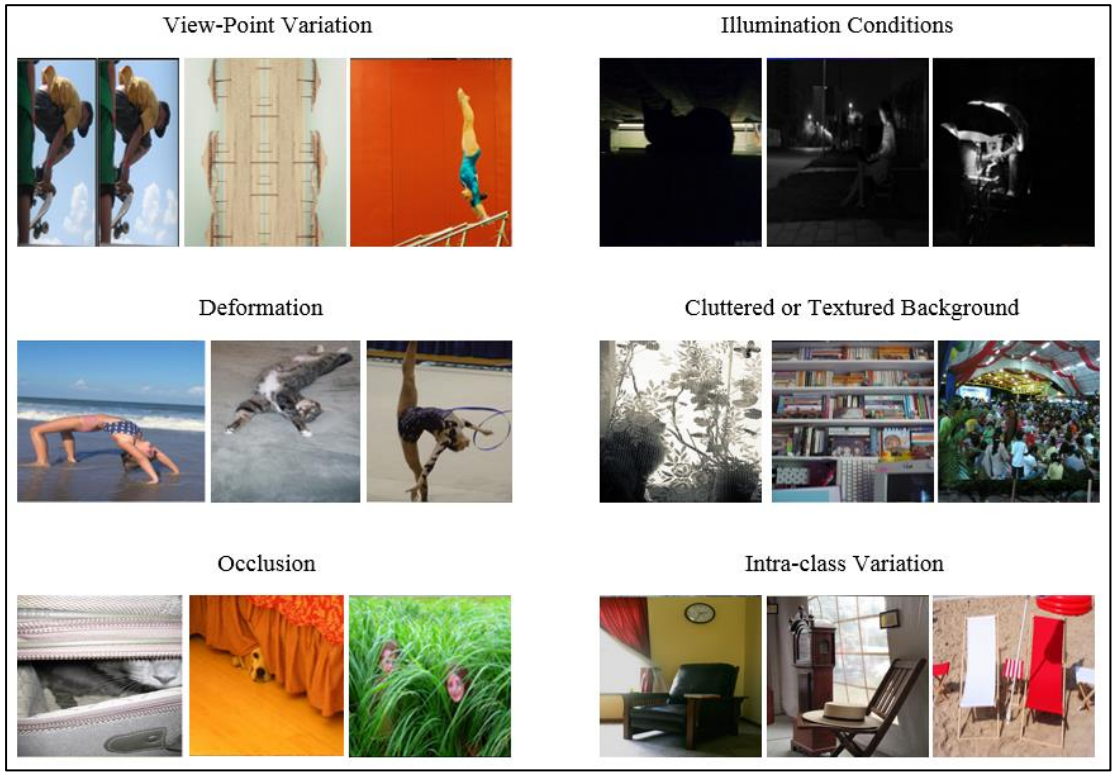
2) the baby sleeps peacefully in the bed				
				
n) ...				
...	...	...	...	...
300) we enjoy our family traditions every year				
				

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

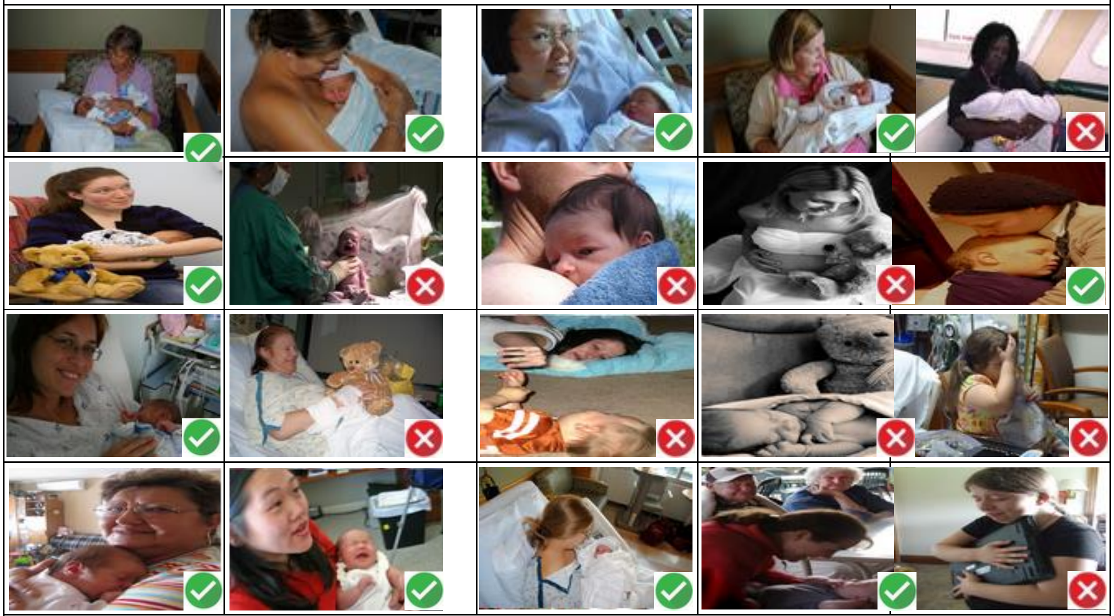
จึงจำเป็นต้องลดความแตกต่างภายในคลาส ขณะเดียวกันก็ขยายความแตกต่างระหว่างคลาส (Inter-class variation) เพื่อเพิ่มประสิทธิภาพการจำแนก

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				
				

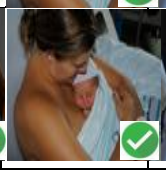
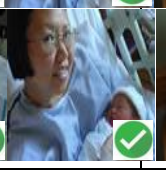
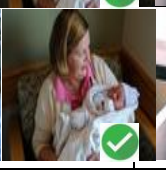
จากตารางที่ 4.9 ผลลัพธ์จากการค้นคืนรูปภาพด้วยข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพที่ผ่านการประเมินความถูกต้องโดยผู้เชี่ยวชาญ พบว่า ค่าความแม่นยำที่ 3 อันดับจากกลุ่มผู้เชี่ยวชาญด้านคอมพิวเตอร์และเทคโนโลยี กลุ่มผู้เชี่ยวชาญด้านการค้นคืนสารสนเทศ และกลุ่มผู้ใช้ทั่วไปอยู่ในระดับดีมาก ค่าเฉลี่ยเท่ากับ 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

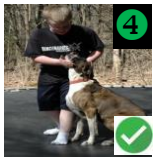
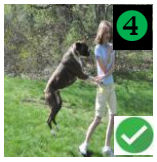
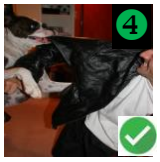
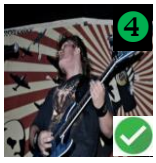
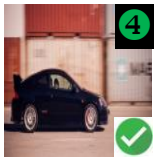
แม้โมเดลทั้งสองจะใช้โครงสร้างพื้นฐานเดียวกัน แต่การปรับค่าน้ำหนักทำให้โมเดลมีความสามารถในการเรียนรู้รายละเอียดเชิงโดเมนที่จำเพาะเจาะจงมากขึ้น เช่น คุณลักษณะบางประการในภาพที่สัมพันธ์กับชุดข้อมูลฝึกฝน หรือการลดความคลาดเคลื่อนในการจับคู่เวกเตอร์คุณลักษณะรูปภาพและคุณลักษณะข้อความค้นหาเชิงความหมายได้ดีขึ้น (Radford et al., 2021) ส่งผลให้คุณลักษณะที่ได้มีความละเอียดและแยกแยะความต่างได้ดีขึ้นในการค้นคืน รูปภาพ 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

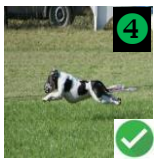
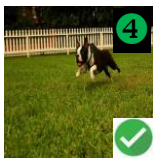
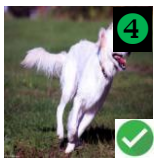
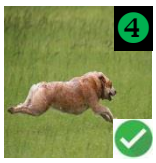
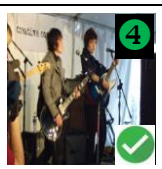
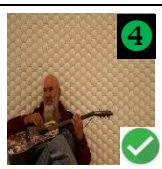
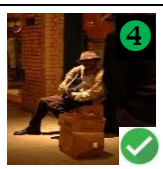
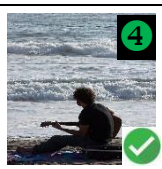
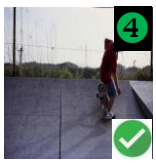
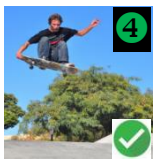
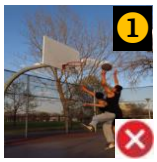
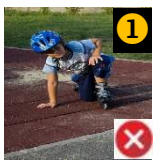
- 0 หมายถึง ไม่มีความเกี่ยวข้องเลย (Non-relevant item)
- 1 หมายถึง มีความเกี่ยวข้องน้อย (Relevant item)
- 2 หมายถึง มีความเกี่ยวข้องปานกลาง (Medium relevant item)
- 3 หมายถึง มีความเกี่ยวข้องมาก (High relevant item)
- 4 หมายถึง มีความเกี่ยวข้องมากที่สุด (Very high relevant item)

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

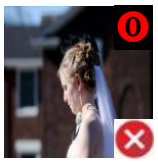
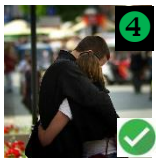
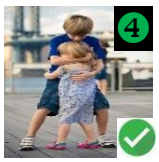
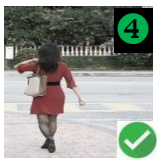
ตารางที่ 4.13 ตัวอย่างผลลัพธ์การค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิมกับการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนัก บนชุดข้อมูล Flickr30k

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

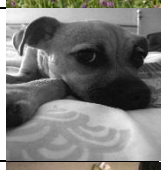
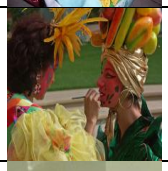
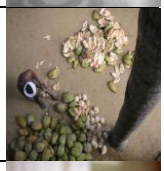
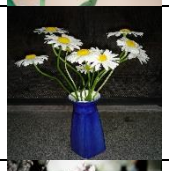
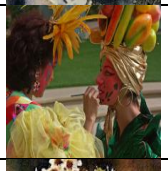
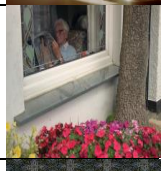
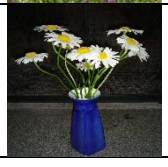
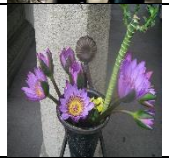
เงื่อนไขการค้นหา	**การค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิม**			**การค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนัก**		
	NDCG@1	**NDCG@3**	**NDCG@5**	**NDCG@1**	**NDCG@3**	**NDCG@5**
1) ข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ ในลักษณะป้ายกำกับของรูปภาพ	0.863	0.899	0.902	0.900	0.902	0.928
2) ข้อความค้นหาด้วยสั้น 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

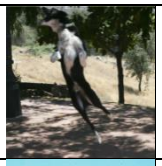
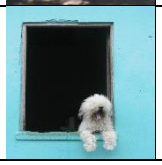
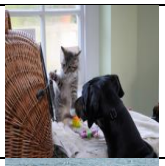
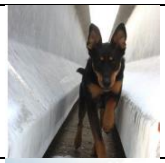
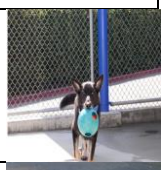
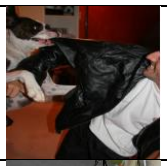
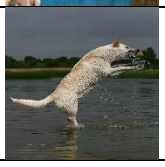
1) ผลการค้นคืนรูปภาพด้วยข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ในลักษณะป้ายกำกับของรูปภาพ พบว่า ค่า NDCG ที่ลำดับ 1, 3 และ 5 จากการค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิม เท่ากับ 0.863, 0.899 และ 0.902 ตามลำดับ ไม่แตกต่างกันมากนักกับค่า NDCG ที่ลำดับ 1, 3 และ 5 จากการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนักเท่ากับ 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

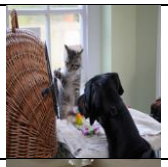
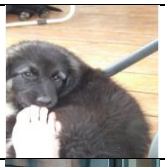
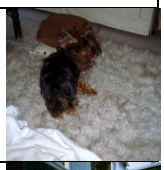
ตารางที่ 4.16 ตัวอย่างผลการเปรียบเทียบผลลัพธ์จากข้อความค้นหาแบบจำแนกรายละเอียดเชิงลึก

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