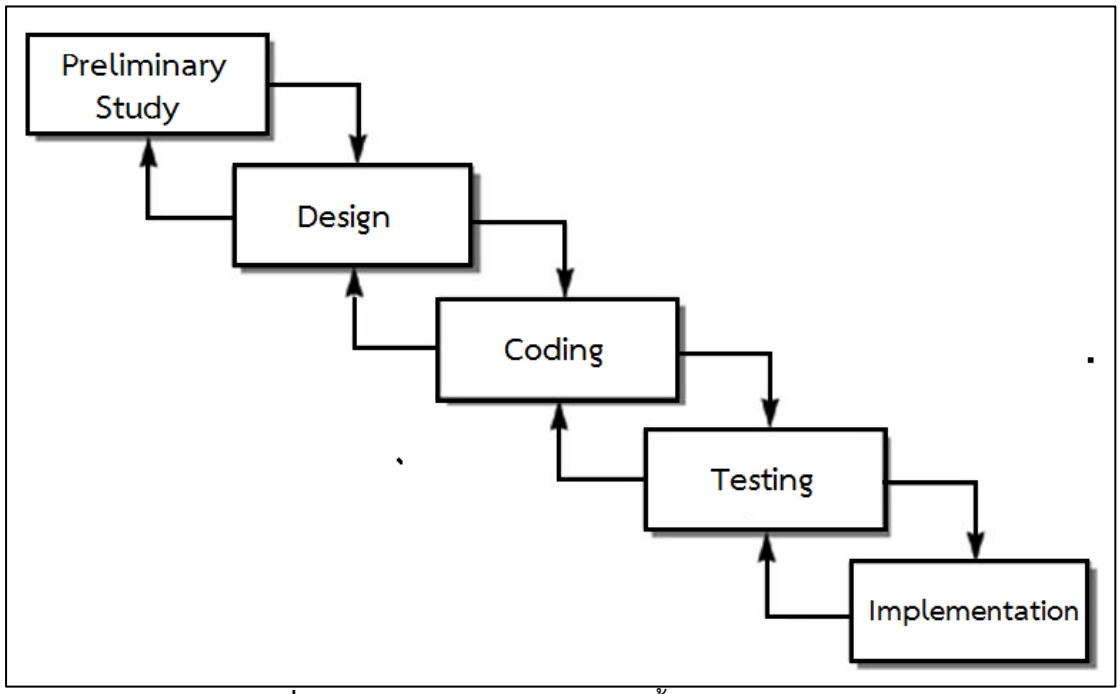
บทที่ 3
วิธีดำเนินการวิจัย

การพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมาย โดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายโดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า และประเมินประสิทธิภาพแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมาย เนื้อหาในบทนี้ประกอบด้วยหัวข้อวิธีวิจัย เครื่องมือที่ใช้ในการวิจัย การรวบรวมข้อมูล และการวิเคราะห์ข้อมูล

3.1 วิธีวิจัย

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



รูปที่ 3.1 วงจรการพัฒนาระบบแบบน้ำตกแบบย้อนกลับได้

3.1.1 การศึกษาข้อมูลเบื้องต้น

งานวิจัยนี้มีการศึกษาข้อมูลเบื้องต้น (Preliminary study) เพื่อกำหนดมาตรฐานสิ่งที่ส่งผลต่อการพัฒนาแบบจำลอง ประกอบด้วย

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

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

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

3.1.2 การออกแบบ

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

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

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

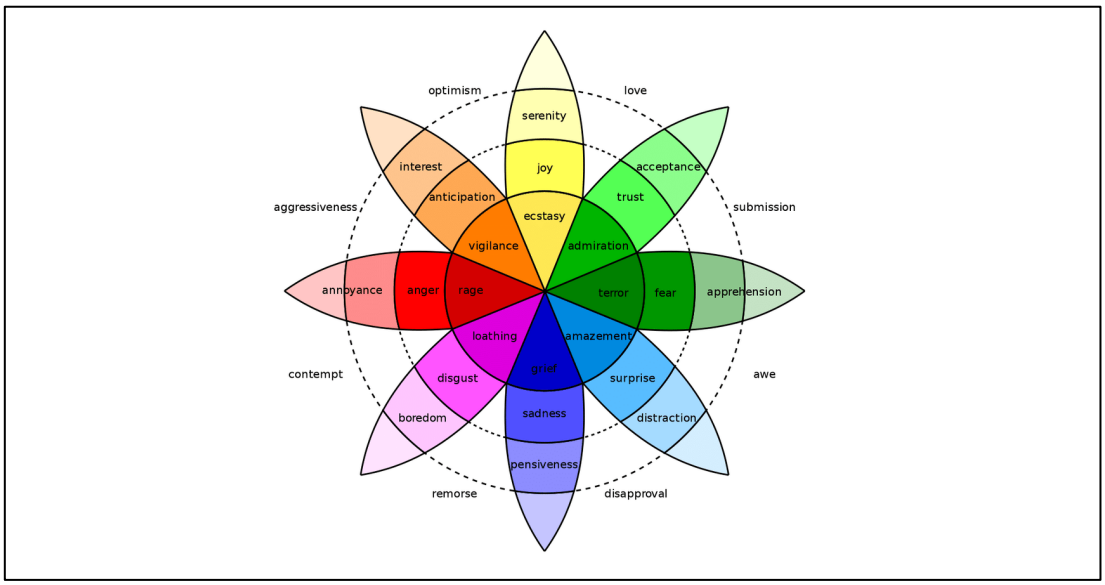
$$n = \frac{N}{1+(N)(e^2)} \quad (3.1)$$

กำหนดให้ n คือ ขนาดของรูปภาพที่สุ่มจากชุดข้อมูล

N คือ ขนาดของรูปภาพทั้งหมดในชุดข้อมูล

e คือ ค่าความคลาดเคลื่อนของการสุ่มตัวอย่างที่ยอมรับได้

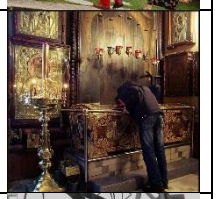
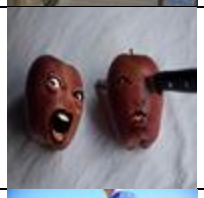
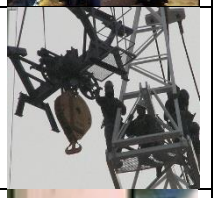
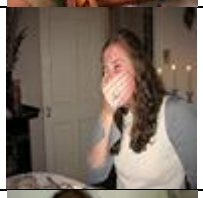
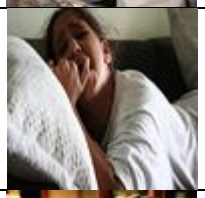
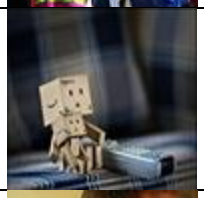
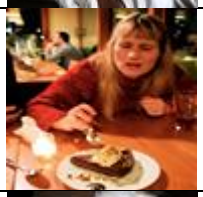
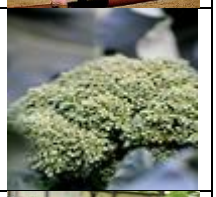
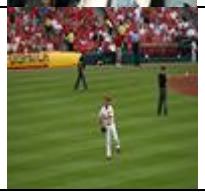
จากสมการที่ 3.1 พบว่า ขนาดของรูปภาพที่สุ่มจากชุดข้อมูลจะเท่ากับ 398.7133 รูปภาพ งานวิจัยนี้กำหนดจำนวนรูปภาพที่สุ่มจากชุดข้อมูลเท่ากับ 400 รูปภาพ แบ่งออกเป็น 8 โดเมน โดเมนละ 50 รูปภาพ รวมทั้งสิ้น 400 รูปภาพ ตามอารมณ์พื้นฐาน 8 รูปแบบ ประกอบด้วย ความรื่นเริง (Joy) ความไว้วางใจ (Trust) ความกลัว (Fear) ความประหลาดใจ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



รูปที่ 3.2 วงล้ออารมณ์

ที่มา: Plutchik (1980)

ตารางที่ 3.1 ตัวอย่างรูปภาพที่สุ่มจากชุดข้อมูล Flickr30k และชุดข้อมูลที่ผู้วิจัยรวบรวมเอง

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

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

1) การนำเข้าโมเดล LLaVA ที่ผ่านการฝึกฝนล่วงหน้าที่รองรับอินพุตแบบมัลติโมดัลพร้อมกันทั้งรูปภาพและข้อความ ซึ่งเป็นคุณสมบัติสำคัญที่ทำให้โมเดลสามารถสร้างข้อความบรรยายรูปภาพที่อิงจากความหมายของภาพได้อย่างมีนัยสำคัญ

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

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

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

5) การคืนผลลัพธ์ภาพพร้อมข้อความคำตอบกลับเป็นข้อความภาษาอังกฤษที่มีลักษณะเป็นคำบรรยายรูปภาพเชิงความหมาย โดยข้อความนี้จะถูกนำมาใช้เป็นตัวเลือกในกระบวนการประเมินของผู้เชี่ยวชาญ เพื่อพิจารณาว่าข้อความใดสื่อความหมายของรูปภาพได้ใกล้เคียงกับการรับรู้ของมนุษย์มากที่สุด

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

ตารางที่ 3.2 ตัวอย่างรูปภาพที่สุ่มจากชุดข้อมูล Flickr30k และชุดข้อมูลที่ผู้วิจัยรวบรวมเอง ต่อข้อความบรรยายรูปภาพที่อธิบายความหมายเชิงคุณภาพของรูปภาพ

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

ตารางที่ 3.2 ตัวอย่างรูปภาพที่สุ่มจากชุดข้อมูล Flickr30k และชุดข้อมูลที่ผู้วิจัยรวบรวมเอง ต่อข้อความบรรยายรูปภาพที่อธิบายความหมายเชิงคุณภาพของรูปภาพ (ต่อ)

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

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

ในการสร้างความหลากหลายเชิงภาษาศาสตร์ และเพื่อส่งเสริมให้แบบจำลองสามารถเข้าใจภาษาในระดับนามธรรมได้ดีขึ้น (Lin et al., 2014) และสามารถนำไปประยุกต์ใช้ในบริบทการค้นคืนรูปภาพได้อย่างมีประสิทธิภาพ

การเตรียมข้อมูลในงานวิจัยนี้แบ่งเป็นการเตรียมรูปภาพก่อนการประมวลผล และการเตรียมข้อความก่อนการประมวลผล มีรายละเอียดดังนี้

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

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

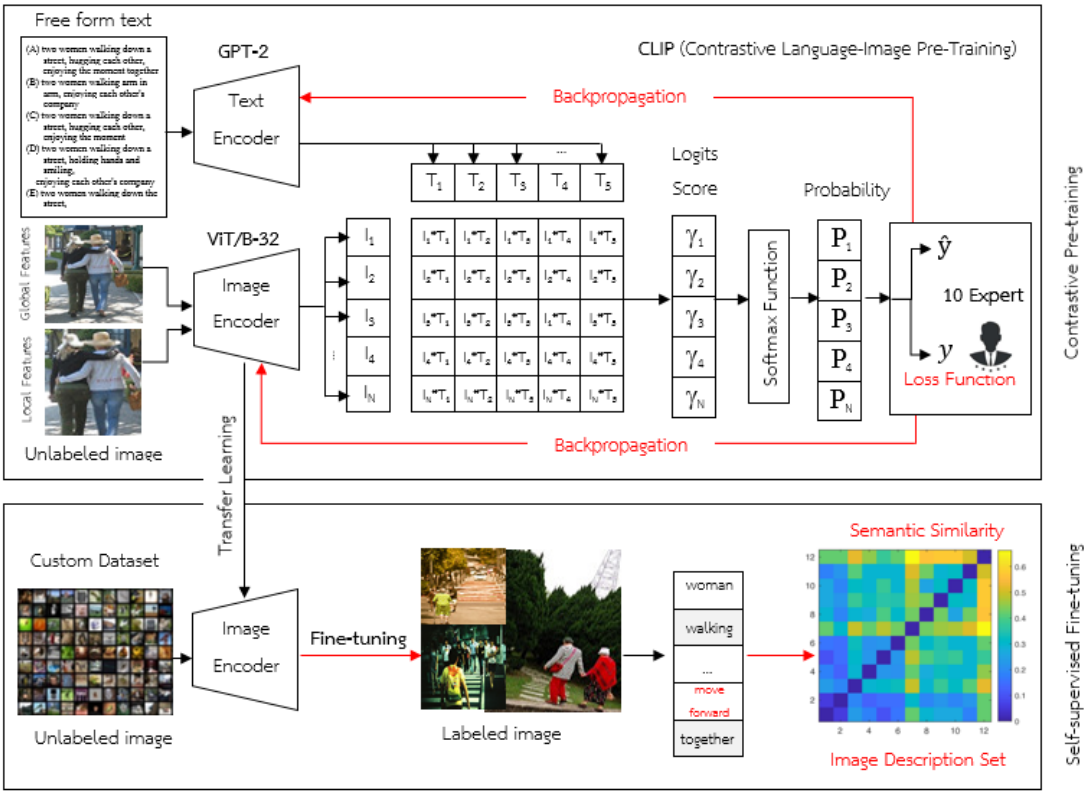
3.1.2.1.2 การฝึกฝนล่วงหน้าแบบคอนทราสต์ระหว่างรูปภาพและข้อความ (Image-text contrastive pre-training) งานวิจัยนี้ใช้ตัวแบบ CLIP เป็นโมเดลฐาน ดังรูปที่ 3.3 มีรายละเอียดดังนี้

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

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

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

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



รูปที่ 3.3 แบบจำลองการค้นคืนรูปภาพเชิงความหมาย โดยฝึกฝนล่วงหน้าแบบคอนทราสต์

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

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

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

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

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

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

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

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

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

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

$$= 0.2467$$

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

ตารางที่ 3.3 ผลการพยากรณ์ความน่าจะเป็นของป้ายกำกับภาษาธรรมชาติ

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

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

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

กระบวนการพัฒนาโมเดลโครงข่ายประสาทเทียมนั้นการตั้งค่าน้ำหนักของพารามิเตอร์ภายในโมเดลเป็นหนึ่งในกระบวนการสำคัญที่มีผลต่อประสิทธิภาพของ

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

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

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

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

แทนการตัดสินแบบสัมบูรณ์ (Absolute judgment) ซึ่งเป็นการเรียนรู้จากสถานการณ์จำลองที่โมเดลอาจเจอกับรูปภาพที่มีคุณลักษณะเชิงความหมายที่ซับซ้อน หรือมีความหมายใกล้เคียงกันมาก และยากที่จะจำแนกด้วยการประเมินแบบถูกหรือผิดเท่านั้น

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

ตารางที่ 3.4 ตัวอย่างการเปรียบเทียบผลการประเมินความหมายของรูปภาพจากผู้เชี่ยวชาญท่านที่ 1 กับผลการพยากรณ์ความน่าจะเป็นของป้ายกำกับภาษาธรรมชาติ

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

ตารางที่ 3.4 ตัวอย่างการเปรียบเทียบผลการประเมินความหมายของรูปภาพจากผู้เชี่ยวชาญท่านที่ 1 กับผลการพยากรณ์ความน่าจะเป็นของป้ายกำกับภาษาธรรมชาติ (ต่อ)

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

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

การเปรียบเทียบกับผลการพยากรณ์ป้ายกำกับจากตัวแบบด้วยการหาค่าการสูญเสียครอสเอนโทรปี เพื่อวัดค่าความน่าจะเป็นของป้ายกำกับ ว่ามีผลการพยากรณ์ใกล้กับค่าจริงที่ต้องการมากน้อยแค่ไหน (Error measurement) ด้วยฟังก์ชันการสูญเสียดังสมการที่ 3.4

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

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

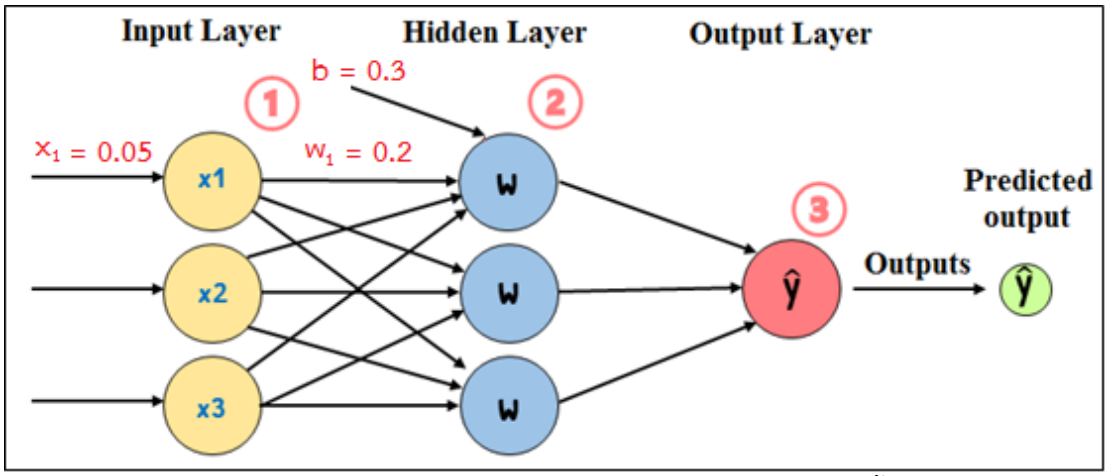
0.1262, 0.2418, 0.2166 และ 0.1687 ตามลำดับ โดยแทนค่าใน q[0.2467, 0.1262, 0.2418, 0.2166, 0.1687] เมื่อแทนค่าในสมการที่ 3.4 มีรายละเอียดดังนี้

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

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

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

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



รูปที่ 3.4 โครงข่ายประสาทเทียมแบบป้อนไปข้างหน้าที่ถูกกำหนดอินพุต ค่าน้ำหนัก และไบแอส

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

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

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

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

เมื่อหาค่าในแต่ละ Term ดังสมการที่ 3.7 และ 3.8

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

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

$$= -1.5066$$

เช่นเดียวกับ

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

$$= 0.2467(1-0.2467)$$

$$= 0.1858$$

เมื่อคำนวณหาค่าอินพุตที่ X_1 ดังสมการที่ 3.9

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

$$= 0.05$$

ดังนั้น

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

$$= -0.014$$

เมื่อกำหนดอัตราการเรียนรู้ (Learning rate) เท่ากับ 0.5 จะสามารถปรับค่าน้ำหนักที่ W_1 ได้ดังนี้

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

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

$$= 0.2070$$

การประเมินความหมายของรูปภาพบนชุดคำถามแบบเลือกคำตอบที่ถูกต้องเพียงข้อเดียวโดยให้ผู้เชี่ยวชาญ 10 ท่านประเมินความหมายรูปภาพแบบเป็นอิสระต่อกัน แล้วเก็บคำตอบไว้ในชุดคำตอบสำหรับเปรียบเทียบกับผลการพยากรณ์ความน่าจะเป็นของป้ายกำกับจากตัวแบบด้วยการคำนวณหาค่าการสูญเสีย เพื่อนำไปปรับปรุงพารามิเตอร์ตามแนวคิดการแผ่กระจายย้อนหลัง

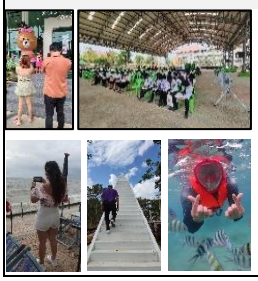
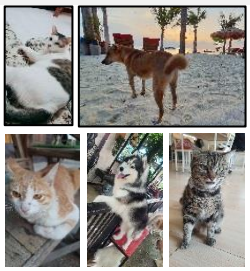
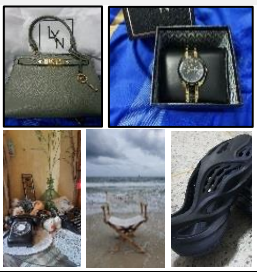
อย่างไรก็ดี การฝึกฝนโมเดลทั่วไปมักอาศัยการเรียนรู้จากข้อมูลจำนวนมากซ้ำหลายรอบ (epoch) ซึ่งหมายถึงหนึ่งรอบเต็มของการนำข้อมูลฝึกทั้งหมดเข้าสู่ระบบโมเดล อย่างไรก็ตาม ในบางกรณี โดยเฉพาะในการพัฒนาโมเดลต้นแบบ อาจมีความจำเป็นต้องกำหนดให้ปรับค่าน้ำหนักเพียงหนึ่งรอบ ซึ่งเป็นแนวทางที่ตั้งใจเพื่อให้ได้ผลลัพธ์ภายใต้เงื่อนไขเฉพาะเจาะจง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

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

นอกจากนี้ การใช้โมเดลเพื่อการสกัดคุณลักษณะเชิงความหมาย การปรับค่าน้ำหนักเพียงรอบเดียว จะช่วยให้ผลลัพธ์สะท้อนคุณสมบัติโครงสร้างเดิมได้ชัดเจนมากกว่า จึงเป็นที่ของงานวิจัยนี้ที่กำหนดรอบการเรียนรู้เพียงรอบเดียว แต่จะดำเนินการซ้ำจนครบตามจำนวนโหนดของตัวแบบ CLIP ที่ถูกฝึกฝนล่วงหน้า

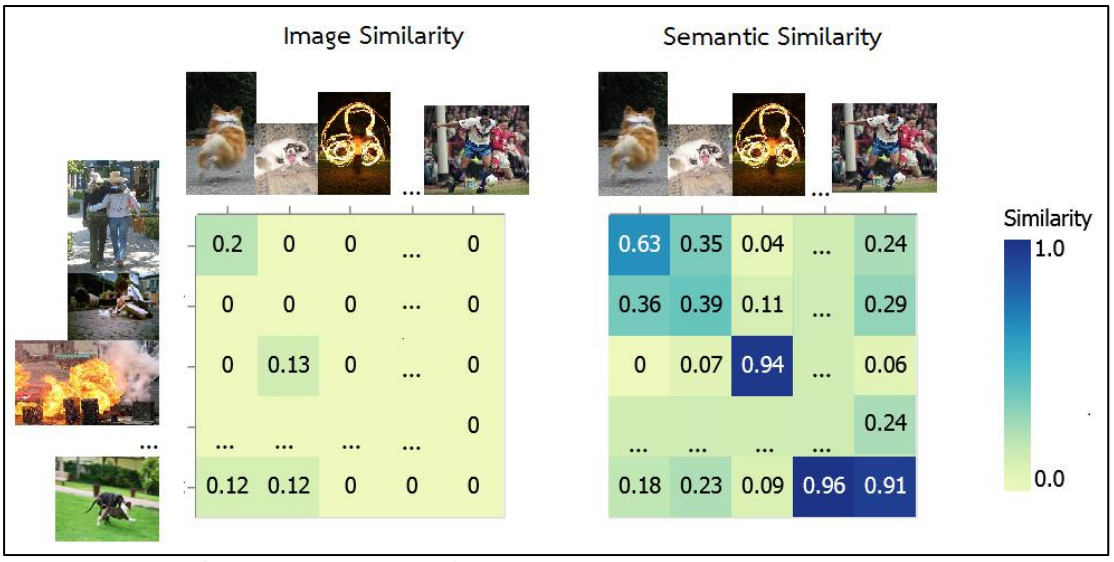
3) การเรียนรู้ซ้ำบนชุดข้อมูลที่กำหนดเอง (Retrain network on custom dataset) เนื่องจากการฝึกฝนแบบจำลองตั้งแต่เริ่มต้น (Training from scratch) จำเป็นต้องใช้ปริมาณข้อมูลจำนวนมากเพื่อให้ได้แบบจำลองที่มีความสามารถในการใช้งานทั่วไป (Generalization) นอกจากนั้น ยังมีข้อจำกัดด้านเวลาและทรัพยากรในการประมวลผล จึงทำให้เทคนิคการถ่ายโอนการเรียนรู้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

ตารางที่ 3.5 ตัวอย่างรูปภาพจากชุดข้อมูลที่ผู้วิจัยรวบรวมเอง

ภาพบุคคล	ภาพสัตว์เลี้ยง	ภาพสิ่งของ	ภาพทิวทัศน์และสิ่งปลูกสร้าง
			

งานวิจัยนี้ให้ความสำคัญกับประเด็นลิขสิทธิ์ของรูปภาพจึงใช้ชุดข้อมูลรูปภาพที่ผู้วิจัยรวบรวมเองมากกว่าร้อยละ 75 ในการเรียนรู้ของตัวแบบ ตลอดจนคำนึงถึงสิทธิส่วนบุคคลที่ปรากฏในภาพ จึงหลีกเลี่ยงการใช้รูปภาพที่ปรากฏใบหน้าของบุคคลหรือส่วนหนึ่งส่วนใดอย่างชัดเจนที่สามารถนำไปสู่การระบุอัตลักษณ์บุคคลได้เป็นเงื่อนไขในการเลือกรูปภาพเพื่อเป็นชุดข้อมูลที่กำหนดเอง

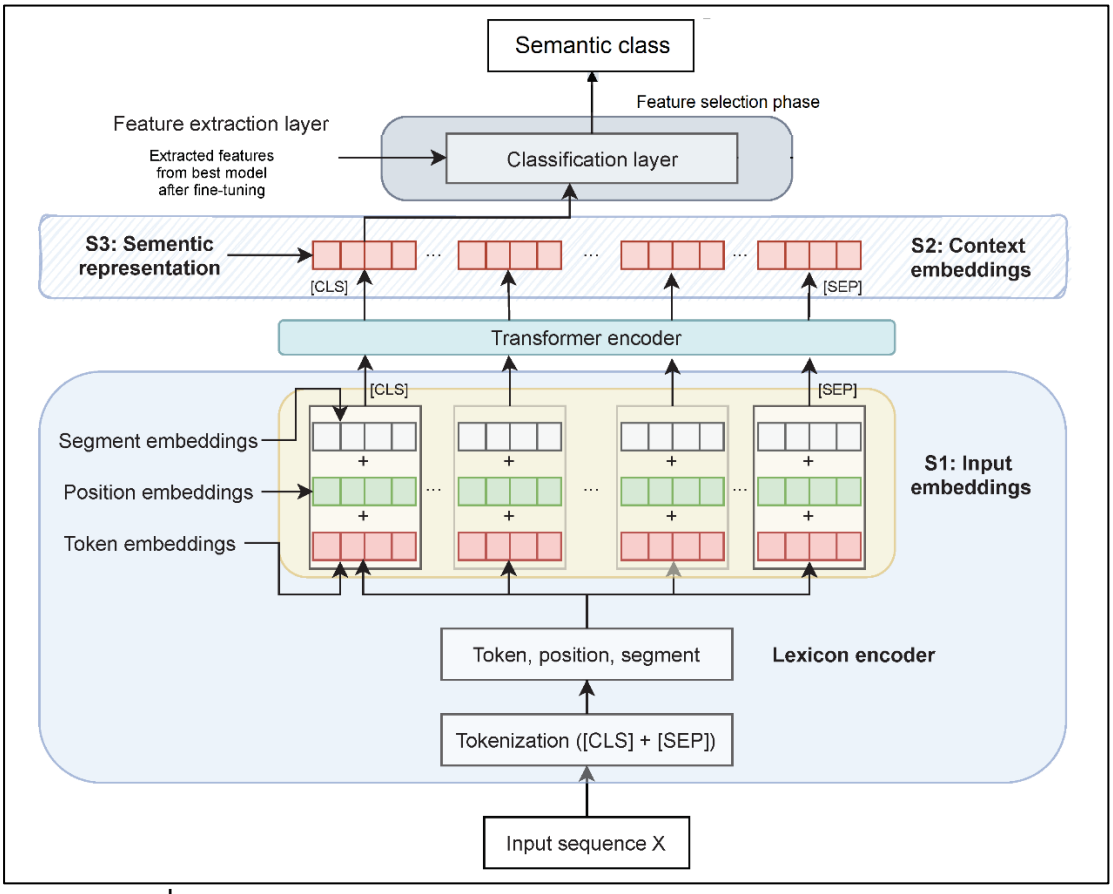
4) การเรียนรู้ความคล้ายคลึงเชิงความหมายด้วยกระบวนการเรียนรู้ด้วยตนเอง โดยบังคับให้แบบจำลองเรียนรู้การแสดงความหมายเกี่ยวกับข้อมูล เปรียบเทียบกับแนวคิดระดับสูงในรูปแบบนามธรรม เพื่ออธิบายว่าแต่ละรูปภาพซ้อนทับกันในเชิงความหมายมากน้อยเพียงใด มีค่าตั้งแต่ 0 ถึง 1 ดังรูปที่ 3.5 ตัวอย่างภาพ “ผู้หญิงสองคนกำลังเดินไปข้างหน้า”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



รูปที่ 3.5 ตัวอย่างรูปภาพที่ผ่านการเรียนรู้ความคล้ายคลึงเชิงความหมาย

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

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



รูปที่ 3.6 กระบวนการประมวลผลข้อความค้นหาด้วยตัวแบบภาษา DistilBERT

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

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

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

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

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

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

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

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

u คือ เวกเตอร์คุณลักษณะข้อความค้นหา

m คือ เวกเตอร์คุณลักษณะเชิงความหมายของรูปภาพ

a_J คือ เกณฑ์ที่ j ของคุณลักษณะเชิงความหมายของรูปภาพ และคุณลักษณะที่ j ของคุณลักษณะข้อความค้นหา

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

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

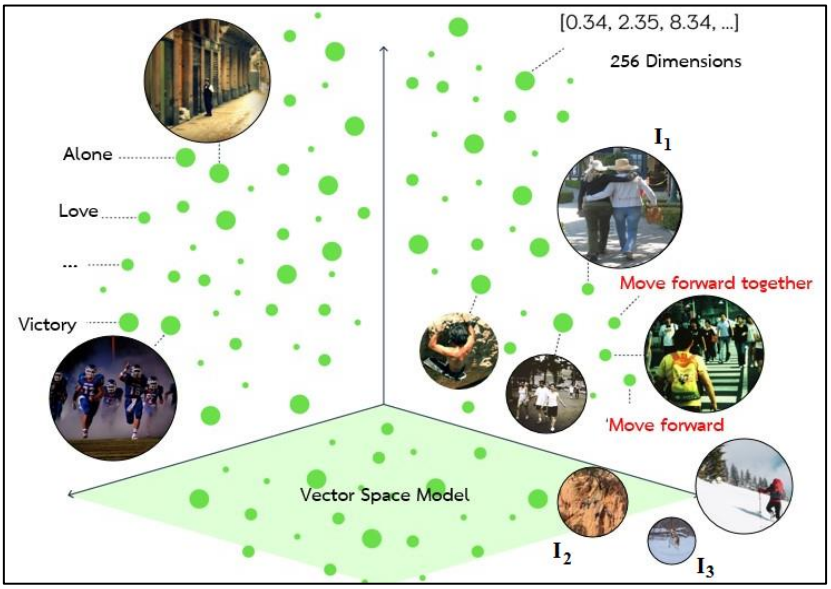
$$= 10.00$$

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

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

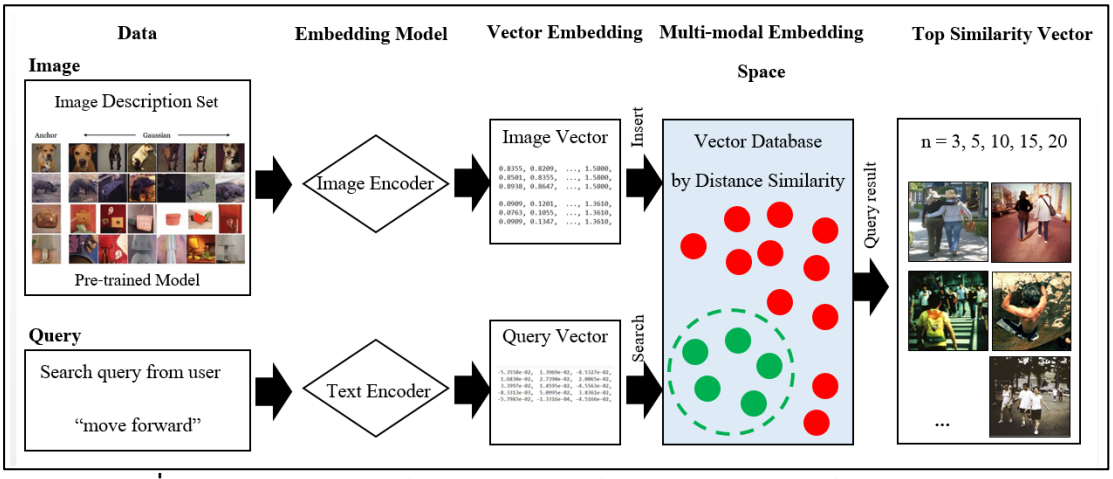
$$= 0.0909$$

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



รูปที่ 3.7 การพิจารณาระยะทางระหว่างเวกเตอร์ตัวแทนรูปภาพและเวกเตอร์ตัวแทนข้อความค้นหา

2) การเรียงลำดับค่าความคล้ายคลึง (Similarity sorting) จากการเปรียบเทียบระหว่างเวกเตอร์คุณลักษณะข้อความค้นหา เรียงลำดับค่าความคล้ายคลึงจากมากไปน้อย



รูปที่ 3.8 การจับคู่เวกเตอร์ ระหว่างเวกเตอร์รูปภาพ และเวกเตอร์ข้อความค้นหา

3.1.3 การเขียนโปรแกรม

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

3.1.4 การทดสอบ

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

ดังนั้นเพื่อเป็นการสะท้อนถึงความพร้อมของแบบจำลองต่อการใช้งานในวงกว้าง โดยเฉพาะในบริบทที่ผู้ใช้งานแต่ละคนมีเงื่อนไขที่แตกต่างกัน จึงเป็นที่มาของการให้ผู้เชี่ยวชาญเข้ามามีบทบาทในการประเมินประสิทธิภาพ เพื่อเพิ่มความน่าเชื่อถือและลดข้อผิดพลาดที่อาจเกิดขึ้นจากกระบวนการอัตโนมัติหรือข้อจำกัดของแบบจำลองในการจำแนกข้อมูลเชิงความหมายในสถานการณ์จริงได้อย่างมีประสิทธิภาพและสอดคล้องกับความต้องการของผู้ใช้งาน นอกจากนั้นการประเมินจากผู้เชี่ยวชาญหลายคนยังช่วยลดความคลาดเคลื่อนของการประเมินที่อาจเกิดขึ้นจากการพิจารณาโดยผู้เชี่ยวชาญเพียงท่านเดียว ทำให้ผลการประเมินมีความน่าเชื่อถือมากขึ้น งานวิจัยนี้สิ่งที่ผู้วิจัยจะต้องคำนึงถึงในการเลือกกลุ่มผู้เชี่ยวชาญ ตามเกณฑ์ในการคัดเลือกผู้เชี่ยวชาญ (Experts)

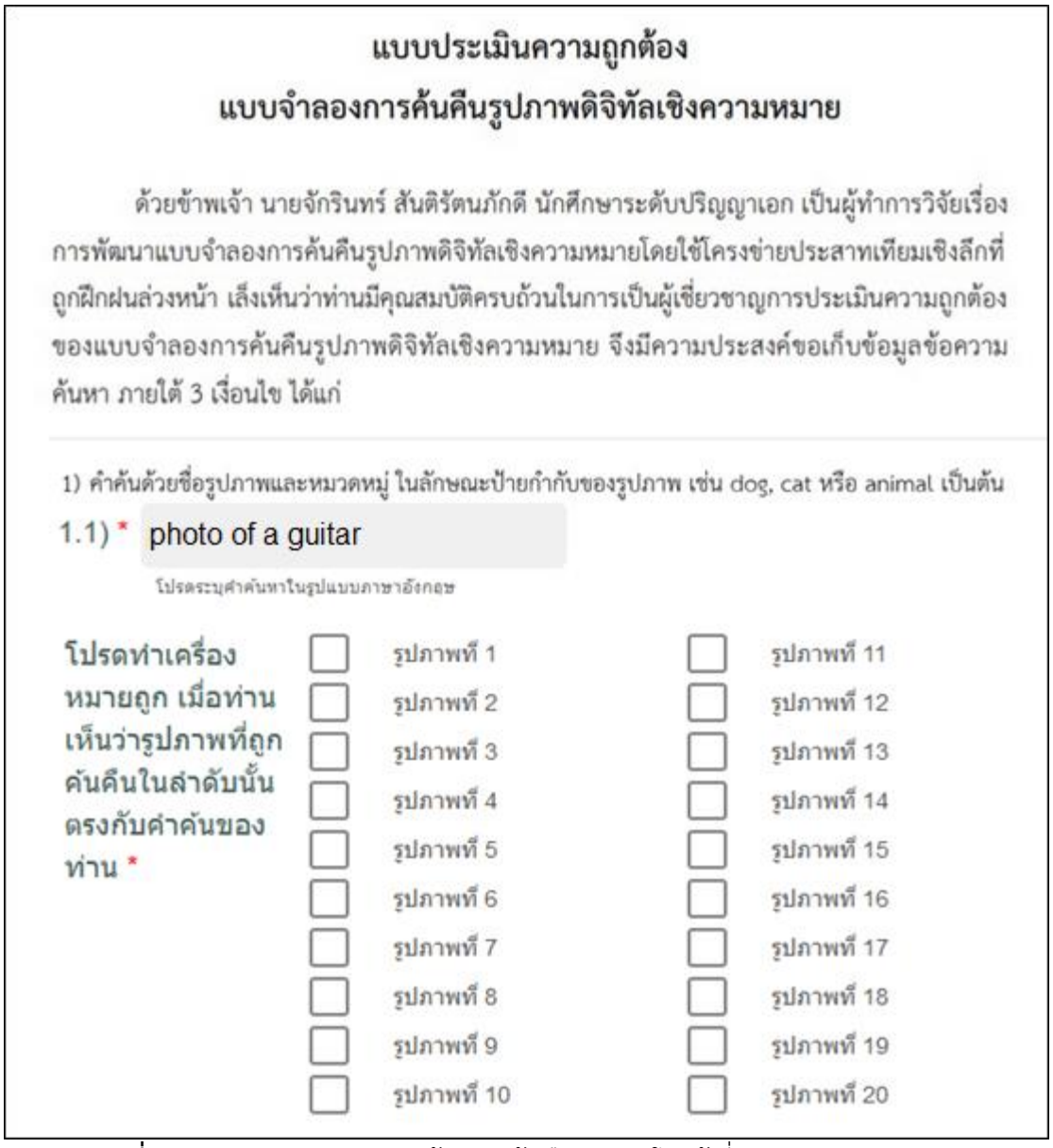
หรือผู้มีประสบการณ์ (Specialist) หรือผู้ทรงคุณวุฒิ (อรนุช ศรีสะอาด, 2538) กำหนดให้ผู้เชี่ยวชาญต้องมีคุณสมบัติดังเงื่อนไขต่อไปนี้ ได้แก่ 1) จบการศึกษาระดับปริญญาโทขึ้นไป หรือมีความรู้ความเชี่ยวชาญเฉพาะด้านอันเป็นที่ประจักษ์ 2) มีประสบการณ์การใช้งานคอมพิวเตอร์และอินเทอร์เน็ตอย่างน้อย 10 ปีขึ้นไป 3) มีประสบการณ์การใช้งานคอมพิวเตอร์และสมาร์ทโฟนอย่างน้อย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

3.1.5 การนำไปใช้งาน

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

3.2 เครื่องมือที่ใช้ในการวิจัย

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



รูปที่ 3.9 แบบประเมินความถูกต้องการค้นคืนรูปภาพโดยผู้เชี่ยวชาญผ่าน Jotform

3.3 การเก็บรวบรวมข้อมูล

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

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

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

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

ตารางที่ 3.6 ตัวอย่างข้อความค้นหาภายใต้ 3 เงื่อนไข ที่ถูกสร้างจาก ChatGPT

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

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

ตารางที่ 3.7 คำศัพท์จากการสุ่มและตัวอย่างข้อความค้นหาที่สร้างจากคำศัพท์ด้วย ChatGPT

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

3.4 การวิเคราะห์ข้อมูล

การวิเคราะห์ผลการค้นคืนรูปภาพแบบไบนารีด้วยโปรแกรมสำเร็จรูปทางสถิติ เพื่อประเมินประสิทธิภาพแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายในทุกมิติ ประกอบด้วย

1) มาตรวัดแบบไม่สนใจลำดับของผลลัพธ์ เป็นมาตรวัดที่แสดงถึงรูปภาพที่ถูกค้นคืนนั้นถูกต้องหรือไม่ โดยไม่คำนึงถึงตำแหน่งที่รูปภาพนั้นปรากฏในลำดับ ประกอบด้วย

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

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

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

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

การตีความผลการวิเคราะห์ข้อมูลในรูปของร้อยละให้ง่ายต่อการทำความเข้าใจ งานวิจัยนี้ใช้เกณฑ์การแปลผลตามระดับคะแนน 5 ระดับตามแนวทางของณัฏฐ์ธนัน ทองดี (2553) โดยพิจารณาความเหมาะสมในการวิเคราะห์ข้อมูลทางสถิติในงานวิจัยเชิงปริมาณ ดังตารางที่ 3.8

ตารางที่ 3.8 การตีความผลการวิเคราะห์ข้อมูลในรูปของร้อยละ ตามระดับคะแนน 5 ระดับ

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