การพัฒนาแบบจำลองการค้นคืนรูปภาพเชิงความหมาย โดยใช้การฝึกฝนตัวแบบล่วงหน้าแบบคอนทราสต์ระหว่างรูปภาพและข้อความ

Development of a Semantic-based Image Retrieval Model Using Contrastive Pre-trained between Image and Text

จักรินทร์ สันติรัตนภักดี[1,2*] และ ศุภกฤษฎิ์ นิวัฒนากูล[1]

[1]สำนักวิชาศาสตร์และศิลป์ดิจิทัล มหาวิทยาลัยเทคโนโลยีสุรนารี นครราชสีมา 30000

[2]สาขาวิชาเทคโนโลยีธุรกิจดิจิทัล คณะบริหารธุรกิจ มหาวิทยาลัยวงษ์ชวลิตกุล นครราชสีมา 30000

Chakkarin Santirattanaphakdi[1,2*] and Suphakit Niwattanakul[1]

[1]Institute of Digital Arts and Science, Suranaree University of Technology, Nakhon Ratchasima, 30000

[2]Department of Digital Business Technology, Faculty of Business Administrator, Vongchavalitkul University, Nakhon Ratchasima, 30000

[*]**Corresponding author:** chakkarin_san@vu.ac.th

Received: 31 May 2023/ Revised: 21 September 2023/ Accepted: 27 September 2023

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาและประเมินประสิทธิภาพแบบจำลองการค้นคืนรูปภาพเชิงความหมาย โดยใช้การฝึกฝนล่วงหน้าแบบคอนทราสต์ ประกอบด้วย 3 โมดูล ได้แก่ 1) โมดูลการสร้างชุดคำอธิบายรูปภาพ เพื่อฝึกฝนตัวเข้ารหัสรูปภาพ และตัวเข้ารหัสข้อความตามความคล้ายคลึงโคไซน์บนพื้นที่การฝังหลายรูปแบบ ก่อนจะแจกแจงความน่าจะเป็นของค่าเอาต์พุตด้วยฟังก์ชันซอฟต์แม็กซ์ จากนั้นจะคำนวณค่าการสูญเสียเพื่อเปรียบเทียบระหว่างผลประเมินความหมายของรูปภาพจากผู้เชี่ยวชาญกับผลการพยากรณ์ความน่าจะเป็นของป้ายกำกับจากตัวแบบ เพื่อปรับพารามิเตอร์ในการเรียนรู้ความหมายของรูปภาพตามแนวคิดการแผ่กระจายย้อนหลังเพื่อเรียนรู้ซ้ำอีกครั้งด้วยการเรียนรู้ด้วยตนเอง และถ่ายโอนการเรียนรู้เพื่อติดป้ายกำกับข้อมูลให้กับรูปภาพด้วยแนวคิดระดับสูงในรูปแบบนามธรรมของรูปภาพที่ได้จากการเรียนรู้ความคล้ายคลึงเชิงความหมาย ก่อนจะสร้างเป็นเวกเตอร์คุณลักษณะรูปภาพ 2) โมดูลการประมวลผลข้อความค้นหาจากผู้ใช้ในรูปแบบภาษาธรรมชาติสำหรับเข้ารหัสข้อความ เพื่อสร้างเวกเตอร์คุณลักษณะข้อความค้นหา 3) โมดูลการจับคู่เวกเตอร์คุณลักษณะรูปภาพและเวกเตอร์คุณลักษณะข้อความค้นหาจากค่าความคล้ายคลึงของเวกเตอร์ ก่อนจะเรียงลำดับตามความเกี่ยวข้อง และแสดงเป็นผลลัพธ์การค้นคืนรูปภาพแก่ผู้ใช้ ผลประเมินประสิทธิภาพการค้นคืนรูปภาพเชิงความหมายพบว่า 1) ค่าเฉลี่ยส่วนกลับของลำดับบนชุดข้อมูล Flickr30k และชุดข้อมูลที่ผู้วิจัยรวบรวมเองมีค่าเท่ากับ 0.628 และ 0.617 ตามลำดับที่ตำแหน่ง k = 5 และ 2) ค่าความครบถ้วนที่ k ลำดับ จำนวน 1, 3 และ 5 รูปภาพบนชุดข้อมูล Flickr30k มีค่าความครบถ้วนเฉลี่ย 0.585, 0.664 และ 0.761 เมื่อเปรียบเทียบกับชุดข้อมูลที่ผู้วิจัยรวบรวมเองพบว่าค่าความครบถ้วนที่ k ลำดับลดลงเล็กน้อย

แต่เป็นไปในทิศทางเดียวกัน ผลลัพธ์จากงานวิจัยนี้จะช่วยลดปัญหาช่องว่างความหมาย และช่วยสนับสนุนผู้ใช้ด้วยคำค้นหาในรูปแบบภาษาธรรมชาติที่ยึดโยงกับความหมายของรูปภาพแทนที่จะยึดตามหลักไวยากรณ์ของภาษา

คำสำคัญ: การค้นคืนรูปภาพ การฝึกฝนล่วงหน้า การเรียนรู้แบบคอนทราสต์ การเรียนรู้เชิงลึก

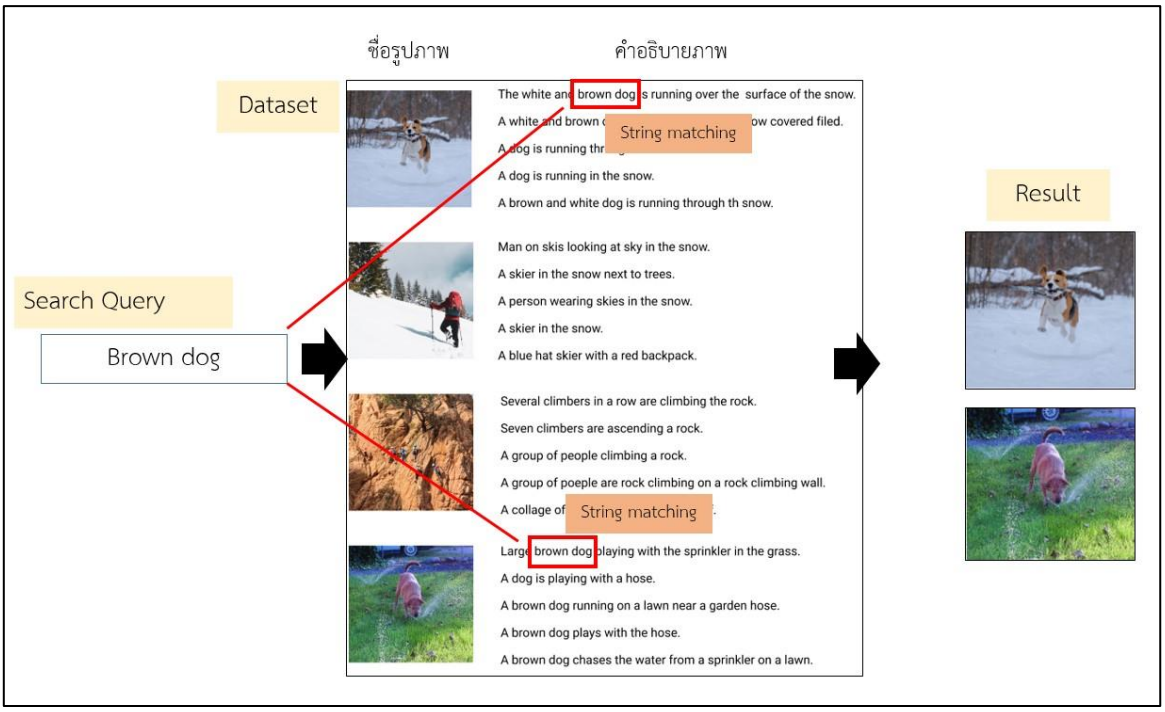
Abstract

The objective of this research is to develop and evaluate the effectiveness of a semantic image retrieval model using contrastive pre-trained. The approach involves 3 modules: 1) the image description set generation module, which trains to encode both image content and textual content in various embedding before estimating the output probability distribution using the softmax function. It then calculates the loss to compare the meaning evaluation of images by experts with the predicted likelihoods of labels from the model. This step involves adjusting parameters for image meaning learning using the concept of retroactive distribution learning and self-paced learning, transferring the learned knowledge to label data for images based on the high-level abstract concept of images obtained from meaning-based similarity learning. This is followed by creating feature vectors for image characteristics, 2) the natural language processing module, which encodes user's natural language queries to generate feature vectors for query characteristics. And 3) the feature matching module, which matches image feature vectors and query feature vectors based on vector similarity values. Then, it ranks the results according to relevance and presents the image retrieval results to the user. The evaluation of semantic image retrieval performance reveals that: The mean reciprocal rank (MRR) values for the top k retrievals on the Flickr30k dataset and the self-collected dataset are 0.628 and 0.617, respectively, at k = 5, and the precision at k $(P@k)$ values for k = 1, 3, and 5 on the Flickr30k dataset are 0.585, 0.664, and 0.761, respectively, when compared to the self-collected dataset. While the precision at k values slightly decrease, the results show a consistent trend. The outcomes of this research will aid in addressing the semantic gap problem and support users in their natural language queries, which are linked to the image semantics rather than following the grammatical rules of the language.

Keyword: Image retrieval, Pre-trained, Contrastive learning, Deep learning

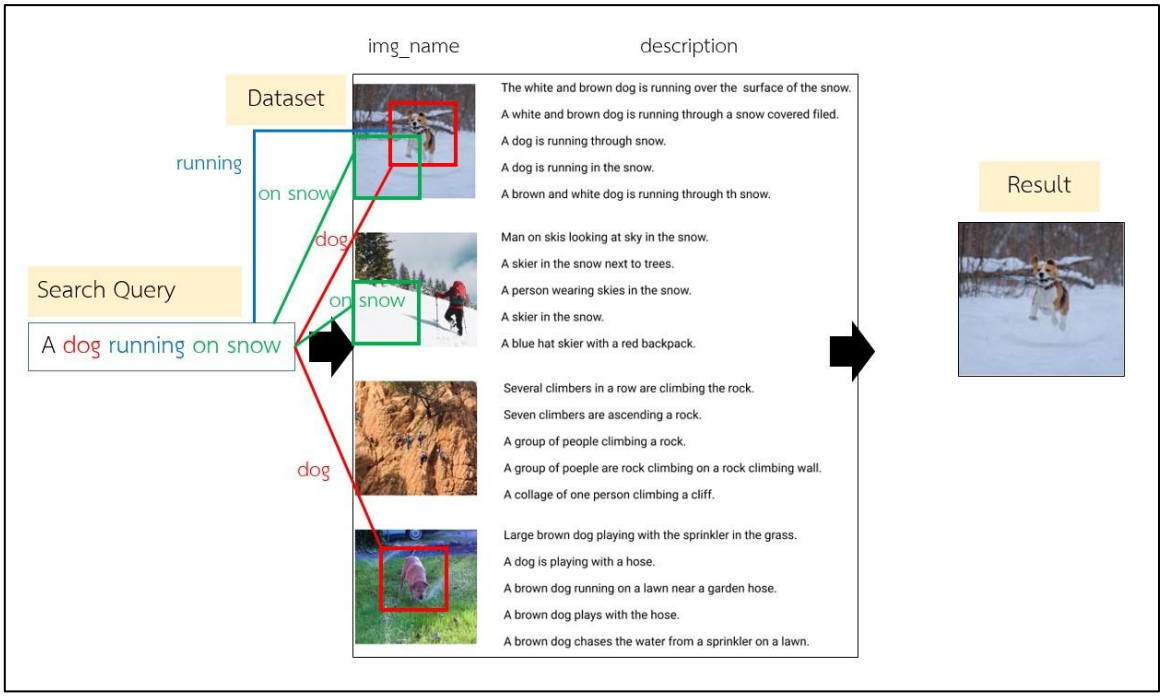
บทนำ

พัฒนาการของอุปกรณ์ถ่ายภาพในรูปแบบของสมาร์ทโฟน ก่อให้เกิดรูปภาพจำนวนมหาศาลบนอินเทอร์เน็ต ตามข้อมูลของเว็บไซต์ Phototutorial [1] รายงานว่าจะมีรูปภาพถึง 1.81 ล้านล้านภาพในปี ค.ศ. 2023 และจะเพิ่มขึ้นมากกว่า 2 ล้านล้านภาพในปี ค.ศ. 2025 โดยในปี ค.ศ. 2030 คาดการณ์ว่าจะมีรูปภาพทั่วโลกถึง 2.33 ล้านภาพ และมีแนวโน้มที่จะเพิ่มมากขึ้นเรื่อย ๆ จากพฤติกรรมของผู้คนที่นิยมใช้สื่อสังคมออนไลน์ในการแบ่งปันเรื่องราวและบันทึกความทรงจำ เกิดเป็นระบบค้นคืนรูปภาพตามความต้องการของผู้ใช้ โดยหลักแล้วมี 2 รูปแบบ คือ 1) การค้นคืนรูปภาพด้วยข้อความ (Text-Based Image Retrieval: TBIR) [2] ดังภาพที่ 1 การค้นคืนรูปภาพด้วยข้อความ “brown dog” เมื่อเปรียบเทียบคำค้นกับชื่อไฟล์รูปภาพหรือคำอธิบายภาพที่อยู่ในชุดข้อมูล เมื่อพบข้อความที่ตรงกันรูปภาพนั้นก็จะถูกดึงมาเป็นผลลัพธ์



ภาพที่ 1 การค้นคืนรูปภาพด้วยข้อความ

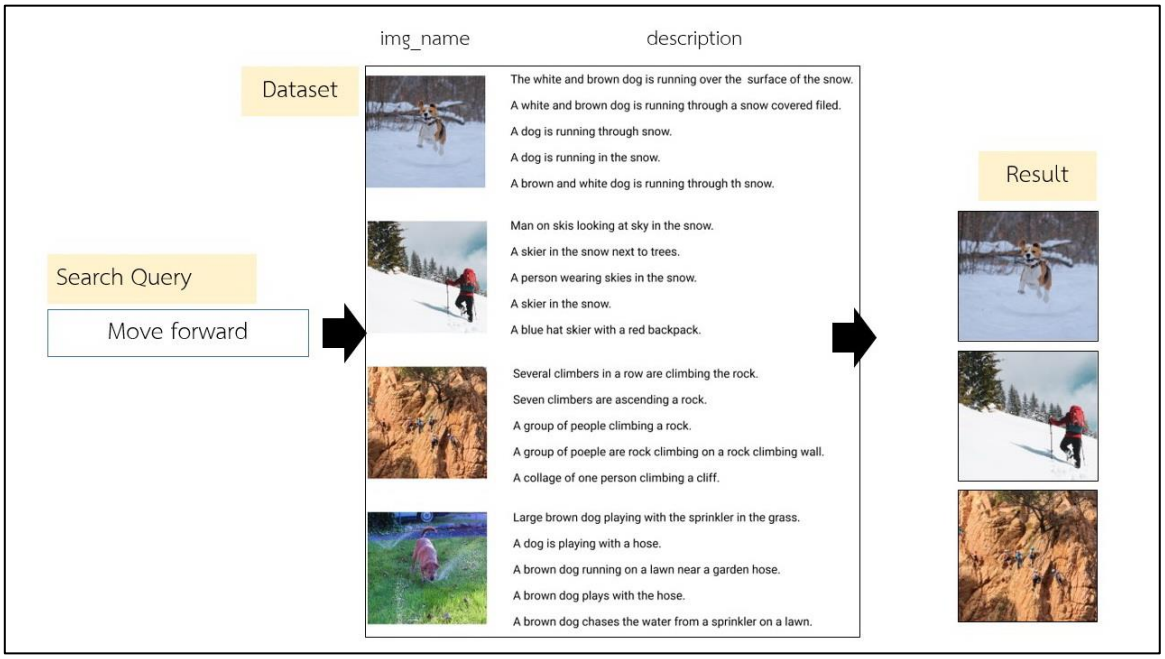
แม้การค้นคืนรูปภาพด้วยข้อความยังเป็นวิธีที่นิยมใช้อยู่ในปัจจุบัน แต่มักพบปัญหาการตั้งชื่อไฟล์ไม่ตรงกับเนื้อหาของรูปภาพ อีกทั้งคำอธิบายภาพที่ใส่โดยมนุษย์อาจไม่สามารถอธิบายได้ครอบคลุมเนื้อหาของรูปภาพได้ทั้งหมด ส่งผลให้รูปภาพจากการค้นคืนมีประสิทธิภาพต่ำ และ 2) การค้นคืนรูปภาพเชิงเนื้อหา (Content-based Image Retrieval: CBIR) [2] จากการเปรียบเทียบระหว่างคำค้นหากับคุณลักษณะระดับต่ำของรูปภาพ (Low-level Features) [3] ซึ่งเป็นข้อมูลที่ไม่มีความหมายในตัวเอง เช่น สี รูปทรง พื้นผิว รวมถึงตำแหน่งในภาพ [4] ร่วมกับคุณลักษณะระดับสูง (High-level Features) เช่น แนวคิด (Concept) วัตถุ (Object) เหตุการณ์ (Event) เพื่อค้นคืนรูปภาพที่มีความคล้ายคลึงกับคำค้นหามากที่สุดมาแสดงเป็นผลลัพธ์ ดังภาพที่ 2 การค้นคืนรูปภาพเชิงเนื้อหาด้วยข้อความ “a dog running on snow”



ภาพที่ 2 การค้นคืนรูปภาพเชิงเนื้อหา

จากภาพที่ 2 การค้นคืนรูปภาพเชิงเนื้อหาจะทำการเปรียบเทียบความคล้ายคลึงระหว่างคุณลักษณะของรูปภาพจากข้อความค้นหาและคุณลักษณะของรูปภาพในชุดข้อมูลบนพื้นที่เวกเตอร์หลายมิติ (Multi-dimensional Vector) เช่น มิติที่หนึ่งคือสี มิติที่สองคือรูปทรง มิติที่สามคือพื้นผิว เป็นต้น คุณลักษณะของรูปภาพใดที่มีค่าความคล้ายคลึงกับคุณลักษณะของรูปภาพจากข้อความค้นหาเกินค่าแบ่ง (Threshold) จะถูกดึงออกมาเป็นผลลัพธ์

อย่างไรก็ตาม “A picture is worth a thousand words” แสดงถึงเนื้อหารูปภาพมีความหมายหลากหลายมากกว่าเมื่อเทียบกับข้อความ ดังนั้นการค้นคืนรูปภาพด้วยคุณลักษณะเชิงประจักษ์เพียงอย่างเดียว อาจส่งผลให้เกิดปัญหาช่องว่างความหมาย (Semantic-gap) เนื่องจากคุณลักษณะเหล่านี้ไม่สื่อถึงความหมายที่อยู่ในรูปภาพได้อย่างถูกต้อง และพฤติกรรมผู้ใช้ที่มักจะใช้คำค้นหาที่เป็นตัวอักษรหรือข้อความมากกว่าการระบุเพียงคุณลักษณะต่าง ๆ ของรูปภาพ จึงเกิดเป็นแนวคิดการค้นคืนรูปภาพเชิงความหมาย (Semantic-based Image Retrieval: SBIR) [5] ที่มุ่งเชื่อมโยงคุณลักษณะของรูปภาพไปสู่ความหมายที่แท้จริงของรูปภาพ ด้วยการแปลงคุณลักษณะให้อยู่ในรูปแบบของความหมายรูปภาพในระดับสูง (High-level Semantics) ดังภาพที่ 3



ภาพที่ 3 การค้นคืนรูปภาพเชิงความหมาย

จากภาพที่ 3 ตัวอย่างคำค้นหาที่อาจทำให้เกิดปัญหาช่องว่างความหมาย เช่น “move forward” เมื่อใช้หลักการค้นคืนรูปภาพด้วยข้อความ หรือการค้นคืนรูปภาพเชิงเนื้อหากับรูปภาพทั้งหมดแล้ว อาจไม่พบผลลัพธ์ที่เกี่ยวข้องกับคำค้นหาเลย แต่เมื่อนำมาเปรียบเทียบกับคุณลักษณะเชิงความหมาย จะพบว่ามีผลลัพธ์จำนวน 3 ภาพ ได้แก่ รูปภาพสุนัขที่กำลังวิ่งไปข้างหน้า รูปภาพนักสกีที่กำลังมุ่งหน้าไปบนยอดเขาหิมะ หรือรูปภาพนักไต่เขาที่กำลังปีนมุ่งหน้าไปยังยอดเขาที่ทั้ง 3 รูปภาพไม่ปรากฏ คำว่า “move forward” ภายในรูปภาพหรือคำบรรยายภาพเลย แต่อาจมีความหมายแฝงว่า “move forward” ที่แปลว่ามุ่งไปข้างหน้า อันเป็นคุณลักษณะที่เกี่ยวข้องกับข้อความค้นหาจากผู้ใช้

ปัจจุบันมีการเสนอแนวคิดเกี่ยวกับการค้นคืนรูปภาพเชิงความหมายมากมาย โดยการเรียนรู้ของเครื่อง (Machine Learning) เป็นหนึ่งในแนวคิดที่ถูกนำมาใช้ในการจำแนกประเภทรูปภาพ (Image Classification) [6] เพื่อสร้างเป็นตัวแปลความหมายระดับสูง (High-level Interpretation) อย่างไรก็ดี อัลกอริทึมการเรียนรู้ของเครื่องทั่วไปจะเหมาะกับข้อมูลแบบมีโครงสร้าง (Structured Data) ในขณะที่รูปภาพ ข้อความ และเสียงนั้นมักจะอยู่ในรูปแบบข้อมูลที่ไม่มีโครงสร้าง (Unstructured Data) เป็นที่มาของการใช้โครงข่ายประสาทเทียม (Neural Network) ตามแนวคิดการแบ่งลำดับชั้น

ในการเรียนรู้ ประกอบด้วย 4 ส่วน ได้แก่ 1) ชั้นอินพุต (Input Layer) คือข้อมูลป้อนเข้าโครงข่าย 2) ชั้นซ่อนเร้น (Hidden Layer) คือชั้นประมวลผลที่ซ่อนอยู่สามารถมีได้หลายชั้น แต่ละชั้นจะมีนิวรอล (Neuron) มีหน้าที่รับข้อมูลมาประมวลผลร่วมกับค่าน้ำหนักของอินพุตแต่ละตัวด้วยฟังก์ชันกระตุ้น (Activation Function) เช่น ฟังก์ชันซอฟต์แม็กซ์ (Softmax Function) ฟังก์ชันซิกมอยด์ (Sigmoid Function) 3) ชั้นเอาต์พุต (Output Layer) คือชั้นที่ประมวลผลอินพุตทั้งหมดจากชั้นก่อนหน้ามาคำนวณร่วมกับค่าน้ำหนัก (Weight) และไบอัส (Bias) ของอินพุตแต่ละตัว และ 4) การพยากรณ์ นำเอาผลลัพธ์จากชั้นเอาต์พุตมาสร้างฟังก์ชันตัดสินใจ (Decision Function) โดยมีผลลัพธ์เป็นค่าพยากรณ์ โดยทั้ง 4 ส่วนนี้ทำงานต่อเนื่องกันแบบแผ่กระจายเดินหน้า (Forward Propagation) อย่างไรก็ดีตัวแบบที่ถูกฝึกฝนนั้นยังขาดค่าน้ำหนัก และไบอัสที่ถูกต้องส่งผลให้ค่าที่พยากรณ์ได้อาจจะไม่ตรงกับความจริง เกิดเป็นแนวคิดการแผ่กระจายย้อนหลัง (Backward Propagation) จากการนำค่าพยากรณ์นั้นมาใส่ในฟังก์ชันการสูญเสีย (Loss Function) เพื่อหาความแตกต่างระหว่างค่าพยากรณ์กับค่าจริงด้วยค่าครอสเอนโทรปี (Cross-entropy) ก่อนนำไปปรับแต่งพารามิเตอร์ (Parameter) ให้ตัวแบบสามารถพยากรณ์ผลลัพธ์ได้ถูกต้องยิ่งขึ้น

การเรียนรู้เชิงลึก (Deep Learning) [7] ถูกพัฒนาต่อยอดมาจากโครงข่ายประสาทเทียม ด้วยการวางโครงข่ายซ้อนกันหลาย ๆ ชั้น เพื่อเพิ่มประสิทธิภาพการวิเคราะห์ ปัจจุบันได้รับการยอมรับว่ามีประสิทธิภาพในการวิเคราะห์และพยากรณ์ผลลัพธ์ได้ดีกว่าโครงข่ายประสาทเทียมแบบเดิมเป็นอย่างมาก

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาและประเมินประสิทธิภาพแบบจำลองการค้นคืนรูปภาพเชิงความหมาย โดยใช้ตัวแบบการเรียนรู้เชิงลึกที่ถูกฝึกฝนล่วงหน้าแบบคอนทราสต์ (Contrastive Pre-training) จากการเรียนรู้ความสัมพันธ์ระหว่างรูปภาพกับข้อความภาษาธรรมชาติ แทนที่จะเรียนรู้ภายใต้คลาสที่จำกัดอย่างสุนัขหรือแมวเช่นเดิม เมื่อตัวแบบได้รับการฝึกฝนจากข้อความทั้งประโยคจะส่งผลให้ตัวแบบสามารถเรียนรู้สิ่งต่าง ๆ ได้มากขึ้น และสามารถค้นหารูปแบบความสัมพันธ์บางอย่างระหว่างรูปภาพกับข้อความด้วยการเรียนรู้ด้วยตนเอง อันจะเป็นแนวทางการค้นคืนสารสนเทศในอนาคตภายใต้คำค้นหาในรูปแบบภาษาธรรมชาติที่ยึดโยงกับความหมายของรูปภาพแทนที่จะยึดตามหลักไวยากรณ์ของภาษา

วิธีดำเนินการวิจัย

งานวิจัยนี้ประยุกต์ใช้การพัฒนาแอปพลิเคชันแบบเร่งรัด (Rapid Application Development model: RAD) [8] มาเป็นกรอบในการพัฒนาแบบจำลองการค้นคืนรูปภาพเชิงความหมาย ก่อนจะพัฒนาเป็นระบบที่สมบูรณ์ มีรายละเอียดดังนี้

1) การวางแผนความต้องการ (Requirement Planning) ปัจจุบันมีแนวคิดเกี่ยวกับการสร้างตัวแปลความหมายระดับสูง โดยประยุกต์หลายศาสตร์มาบูรณาการร่วมกัน สามารถจำแนกตามเทคนิคออกเป็น 6 กลุ่ม ได้แก่ 1) เทคนิคการตอบกลับของผู้ใช้ (Relevance Feedback) [9] ว่ามีข้อมูลใดบ้างที่ตรงกับสิ่งที่กำลังค้นหาอยู่เป็นผลลัพธ์ในการค้นหาครั้งแรก เพื่อปรับปรุงผลลัพธ์ในครั้งต่อไปให้ตรงกับความต้องการมากยิ่งขึ้น 2) เทคนิคการค้นคืนรูปภาพแบบหลายระดับ (Multi-Level Image Retrieval) โดยแทนที่คุณลักษณะของภาพด้วยคำอธิบายที่เป็นลำดับชั้น (Hierarchical Image Analysis) [10] ที่ง่ายต่อการทำความเข้าใจเช่นเดียวกับการรับรู้ของมนุษย์ 3) เทคนิคการใช้คำอธิบายภาพ (Annotation) [11] หรือออนโทโลยี (Ontology) ที่พิจารณาความสัมพันธ์ของรูปภาพที่มีความเกี่ยวข้องกัน เช่น การนิยามรูปภาพนกตามคุณลักษณะเฉพาะ เช่น ลำตัว ปีก และจะงอยปากที่แตกต่างกัน และนำไปสร้างความสัมพันธ์กับสี ลวดลาย ขนาด ถิ่นที่อาศัย และอาหาร เป็นต้น 4) เทคนิคการใช้แม่แบบเชิงความหมาย (Semantic Template: ST) [12] เพื่อกำหนดคุณลักษณะของรูปภาพเป็นตัวแทน (Representative Feature) ของกลุ่มข้อมูลที่แสดงเนื้อหาเกี่ยวกับคุณลักษณะต่าง ๆ 5) เทคนิคการใช้ข้อความเพื่อเรียนรู้การค้นคืนรูปภาพ (Using Text to Teach Image Retrieval) [13] จากข้อมูลต่าง ๆ ที่ปรากฏบนเว็บเพจ เช่น ข้อความโดยรอบแท็ก คำอธิบาย ลิงก์เชื่อมโยง หรือส่วนหัวของเว็บเพจ ร่วมกับคุณลักษณะของรูปภาพในการแปลความหมาย และ 6) เทคนิคการจัดกลุ่มข้อมูลด้วยการเรียนรู้ของเครื่อง (Machine Learning Classification) ปัจจุบันพัฒนาเป็นการเรียนรู้เชิงลึก (Deep

Learning) ซึ่งมีจุดเด่นในการวิเคราะห์และจำแนกคุณลักษณะรูปภาพได้อย่างหลากหลาย และมีประสิทธิภาพการพยากรณ์ผลลัพธ์ได้ดีกว่าการเรียนรู้ของเครื่องแบบเดิมเป็นอย่างมาก

แม้การเรียนรู้เชิงลึกมีบทบาทสำคัญต่อการประมวลผลรูปภาพ แต่ยังพบข้อจำกัดหลายประการ ได้แก่ 1) ต้องการชุดข้อมูลฝึกฝนในปริมาณที่เพียงพอ [14] เพื่อป้องกันปัญหา Overfitting ที่ตัวแบบมีประสิทธิภาพสูงในการจำแนกข้อมูลกับชุดข้อมูลทดสอบ แต่ประสิทธิภาพจะลดลงเมื่อทำงานกับข้อมูลที่ไม่เคยพบมาก่อน 2) การสร้างชุดข้อมูลนั้นมีต้นทุนมหาศาล [15] เนื่องจากต้องใช้มนุษย์ในการติดป้ายกำกับ อย่างไรก็ดีการรวบรวมข้อมูลต้องเผชิญกับความท้าทายมากมาย ทั้งสิทธิส่วนบุคคล รวมถึงนโยบายทางกฎหมายและวัฒนธรรม [16] และ 3) ตัวแบบจะถูกฝึกฝนบนชุดข้อมูลที่มีลักษณะตายตัว ไม่สามารถเปลี่ยนแปลงได้ระหว่างการเรียนรู้ หากมีการเปลี่ยนแปลงต้องเริ่มเรียนรู้ใหม่ตั้งแต่ต้น

จากปัญหาดังกล่าวเป็นที่มาของการใช้งานตัวแบบที่ถูกฝึกฝนล่วงหน้าบนชุดข้อมูลขนาดใหญ่ (Pre-training) จึงสามารถถ่ายโอนความรู้ (Transfer Learning) ไปใช้ได้เลยโดยไม่จำเป็นต้องฝึกฝนตัวแบบใหม่ตั้งแต่ต้น เพียงปรับแต่ง (Fine-tuning) ตัวแบบเพิ่มเติมด้วยชุดข้อมูลที่เฉพาะเจาะจงในงานที่ต้องการ เนื่องจากตัวแบบได้เรียนรู้คุณลักษณะที่สำคัญของรูปภาพเอาไว้ ดังนั้นการประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) จึงได้นำแนวคิดดังกล่าวมาปรับใช้ในการฝึกฝนตัวแบบบริบททางภาษา (Language Model) บนข้อความดิบขนาดใหญ่ด้วยการเรียนรู้ด้วยตนเอง (Self-supervised) เกิดเป็นสถาปัตยกรรม Transformer [17] ด้วยแนวคิดที่ว่าคำแต่ละคำมีความหมายในตัวเอง โดยความหมายเหล่านี้มีผลมาจากบริบทของคำที่อยู่รอบ ๆ ก่อนจะถูกนำมาประยุกต์ใช้กับการสกัดคุณลักษณะรูปภาพที่มีหลักการเช่นเดียวกับการทำงานกับข้อความแต่เปลี่ยนมาใช้กับรูปภาพ เรียกว่า Vision in Transformer (ViT) [18]

2) การออกแบบ (Design) แบบจำลองการค้นคืนรูปภาพเชิงความหมาย โดยใช้การฝึกฝนตัวแบบล่วงหน้าแบบคอนทราสต์ระหว่างรูปภาพและข้อความ ประกอบด้วย 3 ขั้นตอน ได้แก่ 1) การเตรียมข้อมูลก่อนการประมวลผล 2) การฝึกฝนล่วงหน้าแบบคอนทราสต์ระหว่างรูปภาพและข้อความ และ 3) การปรับแต่งพารามิเตอร์ด้วยการเรียนรู้ด้วยตนเอง มีรายละเอียดดังนี้

2.1) การเตรียมข้อมูลก่อนการประมวลผล (Dataset Pre-processing) งานวิจัยนี้ใช้ชุดข้อมูล Flickr30k [19] ที่เป็นชุดข้อมูลรูปภาพที่รวบรวมจากเว็บไซต์ Flickr จำนวน 31,783 รูปภาพและชุดข้อมูลที่ผู้วิจัยรวบรวมเองจำนวน 92,168 รูปภาพ รวมทั้งสิ้น 123,951 รูปภาพ สำหรับฝึกฝนให้คอมพิวเตอร์เรียนรู้จากรูปภาพและข้อความภาษาธรรมชาติ กำหนดให้รูปภาพทั้งหมดที่นำมาใช้คือประชากร จึงเลือกกลุ่มตัวอย่างที่เป็นไปตามโอกาสทางสถิติ (Probability Sampling) ด้วยการสุ่มอย่างมีระบบ (Systematic Random Sampling) [20] เป็นกลุ่มตัวอย่างรวมทั้งสิ้น 400 รูปภาพ แบ่งเป็นรูปภาพบนชุดข้อมูล Flickr30k จำนวน 100 รูปภาพ และอีก 300 รูปภาพจากชุดข้อมูลที่ผู้วิจัยรวบรวมเอง (Custom Dataset) ตามสัดส่วนของจำนวนรูปภาพที่ใช้ในงานวิจัย มาสร้างข้อความบรรยายรูปภาพที่อธิบายความหมายเชิงคุณภาพของรูปภาพในรูปแบบภาษาอังกฤษที่กำหนดขึ้นโดยผู้วิจัย กำหนด 1 รูปภาพต่อ 5 ข้อความบรรยายรูปภาพ รวมทั้งสิ้น 2,000 รายการ ดังตารางที่ 1

จากตารางที่ 1 กำหนดรูปภาพจากการสุ่มเป็น 4 โดเมน ได้แก่ 1) ภาพบุคคล 2) ภาพสัตว์เลี้ยง 3) ภาพสิ่งของ และ 4) ภาพสิ่งปลูกสร้างและทิวทัศน์ โดยงานวิจัยนี้ให้ความสำคัญกับประเด็นลิขสิทธิ์ของรูปภาพจึงใช้ชุดข้อมูลรูปภาพที่ผู้วิจัยรวบรวมเองมากกว่าร้อยละ 75 ในการเรียนรู้ของตัวแบบ ตลอดจนคำนึงถึงสิทธิส่วนบุคคลที่ปรากฏในภาพ จึงหลีกเลี่ยงการใช้รูปภาพที่ปรากฏใบหน้าของบุคคลหรือส่วนหนึ่งส่วนใดอย่างชัดเจนที่สามารถนำไปสู่การระบุอัตลักษณ์บุคคลได้เป็นเงื่อนไขในการสุ่มรูปภาพเพื่อเป็นข้อมูลก่อนการประมวลผล มีรายละเอียดดังนี้

2.1.1) การสร้างรูปภาพใหม่จากการดัดแปลงรูปภาพต้นฉบับ (Image Augmentation) [21] เพื่อแก้ปัญหา Overfitting เนื่องจากประสิทธิภาพของตัวแบบการเรียนรู้เชิงลึกนั้นขึ้นกับปริมาณข้อมูลเป็นปัจจัยสำคัญอันดับหนึ่ง เกิดเป็นแนวคิดการปรับแต่งรูปภาพ เช่น การบิด ตัด หมุน เปลี่ยนสี ทำภาพให้มืดลงหรือสว่างขึ้น หรือใส่ตัวรบกวน (Noise) ลงไปให้ได้รูปภาพใหม่ออกมา [22] เพื่อให้แบบจำลองจดจำสิ่งที่ไม่จำเป็นได้ยากขึ้น และจดจำคุณลักษณะสำคัญของรูปภาพได้ดียิ่งขึ้น

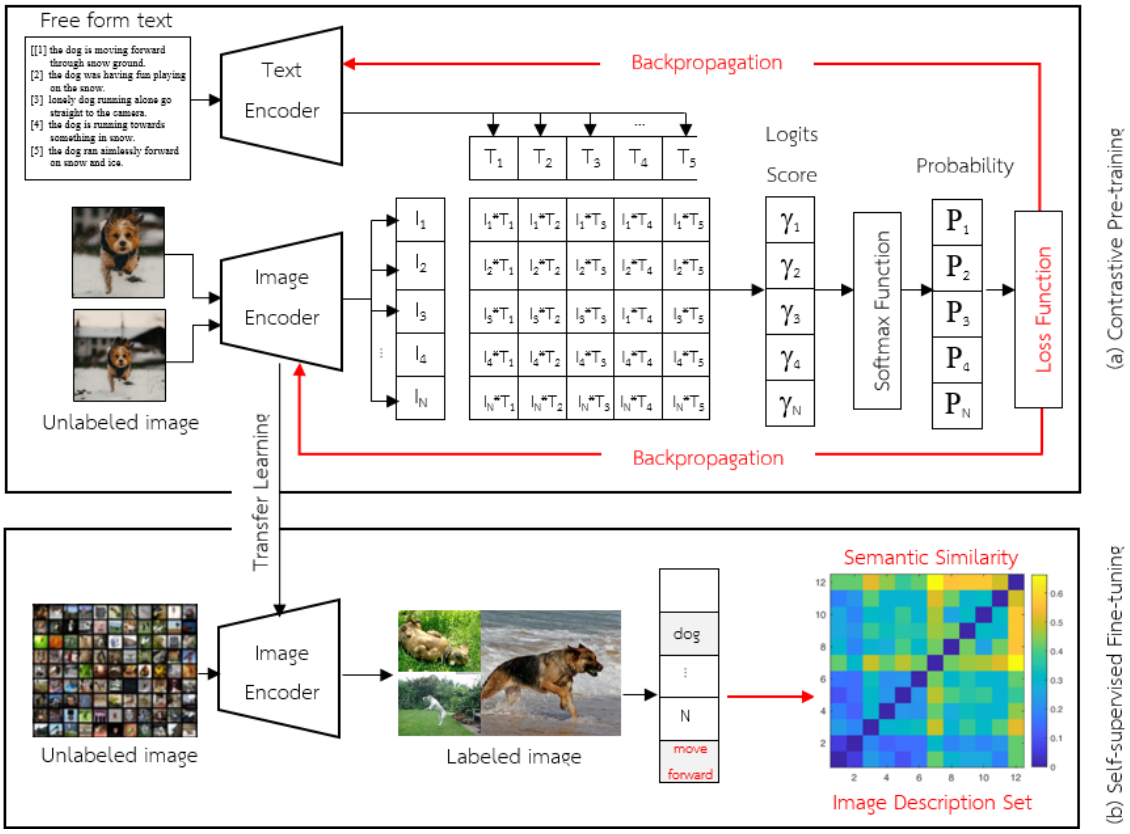
ตารางที่ 1 ตัวอย่างรูปภาพที่สุ่มจากชุดข้อมูล Flickr30k และชุดข้อมูลที่ผู้วิจัยรวบรวมเอง

ลำดับ	ชุดข้อมูล	ข้อความบรรยายรูปภาพ	รูปภาพ	โดเมน
1	Flickr30k	(1) two women were slowly walking forward. (2) two women are heading somewhere. (3) two women walking on the sidewalk alone. (4) the two women walked aimlessly forward. (5) the two women stepped forward determinedly.		ภาพบุคคล
2	Flickr30k	(1) the dog is moving forward through snow ground. (2) the dog was having fun playing on the snow. (3) the dog running alone go straight to the camera. (4) the dog is running towards something in snow. (5) the dog ran aimlessly forward on snow and ice.		ภาพสัตว์เลี้ยง
...	...		...	...
400	Custom	(1) the towering skyscrapers and the urban landscape stretches as far as the eye can see. (2) the cityscape is a mesmerizing blend of sleek architecture and vibrant neighborhoods. (3) the hustle and bustle of the streets below echoes the pulsating rhythm of urban life. (4) this metropolis is a testament to human ingenuity and urban design. (5) the mix of vibrant neighborhoods and bustling streets is absolutely mind-blowing.		ภาพสิ่งปลูกสร้างและทิวทัศน์

2.1.2) การทำความสะอาดข้อความและการกำหนดมาตรฐานข้อความ (Text Cleansing and Text Normalization) [23] งานวิจัยนี้ไม่มีการกำจัดคำหยุด เนื่องจากถือว่าคำทั้งหมดมีผลต่อการกำหนดคุณลักษณะของรูปภาพเป็นไปในทิศทางเดียวกับแนวโน้มการกำจัดคำหยุดที่มีจำนวนคำลดลงเรื่อย ๆ จนในปัจจุบันการเรียนรู้เชิงลึกไม่มีการกำจัดคำหยุดเลย อย่างไรก็ตามยังคงกำจัดเครื่องหมายต่าง ๆ และเปลี่ยนเป็นตัวอักษรพิมพ์เล็กทั้งหมดก่อนจะนำไปประมวลผล

2.2) การฝึกฝนล่วงหน้าแบบคอนทราสต์ระหว่างรูปภาพและข้อความ งานวิจัยนี้ใช้ตัวแบบ CLIP (Contrastive Language–Image Pre-training) [24] เป็นโมเดลฐาน (Baseline Model) ดังภาพที่ 4 เพื่อลดภาระการฝึกฝนตัวแบบจากเดิมที่ต้องการข้อมูลจำนวนมากในการเรียนรู้ อีกทั้งมีประสิทธิภาพดีกว่าตัวแบบที่ถูกสร้างขึ้นมาเองตั้งแต่ต้น

2.2.1) การฝึกฝนตัวเข้ารหัสรูปภาพด้วยตัวแบบ ViT-B/32 สำหรับการสกัดคุณลักษณะรูปภาพตามแนวคิดของสถาปัตยกรรม Transformer ที่ทำงานกับข้อความ แต่นำมาปรับใช้กับรูปภาพ (Vision in Transformer) จากการเปรียบเทียบความสัมพันธ์ระหว่างแพทช์เพื่อคำนวณหาผังคุณลักษณะในการหาความสัมพันธ์กับบริเวณทั้งภาพ โดยผลลัพธ์จากการฝังรูปภาพ (Image Embedding) จะแปลงเป็นเวกเตอร์คุณลักษณะรูปภาพ (Image Feature Vector) ขนาด 2,048



ภาพที่ 4 แบบจำลองการค้นคืนรูปภาพเชิงความหมาย โดยใช้การฝึกฝนตัวแบบล่วงหน้าแบบคอนทราสต์

2.2.2) การฝึกฝนตัวเข้ารหัสข้อความด้วยตัวแบบ GPT-2 บนพื้นฐานสถาปัตยกรรม Transformer เพื่อเรียนรู้ความหมายโดยรวมของข้อความบรรยายรูปภาพแต่ละประโยค ผลลัพธ์จากการฝังข้อความ (Text Embedding) แปลงเป็นเวกเตอร์คุณลักษณะข้อความ (Text Feature Vector) ขนาด 768

2.2.3) การเรียนรู้แบบคอนทราสต์บนพื้นที่การฝังหลายรูปแบบ (Contrastive Learning on Multi-modal Embedding Space) [25] เนื่องจากเวกเตอร์คุณลักษณะข้อความและเวกเตอร์คุณลักษณะรูปภาพมีขนาดต่างกัน จึงต้องเรียกใช้ Projection Head [26] ในการเปลี่ยนขนาดเวกเตอร์บนพื้นที่การฝังให้มีมิติเท่ากับ 256 เพื่อฝึกฝนร่วมกันระหว่างตัวเข้ารหัสรูปภาพและตัวเข้ารหัสข้อความด้วยการพยายามขยับเวกเตอร์ตัวแทนด้วยค่าความคล้ายคลึงโคไซน์ (Cosine Similarity) จากการคำนวณดอตโปรดัค (Dot Product) ของเวกเตอร์แต่ละคู่ หากเป็นเวกเตอร์คู่เดียวกันจะมีค่าความคล้ายคลึงสูง ตรงกันข้ามหากเป็นเวกเตอร์คู่ที่แตกต่างกันจะมีค่าความคล้ายคลึงต่ำด้วย ดังสมการที่ 1 [27]

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

กำหนดให้ A และ B คือ เวกเตอร์

i คือ องค์ประกอบของเวกเตอร์ (Element of Vector)

n คือ จำนวนขององค์ประกอบ (Number of Element)

$.$ คือ การคูณดอทยุคลิด (Euclidean Dot Product)

$\|\ \|$ คือ ขนาดของเวกเตอร์ (Magnitude)

จากภาพที่ 4 ยกตัวอย่างการคำนวณเป็นอาร์เรย์ 1 มิติ เพื่อง่ายต่อการทำความเข้าใจ กำหนดให้รูปภาพแทนด้วย I_1 ประกอบด้วยชุดตัวเลข $I_1 = \{0.45, 0.32, 0.46, 0.23\}$ และข้อความบรรยายรูปภาพในลำดับที่ (1) “the dog is moving forward through snow ground.” ลำดับที่ (2) “the dog was having fun playing on the snow.” ลำดับที่ (3) “lonely dog running alone go straight to the camera.” ลำดับที่ (4) “the dog is running towards something in snow” และลำดับที่ (5) “the dog ran aimlessly forward on snow and ice.” 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 $T_5\{-0.22, 0.64, 0.53, 0.22\}$ ตามลำดับ แทนค่าในสมการที่ 1 มีรายละเอียดดังนี้

$$sim(I_1, T_1) = \frac{(I_{1,1} * T_{1,1}) + (I_{1,2} * T_{1,2}) + (I_{1,3} * T_{1,3}) + (I_{1,4} * T_{1,4})}{\sqrt{I_1^2 + I_2^2 + I_3^2 + I_4^2} * \sqrt{T_1^2 + T_2^2 + T_3^2 + T_4^2}}$$
$$= \frac{(0.45*0.24) + (0.32*0.46) + (0.46*0.44) + (0.23*0.35)}{\sqrt{0.45^2 + 0.32^2 + 0.46^2 + 0.23^2} * \sqrt{0.24^2 + 0.46^2 + 0.44^2 + 0.35^2}}$$
$$= \frac{0.54}{0.58} = 0.93$$

ดังนั้นค่าความคล้ายคลึงโคไซน์ระหว่างรูปภาพ 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

2.2.4) การคำนวณความน่าจะเป็นของป้ายกำกับ (Label Probability) เนื่องจากเอาต์พุตจากการฝึกฝนล่วงหน้าแบบคอนทราสต์นั้นจะอยู่ในรูปแบบค่า Logits Score ที่เกิดจากค่าความคล้ายคลึงของเวกเตอร์ จึงต้องใช้ฟังก์ชันซอฟต์แม็กซ์ [28] มาแปลงให้เป็นมาตรฐานการแจกแจงความน่าจะเป็นตามสัดส่วนของค่าเอาต์พุตแต่ละคลาส ดังสมการที่ 2

$$f(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \quad (2)$$

กำหนดให้ x_i คือ เวกเตอร์อินพุตของฟังก์ชันซอฟต์แม็กซ์

$\exp(x_i)$ คือ ฟังก์ชันเอกซ์โพเนนเชียล (Exponential Function) มีค่าคงที่เท่ากับ 2.71828 ยกกำลังด้วยเวกเตอร์อินพุต

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

จากภาพที่ 4 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 เมื่อแทนค่าเวกเตอร์อินพุตของคำบรรยายรูปภาพลำดับที่ T_1 เท่ากับ 0.93 ในสมการที่ 2 มีรายละเอียดดังนี้

$$f(0.93) = \frac{\exp(0.93)}{\exp(0.93) + \exp(0.26) + \exp(0.91) + \exp(0.80) + \exp(0.55)}$$
$$= \frac{2.5345}{2.5345 + 1.2969 + 2.4843 + 2.2255 + 1.7333}$$
$$= 0.2467$$

ผลลัพธ์จากการแปลงค่าให้เป็นมาตรฐานการแจกแจงความน่าจะเป็นตามสัดส่วนของค่าเอาต์พุตของคำบรรยายรูปภาพลำดับที่ T_1 กำหนดทศนิยม 4 ตำแหน่ง จะเท่ากับ 0.2467 จากนั้นคำนวณต่อกับข้อความบรรยายรูปภาพลำดับที่ T_2 ถึง T_5 จะได้ผลลัพธ์เท่ากับ 0.1262, 0.2418, 0.2166 และ 0.1687 ตามลำดับ ซึ่งมีผลรวมเท่ากับ 1

การประยุกต์ใช้งานตัวแบบ CLIP นั้นส่วนใหญ่จะเน้นไปที่การปรับปรุงพารามิเตอร์ของตัวแบบเพื่อใช้ในการจำแนกรูปภาพ โดยการสร้างตัวแยกประเภทชุดข้อมูลจากข้อความป้ายกำกับ (Create Dataset Classifier from Label Text) โดยฝังอ็อบเจกต์คลาสลงในข้อความบรรยายรูปภาพ (Object Classes into Captions) ในขั้นตอนสุดท้าย เมื่อใส่รูปภาพเข้าไปในตัวแบบจะสร้างป้ายกำกับจากผลการพยากรณ์ข้อมูลแบบไม่มีตัวอย่าง (Use for Zero-shot Prediction) อย่างไรก็ดีมีบางงานวิจัยที่นำมาประยุกต์ใช้กับการค้นคืนรูปภาพเชิงเนื้อหา (Content-based Image Retrieval: CBIR) โดยงานวิจัยของ Le และคณะ [29] ประยุกต์ใช้ตัวแบบ CLIP มาเรียนรู้คำอธิบายภาษาธรรมชาติเพื่อการจำแนกรถยนต์ในมุมมองที่แตกต่างกันจากการฝึกฝนตัวเข้ารหัสรูปภาพด้วยคุณลักษณะแบบโกลบอลและคุณลักษณะแบบโลคอล เช่นเดียวกับการฝึกฝนตัวเข้ารหัสข้อความภาษาธรรมชาติตามแนวคิดการเรียนรู้ด้วยตนเองระหว่างรูปภาพและข้อความ เพื่อติดป้ายกำกับรถยนต์ในรูปภาพประกอบด้วยสี รุ่น และทิศทาง เช่นเดียวกับงานวิจัยของ Xie และคณะ [30] นำเสนอ RA-CLIP ที่นำตัวแบบ CLIP มาปรับปรุงการเรียนรู้แบบคอนทราสต์แบบดั้งเดิม โดยเมื่อมีอินพุตของรูปภาพตัวแบบจะดึงคู่รูปภาพและข้อความที่มีความคล้ายคลึงกันมากที่สุดจากชุดข้อมูลอ้างอิง (Reference Set) ซึ่งมีคำอธิบายข้อมูลสำหรับรูปภาพที่อินพุตเข้ามาร่วมอธิบายความหมายของรูปภาพบนโมดูลสนับสนุนการค้นคืน (Retrieval Augmented Module) เพื่อเพิ่มความแม่นยำในการพยากรณ์หมวดหมู่ป้ายกำกับ (Category Label) สำหรับการจำแนกรูปภาพที่มีการลงรายละเอียด (Fine-grained Classification) แต่อย่างไรก็ดีแม้งานวิจัยที่กล่าวมานั้นจะอยู่บนแนวคิดการเรียนรู้แบบคอนทราสต์ แต่ก็ยังมิได้มุ่งค้นคืนรูปภาพเชิงความหมายโดยใช้ข้อความค้นหาภาษาธรรมชาติ อันจะเป็น Contribution ของงานวิจัยนี้

2.3) การปรับแต่งพารามิเตอร์ด้วยการเรียนรู้ด้วยตนเอง (Self-supervised Fine-tuning) แม้ปัจจุบันการค้นคืนรูปภาพเชิงเนื้อหาจะทำงานได้อย่างมีประสิทธิภาพด้วยการระบุแนวคิดระดับสูง (High-level Semantics) ที่เป็นรูปธรรม (Concrete Concept) อย่างวัตถุ (Object) กิจกรรม (Activities) หรือเหตุการณ์ (Event) แต่ในความเป็นจริงพฤติกรรมผู้ใช้โดยมากจะค้นคืนรูปภาพโดยใช้ความสัมพันธ์ระหว่างแนวคิดระดับสูงที่เป็นรูปธรรม และแนวคิดที่เป็นนามธรรม (Abstract Concept) อย่างอารมณ์และความรู้สึก (Mood and Feelings) [31] รวมเข้าด้วยกันเป็นข้อความค้นหาภาษาธรรมชาติ ในทางกลับกันคุณลักษณะที่ถูกแยกอัตโนมัติด้วยคอมพิวเตอร์มักเป็นแนวคิดระดับสูงที่เป็นรูปธรรมเท่านั้น เป็นที่มาของงานวิจัยชิ้นนี้ที่มุ่งเชื่อมโยงระหว่างแนวคิดระดับสูงที่เป็นรูปธรรมและนามธรรมเข้าด้วยผ่านแบบจำลองการค้นคืนรูปภาพ เชิงความหมายอันเป็นผลลัพธ์จากงานวิจัยนี้

2.3.1) การปรับปรุงการเรียนรู้ของตัวแบบจากการนำแนวคิดการแผ่กระจายย้อนหลัง (Backpropagation) มาปรับปรุงการอัลกอริทึมการเรียนรู้ จากการหาค่าความผิดพลาดที่ได้จากการเปรียบเทียบระหว่างผลลัพธ์จากตัวแบบกับผลลัพธ์ที่ถูกประเมินความหมายโดยผู้เชี่ยวชาญด้วยการหาค่าการสูญเสียจากฟังก์ชันการสูญเสีย (Loss Function) จากนั้นนำค่าการสูญเสียมาใช้กับฟังก์ชันเพิ่มประสิทธิภาพ (Optimize Function) สำหรับปรับค่าพารามิเตอร์ที่ใช้ในการเรียนรู้ของตัวแบบเพื่อการค้นคืนรูปภาพจากการเปรียบเทียบความคล้ายคลึงเชิงความหมาย

อย่างไรก็ดี แนวคิดระดับสูงที่เป็นนามธรรมนั้นยากที่จะประเมินผลด้วยค่าความถูกต้อง หรือค่าความแม่นยำ จึงเป็นที่มาของการให้ผู้เชี่ยวชาญ 30 ท่าน เข้ามาร่วมเป็นส่วนหนึ่งในการประเมินผลความหมายของแต่ละรูปภาพให้สอดคล้องกับหลักคอมพิวเตอร์วิทัศน์ (Computer Vision) กำหนดคุณสมบัติผู้เชี่ยวชาญ [32] ว่าเป็นผู้สำเร็จการศึกษาด้านการจัดเก็บและเรียกค้นสารสนเทศ (Information Storage and Retrieval) หรือสาขาที่เกี่ยวข้อง และมีประสบการณ์การใช้งานคอมพิวเตอร์อินเทอร์เน็ต และเสิร์ชเอนจินในการเรียกค้นรูปภาพอย่างน้อย 10 ปีขึ้นไป ตามหลักการการทดสอบทัวริง (Turing Test) [33] ที่มุ่งทดสอบการทำงานของปัญญาประดิษฐ์เปรียบเทียบกับการทำงานของมนุษย์ โดยเปรียบเทียบระหว่างผลการประเมินความหมายของรูปภาพจากผู้เชี่ยวชาญกับผลการพยากรณ์ความน่าจะเป็นของป้ายกำกับจากตัวแบบว่าแปลความหมายของ

รูปภาพได้ใกล้เคียงกับมนุษย์มากน้อยเพียงใด โดยผู้วิจัยสร้างข้อความบรรยายรูปภาพที่อธิบายความหมาย เชิงคุณภาพของรูปภาพที่ได้จากการสุ่มเป็นชุดคำถาม โดยข้อความบรรยายรูปภาพทั้งหมดอยู่ในรูปแบบภาษาอังกฤษ จากนั้นให้ผู้เชี่ยวชาญแต่ละท่านประเมินความหมายของแนวคิดระดับสูงเชิงนามธรรมจำนวน 400 รูปภาพ บนชุดคำถามแบบเลือกคำตอบที่ถูกต้องเพียงข้อเดียว เพื่อเปรียบเทียบกับผลการพยากรณ์ป้ายกำกับด้วยการหาค่าครอสเอนโทรปี เพื่อวัดค่าความน่าจะเป็นของป้ายกำกับว่ามีผลการพยากรณ์ใกล้กับค่าจริงที่ต้องการมากน้อยแค่ไหน (Error Measurement) ด้วยฟังก์ชัน การสูญเสียที่นิยมใช้ในการจำแนกรูปภาพด้วย [34] ดังสมการที่ 3

$$H(p,q) = -\sum_{x \in classes} p(x) \log q(x) \quad (3)$$

กำหนดให้ $p(x)$ คือ ป้ายกำกับจริง (True Label) จากการประเมินของผู้เชี่ยวชาญแบบ one-hot ตัวเลือกที่ผู้เชี่ยวชาญเลือก เท่ากับ 1 ในทางตรงกันข้ามตัวเลือกที่ไม่ถูกเลือกจะเท่ากับ 0

$q(x)$ คือ ผลการพยากรณ์ความน่าจะเป็นของป้ายกำกับ

จากภาพที่ 4 ตัวอย่างป้ายกำกับที่ผู้เชี่ยวชาญเลือกคือลำดับที่ 1 ถือว่าเป็นป้ายกำกับจริง ดังนั้นตัวเลือกลำดับที่ 2 ถึง 5 จะเท่ากับ 0 ทั้งหมด แทนค่า $p[1,0,0,0,0]$ ในทางกลับกันเมื่อพิจารณาผลการพยากรณ์ความน่าจะเป็นของป้ายกำกับจากฟังก์ชันซอฟต์แม็กซ์ในตัวเลือกที่ 1 ถึง 5 เท่ากับ 0.2467, 0.1262, 0.2418, 0.2166 และ 0.1687 ตามลำดับ แทนค่าใน $q[0.2467, 0.1262, 0.2418, 0.2166, 0.1687]$ เมื่อแทนค่าในสมการที่ 3 มีรายละเอียดดังนี้

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

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

จากภาพที่ 4 ผลการเปรียบเทียบระหว่างผลการประเมินความหมายของรูปภาพจากผู้เชี่ยวชาญคือตัวเลือกที่ 1 ตรงกับผลการพยากรณ์ความน่าจะเป็นของป้ายกำกับที่สูงที่สุดในตัวเลือกที่ 1 ส่งผลให้ค่าครอสเอนโทรปี เท่ากับ 2.0192 ซึ่งอยู่ระหว่าง 0 ถึงไม่จำกัด (Infinity) โดย 0 คือไม่มีค่าสูญเสียเลย แสดงถึงความมั่นใจในผลการพยากรณ์ของตัวแบบ หากค่าครอสเอนโทรปีมีค่าน้อย แสดงถึงตัวแบบพยากรณ์ความน่าจะเป็นของป้ายกำกับถูกต้องตรงกับผลประเมินจากผู้เชี่ยวชาญได้อย่างมั่นใจ ตรงกันข้ามแม้ผลพยากรณ์ความน่าจะเป็นของป้ายกำกับถูกต้องแต่มีค่าความน่าจะเป็นน้อย ค่าครอสเอนโทรปีก็จะมีค่ามาก เช่นเดียวกับผลพยากรณ์ความน่าจะเป็นของป้ายกำกับที่ไม่ถูกต้องแม้จะมีค่าความน่าจะเป็นสูงก็จะทำให้ค่าครอสเอนโทรปีมากตามไปด้วย

2.3.2) การเรียนรู้ซ้ำบนชุดข้อมูลที่กำหนดเอง (Retrain Network on Custom Dataset) โดยค่าการสูญเสียที่คำนวณได้จากการเปรียบเทียบระหว่างผลการประเมินความหมายของรูปภาพจากผู้เชี่ยวชาญกับผลการพยากรณ์ความน่าจะเป็นของป้ายกำกับ และผลการประเมินความหมายของรูปภาพจากผู้เชี่ยวชาญคำนวณจากพารามิเตอร์ค่าน้ำหนัก (Weight) ดังนั้นจะทำให้สามารถปรับพารามิเตอร์ที่มีผลกระทบกับค่าการสูญเสีย เพื่อให้ค่าการสูญเสียในการพยากรณ์น้อยที่สุดตามแนวคิดการแผ่กระจายย้อนหลัง (Backpropagation) ก่อนจะนำไปเรียนรู้ซ้ำอีกครั้งด้วยการเรียนรู้ด้วยตนเองบนชุดข้อมูล Flickr30k จำนวน 31,783 รูปภาพ และชุดข้อมูลที่ผู้วิจัยรวบรวมเองจำนวน 92,168 รูปภาพ รวมทั้งสิ้น 123,951 รูปภาพ

2.3.3) การเรียนรู้ความคล้ายคลึงเชิงความหมายด้วยตนเอง (Semantic Similarity Self-supervised) [35] เมื่อตัวแบบได้เรียนรู้ซ้ำเพื่อจดจำคุณลักษณะเชิงความหมายในรูปแบบแนวคิดที่เป็นนามธรรมของรูปภาพบนชุดข้อมูลแบบติดป้ายกำกับอัตโนมัติ (Automatically Labeled) เนื่องจากแบบจำลองมีการเรียนรู้จากรูปภาพที่ไม่จำเป็นต้องมีป้ายกำกับล่วงหน้า (Unlabeled Image) แต่ต้องมีป้ายกำกับสำหรับการค้นหา และใช้ประโยชน์จากความสัมพันธ์ (Relations) หรือความเกี่ยวข้อง (Correlations) ระหว่างคุณสมบัติของข้อมูลอินพุตที่แตกต่างกัน โดยบังคับให้แบบจำลองเรียนรู้การแสดงความหมายเกี่ยวกับข้อมูล จากนั้นชุดความรู้ (Knowledge) ที่เป็นผลลัพธ์จากการเรียนรู้ด้วยตนเองจะถูกถ่ายโอนการเรียนรู้ไป

ยังแบบจำลองเพื่อใช้สำหรับงานหลัก (Main Task) เพื่อติดป้ายกำกับข้อมูลให้กับรูปภาพ (Labeled Image) เปรียบเทียบกับแนวคิดระดับสูงในรูปแบบนามธรรมของรูปภาพที่ได้จากการเรียนรู้ความคล้ายคลึงเชิงความหมาย (Semantic Similarity) ก่อนจะสร้างเป็นเวกเตอร์คุณลักษณะรูปภาพจัดเก็บไว้ในชุดคำอธิบายรูปภาพ (Image Description Set) ต่อไป

3) การสร้าง (Construction) แบบจำลองการค้นคืนรูปภาพเชิงความหมายพัฒนาขึ้นด้วยการเขียนโปรแกรมภาษาไพธอนบน Google Colaboratory เพื่อลดระยะเวลาการประมวลผลของเครื่อง เนื่องจากคอมพิวเตอร์ทั่วไปการเรียนรู้ของตัวแบบ 1 ครั้งอาจใช้เวลาถึง 2-3 ชั่วโมง แต่บน Google Colaboratory นั้นจะใช้เวลาประมาณ 30-35 นาที อีกทั้งการเรียนรู้นั้นอาจต้องทำหลายสิบครั้งไปจนถึงหลายร้อยครั้ง ดังนั้นการเรียนรู้ของตัวแบบบนคอมพิวเตอร์ทั่วไปจะต้องใช้เวลานานมาก

4) การทดสอบและการกลับมาตรวจสอบซ้ำ (Testing and Turnover) เพื่อประเมินประสิทธิภาพแบบจำลองการค้นคืนรูปภาพเชิงความหมายแบบไม่ทราบจำนวนผลลัพธ์ที่ถูกต้อง (Order-Unaware Metrics) [36] ที่ใช้ค้นคืนเนื้อหาที่เกี่ยวข้อง เช่น ค้นคืนเนื้อหาหรือวิดีโอบนเว็บไซต์ที่มีผลลัพธ์จำนวนมากและไม่สามารถระบุจำนวนผลลัพธ์ที่ถูกต้องทั้งหมดได้ โดยผู้วิจัยระบุข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ (Qualitative semantic concepts of the image) ในรูปแบบภาษาอังกฤษจำนวน 30 รายการ ภายใต้แนวคิดระดับสูงของรูปภาพที่อยู่ในลักษณะนามธรรมบนชุดข้อมูล Flickr30k และชุดข้อมูลที่ผู้วิจัยรวบรวมเอง แต่ละรายการแสดงผลลัพธ์ 5 รูปภาพ แล้วให้ผู้เชี่ยวชาญ 3 ท่าน ร่วมกันประเมินผลลัพธ์จากการค้นคืนแต่ละครั้งแบบเป็นอิสระต่อกัน ประกอบด้วย

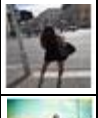
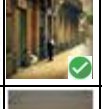
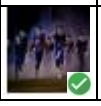
4.1) ค่าเฉลี่ยส่วนกลับของลำดับ (Mean Reciprocal Rank: MRR) เพื่อประเมินความสามารถในการค้นคืนรูปภาพอันดับแรกที่เกี่ยวข้องกับข้อความค้นหาจากจำนวนผลการค้นคืนทั้งหมด [36, 37] ดังสมการที่ 4 เมื่อ Q คือจำนวนการสืบค้นทั้งหมด และ $Rank_q$ คือ ลำดับของผลลัพธ์ที่เกี่ยวข้องอันดับแรก ผู้เชี่ยวชาญจะประเมินผลรูปภาพตามความเกี่ยวข้องลำดับที่ 1 ที่พบจาก 5 รูปภาพต่อข้อความค้นหาแต่ละรายการ

$$MRR = \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{Rank_q} \tag{4}$$

ตัวอย่างข้อความค้นหา 3 รายการได้แก่ 1) move forward together 2) feel alone in the city และ 3) strive for victory ที่ผ่านการประเมินผลโดยผู้เชี่ยวชาญท่านที่ 1 ดังตารางที่ 2 มีค่าเฉลี่ยส่วนกลับของลำดับ เท่ากับ 0.567

จากตารางที่ 2 จะดำเนินการกับข้อความค้นหาไปจนครบทั้ง 30 รายการ แล้วให้ผู้เชี่ยวชาญ 3 ท่าน ร่วมกันประเมินลำดับของผลลัพธ์ที่เกี่ยวข้องอันดับแรก โดยในแต่ละรายการจะยึดลำดับของรูปภาพที่ผู้เชี่ยวชาญประเมินแบบ 2 ใน 3 ตรงกัน ข้ามหากผลลัพธ์จากรายการค้นหาใดที่นอกเหนือจากเงื่อนไขที่กำหนดจะถือว่าค่าเฉลี่ยส่วนกลับของลำดับในรายการนั้นจะเท่ากับ 0 คะแนน ก่อนจะคำนวณเป็นค่าเฉลี่ยส่วนกลับของลำดับรวมของตัวแบบจากผู้เชี่ยวชาญ 3 ท่าน

ตารางที่ 2 การประเมินค่าเฉลี่ยส่วนกลับของลำดับจากข้อความค้นหา 3 รายการ ผ่านการประเมินผลโดยผู้เชี่ยวชาญท่านที่ 1

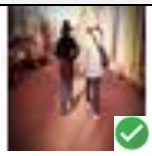
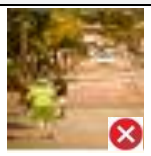
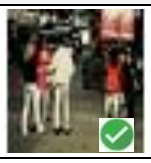
Query	Reciprocal Rank					MRR
	1)	2)	3)	4)	5)	
1) move forward together						$Query1: \frac{1}{Rank_1} = \frac{1}{1} = 1.0$
2) feel alone in the city						$Query2: \frac{1}{Rank_2} = \frac{1}{5} = 0.2$
3) strive for victory						$Query3: \frac{1}{Rank_3} = \frac{1}{2} = 0.5$
						$\sum_{q=1}^{Q}\frac{1}{Rank_q} = \frac{(1.0+0.2+0.5)}{3} = 0.567$

4.2) ค่าความครบถ้วน (Recall) เป็นอัตราส่วนของรูปภาพที่เกี่ยวข้อง และถูกค้นคืนออกมาจากจำนวนรูปภาพที่เกี่ยวข้องทั้งหมด แทนจำนวนผลลัพธ์จากการค้นคืนด้วย k [36, 37] ที่ถูกใช้ในการประเมินว่าตัวแบบนั้นมีประสิทธิภาพในการค้นคืนรูปภาพที่เกี่ยวข้องกับข้อความค้นหาจากผู้ใช้งานหรือไม่ โดยมุ่งพิจารณาเฉพาะรูปภาพที่เกี่ยวข้องกับข้อความค้นหาเท่านั้น กำหนดให้ $@k$ เป็นจำนวนเต็มของรายการที่พิจารณาตามลำดับที่กำหนด เพื่อลดข้อจำกัดของการค้นคืนที่มีผลลัพธ์จำนวนมากและไม่สามารถระบุจำนวนผลลัพธ์ที่ถูกต้องทั้งหมดได้ ดังสมการที่ 5

$$\mathrm{Re}call@k = \frac{TP@k}{(TP@k)+(FN@k)} \tag{5}$$

ผลลัพธ์จากข้อความค้นหา “move forward together” จำนวน 5 รูปภาพ ที่ผู้เชี่ยวชาญท่านที่ 1 ประเมินว่าถูกต้องจำนวน 3 รูปภาพ ดังตารางที่ 3

ตารางที่ 3 การประเมินค่าความครบถ้วนจำนวน 5 รูปภาพ ที่ผ่านการประเมินผลโดยผู้เชี่ยวชาญท่านที่ 1

Rank	1)	2)	3)	4)	5)
k					
$\mathrm{Re}call@k$	$\frac{1}{(1+2)}=\frac{1}{3}=0.33$	$\frac{1}{(1+2)}=\frac{1}{3}=0.33$	$\frac{2}{(2+1)}=\frac{2}{3}=0.67$	$\frac{2}{(2+1)}=\frac{2}{3}=0.67$	$\frac{3}{(3+0)}=\frac{3}{3}=1.00$

จากตารางที่ 3 จะดำเนินการกับข้อความค้นหาไปจนครบทั้ง 30 รายการแล้วให้ผู้เชี่ยวชาญ 3 ท่าน ร่วมกันประเมินความถูกต้องของผลลัพธ์ หากรูปภาพใดมีผลการประเมินจากผู้เชี่ยวชาญถูกต้องทั้งหมดจะถือว่ารูปภาพนั้นถูกต้องตามข้อความค้นหา เท่ากับ 1 คะแนน ตรงกันข้ามหากรูปภาพใดที่ผู้เชี่ยวชาญประเมินว่าไม่ถูกต้องแม้เพียงท่านเดียว จะถือว่ารูปภาพนั้นไม่ถูกต้อง เท่ากับ 0 คะแนน ก่อนจะคำนวณเป็นค่าเฉลี่ยความครบถ้วนของตัวแบบจากผู้เชี่ยวชาญ 3 ท่าน

ผลการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาและประเมินประสิทธิภาพแบบจำลองการค้นคืนรูปภาพเชิงความหมาย โดยใช้ตัวแบบการเรียนรู้เชิงลึกที่ถูกฝึกฝนล่วงหน้าแบบคอนทราสต์ ผลการประเมินประสิทธิภาพการค้นคืนรูปภาพเชิงความหมายบนชุดข้อมูลรูปภาพ Flickr30k เปรียบเทียบกับชุดข้อมูลที่ผู้วิจัยรวบรวมเองด้วยค่าเฉลี่ยส่วนกลับของลำดับ (MRR) และค่าความครบถ้วนที่ k ลำดับ ดังตารางที่ 4 ผลลัพธ์จากการค้นคืนรูปภาพเชิงความหมายบนชุดข้อมูล Flickr30k ที่ผ่านการประเมินผลโดยผู้เชี่ยวชาญท่านที่ 1

งานวิจัยนี้ให้ความสำคัญความคลาดเคลื่อน (Error) คือ ผลต่างระหว่างค่าที่วัดได้กับค่าที่แท้จริง ถ้าค่าที่วัดได้ใกล้เคียงกับค่าจริงมากแสดงว่าการวัดนั้นมีความถูกต้องสูง [38] โดยการวัดทุกครั้งมักมีค่าความคลาดเคลื่อนเกิดขึ้นเสมอ แบ่งออกเป็น 3 ชนิด ได้แก่ 1) ความคลาดเคลื่อนที่เกิดจากผู้วัด (Human Error) อันเกิดจากความประมาทเลินเล่อในการวัด2) ความคลาดเคลื่อนเชิงระบบ (Systematic Error) มักเกิดจากเครื่องมือวัดไม่ได้ประสิทธิภาพ และ 3) ความคลาดเคลื่อนเชิงสถิติ (Statistical Error) ซึ่งความคลาดเคลื่อนประเภทที่ 1 และ 2 สามารถจัดการได้ด้วยความรอบคอบ ระมัดระวัง และการพัฒนาเครื่องมือวัดให้มีประสิทธิภาพ แต่ความคลาดเคลื่อนประเภทที่ 3 นั้นไม่มีทางกำจัดให้หมดไปได้ เนื่องจากอยู่นอกเหนือการควบคุม จึงมักใช้การวัดซ้ำหลาย ๆ ครั้งเพื่อให้ความคลาดเคลื่อนลดน้อยลง ดังนั้นเมื่อครบรอบจะทำการประเมินอีกจนครบ 3 รอบด้วยข้อความค้นหาเดิมแบบผลลัพธ์เป็นอิสระต่อกัน แล้วนำผลการประเมินมาหาค่าเฉลี่ยเพื่อให้ความคลาดเคลื่อนลดน้อยลงจนถึงในระดับที่ยอมรับได้ ดังตารางที่ 5

ตารางที่ 4 ผลลัพธ์จากการค้นคืนรูปภาพเชิงความหมายที่ผ่านการประเมินผลโดยผู้เชี่ยวชาญท่านที่ 1

1) move forward together				
				
2) feel alone in the city				
				
n) ...				
...	...	...	...	...
30) strive for victory				
				

ตารางที่ 5 ค่าเฉลี่ยส่วนกลับของลำดับ และค่าความครบถ้วนที่ k ลำดับจากการค้นคืนรูปภาพ

Dataset	MRR	$\mathrm{Re}call@k$		
		$k=1$	$k=3$	$k=5$
Flickr30k	0.628	0.585	0.664	0.761
Custom	0.617	0.572	0.632	0.731

จากตารางที่ 5 ผลการประเมินประสิทธิภาพการค้นคืนรูปภาพเชิงความหมายด้วยค่าเฉลี่ยส่วนกลับของลำดับและค่าความครบถ้วนที่ k ลำดับจากการค้นคืนรูปภาพจากผู้เชี่ยวชาญทั้ง 3 ท่าน เมื่อพิจารณาลงไปในรายละเอียด พบว่า

1) ผลการวัดประสิทธิภาพในการจัดอันดับด้วยค่าเฉลี่ยส่วนกลับของลำดับ (MRR) ซึ่งจะพิจารณาความแม่นยำในการค้นคืนรูปภาพใน k อันดับแรกแล้ว จะต้องพิจารณาด้วยว่ารูปภาพที่คาดว่าจะเป็นรูปภาพที่เกี่ยวข้องกับข้อความค้นหานั้นถูกค้นคืนมาในอันดับที่ใดในจำนวน k รูปภาพที่ถูกค้นคืน หากมีค่าเข้าใกล้ 1 แสดงว่ารูปภาพที่ค้นคืนได้ใน k อันดับแรกยอมรับได้ ผลลัพธ์จากงานวิจัยนี้พบว่าค่าเฉลี่ยส่วนกลับของลำดับบนชุดข้อมูล Flickr30k และชุดข้อมูลที่ผู้วิจัยรวบรวมเองมีค่าเท่ากับ 0.628 และ 0.617 ตามลำดับ ที่ตำแหน่ง k = 5 โดยค่าเฉลี่ยส่วนกลับของลำดับจะสะท้อนถึงพฤติกรรมของผู้ใช้ต้องการเพียงผลลัพธ์ที่ตนเองสนใจเท่านั้น [39] ดังนั้นผลลัพธ์จำนวนมากจากค้นคืนอาจทำให้ผู้ใช้เสียเวลาคัดเลือกรูปภาพที่ตรงกับความต้องการอย่างแท้จริง โดย Jansen และ Spink [40] พบว่า ผู้ใช้จะเรียกดูเอกสารไม่เกิน 5 รายการที่ค้นคืน มีเพียงผู้ใช้ร้อยละ 28.3 เท่านั้นที่เรียกดูเพียง 2-3 รายการ อีกทั้งร้อยละ 55 จะเรียกดูผลลัพธ์เพียง 1 รายการต่อ 1 คำค้นหา และมีแนวโน้มจะลดลงเรื่อย ๆ ซึ่งนับเป็นความสูญเปล่าทางสารสนเทศ เนื่องจากเป็นไปได้น้อยมากที่ผู้ใช้จะเลือกดูรูปภาพจากผลการค้นคืนทั้งหมด จึงเป็นการเสียโอกาสสำหรับผู้ใช้หากมีรูปภาพบางส่วนที่ไม่ถูกเรียกดู ดังนั้นระบบค้นคืนรูปภาพจึงต้องมุ่งเพิ่มค่าความครบถ้วนอันดับ 1, 3 และ 5 ที่เกี่ยวข้องกับข้อความค้นหาเป็นหลักเพื่อตอบสนองต่อความต้องการของผู้ใช้

2) ผลการค้นคืนรูปภาพเชิงความหมายจากคำค้นที่อธิบายความหมายเชิงคุณภาพของรูปภาพที่ค่าความครบถ้วนที่ k ลำดับ จำนวน 1, 3 และ 5 รูปภาพบนชุดข้อมูล Flickr30k นั้นมีค่าความครบถ้วนเฉลี่ย 0.585, 0.664 และ 0.761 ตามลำดับ แสดงถึงการฝึกฝนล่วงหน้าแบบคอนทราสต์ระหว่างรูปภาพและข้อความของแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายนั้นสามารถแก้ปัญหาช่องว่างความหมายได้อย่างมีประสิทธิภาพ และช่วยสนับสนุนผู้ใช้ด้วยคำค้นในรูปแบบภาษาธรรมชาติที่ยึดโยงกับความหมายของรูปภาพ แทนที่จะยึดตามหลักไวยากรณ์ของภาษา อย่างไรก็ดีแนวคิดระดับสูงในรูปแบบนามธรรมนั้นมีความซับซ้อน เนื่องจากมีอารมณ์และความรู้สึกเข้ามาเกี่ยวข้อง ดังนั้นการแปลความหมายจึงเป็นไปตามหลักการรับรู้ของมนุษย์ (Human Perception) [41] ซึ่งมีที่มาจากประสบการณ์หรือความรู้เดิมของแต่ละบุคคลจึงแตกต่างกัน

3) ผลการเปรียบเทียบประสิทธิภาพการค้นคืนรูปภาพเชิงความหมายบนชุดข้อมูลที่ผู้วิจัยรวบรวมเองนั้น พบว่าค่าความครบถ้วนที่ k ลำดับ จำนวน 1, 3 และ 5 รูปภาพนั้นเท่ากับ 0.572, 0.632 และ 0.731 ตามลำดับ เมื่อเปรียบเทียบกับชุดข้อมูล Flickr30k ที่นำมาใช้ฝึกฝนตัวแบบ พบว่า ติดกันค่าความครบถ้วนที่ k จะลดลงเล็กน้อยและเป็นไปในทิศทางเดียวกัน ซึ่งค่าความถูกต้องและค่าความครบถ้วนเป็นประโยชน์ต่อผู้ใช้ต่างกลุ่มกันไป โดยผู้ใช้จะให้ความสำคัญกับค่าความถูกต้องมากกว่าค่าความครบถ้วน เนื่องจากผู้ใช้ส่วนใหญ่ต้องการให้ข้อมูลที่เป็นผลลัพธ์ที่ตรงกับสิ่ง ที่ต้องการค้นหา แต่ในทางกลับกันระบบค้นคืนรูปภาพนั้นจะเน้นไปที่ค่าความครบถ้วน เนื่องจากต้องการให้ระบบค้นคืนรูปภาพที่ถูกต้องจากชุดข้อมูลให้ได้มากที่สุด ดังนั้นงานวิจัยต่อไปจึงควรวัดประสิทธิภาพของตัวแบบด้วยค่าความถูกต้องมาประกอบด้วย

อภิปรายและสรุปผลการวิจัย

ผลการประเมินประสิทธิภาพการค้นคืนรูปภาพเชิงความหมาย พบว่า 1) ค่าเฉลี่ยส่วนกลับของลำดับ (MRR) บนชุดข้อมูล Flickr30k และชุดข้อมูลที่ผู้วิจัยรวบรวมเองมีค่าเท่ากับ 0.628 และ 0.617 ตามลำดับ ที่ตำแหน่ง k = 5 ซึ่งถือว่าอยู่ในระดับที่ยอมรับได้ เมื่อเปรียบเทียบกับงานวิจัยของ Le และคณะ [29] ที่ประยุกต์ใช้ตัวแบบ CLIP มาเรียนรู้คำอธิบายภาษาธรรมชาติเพื่อการจำแนกรถยนต์ในมุมมองที่แตกต่างกัน มีค่าเฉลี่ยส่วนกลับของลำดับเท่ากับ 0.477 งานวิจัยของ Bianchi และ คณะ [42] ที่ประยุกต์ใช้ตัวแบบ CLIP มาใช้ค้นคืนรูปภาพจากการฝึกฝนตัวแบบให้เรียนรู้ข้อความบรรยายรูปภาพที่แปลงจากภาษาอังกฤษเป็นภาษาอิตาลีที่มีค่าเฉลี่ยส่วนกลับของลำดับเท่ากับ 0.504 2) ค่าความครบถ้วนที่ k ลำดับจำนวน 1, 3 และ 5 รูปภาพบนชุดข้อมูล Flickr30k นั้นมีค่าความครบถ้วนเฉลี่ย 0.585, 0.664 และ 0.761 ตามลำดับ อย่างไรก็ดีผลลัพธ์จำนวนมากอาจทำให้ผู้ใช้เสียเวลาคัดเลือกรูปภาพที่ตรงกับความต้องการอย่างแท้จริง ดังนั้นระบบค้นคืนรูปภาพจึงต้องมุ่งเพิ่มค่าความครบถ้วนอันดับ 1, 3 และ 5 เป็นหลักเพื่อตอบสนองต่อความต้องการของผู้ใช้ที่มีแนวโน้มการเรียกดูจำนวนผลลัพธ์จากการค้นคืนที่ลดลงเรื่อย ๆ และ 3) ค่าความครบถ้วนที่ k ลำดับจำนวน 1, 3 และ 5 รูปภาพบนชุดข้อมูลที่ผู้วิจัยรวบรวมเอง เท่ากับ 0.572, 0.632 และ 0.731 ตามลำดับ เมื่อเปรียบเทียบกับผลการค้นคืนบนชุดข้อมูล Flickr30k นั้นจะลดลงเล็กน้อยแต่เป็นไปในทิศทางเดียวกัน และไม่เกิดปัญหาที่ตัวแบบจะมีประสิทธิภาพลดลงเมื่อทำงานกับข้อมูลที่ไม่เคยพบมาก่อน อย่างไรก็ดีปัญหาช่องว่างความหมายของรูปภาพนั้นยากที่จะทำให้หมดไป ทำได้เพียงลดช่องว่างความหมายระหว่างคอมพิวเตอร์ ผู้ใช้ และรูปภาพลงด้วยเทคนิคและวิธีการต่าง ๆ เท่านั้น [43] งานวิจัยต่อไปจึงควรให้ความสำคัญกับการลดช่องว่างความหมายของข้อความค้นหาจากผู้ใช้ควบคู่ไปกับการวิเคราะห์ความหมายของรูปภาพ เพื่อเพิ่มประสิ ทธิภาพในการค้นคืนให้ตรงกับความต้องการของผู้ใช้มากที่สุด

เอกสารอ้างอิง

1. Broz M. Number of Photos (2023): Statistics, Facts, & Predictions. [Internet]. 2023 [cited 2023 May 28]. Available from: https://photutorial.com/photos-statistics/

2. Tyagi V. Content-Based Image Retrieval Ideas, Influences, and Current Trends. Gateway East: Springer; 2017.
3. Alkhawlani M, Elmogy M, Elbakry H. Content-Based Image Retrieval using Local Features Descriptors and Bag-of-Visual Words. Int J Adv Comput Sci Appl 2015;6(9):212-9.
4. Marques O, Furht B. Content-based image and video retrieval. New York: Springer; 2002.
5. Barz B. Semantic and Interactive Content-based Image Retrieval. Ph.D. Dissertation, Friedrich Schiller University Jena. Germany; 2020.
6. Goodfellow I, Bengio Y, Courville A. Deep Learning. Massachusetts: MIT Press; 2016.
7. Aggarwal CC. Neural Networks and Deep Learning A Textbook. Cham: Springer; 2018.
8. McConnell S. Rapid Development: Taming Wild Software Schedules. Washington, D.C.: Microsoft Press; 1996.
9. Manning CD, Raghavan P, Schütze H. Introduction to information retrieval. Cambridge: Cambridge University Press; 2008.
10. Perret B. Hierarchical image analysis: theory, algorithms, and applications. [Internet]. 2021 [cited 2023 May 28]. Available from: https://hal.science/tel-03231061/file/HDR_ BP.pdf
11. Liu Y, Zhang J, Tjondronegoro D, Geve S. A Shape Ontology Framework for Bird Classification. In: proceedings of the 9th Biennial Conference of the Australian Pattern Recognition Society on Digital Image Computing Techniques and Applications, December 3-5, 2007; South Australia, Australia; 2007. p. 478-84.
12. Liu Y, Huang Y, Zhang S, Zhang D, Ling N. Integrating object ontology and region semantic template for crime scene investigation image retrieval. In: proceedings of the 12th IEEE Conference on Industrial Electronics and Applications (ICIEA), June 18-20, 2017; Siem Reap, Cambodia; 2017. p. 149-53.
13. Dong H, Wang Z, Qiu Q, Sapiro G. Using Text to Teach Image Retrieval. [Internet]. 2020 [cited 2023 May 28]. Available from: https://arxiv.org/pdf/2011.09928.pdf
14. Mikolajczyk A, Grochowski M. Data augmentation for improving deep learning in image classification problem. In: proceedings of the 2018 International Interdisciplinary PhD Workshop (IIPhDW), May 9-12, 2018; Swinoujscie, Poland; 2018. p. 117-22.
15. Roh Y, Heo G, Whang SE. A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective. IEEE Trans Knowl Data Eng 2019;33(4):1328-47.
16. Althnian A, AlSaeed D, Al-Baity H, Samha A, Dris AB, Alzakari N, Elwafa AA, Kurdi H. Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain. Appl Sci 2021;11(2):796-813.
17. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, Kaiser L, Polosukhin I. Attention Is All You Need. In: proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), December 4 - 9, 2017; NY, USA; 2017. p. 6000-10.

18. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: proceedings of the 9th International Conference on Learning Representations 2021 (ICLR 2021), May 3 - 7, 2021; Virtual Event, Austria; 2021. p. 1-21.
19. Plummer BA, Wang L, Cervantes CM, Caicedo JC, Hockenmaier J, Lazebnik S. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. [Internet]. 2015 [cited 2023 May 28]. Available from: https://arxiv.org/pdf/1505.04870.pdf
20. Brase CH, Brase CP. Understanding Basic Statistics. Boston: Cengage Learning; 2018.
21. Mumuni A, Mumuni F. Data augmentation: A comprehensive survey of modern approaches. Array 2022;16(2022):100258.
22. Xu M, Yoon S, Fuentes A, Park DS. A Comprehensive Survey of Image Augmentation Techniques for Deep Learning. Pattern Recognition 2023;137(2023):109347.
23. Mitchell R. Web Scraping with Python: Collecting Data from the Modern Web. 2nd ed. Sebastopol, CA: O'Reilly Media Inc.; 2018.
24. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. Learning Transferable Visual Models From Natural Language Supervision. [Internet]. 2021 [cited 2023 May 28]. Available from: https://arxiv.org/pdf/2103.00020.pdf
25. Baltrusaitis T, Ahuja C, Morency L. Multimodal Machine Learning: A Survey and Taxonomy. IEEE Trans Pattern Anal Mach Intell 2019;41(2):423-33.
26. Weers F, Shankar V, Katharopoulos A, Yang Y, Gunte T. Masked Autoencoding Does Not Help Natural Language Supervision at Scale. In: proceedings of the 2023 Conference on Computer Vision and Pattern Recognition (CVPR), June 18-22, 2023; Vancouver, Canada; 2023. p. 1-19.
27. Zizka J, Darena F, Svoboda A. Text Mining with Machine Learning Principles and Techniques. Boca Raton: CRC Press; 2020.
28. Nwankpa C, Ijomah W, Gachagan A, Marshall S. Activation Functions: Comparison of trends in Practice and Research for Deep Learning. [Internet]. 2018 [cited 2023 May 28]. Available from: https://arxiv.org/pdf/1811.03378.pdf
29. Le HD, Nguyen QQ, Nguyen VA, Nguyen TD, Chung NM, Thái T, Ha SV. Tracked-Vehicle Retrieval by Natural Language Descriptions with Domain Adaptive Knowledge. In: proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 19-20, 2022; Los Angeles, USA; 2022. p. 3300-9.
30. Xie C, Sun S, Xiong X, Zheng Y, Zhao D, Zhou J. RA-CLIP: Retrieval Augmented Contrastive Language-Image Pre-training. In: proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 17-24, 2023; Vancouver, Canada; 2023. p. 19265-74.

31. Barz B. Denzler J. Content-based Image Retrieval and the Semantic Gap in the Deep Learning Era. In: proceedings of the International Workshop on Content-Based Image Retrieval: where have we been, and where are we going (CBIR 2020), January 10, 2021; Milan, Italy; 2021. p. 2-19.
32. Srisa-ard O. Validation of measurement tools by experts. JEM-MSU 2018;1(1):45-9.
33. Christian B. The Most Human Human: What Talking with Computers Teaches Us About What It Means to Be Alive. New York: Doubleday; 2011.
34. Rosebrock A. Deep Learning for Computer Vision with Python. New York: PYIMAGESEARCH; 2017.
35. Sawarka K. Deep Learning with PyTorch Lightning Swiftly build high-performance Artificial Intelligence (AI) models using Python. Birmingham: Packt Publishing; 2022.
36. Chaudhary A. Evaluation Metrics For Information Retrieval. [Internet]. 2023 [cited 2023 May 28]. Available from: https://amitness.com/2020/08/information-retrieval-evaluation/
37. Carnevali L, Briggs J. Metrics in Information Retrieval. [Internet]. 2023 [cited 2023 May 28]. Available from: https://www.pinecone.io /learn/offline-evaluation/
38. Sangsuriyong R. Risks of Error in the Quantitative Sociology Research. JHSS BUU 2022;30(1):158-85.
39. Jansen BJ, Spink A. How are we searching the world wide web? A comparison of nine search engine transaction logs. Inform Process Manag 2006;42(1):248-63.
40. Jansen BJ, Spink A. An Analysis of Web Documents Retrieved and Viewed. In: proceedings of the 4th International Conference on Internet Computing, April 26-28, 2003; Chennai, India; 2003. p. 65-9.
41. Dix A, Finlay J, Abowd GD, Beale R. Human–Computer Interaction. 3rd ed. Harlow: Pearson; 2004.
42. Bianchi F, Attanasio G, Pisoni R, Terragni S, Sarti G, Lakshmi S. Contrastive Language-Image Pre-training for the Italian Language. [Internet]. 2021 [cited 2023 May 28]. Available from: https://arxiv.org/pdf/2108.08688.pdf
43. Liu C, Song G. A Method of Measuring the Semantic Gap in Image Retrieval: Using the Information Theory. In: proceedings of the 2011 International Conference on Image Analysis and Signal Processing (IASP 2011), October 21-23, 2011; Hubei, China; 2011. p. 1-5.