การค้นคืนรูปภาพเชิงเนื้อหาด้วยคุณลักษณะระดับสูงจากการเรียนรู้ด้วยตนเอง ของตัวแบบโครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า

A Content-based Image Retrieval by High-level Features from Self-supervised Learning of Pre-trained Deep Neural Networks Model

จักรินทร์ สันติรัตนภักดี (Chakkarin Santirattanaphakdi)*,**
และศุภกฤษฎิ์ นิวัฒนากูล (Suphakit Niwattanakul)*

Received: May 18, 2023
Revised: September 9, 2023
Accepted: September 14, 2023

* ผู้นิพนธ์ประสานงาน: จักรินทร์ สันติรัตนภักดี (Chakkarin Santirattanaphakdi) อีเมล: chakkarin_san@vu.ac.th

DOI:10.14416/j.it.2024.v1.006

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาตัวแบบการค้นคืนรูปภาพเชิงเนื้อหา เพื่อแก้ปัญหาช่องว่างความหมายที่คุณลักษณะระดับต่ำไม่สามารถสื่อความหมายของรูปภาพได้อย่างถูกต้อง ตัวแบบที่พัฒนาขึ้นประกอบด้วย 3 มอดูล ได้แก่ 1) มอดูลการสร้างชุดคำอธิบายรูปภาพ โดยใช้ตัวแบบ CLIP เพื่อเรียนรู้ความหมายของรูปภาพด้วยการเรียนรู้ด้วยตนเองจากความสัมพันธ์ระหว่างรูปภาพกับข้อความบรรยายรูปภาพจากการฝึกฝนตัวเข้ารหัสรูปภาพ และตัวเข้ารหัสข้อความด้วยค่าความคล้ายคลึงโคไซน์ ก่อนจะจัดเก็บไว้ที่ชุดคำอธิบายรูปภาพ เพื่อสร้างเป็นเวกเตอร์คุณลักษณะรูปภาพ 2) มอดูลการประมวลผลข้อความค้นหา เพื่อเรียนรู้ความหมายของข้อความ และสร้างเป็นเวกเตอร์คุณลักษณะข้อความค้นหา 3) มอดูลการจับคู่เวกเตอร์ด้วยค่าความคล้ายคลึงระหว่างเวกเตอร์คุณลักษณะรูปภาพ และเวกเตอร์คุณลักษณะข้อความค้นหา ก่อนจะเรียงลำดับตามความเกี่ยวข้อง และแสดงเป็นผลลัพธ์แก่ผู้ใช้ ผลการค้นคืนรูปภาพบนชุดข้อมูล Flickr30k แบบไม่ทราบจำนวนผลลัพธ์ที่ถูกต้องมีค่าความครบถ้วนเฉลี่ยถึง 0.93 เมื่อผลลัพธ์เป็น 10 อันดับอยู่ในระดับสูงมาก อย่างไรก็ดี ยังพบว่าอุปสรรคสำคัญต่อความถูกต้องของผลลัพธ์คือความแปรผันของรูปภาพ เมื่อเปรียบเทียบผลการค้นคืนรูปภาพกับชุดข้อมูลรูปภาพที่ผู้วิจัยรวบรวมเอง พบว่าค่าความครบถ้วนเฉลี่ยเป็นไปในทิศทางเดียวกัน และไม่เกิดปัญหาที่แบบจำลองมีประสิทธิภาพจะลดลงเมื่อทำงานกับข้อมูลที่ไม่เคยพบมาก่อน แสดงถึงตัวแบบสามารถค้นคืนรูปภาพเชิงเนื้อหาได้อย่างมีประสิทธิภาพ อีกทั้งยังสนับสนุนผู้ใช้ด้วยข้อความค้นหาในรูปแบบของภาษาธรรมชาติที่ยึดโยงกับความหมายของรูปภาพแทนที่จะยึดตามหลักไวยากรณ์ของภาษา อันจะเป็นแนวทางการค้นคืนสารสนเทศในอนาคต

คำสำคัญ: การค้นคืนรูปภาพ การเรียนรู้ด้วยตนเอง คุณลักษณะระดับสูง โครงข่ายประสาทเทียมเชิงลึก

Abstract

This research aims to developed a Content-based image retrieval model for resolve semantic gaps problem where low-level features cannot correctly convey the meaning of images. The result of developed model consists of 3 modules: 1) build the image description set module, it applies a CLIP (Contrastive Language-Image Pre-training) to learn the meaning of images by self-supervised learning from the relationship between images and caption on image encoder and text encoder with cosine similarity before collecting to the image description set and create to an image feature vector. 2) query processing module to learn the meaning of the text and constructs it as a query feature vector, and 3) vectors matching module with similar values between image feature vectors and query feature vectors before sorting by relevance and display the result to the user. The result of image retrieval on the Flickr30k dataset with order-unaware metric had a mean of recall is 0.93

* สำนักวิชาศาสตร์และศิลป์ดิจิทัล มหาวิทยาลัยเทคโนโลยีสุรนารี
* Institute of Digital Arts and Science, Suranaree University of Technology.
** สาขาวิชาเทคโนโลยีธุรกิจดิจิทัล คณะบริหารธุรกิจ มหาวิทยาลัยวงษ์ชวลิตกุล
** Department of Digital Business Technology, Faculty of Business Administrator, Vongchavalitkul University.

when the result was in the top 10 is very high, but anyway it also found that the main barrier to the accuracy of the results was image variation. When comparing the image retrieval results with the image custom dataset, it was found that the average of recall was in the same direction. And there is no problem that the model's performance is compromised when working with previously unseen data. Demonstrate that the model can retrieve content-based images effectively. It also supports users with search terms in the form of natural language that are based on the meaning of the image rather than the grammar of the language. This impact of results is a guideline for information retrieval in the future.

Keywords: Image Retrieval, Self-supervised Learning, High-Level Features, Deep Neural Network.

1. บทนำ

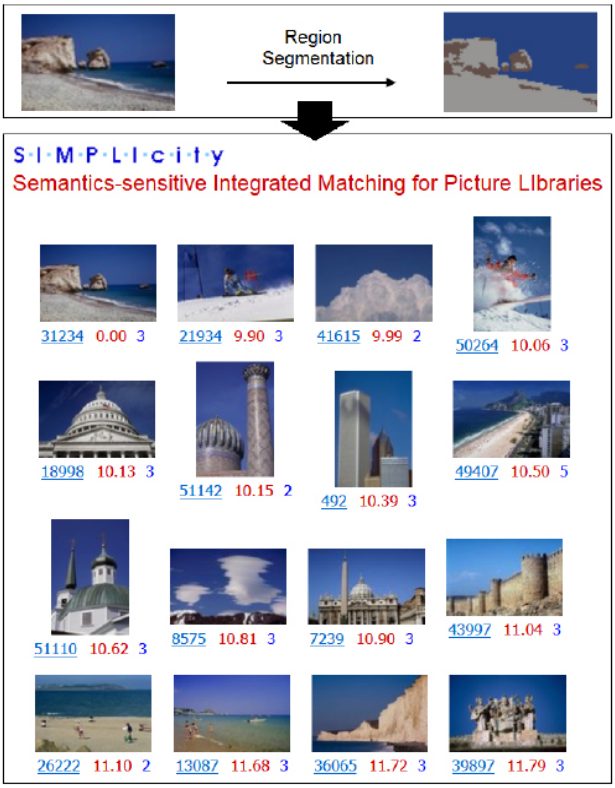
พัฒนาการของอุปกรณ์ถ่ายภาพดิจิทัลที่มีราคาถูกและสะดวกต่อการใช้งานในลักษณะอุปกรณ์พกพาอย่างสมาร์ทโฟน ตลอดจนพฤติกรรมของผู้คนที่นิยมใช้สื่อสังคมออนไลน์ในการแบ่งปันเรื่องราวและบันทึกความทรงจำ ก่อให้เกิดรูปภาพจำนวนมหาศาลบนอินเทอร์เน็ต ตามข้อมูลของเว็บไซต์ Phototutorial [1] รายงานว่ามีรูปภาพถึง 1.81 ล้านล้านภาพในปี 2023 และจะเพิ่มขึ้นมากกว่า 2 ล้านล้านภาพในปี 2025 โดยในปี 2030 คาดการณ์ว่าจะมีรูปภาพทั่วโลกถึง 2.33 ล้านภาพ และมีแนวโน้มที่จะเพิ่มมากขึ้นอีกในอนาคต รูปภาพจำนวนมหาศาลจึงเป็นอุปสรรคต่อการค้นคืนรูปภาพ

แนวคิดการค้นคืนรูปภาพเริ่มจากการเปรียบเทียบคำสำคัญ (Keyword) กับชื่อไฟล์หรือคำอธิบายภาพ (String Matching) [2] แต่บางครั้งการตั้งชื่อไฟล์ไม่ตรงกับเนื้อหาของรูปภาพ หรือการใส่คำอธิบายภาพโดยบุคคลนั้นไม่อาจอธิบายครอบคลุมเนื้อหารูปภาพได้ทั้งหมด ทำให้ผลการค้นคืนมีประสิทธิภาพต่ำ ต่อมาจึงมีแนวคิดการดึงเอาคุณลักษณะระดับต่ำ (Low-level Features) [2] เช่น รูปร่าง สี และพื้นผิวในบริเวณต่าง ๆ ของรูปภาพออกมา เพื่อใช้เปรียบเทียบกับคำค้นโดยตรง โดยไม่มีการแปลงให้อยู่ในรูปแบบของข้อมูลที่มนุษย์สามารถเข้าใจได้

งานวิจัยที่เกี่ยวกับการค้นคืนรูปภาพจากฐานข้อมูลโดยทั่วไปจะระบุภาพคำถาม (Query Image) [2] เพื่อคัดแยกคุณลักษณะระดับต่ำ จากนั้นจะนำไปเปรียบเทียบกับคุณลักษณะของแต่ละภาพที่เก็บไว้ในฐานข้อมูล โดยอาจจะพิจารณาจากค่าสี พื้นผิว รูปทรง หรือจากหลาย ๆ ค่าร่วมกัน หากภาพใดมีคุณลักษณะที่ใกล้เคียงกับภาพคำถามจะถูกค้นคืนเป็นผลลัพธ์ให้แก่ผู้ใช้

อย่างไรก็ตาม ยังคงพบว่ามีข้อจำกัดหลายประการที่ส่งผลต่อประสิทธิภาพของการค้นคืนรูปภาพให้ตรงความต้องการของผู้ใช้ ได้แก่

1) การสืบค้นด้วยภาพคำถามนั้นผลลัพธ์ส่วนใหญ่จะเป็นภาพที่มีคุณลักษณะไม่ซับซ้อนมากนัก เช่น ภาพชายทะเล เมื่อแบ่งส่วนภาพ (Region Segmentation) ตามโทนสีของพื้นที่ [3] จะได้ 2 สีหลักคือ สีฟ้าของน้ำทะเล และสีเทาของชายหาด ดังภาพที่ 1



ภาพที่ 1 *ผลลัพธ์จากการค้นคืนรูปภาพจากคุณลักษณะระดับต่ำของตัวแบบ* SIMPLIcity [3]

จากภาพที่ 1 เมื่อเปรียบเทียบภาพคำถามกับคุณลักษณะรูปภาพในฐานข้อมูล ผลลัพธ์ไม่เพียงจะพบรูปภาพที่มีสีฟ้าของน้ำทะเล และสีเทาของชายหาดเท่านั้น แต่ยังพบรูปภาพอื่น เช่น หิมะ ก้อนเมฆ หรือสิ่งปลูกสร้างที่มีสีเทา ร่วมกับสีฟ้าของท้องฟ้าร่วมเป็นผลลัพธ์ เนื่องจากตัวแบบไม่สามารถพิจารณาความเหมือนกับภาพคำถามด้วยคุณลักษณะระดับต่ำเพียงอย่างเดียวได้

2) ผู้ใช้จำนวนไม่น้อยที่ขาดความชำนาญในการใช้คำค้น ส่วนใหญ่มักใช้คำหลัก (Keyword) เพียง 1-2 คำในการค้นหาในแต่ละครั้ง ประกอบกับบางอัลกอริทึมยังไม่เอื้อต่อการค้นหาด้วยข้อความในรูปแบบภาษาธรรมชาติ (Natural Language) เกิดเป็นผลลัพธ์จำนวนมากทั้งที่ตรงกับความต้องการและไม่ตรงกับความต้องการของผู้ใช้

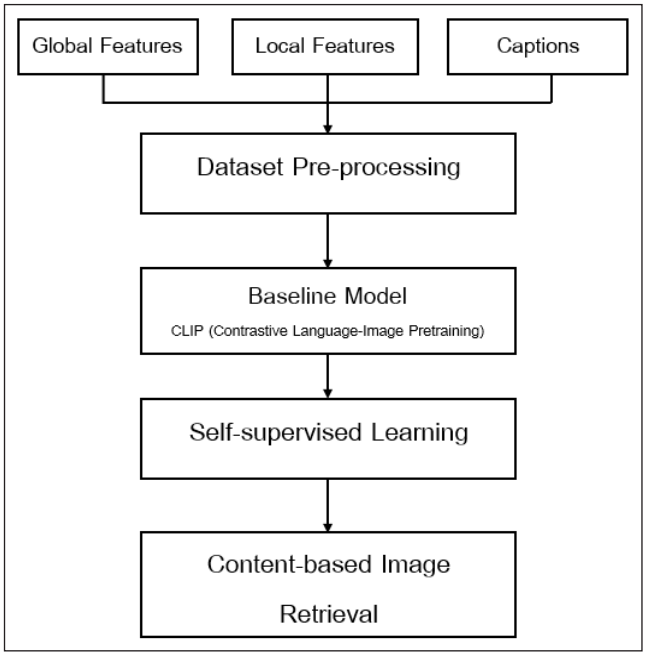
3) ปริมาณผลลัพธ์ที่มากเกินไปอาจทำให้ผู้ใช้เสียเวลาในการคัดเลือกรูปภาพที่ตรงกับความต้องการ ซึ่งเป็นได้น้อยมากที่ผู้ใช้จะเลือกดูรูปภาพทั้งหมดจากผลการค้นคืน จึงเป็นการเสียโอกาสสำหรับผู้ใช้หากมีรูปภาพบางส่วนที่ไม่ถูกเรียกดู

เนื่องจากพฤติกรรมการค้นหาของผู้ใช้นั้นจะใช้เนื้อหาหรือคอนเซ็ปต์ในรูปภาพเป็นหลัก จึงอาจเกิดปัญหาช่องว่างความหมาย (Semantic Gap) เนื่องจากคุณลักษณะเหล่านี้มิได้สื่อความหมายที่แท้จริงของรูปภาพ เกิดเป็นแนวคิดที่มุ่งแปลความหมายจากคุณลักษณะระดับต่ำเป็นคุณลักษณะระดับสูง (High-level Features) เพื่อระบุถึงวัตถุ (Object) หรือเหตุการณ์ (Event) ของรูปภาพ เป็นที่มาของการค้นคืนรูปภาพเชิงเนื้อหา (Content-based Image Retrieval: CBIR) [2]

ปัจจุบันมีแนวคิดเกี่ยวกับการสร้างตัวแปลความหมายระดับสูง (High-level Interpretation) เพื่อเชื่อมโยงคุณลักษณะระดับต่ำของรูปภาพไปสู่คุณลักษณะระดับสูง โดยประยุกต์หลายศาสตร์บูรณาการร่วมกัน เพื่อแก้ปัญหาช่องว่างความหมายและสร้างดัชนีรองรับการแปลความหมายระดับสูง แต่พบว่ายังไม่มีแนวคิดใดที่สามารถให้ความหมายของภาพได้อย่างเพียงพอ และมีข้อจำกัดมากมายเมื่อต้องรับมือกับเนื้อหาจำนวนมาก ดังนั้นผลลัพธ์ของการค้นคืนรูปภาพจึงยังห่างไกลจากความคาดหวังของผู้ใช้

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

งานวิจัยนี้ได้ศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้องตามกรอบการพัฒนาแบบจำลองการค้นคืนรูปภาพเชิงเนื้อหาจากเรียนรู้ด้วยตนเอง โดยใช้ตัวแบบโครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า ดังภาพที่ 2



ภาพที่ 2 *กรอบการพัฒนาแบบจำลองการค้นคืนรูปภาพเชิงเนื้อหาจากเรียนรู้ด้วยตนเอง โดยใช้ตัวแบบโครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า*

2.1 คุณลักษณะระดับต่ำของรูปภาพ

คุณลักษณะระดับต่ำของรูปภาพ ที่มักนำมาใช้ในการค้นคืนรูปภาพเชิงเนื้อหา แบ่งออกเป็น 2 ประเภทคือ

2.1.1 คุณลักษณะแบบโกลบอล (Global Features) [4] เป็นคุณลักษณะแบบรวม ๆ ไม่เจาะจง จุดเด่นคือความง่าย ไม่ซับซ้อน และประมวลผลได้รวดเร็ว ประกอบด้วย

2.1.1.1 คุณลักษณะสี (Color Features) เป็นคุณลักษณะสำคัญในการค้นคืนรูปภาพ คุณลักษณะสีคำนวณตามช่องว่างสี (Color Spaces) ระบบสีที่นิยม ได้แก่ ระบบสี RGB (Red Green Blue) ใช้แสดงผลทางจอภาพ ระบบสี CMYK (Cyan Magenta Yellow Black) ที่ใช้ในงานพิมพ์ และระบบสี HSV (Hue Saturation Value) ซึ่งสามารถแยกองค์ประกอบสีออกจากความเข้มของแสงได้ จึงใกล้เคียงกับการรับรู้ของมนุษย์มากที่สุด

2.1.1.2 คุณลักษณะรูปทรง (Shape Features) ใช้การแยกรูปทรงออกจากพื้นที่หรือขอบเขต (Region or Boundary) เพื่อระบุวัตถุ การแปลความหมายจะแตกต่างกันไปตามมาตราส่วน

2.1.1.3 คุณลักษณะลักษณะพื้นผิว (Texture Features) เป็นคุณลักษณะที่เห็นได้ชัดเจน แต่ไม่สามารถระบุคุณลักษณะได้ด้วยความเข้ม หรือรหัสสี มักใช้ในการค้นคืนรูปภาพและการจดจำรูปแบบ แต่ยังมีข้อจำกัดในการวิเคราะห์พื้นผิว คือปัญหาความซับซ้อนในการคำนวณ และความไวต่อสัญญาณรบกวน

2.1.1.4 คุณลักษณะข้อมูลเชิงพื้นที่ (Spatial Information Feature) ซึ่งเกี่ยวข้องกับตำแหน่งของวัตถุในภาพสองมิติ ใช้ร่วมกับคุณลักษณะอื่น ๆ เพื่อแปลความหมายเป็นคุณลักษณะระดับสูง

2.1.2 คุณลักษณะแบบโลคอล (Local Features) [5] เป็นคุณลักษณะเฉพาะของแต่ละส่วนในรูปภาพ ได้แก่

2.1.2.1 SIFT (Scale Invariant Feature Transform) เป็นคุณลักษณะที่ไม่เปลี่ยนแปลงตามมาตราส่วน จากการหาจุดสำคัญในรูปภาพ (Keypoints) แสดงเป็นเว็กเตอร์ขนาด 128 มักใช้จำแนกประเภทข้อมูลภาพ เนื่องจากมีประสิทธิภาพต่อการหมุนภาพ และการปรับขนาดภาพ อย่างไรก็ดี ยังมีข้อเสียคือใช้หน่วยความจำจำนวนมาก และมีค่าใช้จ่ายในการคำนวณสูง

2.1.2.2 SURF (Speeded-up Robust Features) เป็นตัวบ่งชี้คุณลักษณะที่มีประสิทธิภาพสูงกว่า SIFT โดยใช้เทคนิคการลดขนาด จึงใช้เวลาน้อยกว่าในการคำนวณคุณลักษณะ และจัดทำดัชนีตามสัญลักษณ์ Laplacian อย่างไรก็ตาม จะมีประสิทธิภาพลดลงเมื่อเกิดการหมุน (Poorly in Rotation)

2.1.2.3 LBP (Local Binary Pattern) ใช้เทคนิคเปรียบเทียบพิกเซลกลางกับ 8 พิกเซลโดยรอบ จึงมีประสิทธิภาพสูงเนื่องจากไม่แปรผันกับการเปลี่ยนแปลงแบบโมโนโทนิกในโทนสีเทา (Monotonic Transformations in the Grayscale) แต่มีข้อจำกัดคือจะสูญเสียข้อมูลเชิงพื้นของคุณลักษณะแบบโกลบอล

2.2 ข้อความบรรยายรูปภาพ (Caption)

การอธิบายเนื้อหาของภาพด้วยข้อความบรรยายรูปภาพนำไปใช้กับการมองเห็นของคอมพิวเตอร์ และการประมวลผลภาษาธรรมชาติ ส่วนใหญ่จะใช้เฟรมเวิร์กตัวเข้ารหัส-ตัวถอดรหัส โดยที่รูปภาพอินพุตจะถูกเข้ารหัสเป็นตัวแทนสื่อกลางของข้อมูลในภาพ จากนั้นจึงถอดรหัสเป็นลำดับข้อความอธิบาย

2.3 การเตรียมข้อมูลก่อนการประมวลผล (Pre-processing)

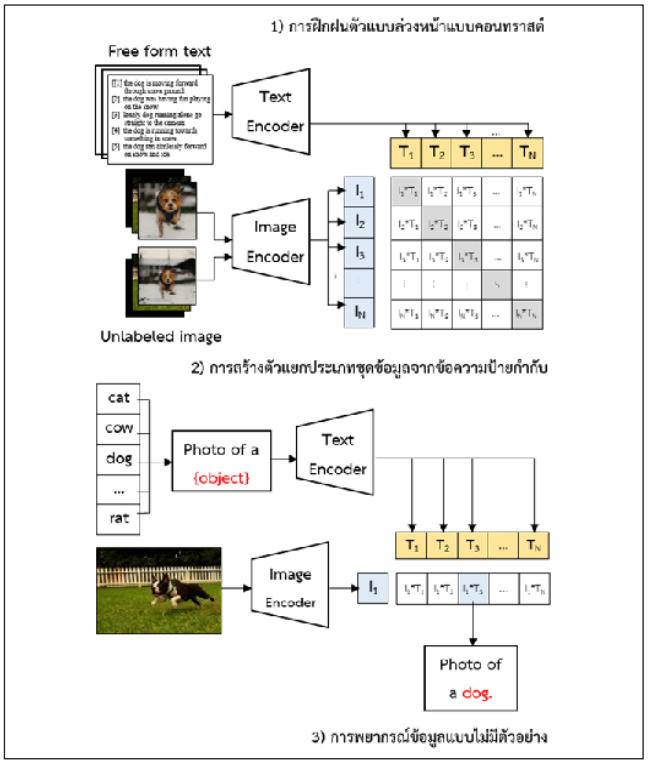
2.3.1 การสร้างรูปภาพใหม่จากการดัดแปลงรูปภาพต้นฉบับ (Image Augmentation) [6] เช่น การบิด ตัด หมุน เปลี่ยนสี ทำภาพให้มืดลงหรือสว่างขึ้น หรือใส่ตัวรบกวน (Noise) ลงไป เช่น เพิ่มรูปสุนัขที่เห็นเพียงช่วงครึ่งของหน้า รูปสุนัขที่ถูกหมุน รูปสุนัขที่ถูกปรับแสง หรือรูปสุนัขที่มีสิ่งปกปิดหรือรบกวน เป็นต้น รูปสุนัขเหล่านี้จะทำให้แบบจำลองจดจำสิ่งที่ไม่จำเป็นได้ยากขึ้น และจดจำคุณลักษณะสำคัญของภาพสุนัขได้ดีขึ้น เช่น รูปทรงของตา จะมีลักษณะทรงกลมหรือทรงรีสีดำ ขนาดใดก็ได้ เป็นต้น เพื่อแก้ปัญหา Overfitting เนื่องจากประสิทธิภาพความแม่นยำของตัวแบบการเรียนรู้เชิงลึกนั้นขึ้นกับปริมาณข้อมูลเป็นปัจจัยสำคัญอันดับหนึ่ง

2.3.2 การทำความสะอาดข้อความ และการกำหนดมาตรฐานข้อความ (Text Cleansing and Text Normalization) เนื่องจากข้อความบรรยายรูปภาพที่นำมาจากอินเทอร์เน็ตมักจะอยู่ในรูปแบบที่ไม่เป็นระเบียบ (Messy Data) [7] ดังนั้นจึงต้องกำหนดรูปแบบอักขระ (Regular Expression) เพื่อให้ข้อมูลอยู่ในรูปแบบที่เหมาะสมต่อการนำไปประมวลผล งานวิจัยนี้ถือว่าข้อความทั้งหมดมีผลต่อการกำหนดคุณลักษณะระดับสูงของรูปภาพ จึงไม่มีกระบวนการกำจัดคำหยุด เป็นไปในทิศทางเดียวกับแนวโน้มการกำจัดคำหยุดที่มีจำนวนคำลดลงเรื่อย ๆ จนในปัจจุบันแนวคิดการเรียนรู้เชิงลึกไม่มีการกำจัดคำหยุดเลย อย่างไรก็ตาม การเตรียมข้อมูลก่อนการประมวลผลยังคงกำจัดเครื่องหมายต่าง ๆ และเปลี่ยนตัวอักษรเป็นตัวพิมพ์เล็กทั้งหมด ตลอดจนปรับโครงสร้างข้อความบรรยายรูปภาพให้อยู่ในเซลล์เดียวกันทั้งประโยคก่อนจะนำไปประมวลผล

2.4 ตัวแบบ CLIP ที่ถูกฝึกฝนล่วงหน้า

Radford และคณะ [8] นำเสนอตัวแบบ CLIP (Contrastive Language–Image Pre-training) ที่เป็นตัวแบบโครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้าจากคู่รูปภาพและข้อความในรูปแบบภาษาธรรมชาติบนอินเทอร์เน็ตจำนวน 400 ล้านคู่ พัฒนาโดย OpenAI เผยแพร่ในลักษณะซอฟต์แวร์โอเพนซอร์ซ (Opensource) ดังภาพที่ 3 การทำงานประกอบด้วย 3 ขั้นตอน ได้แก่

1) การฝึกฝนตัวแบบล่วงหน้าแบบคอนทราสต์ [9] เพื่อฝึกฝนตัวเข้ารหัสรูปภาพ และตัวเข้ารหัสข้อความด้วยค่าความคล้ายคลึงเชิงมุมโคไซน์ [10] ระหว่างรูปภาพและข้อความบรรยายรูปภาพคู่เดียวกัน ให้มีค่าเข้าใกล้กับ 1 มากขึ้น และขยับตัวแทนของรูปภาพ และข้อความบรรยายรูปภาพคู่ที่แตกต่างกันให้ค่าความคล้ายคลึงลดลงเรื่อย ๆ จนมีค่าเข้าใกล้กับ 0



ภาพที่ 3 *การทำงานของตัวแบบ CLIP ที่ถูกฝึกฝนล่วงหน้า [8]*

2) การสร้างตัวแยกประเภทชุดข้อมูลจากข้อความป้ายกำกับ (Create Dataset Classifier from Label Text) โดยฝังอ็อบเจกต์คลาสลงในข้อความบรรยายรูปภาพ (Object Classes into Captions) จากภาพที่ 3 อ็อบเจกต์คลาสที่ฝังเป็นข้อความบรรยายรูปภาพ คือ "a photo of a dog" เพื่อนำไปใช้ในการค้นคืนรูปภาพต่อไป

3) การพยากรณ์ข้อมูลแบบไม่มีตัวอย่าง (Use for Zero-shot Prediction) เป็นการฝึกฝนแบบจำลองให้สามารถจัดหมวดหมู่วัตถุที่ไม่เคยเรียนรู้มาก่อนคล้ายกับมนุษย์ที่สามารถแยกประเภทสิ่งของที่แตกต่างกันได้โดยที่ไม่เคยเห็นมาก่อน [11] โดยนำรูปภาพที่ผ่านตัวเข้ารหัสรูปภาพ เพื่อสร้างเวกเตอร์ฝังตัว (Embedding Vector) เพื่อวัดความคล้ายคลึงเชิงมุมโคไซน์ระหว่างเวกเตอร์การฝังของรูปภาพ และเวกเตอร์การฝังข้อความ โดยค่าที่มีความคล้ายคลึงโคไซน์สูงสุดคือระดับที่คาดการณ์ไว้ จากภาพที่ 3 ผลลัพธ์ที่มีความคล้ายคลึงสูงสุดจากการพยากรณ์จะเป็น "รูปภาพสุนัข (a photo of a dog)"

งานวิจัยนี้ประยุกต์ใช้ตัวแบบ CLIP ที่ถูกฝึกฝนล่วงหน้าเป็นโมเดลพื้นฐาน (Baseline Model) ในการสกัดคุณลักษณะโกลบอล และคุณลักษณะโลคอลจากรูปภาพ ร่วมกับข้อความบรรยายรูปภาพผ่านการเตรียมข้อมูลก่อนการประมวลผลด้วยการฝึกฝนแบบคอนทราสต์ตามแนวคิดเรียนรู้ด้วยตนเอง เพื่อแปลความหมายเป็นคุณลักษณะระดับสูงและสร้างดัชนีคุณลักษณะ (Index Object) สำหรับการค้นคืนรูปภาพเชิงเนื้อหาก่อนจะเก็บไว้ในชุดคำอธิบายรูปภาพ (Image Description Set)

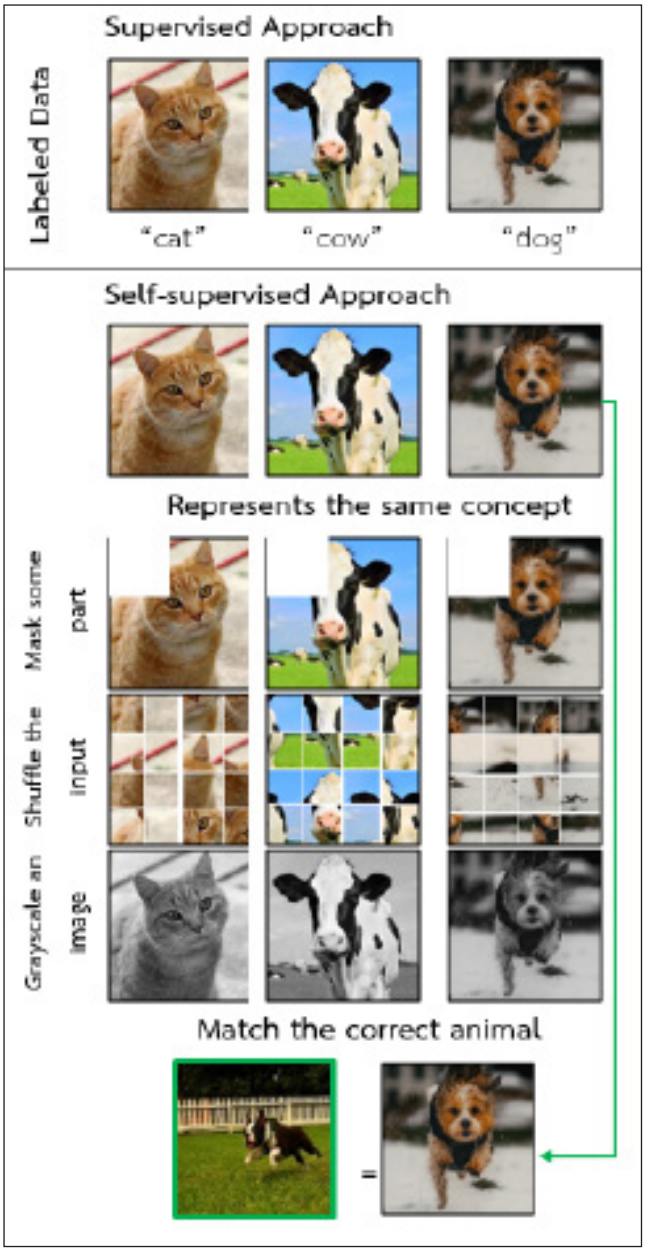
2.5 การเรียนรู้ด้วยตนเอง

การเรียนรู้ด้วยตนเอง (Self-supervised Learning) [12] เป็นรูปแบบการเรียนรู้แบบไม่มีผู้สอนบนชุดข้อมูลแบบติดป้ายกำกับอัตโนมัติ (Automatically Labeled) เนื่องจากแบบจำลองมีการเรียนรู้จากข้อมูลที่ไม่จำเป็นต้องมีป้ายกำกับล่วงหน้า แต่ต้องมีป้ายกำกับสำหรับการค้นหาและใช้ประโยชน์จากความสัมพันธ์ (Relations) หรือความเกี่ยวข้อง (Correlations) ระหว่างคุณสมบัติของข้อมูลอินพุตที่แตกต่างกัน โดยบังคับให้แบบจำลองเรียนรู้การแสดงความหมายเกี่ยวกับข้อมูลด้วยการเรียนรู้ความสัมพันธ์ระหว่างรูปภาพกับข้อความภาษาธรรมชาติ แทนที่จะเรียนรู้ภายใต้คลาสที่จำกัดอย่างสุนัขหรือแมวเช่นเดิม เมื่อตัวแบบได้รับการฝึกฝนจากข้อความทั้งประโยคจะส่งผลให้ตัวแบบสามารถเรียนรู้สิ่งต่าง ๆ ได้มากขึ้นและสามารถค้นหารูปแบบความสัมพันธ์บางอย่างระหว่างรูปภาพกับข้อความด้วยการเรียนรู้ด้วยตนเอง จากนั้นชุดความรู้ (Knowledge) ที่เป็นผลลัพธ์จากการเรียนรู้จะถูกถ่ายโอนการเรียนรู้ไปยังแบบจำลองเพื่อใช้สำหรับงานหลัก (Main Task)

จากภาพที่ 4 แนวคิดการทำความเข้าใจรูปภาพจากการเรียนรู้คุณลักษณะเฉพาะของแต่ละรูปภาพ เช่น รูปภาพสุนัข จากนั้นจะฝึกฝนตัวแบบให้เรียนรู้เพื่อจดจำรูปภาพสุนัขอื่น ๆ ที่มีรูปแบบหรือโครงสร้างที่เหมือนกัน เช่นเดียวกับเรียนรู้ของเด็กที่แม้ยังไม่สามารถสื่อสาร หรือเข้าใจภาษาได้ดูรูปสุนัขแล้วให้เลือกว่าภาพใดคล้ายคลึงมากที่สุดเมื่อเปรียบเทียบกับรูปภาพสัตว์ต่าง ๆ เป็นไปได้มากที่เด็กจะเลือกรูปภาพสุนัขได้อย่างถูกต้อง จะเห็นได้ว่าการเรียนรู้ด้วยตนเองนั้นไม่จำเป็นต้องติดป้ายกำกับเพื่อแบ่งคลาสใด ๆ เหมือนเรียนรู้แบบมีผู้สอน (Supervised Learning) ดังนั้นจึงถูกนำมาใช้กับการฝึกฝนแบบคอนทราสต์ เพื่อฝึกฝนแบบจำลองให้สามารถจัดหมวดหมู่วัตถุที่ไม่เคยเห็นมาก่อน (Zero-shot Prediction) [13] คล้ายกับมนุษย์ที่สามารถแยกประเภทสิ่งของที่แตกต่างกันได้โดยที่ไม่มีการสอนหรือเรียนรู้มาก่อน

2.6 การค้นคืนรูปภาพเชิงเนื้อหา

การค้นคืนรูปภาพด้วยคุณลักษณะระดับต่ำ เช่น สี รูปทรง พื้นผิว และพื้นที่ ซึ่งเป็นคุณลักษณะแบบรวม ๆ ไม่เจาะจง

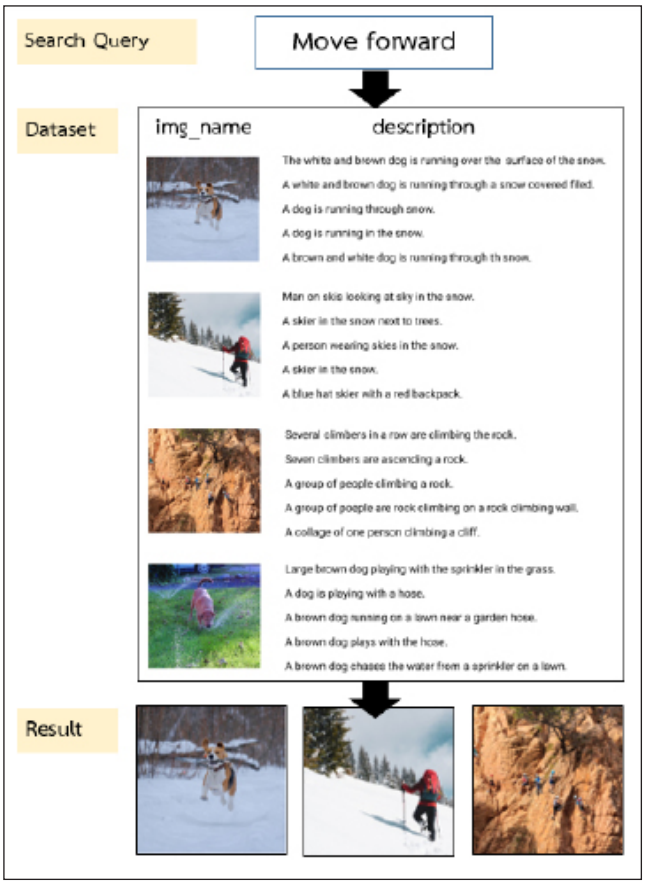


ภาพที่ 4 *แนวคิดการเรียนรู้ด้วยตนเองเปรียบเทียบกับแนวคิดการเรียนรู้แบบมีผู้สอน*

หรือคุณลักษณะแบบโกลบอลร่วมกับคุณลักษณะแบบโลคอลนั้น การแปลความหมายจึงต้องใช้หลายคุณลักษณะร่วมกัน เช่น รูปทรงกลมสีส้ม หากอยู่บริเวณด้านบนของรูปภาพอาจถูกตีความว่าเป็นดวงอาทิตย์ ซึ่งความเป็นจริงแล้วอาจเป็นไข่แดงของไข่ดาว หรือแม้แต่ลูกบาสเก็ตบอลก็เป็นได้ ดังนั้น ผลลัพธ์จึงมีประสิทธิภาพไม่สูงเท่าที่ควร เนื่องจากคุณลักษณะระดับต่ำเหล่านี้ ไม่สามารถสื่อความหมายของรูปภาพได้อย่างถูกต้องเรียกว่าปัญหาช่องว่างความหมาย

อย่างไรก็ดี พฤติกรรมการค้นหาของผู้ใช้จะใช้เนื้อหาหรือคอนเซ็ปต์ในรูปภาพเป็นหลัก เป็นที่มาของแนวคิดการค้นคืนรูปภาพจากเนื้อหาที่มุ่งเชื่อมโยงคุณลักษณะระดับต่ำของรูปภาพให้กลายเป็นคุณลักษณะระดับสูง เพื่อระบุถึงวัตถุหรือเหตุการณ์ของรูปภาพในรูปแบบของข้อมูลที่มนุษย์สามารถเข้าใจได้ [14]

ปัจจุบันการค้นคืนรูปภาพเชิงเนื้อหาทำงานได้อย่างมีประสิทธิภาพ จากการค้นหาภาพในฐานข้อมูลที่มีคุณลักษณะคล้ายคลึงกับคำค้นหามาแสดงเป็นผลลัพธ์ ดังภาพที่ 5 ผลลัพธ์จากการค้นคืนด้วยข้อความ "a dog running on a snow" จะเป็นรูปภาพที่ระบุถึงวัตถุหรือเหตุการณ์ที่เป็นเนื้อหาหลักของรูปภาพ



ภาพที่ 5 *การค้นคืนรูปภาพจากคุณลักษณะระดับสูง*

2.7 งานวิจัยที่เกี่ยวข้อง

แนวคิดที่มุ่งเชื่อมโยงคุณลักษณะระดับต่ำของรูปภาพให้กลายเป็นคุณลักษณะระดับสูง จำแนกตามเทคนิค และวิธีการเป็น 6 กลุ่ม ได้แก่

1) การตอบกลับของผู้ใช้ (Relevance Feedback) [15] ว่ามีข้อมูลใดบ้างที่ตรงกับสิ่งที่กำลังค้นหาอยู่เป็นผลลัพธ์

ในการค้นหาครั้งแรก เพื่อปรับปรุงผลลัพธ์ในครั้งต่อไปให้ตรงกับความต้องการของผู้ใช้มากยิ่งขึ้น

2) การค้นคืนรูปภาพหลายระดับ (Multi-Level Image Retrieval) โดยแทนที่คุณลักษณะของภาพด้วยคำอธิบายที่เป็นลำดับชั้น (Hierarchical Image Analysis) [16] ที่ง่ายต่อการทำความเข้าใจเช่นเดียวกับการรับรู้ของมนุษย์

3) การใช้คำอธิบายภาพ (Annotation) [17] ที่พิจารณาความสัมพันธ์ของรูปภาพที่มีความเกี่ยวข้องกัน แม้ว่าจะมีคุณลักษณะระดับต่ำที่แตกต่างกัน เช่น รูปภาพของรถและถนนที่มีสีหรือพื้นผิวที่ต่างกัน แต่ถือว่ามีความสัมพันธ์กัน เนื่องจากรถเป็นพาหนะที่ใช้บนท้องถนน เป็นต้น หรือเรียกว่าออนโทโลยี (Ontology) [18]

4) การใช้แม่แบบเชิงความหมาย (Semantic Template: ST) [19] เพื่อสร้างตัวแทนคุณลักษณะ (Representative Feature) ต่าง ๆ ของรูปภาพ ซึ่งผู้ใช้ต้องมีความรู้เกี่ยวกับคุณลักษณะรูปภาพอย่างลึกซึ้ง จึงยากต่อการใช้งานปกติ

5) การใช้ข้อความเพื่อเรียนรู้การค้นคืนรูปภาพ (Using Text to Teach Image Retrieval) [20] จากข้อมูลที่ปรากฏบนเว็บเพจ เช่น ข้อความโดยรอบ ลิงก์ แท็ก หรือคำอธิบายร่วมกันแปลความหมายคุณลักษณะรูปภาพ

6) การจัดกลุ่มข้อมูลด้วยการเรียนรู้ของเครื่อง (Machine Learning Classification) ปัจจุบันพัฒนาเป็นการเรียนรู้เชิงลึก (Deep Learning) [21] ซึ่งวิเคราะห์และจำแนกคุณลักษณะรูปภาพได้อย่างมีประสิทธิภาพกว่าการเรียนรู้ของเครื่องแบบเดิมเป็นอย่างมาก

3. วิธีดำเนินการวิจัย

งานวิจัยนี้ประยุกต์ใช้แนวคิดการพัฒนาแอปพลิเคชันแบบเร่งรัด (Rapid Application Development model: RAD model) [22] มาเป็นกรอบในการพัฒนาตัวแบบ ประกอบด้วย 4 ขั้นตอน ได้แก่

1) การวางแผนความต้องการ (Requirement Planning) แม้การเรียนรู้เชิงลึกมีบทบาทสำคัญต่อการประมวลผลรูปภาพ แต่ปัจจุบันยังพบปัญหาหลายประการ ได้แก่

1.1) ต้องการชุดข้อมูลฝึกฝนในปริมาณที่เพียงพอ [23] เพื่อป้องกันปัญหา Overfitting ที่ตัวแบบมีประสิทธิภาพสูงในการจำแนกข้อมูลกับชุดข้อมูลทดสอบ แต่ประสิทธิภาพจะลดลงเมื่อทำงานกับข้อมูลที่ไม่เคยพบมาก่อน (Poor real-world Performance)

1.2) การสร้างชุดข้อมูลฝึกฝนที่มีคุณภาพนั้นมีต้นทุนมหาศาล (Costly Datasets) [24] เนื่องจากต้องใช้มนุษย์ในการติดป้ายกำกับ (Data Labeling)

1.3) ตัวแบบจะถูกฝึกฝนบนชุดข้อมูลที่มีลักษณะตายตัว ไม่สามารถเปลี่ยนแปลงได้ระหว่างการเรียนรู้หากมีการเปลี่ยนแปลงต้องเริ่มเรียนรู้ใหม่ตั้งแต่ต้น

จากปัญหาดังกล่าวเป็นที่มาของการใช้งานตัวแบบที่ถูกฝึกฝนล่วงหน้า (Pre-trained) ด้วยเทคนิคถ่ายโอนการเรียนรู้ (Transfer Learning) และการปรับแต่งพารามิเตอร์ (Fine-tuning) [25] แล้วเรียนรู้ซ้ำ (Retrain Network) บนชุดข้อมูลที่กำหนดเอง เพื่อลดภาระการฝึกฝนตัวแบบจากเดิมที่ต้องการข้อมูลจำนวนมาก อีกทั้งมีประสิทธิภาพดีกว่าตัวแบบที่ถูกสร้างขึ้นมาเองตั้งแต่ต้น การพัฒนาตัวแบบการค้นคืนรูปภาพเชิงเนื้อหาด้วยคุณลักษณะระดับสูงจากการเรียนรู้ด้วยตนเองของตัวแบบโครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า เพื่อแก้ปัญหาช่องว่างความหมาย และสนับสนุนผู้ใช้ด้วยคำค้นหาในรูปแบบภาษาธรรมชาติที่ยึดโยงกับความหมายของรูปภาพแทนที่จะยึดตามไวยากรณ์ของภาษา อันจะเป็นแนวทางการค้นคืนสารสนเทศในอนาคต

2) การออกแบบ (Design) งานวิจัยนี้ประยุกต์ใช้ตัวแบบ CLIP เพื่อเรียนรู้ความสัมพันธ์ระหว่างรูปภาพกับข้อความบรรยายรูปภาพที่อยู่ในลักษณะภาษาธรรมชาติ แทนที่จะเรียนรู้ภายใต้คลาสที่จำกัดเช่นเดิมอย่างสุนัขหรือแมว ดังนั้น เมื่อตัวแบบได้รับการฝึกฝนทั้งประโยค จะส่งผลให้ตัวแบบสามารถเรียนรู้สิ่งต่าง ๆ ได้มากขึ้น และค้นหารูปแบบความสัมพันธ์บางอย่างระหว่างรูปภาพกับข้อความด้วยการเรียนรู้ด้วยตนเอง ดังภาพที่ 6

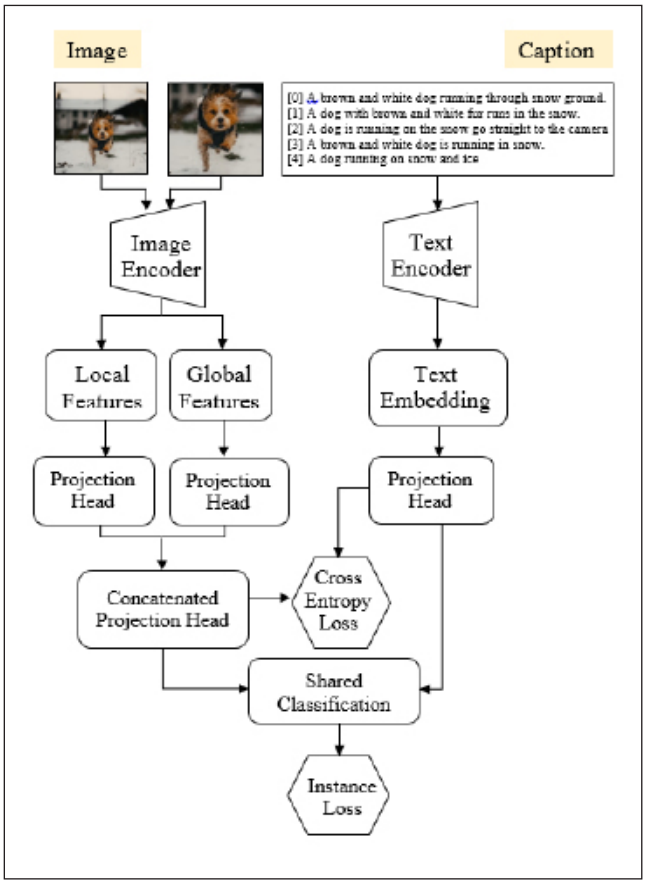
การออกแบบโครงข่ายงานหลัก เพื่อสร้างชุดคำอธิบายรูปภาพแบ่งออกเป็น 3 มอดูล ได้แก่

1) มอดูลการจัดการชุดข้อมูล (Dataset Management Modules) งานวิจัยนี้เรียกใช้ชุดข้อมูล Flickr30k บน Google Colaboratory ประกอบด้วย

1.1) การจัดการกับรูปภาพด้วยการสร้างรูปภาพใหม่จากการดัดแปลงรูปภาพต้นฉบับ เช่น การบดบังรูปภาพบางส่วน การแยกภาพออกเป็นส่วนย่อย ๆ หรือการปรับภาพให้อยู่ในโหมดสีเทา เป็นต้น เพื่อแก้ปัญหา Overfitting ด้วยการเพิ่มจำนวนชุดข้อมูลฝึกสอนให้ตัวแบบจดจำคุณลักษณะสำคัญของรูปภาพได้ดียิ่งขึ้น

1.2) การจัดการกับข้อความบรรยายรูปภาพด้วย

การกำจัดเครื่องหมายต่าง ๆ และเปลี่ยนเป็นตัวอักษรพิมพ์เล็กทั้งหมด ตลอดจนปรับข้อความให้อยู่ในรูปแบบที่เหมาะสมต่อการใช้งาน



ภาพที่ 6 *การออกแบบโครงข่ายงานหลักเพื่อสร้างชุดคำอธิบายรูปภาพ*

2) มอดูลการสร้างตัวเข้ารหัส (Build the Encoder Modules) ประกอบด้วย

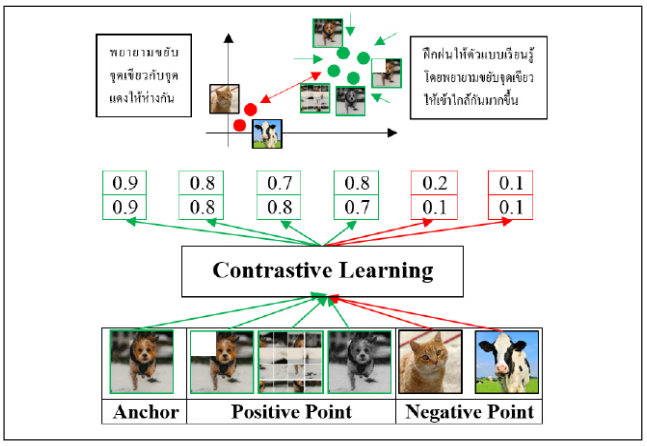
2.1) ตัวเข้ารหัสรูปภาพ (Image Encoder) ด้วยตัวแบบ ViT-B/32 [26] สำหรับการสกัดคุณลักษณะรูปภาพตามแนวคิดของสถาปัตยกรรม Transformer ที่ทำงานกับข้อความแต่นำมาปรับใช้กับรูปภาพ (Vision Transformer) จากการเปรียบเทียบความสัมพันธ์ระหว่างแพทช์ (Patches) เพื่อคำนวณหาผังคุณลักษณะในการหาความสัมพันธ์กับบริเวณทั้งภาพจากการสร้างคอนโวลูชันฟิลเตอร์ 768 ตัวขนาด 32*32*3 ใช้แพทช์ขนาด 32*32 หากอินพุตเป็นรูปภาพสี ต้องคูณ 3 เข้าไปเพราะแต่ละพิกเซลจะมีข้อมูลความสว่าง 3 สี คือ สีแดง เขียว และน้ำเงิน และเข้ารหัสขนาด 768 เป็นอาร์เรย์ 7*7*3 โดยผลลัพธ์จากการเข้ารหัสรูปภาพแต่ละภาพจะประกอบด้วยคุณลักษณะโกลบอล และคุณลักษณะโลคอล โดยจะแปลงเป็นเวกเตอร์รูปภาพขนาด 2048

2.2) ตัวเข้ารหัสข้อความ (Text Encoder) ด้วยโมเดลภาษา DistilBERT [27] เริ่มจากนำข้อความบรรยายรูปภาพมาแยกเป็นโทเค็น (Token) กำหนดรหัสโทเค็น และสร้างมาสก์ระบุความสนใจ (Attention Masks) ที่เป็นเทนเซอร์ขนาดเท่ากับรหัสโทเค็น ประกอบด้วยเลข 1 ที่แปลว่าโทเค็นในตำแหน่งนั้น ๆ จำเป็นต้องพิจารณา ปกติจะเป็นคำหลักที่ใช้ในการสื่อความหมายของประโยค ตรงกันข้ามกับโทเค็นในลักษณะคำที่พบบ่อย ๆ ในประโยคแต่ไม่ค่อยช่วยในการสื่อความหมายของข้อความจะถูกแทนด้วยเลข 0 จะไม่ถูกพิจารณาความหมายบน Attention Layers ของโมเดล จากนั้นจะเพิ่มโทเค็น CLS เพื่อใช้แทนตำแหน่งเริ่มต้นของข้อมูล และโทเค็น SEP คั่นกลางระหว่างประโยค เพื่อกำหนดจุดเริ่มต้น และจุดสิ้นสุดของประโยค สำหรับสร้างเวกเตอร์ตัวแทน (Vector Representation) [28] ในการฝึกฝนตัวเข้ารหัสข้อความเพื่อเรียนรู้ความหมายโดยรวมของข้อความบรรยายรูปภาพแต่ละประโยค จากนั้นทำการฝังข้อความ (Text Embedding) เพื่อแปลงให้เป็นเวกเตอร์ขนาด 768

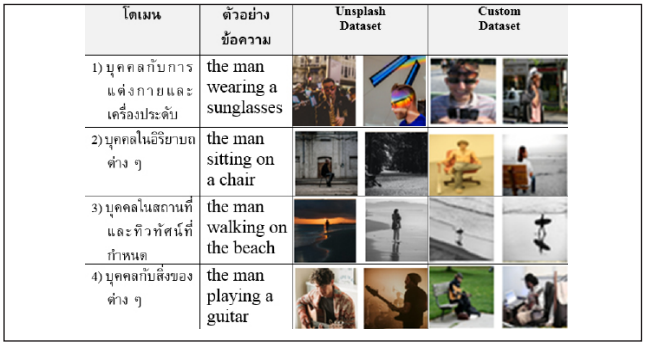
3) มอดูลการเรียนรู้แบบคอนทราสต์บนพื้นที่การฝังหลายรูปแบบ (Contrastive Learning on Multi-modal Embedding Space Modules) [29] โดยเรียกใช้ Projection Head ในการเปลี่ยนขนาดเวกเตอร์เป็น 256 เพื่อเรียนรู้แบบคอนทราสต์ระหว่างตัวเข้ารหัสรูปภาพ และตัวเข้ารหัสข้อความด้วยค่าความคล้ายคลึงโคไซน์ (Cosine Similarity) เมื่อคู่ของรูปภาพ และข้อความบรรยายรูปภาพที่ถูกต้องจะมีค่าเข้าใกล้กับ 1 มากขึ้น ในขณะที่คู่ที่ไม่ถูกต้องมีค่าเข้าใกล้กับ 0 จากนั้นกำหนดคีย์ของรูปภาพเก็บไว้ในชุดคำอธิบายรูปภาพ (Image Description Set) ร่วมกับคีย์ของข้อความบรรยายรูปภาพ เพื่อใช้ในการเรียกรูปภาพมาแสดงผล ดังภาพที่ 7

3) การสร้าง (Construction) แบบจำลองการค้นคืนรูปภาพเชิงเนื้อหาพัฒนาขึ้นด้วยการเขียนโปรแกรมภาษาไพธอน แสดงผลการทำงาน และปฏิสัมพันธ์กับผู้ใช้บน Google Colaboratory สำหรับเขียน และเรียกใช้ภาษาไพธอนบนคลาวด์ งานวิจัยนี้เรียกใช้งานตัวแบบ CLIP ที่ถูกฝึกฝนล่วงหน้ารุ่น ViT-B/32 แบบดั้งเดิม ตั้งค่าแบบ Zero-shot เรียนรู้ด้วยตนเองแบบคอนทราสต์บนชุดข้อมูลรูปภาพ Flickr30k เพื่อประเมินประสิทธิภาพการค้นคืนรูปภาพเชิงเนื้อหาเปรียบเทียบกับ

ชุดข้อมูลที่ผู้วิจัยรวบรวมเอง (Custom Dataset) เพื่อป้องกันปัญหา Overfitting กำหนดเนื้อหาเป็นรูปภาพบุคคลตามคำค้นหาใน 4 โดเมน ได้แก่ 1) บุคคลกับการแต่งกายและเครื่องประดับต่าง ๆ 2) บุคคลในอิริยาบถต่าง ๆ 3) บุคคลในสถานที่ และทิวทัศน์ต่าง ๆ และ 4) บุคคลกับสิ่งของต่าง ๆ ดังภาพที่ 8



***ภาพที่** 7 การเรียนรู้แบบคอนทราสต์*

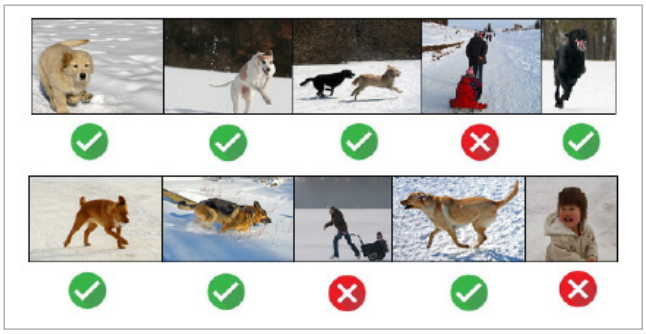


***ภาพที่ 8** ตัวอย่างรูปภาพบุคคลและคำค้นหาใน 4 โดเมน*

4) การทดสอบและการกลับมาตรวจสอบซ้ำ (Testing and Turnover) งานวิจัยนี้วัดประสิทธิภาพการค้นคืนรูปภาพแบบไม่เรียงลำดับความเกี่ยวข้องด้วยมาตรวัดออฟไลน์ (Offline Metric) แบบไม่ทราบจำนวนผลลัพธ์ที่ถูกต้อง (Order-Unaware Metric) ที่มักใช้ประเมินผลการค้นคืนสารสนเทศขนาดใหญ่บนเว็บไซต์ [30] ด้วยค่าความครบถ้วน (Recall) ที่เป็นอัตราส่วนของรูปภาพที่เกี่ยวข้องที่ถูกค้นคืนออกมาจากจำนวนรูปภาพที่เกี่ยวข้องทั้งหมด กำหนดจำนวนผลลัพธ์จากการค้นคืนเป็น 3, 5 และ 10 รูปภาพ แทนด้วย k ดังสมการที่ 1

$$\text{Recall@k} = \frac{\text{true positives@k}}{(\text{true positives@k})+(\text{false negatives@k})}$$

อย่างไรก็ดี การค้นคืนรูปภาพเชิงเนื้อหานั้นยากที่จะประเมินผลความถูกต้อง จึงเป็นที่มาของการให้ผู้เชี่ยวชาญเข้ามาร่วมเป็นส่วนหนึ่งในการประเมินความถูกต้องของการแสดงผลลัพธ์แต่ละครั้ง กำหนดข้อความค้นหาภายใต้เงื่อนไขคำค้นสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของภาพ (Query by high level concepts of the images) จำนวน 100 รายการ จากนั้นให้ผู้เชี่ยวชาญ 4 ท่านร่วมประเมินความถูกต้องทีละรูปภาพแบบเป็นอิสระต่อกัน ดังภาพที่ 9 ตัวอย่างผลลัพธ์จากข้อความค้นหา "a dog running on snow" จำนวน 10 รูปภาพ ที่ผู้เชี่ยวชาญท่านที่ 1 ประเมินว่าถูกต้องจำนวน 7 จาก 10 รูปภาพ



***ภาพที่ 9** ผลลัพธ์จากการค้นคืนรูปภาพ 10 รายการที่ผ่านการประเมินความถูกต้อง*

จากภาพที่ 9 ผลการประเมินด้วยค่าความครบถ้วนจากการค้นคืนรูปภาพ 10 รายการ ดังตารางที่ 1

***ตารางที่ 1** ผลการประเมินด้วยค่าความครบถ้วนจากการค้นคืนรูปภาพ 10 รายการ*

ลำดับ	ค่าความครบถ้วน
รูปภาพที่ 1 ☑	1/7 = 0.14
รูปภาพที่ 2 ☑	2/7 = 0.29
รูปภาพที่ 3 ☑	3/7 = 0.43
รูปภาพที่ 4 ☒	3/7 = 0.43
รูปภาพที่ 5 ☑	4/7 = 0.57
รูปภาพที่ 6 ☑	5/7 = 0.71
รูปภาพที่ 7 ☑	6/7 = 0.86
รูปภาพที่ 8 ☒	6/7 = 0.86
รูปภาพที่ 9 ☑	7/7 = 1.00
รูปภาพที่ 10 ☒	7/7 = 1.00

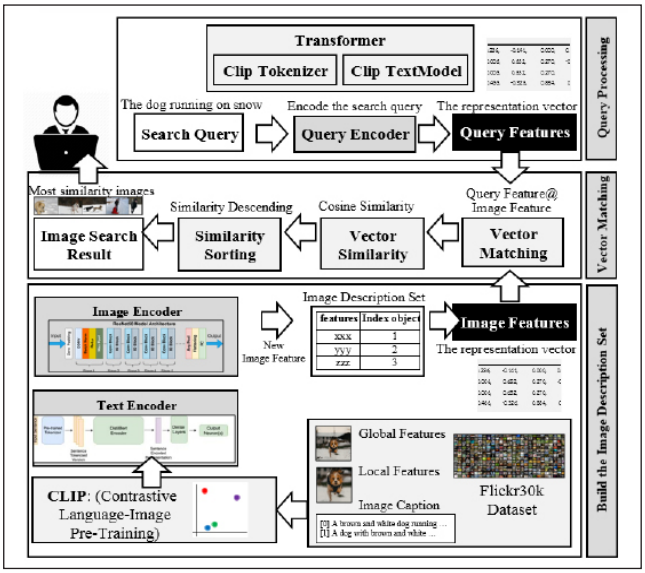
จากตารางที่ 1 ผลการประเมินผลลัพธ์จากการค้นคืนรูปภาพ พบว่า ค่าความครบถ้วนที่ R@1, R@5 และ R@10 เท่ากับ 0.14, 0.57 และ 1.00 ตามลำดับ โดยตัวเลขเหล่านี้ คือเปอร์เซ็นต์ของคำตอบที่ถูกต้องที่อยู่ในภาพจากการค้นคืนอันดับ 3, 5 และ 10 อันดับแรกที่เกี่ยวข้องกับข้อความค้นหาตามลำดับ อย่างไรก็ดี ค่าความครบถ้วนจะเพิ่มขึ้นเรื่อย ๆ ตามจำนวนการค้นคืน (k) ที่มากขึ้น และขอบเขตของการค้นคืนที่เพิ่มมากขึ้น ไม่ว่าผลลัพธ์นั้นจะถูกต้องหรือไม่ถูกต้องก็ตาม จากนั้นดำเนินการทดสอบจนครบทั้ง 100 รายการทั้งสองชุดข้อมูล รวมเป็นค่าความครบถ้วนเฉลี่ยของตัวแบบ

4. ผลการดำเนินงาน

งานวิจัยนี้แบ่งผลการดำเนินงานออกเป็น 2 ส่วน ตามวัตถุประสงค์ ได้แก่ 1) เพื่อพัฒนาตัวแบบการค้นคืนรูปภาพเชิงเนื้อหา และ 2) เพื่อประเมินประสิทธิภาพการค้นคืนรูปภาพ มีรายละเอียดดังนี้

4.1 ผลการพัฒนาตัวแบบการค้นคืนรูปภาพเชิงเนื้อหา

ตัวแบบการค้นคืนรูปภาพเชิงเนื้อหาในงานวิจัยนี้ ดังภาพที่ 10 ประกอบด้วย 3 มอดูล ได้แก่ มอดูลการสร้างชุดคำอธิบายรูปภาพ มอดูลการประมวลผลคำค้นหา และมอดูลการจับคู่เวกเตอร์ มีรายละเอียดดังนี้



ภาพที่ 10 *ภาพรวมกระบวนการทำงานของแบบจำลองการค้นคืนรูปภาพเชิงเนื้อหา*

1) มอดูลการสร้างชุดคำอธิบายรูปภาพ (Build the Image Description Set Module) จากการใช้ตัวแบบ CLIP ที่ถูกฝึกฝนล่วงหน้ามาเรียนรู้ความหมายของรูปภาพบนชุดข้อมูล Flickr30k กำหนด 1 รูปภาพต่อ 5 ข้อความบรรยายรูปภาพ ดังนั้นจึงมีความเป็นไปได้ที่จะมีรูปภาพที่เหมือนกันมากกว่า 1 รูปพร้อมคำบรรยายที่คล้ายกันอยู่ในชุดข้อมูลนี้ จึงต้องฝึกฝนตัวแบบด้วยค่าความคล้ายคลึงโคไซน์ โดยคำนึงถึงการเรียนรู้ที่จะแยกการเป็นตัวแทนแต่ละตัวออกจากกัน ด้วยการคำนวณดอตโปรดัค (Dot Product) ตามหลักพีชคณิตเชิงเส้น (Linear Algebra) จากการคูณรายการที่ตรงกันและหาผลรวมของเวกเตอร์รูปภาพ และเวกเตอร์ข้อความบรรยายภาพแต่ละคู่ ดังนั้นหากเป็นเวกเตอร์คู่เดียวกันจะมีค่าสูง ตรงกันข้ามเวกเตอร์คู่ที่แตกต่างกันจะมีค่าความคล้ายคลึงต่ำนั่นเอง ผลลัพธ์คือค่า Logits Score ที่เกิดจากการแทนค่าความคล้ายคลึงของเวกเตอร์ (Similar Vectors Representations) กรณีที่ดีที่สุดจะมีค่า 1.0 ก่อนจะนำมาผสานเป็นคุณลักษณะใหม่ของรูปภาพแล้วสร้างเป็นดัชนีคุณลักษณะ (Index Object) จัดเก็บไว้ที่ชุดคำอธิบายรูปภาพ เพื่อสร้างเป็นเวกเตอร์คุณลักษณะรูปภาพ (Image Feature Vector) ต่อไป

2) มอดูลการประมวลผลข้อความค้นหา (Search Query Processing Module) โดยใช้ตัวเข้ารหัสข้อความค้นหาเพื่อสร้างเวกเตอร์ข้อความค้นหา (Query Feature Vector) นำไปเปรียบเทียบกันด้วยค่าความคล้ายคลึงในขั้นตอนต่อไป

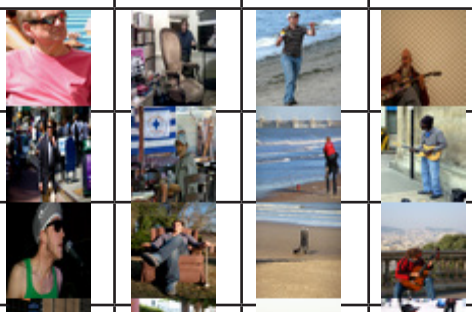
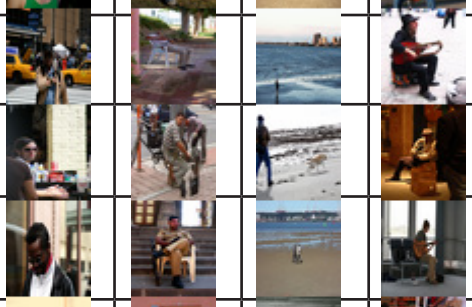
3) มอดูลการจับคู่เวกเตอร์ (Vector Matching Module) ระหว่างเวกเตอร์คุณลักษณะรูปภาพ และเวกเตอร์คุณลักษณะข้อความค้นหา เพื่อนำมาเปรียบเทียบค่าความคล้ายคลึงของเวกเตอร์บนพื้นที่การฝังหลายรูปแบบ (Multi-modal Embedding Space) สำหรับเปลี่ยนขนาดเวกเตอร์อินพุตให้มีมิติเท่ากัน ก่อนจะเรียงลำดับตามความเกี่ยวข้อง ก่อนจะนำมาคิวรีกับดัชนีคุณลักษณะของแต่ละรูปภาพที่ถูกจัดเก็บอยู่ในชุดคำอธิบายรูปภาพ และแสดงเป็นผลลัพธ์การค้นคืนรูปภาพตามบริบทของเนื้อหาแก่ผู้ใช้

4.2) ประเมินประสิทธิภาพการค้นคืนรูปภาพ

งานวิจัยนี้ประเมินประสิทธิภาพการค้นคืนรูปภาพด้วยค่าความครบถ้วนแบบไม่ทราบจำนวนผลลัพธ์ที่ถูกต้อง โดยสุ่มคำบรรยายรูปภาพที่มาพร้อมกับชุดข้อมูลภายใต้คำค้นสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของภาพ จำนวน 100 รายการ กำหนดรูปภาพที่แสดงผลเป็น 3, 5, และ 10 รูปภาพต่อ 1 ข้อความค้นหา

จากนั้นให้ผู้เชี่ยวชาญ 4 ท่านร่วมประเมินความถูกต้องทีละรูปภาพแบบเป็นอิสระต่อกัน ดังตารางที่ 2 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพทั้ง 4 โดเมนจากชุดข้อมูล Flickr30k ที่ผ่านการประเมินความถูกต้องโดยผู้เชี่ยวชาญท่านที่ 1

ตารางที่ 2 *ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพทั้ง 4 โดเมนจากชุดข้อมูล Flickr30k*

จำนวนรูปภาพ	ข้อความค้นหา			
	the man wearing a sunglasses	the man sitting on a	the man walking on the beach	the man playing a guitar
3 ลำดับ				
5 ลำดับ				
10 ลำดับ				

จากตารางที่ 2 เมื่อตรวจสอบผลการประเมินจากผู้เชี่ยวชาญทั้ง 4 ท่าน กำหนดเงื่อนไขหากรูปภาพใดมีผลการประเมินจากผู้เชี่ยวชาญเป็นถูกต้องทั้งหมด จะถือว่ารูปภาพนั้นถูกต้องเท่ากับ 1 คะแนน ในทางกลับกัน หากรูปภาพใดถูกผู้เชี่ยวชาญประเมินว่าไม่ถูกต้องแม้เพียงท่านเดียว จะถือว่ารูปภาพนั้นไม่ถูกต้อง เท่ากับ 0 คะแนน ก่อนจะคำนวณเป็นค่าความครบถ้วน ดังตารางที่ 3 ผลการประเมินประสิทธิภาพการค้นคืนรูปภาพด้วยค่าความครบถ้วนจำนวน 3, 5 และ 10 รูปภาพ จากชุดข้อมูล Flickr30k จำนวน 31,783 รูปภาพ เปรียบเทียบกับชุดข้อมูลที่ผู้วิจัยรวบรวมเอง จำนวน 24,999 รูปภาพ

ตารางที่ 3 *ผลการประเมินประสิทธิภาพการค้นคืนรูปภาพด้วยค่าความครบถ้วน*

ผู้เชี่ยวชาญ	Flickr30k Dataset			Custom Dataset		
	k@3	k@5	k@10	k@3	k@5	k@10
ท่านที่ 1	0.53	0.81	0.87	0.40	0.70	0.80
ท่านที่ 2	0.68	0.90	0.96	0.49	0.78	0.85
ท่านที่ 3	0.70	0.92	0.95	0.50	0.79	0.86
ท่านที่ 4	0.72	0.91	0.95	0.49	0.76	0.86
ค่าเฉลี่ย	**0.66**	**0.88**	**0.93**	**0.47**	**0.76**	**0.84**

จากตารางที่ 3 ผลการประเมินประสิทธิภาพการค้นคืนรูปภาพเชิงเนื้อหาด้วยค่าความครบถ้วน เมื่อพิจารณาในรายละเอียด พบว่า

1) ผลการค้นคืนรูปภาพเชิงเนื้อหาด้วยคำค้นที่อธิบายความหมายเชิงคุณภาพของรูปภาพ จำนวน 3, 5 และ 10 รูปภาพบนชุดข้อมูล Flickr30k นั้น มีค่าความครบถ้วนเฉลี่ยถึง 0.66, 0.88 และ 0.93 ตามลำดับ แสดงถึงคุณลักษณะระดับสูงจากการเรียนรู้ด้วยตนเองของตัวแบบโครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้านั้นสามารถแก้ปัญหาช่องว่างความหมายได้อย่างมีประสิทธิภาพ และช่วยสนับสนุนผู้ใช้ด้วยคำค้นในรูปแบบภาษาธรรมชาติที่ยึดโยงกับความหมายของรูปภาพแทนที่จะยึดตามหลักไวยากรณ์ของภาษา

2) ผลการค้นคืนรูปภาพเชิงเนื้อหาบนชุดข้อมูลที่ผู้วิจัยรวบรวมเองนั้น พบว่า ค่าความครบถ้วนเฉลี่ย จำนวน 3, 5 และ 10 รูปภาพนั้นเท่ากับ 0.47, 0.76 และ 0.84 ตามลำดับ เมื่อเปรียบเทียบกับชุดข้อมูล Flickr30k ที่นำมาใช้ฝึกฝนตัวแบบ พบว่า ค่าความครบถ้วนเฉลี่ยจากผู้เชี่ยวชาญทั้งหมดลดลงเล็กน้อยและเป็นไปในทิศทางเดียวกัน โดยผลลัพธ์ดังกล่าวนั้นถือว่าอยู่ในระดับที่ยอมรับได้ และไม่เกิดปัญหาที่ตัวแบบจะประสิทธิภาพจะลดลงเมื่อทำงานกับข้อมูลที่ไม่เคยพบมาก่อน

3) ค่าความครบถ้วนเฉลี่ยของรูปภาพนั้นสูงขึ้นอย่างก้าวกระโดดเมื่อจำนวนผลลัพธ์จากการค้นคืนเปลี่ยนเป็น 3 เป็น 5 รูปภาพ

และค่าความครบถ้วนเฉลี่ยจะค่อย ๆ คงที่ และจะเข้าใกล้ 1 มากที่สุดเมื่อจำนวนผลลัพธ์เท่ากับ 10 รูปภาพ

5. สรุป

ผลการพัฒนาแบบจำลองการค้นคืนรูปภาพเชิงเนื้อหาด้วยคุณลักษณะระดับสูงจากการเรียนรู้ด้วยตนเองของตัวแบบโครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้าประกอบด้วย 3 มอดูล ได้แก่ 1) มอดูลการสร้างชุดคำอธิบายรูปภาพโดยใช้ตัวแบบ CLIP เพื่อเรียนรู้ความหมายของรูปภาพจากความสัมพันธ์ระหว่างรูปภาพกับข้อความบรรยายรูปภาพด้วยการพยายามฝึกฝนตัวเข้ารหัสรูปภาพ และตัวเข้ารหัสข้อความด้วยค่าความคล้ายคลึงโคไซน์ ก่อนจะผสานเป็นคุณลักษณะใหม่แล้วจัดเก็บไว้ที่ชุดคำอธิบายรูปภาพเพื่อสร้างเป็นเวกเตอร์คุณลักษณะรูปภาพ 2) มอดูลการประมวลผลข้อความค้นหาในรูปแบบภาษาธรรมชาติ และสร้างเป็นเวกเตอร์คุณลักษณะข้อความค้นหา 3) มอดูลการจับคู่เวกเตอร์ระหว่างเวกเตอร์คุณลักษณะรูปภาพ และเวกเตอร์คุณลักษณะข้อความค้นหา ก่อนจะนำมาเปรียบเทียบค่าความคล้ายคลึงระหว่างเวกเตอร์ก่อนจะเรียงลำดับตามความเกี่ยวข้องและแสดงเป็นผลลัพธ์การค้นคืนรูปภาพตามบริบทของเนื้อหาแก่ผู้ใช้

ผลการค้นคืนรูปภาพบนชุดข้อมูล Flickr30k แบบไม่ทราบจำนวนผลลัพธ์ที่ถูกต้องมีค่าความครบถ้วนเฉลี่ยถึง 0.93 เมื่อผลลัพธ์เป็น 10 อันดับ อยู่ในระดับสูงมาก แสดงถึงตัวแบบสามารถค้นคืนรูปภาพเชิงเนื้อหาได้อย่างมีประสิทธิภาพ อย่างไรก็ดี เมื่อพิจารณาในรายละเอียดของผลลัพธ์ที่ไม่ถูกต้องพบว่า ความแปรผันของรูปภาพเป็นอุปสรรคสำคัญในการจำแนกข้อมูลภาพ

การเปรียบเทียบผลการค้นคืนรูปภาพบนชุดข้อมูล Flickr30k กับชุดข้อมูลรูปภาพที่ผู้วิจัยรวบรวมเอง พบว่าค่าความครบถ้วนเฉลี่ยนั้นเป็นไปในทิศทางเดียวกัน และไม่เกิดปัญหาประสิทธิภาพของแบบจำลองจะลดลงเมื่อทำงานกับข้อมูลที่ไม่เคยพบมาก่อน

นอกจากนี้ ผลลัพธ์จำนวนมากอาจทำให้ผู้ใช้เสียเวลาคัดเลือกรูปภาพที่ตรงกับความต้องการอย่างแท้จริง เนื่องจากเป็นได้น้อยมากที่ผู้ใช้จะเลือกดูรูปภาพจากผลการค้นคืนทั้งหมด จึงเป็นการเสียโอกาสสำหรับผู้ใช้หากมีรูปภาพบางส่วนที่ไม่ถูกเรียกดู ดังนั้นการค้นคืนรูปภาพจึงต้องมุ่งเพิ่มค่าความครบถ้วนอันดับ 1, 3 และ 5 เป็นหลักเพื่อตอบสนองต่อความต้องการของผู้ใช้

ข้อเสนอแนะการวิจัยต่อไปในการพัฒนาแบบจำลองการค้นคืนรูปภาพนั้นควรให้ความสำคัญที่การปรับค่าไฮเปอร์พารามิเตอร์ (Hyperparameter) ก่อนที่ตัวแบบจะทำการเรียนรู้ เนื่องจากการกำหนดค่าไฮเปอร์พารามิเตอร์ที่เหมาะสมจะส่งผลให้ตัวแบบมีความแม่นยำที่สูงขึ้น หรือลดค่าการสูญเสียให้ต่ำลงได้

6. เอกสารอ้างอิง

[1] M. Broz, *Number of Photos (2023): Statistics, Facts, & Predictions*. Available Online at https://photutorial.com/photos-statistics/, accessed on 29 March 2023.

[2] V. Tyagi. *Content-Based Image Retrieval Ideas, Influences, and Current Trends.* Published by Springer Nature, ISBN 978-981-10-6758-7, 2017.

[3] J. Z. Wang, J. Li, and G. Wiederhold. "SIMPLIcity: semantics-sensitive integrated matching for picture libraries." *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 23, No. 9, pp. 947-963, 2001.

[4] D. Zhang. *Fundamentals of Image Data Mining: Analysis, Features, Classification and Retrieval.* Published by Springer Nature, ISBN 978-3-030-69251-3, 2021.

[5] R. Scherer. *Computer Vision Methods for Fast Image Classification and Retrieval.* Published by Springer Nature, ISBN 978-3-030-12195-2, 2020.

[6] C. Shorten and T. M. Khoshgoftaar. "A survey on Image Data Augmentation for Deep Learning." *Journal of Big Data,* Vol. 6, No. 1, pp. 1-48, 2019.

[7] R. Mitchell. *Web Scraping with Python: Collecting Data from the Modern Web,* Published by O'Reilly Media Inc, ISBN 978-149-19-8557-1, 2018.

[8] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. "Learning Transferable Visual Models From Natural Language Supervision." *In 38th International Conference on*

Machine Learnin, pp. 8748-8763, 2021.

[9] Y. Bastanlar and S. Orhan. "Self-Supervised Contrastive Representation Learning in Computer Vision." *Artificial Intelligence Annual Volume 2022,* Published by IntechOpen, ISBN 978-1-83768-948-4, 2022.

[10] J. Zizka, F. Darena, and A. Svoboda. *Text Mining with Machine Learning Principles and Techniques*. Published by CRC Press, ISBN 978-1-13860-182-6, 2019.

[11] K. Ueki. "Survey of Visual-Semantic Embedding Methods for Zero-Shot Image Retrieval." *In 20th IEEE International Conference on Machine Learning and Applications,* pp. 628-634, 2021.

[12] K. Sawarkar. *Deep Learning with PyTorch Lightning,* Published by Packt Publishing, ISBN 978-1-80056-161-8, 2022.

[13] F. Pourpanah, M. Abdar, Y. Luo, X. Zhou, R. Wang, C. P. Lim, X. Wang, and Q. M. J. Wu. "A Review of Generalized Zero-Shot Learning Methods." *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 45, No. 4, pp. 4051-4070, 2023.

[14] B. Barz. *Semantic and Interactive Content-based Image Retrieval.* A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Mathematics and Computer Science, Friedrich Schiller University Jena, 2020.

[15] H. Xu, J. Wang, and L. Mao. "Relevance feedback for Content-based Image Retrieval using deep learning." *In 2nd International Conference on Image, Vision and Computing (ICIVC),* pp. 629-633, 2017.

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

[17] A. Vatani, M. T. Ahvanooey, and M. Rahim. "An Effective Automatic Image Annotation Model ViaAttention Model and Data Equilibrium." *International Journal of Advanced Computer Science and Applications,* Vol. 9, No. 3, pp. 269-277, 2018.

[18] F. Neuhaus. *"What is an Ontology?."* arXiv preprint arXiv:1810.09171v1., pp. 1-18, 2018.

[19] Y. Liu, Y. Huang, S. Zhang, D. Zhang, and N. Ling. "Integrating object ontology and region semantic template for crime scene investigation image retrieval." *In 12th IEEE Conference on Industrial Electronics and Applications (ICIEA),* pp. 149-153, 2017.

[20] H. Dong, Z. Wang, Q. Qiu, and G. Sapiro. "Using Text to Teach Image Retrieval." *In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* pp. 1-10, 2021.

[21] E. Charniak. *Introduction to Deep Learning,* ISBN 978-0-262-03951-2, 2019.

[22] S. Tilley and H. J. Rosenblatt. *Systems Analysis and Design.* Published by Cengage Learning, ISBN 978-1-305-49460-2, 2017.

[23] A. Mikolajczyk and M. Grochowski. "Data augmentation for improving deep learning in image classification problem." *In 2018 International Interdisciplinary PhD Workshop (IIPhDW),* pp. 117-122, 2018.

[24] Y. Roh, G. Heo, and S. E.Whang. "A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective." *IEEE Transactions on Knowledge and Data Engineering,* Vol. 33, No 4, pp. 1328-1347, 2019.

[25] C. C. Aggarwal. *Neural Networks and Deep Learning A Textbook,* Published by Springer, ISBN 978-3-319-94463-0, 2018.

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

[27] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. "Dis tilBERT, a distilled version of BERT: smaller, faster,

cheaper and lighter." *arXiv preprint arXiv: 1910.01108.,* pp. 1-5, 2019.

[28] F. Weers, V. Shankar, A. Katharopoulos, Y. Yang, and T. Gunte. "Masked Autoencoding Does Not Help Natural Language Supervision at Scale." *In 2023 Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 1-19, 2023.

[29] T. Baltrusaitis, C. Ahuja, and L. Morency. "Multimodal Machine Learning: A Survey and Taxonomy." *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 41, No. 2, pp. 423-443, 2019.

[30] A. Chaudhary, *Evaluation Metrics For Information Retrieval*. Available Online at https://amitness.com/2020/08/information-retrieval-evaluation//, accessed on 1 April 2023.