แบบจำลองการจำแนกรูปภาพตามความหมาย ได้รับการฝึกฝนสำหรับการเรียกค้นรูปภาพโดยใช้ภาษาธรรมชาติ

The Semantic-Based Image Classification Model Trained for Image Retrieval Using Natural Language

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

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

[1]สาขาวิชาเทคโนโลยีธุรกิจดิจิทัล คณะบริหารธุรกิจ มหาวิทยาลัยวงษ์ชวลิตกุล

[2]สำนักวิชาศาสตร์และศิลป์ดิจิทัล มหาวิทยาลัยเทคโนโลยีสุรนารี

[1]Department of Digital Business Technology, Faculty of Business Administrator, Vongchavalitkul University

[2]Institute of Digital Arts and Science, Suranaree University of Technology

Submitted 25/06/2023 ; Revised 20/08/2023 ; Accepted 05/10/2023

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองการจำแนกรูปภาพตามความหมายโดยใช้ตัวแบบที่ถูกฝึกฝนล่วงหน้าแบบคอนทราสต์ด้วยข้อความและรูปภาพ และเรียนรู้ซ้ำบนชุดข้อมูลที่กำหนดเอง ผลการประเมินตัวแบบที่ได้รับการฝึกฝนสำหรับการเรียกค้นรูปภาพโดยใช้ภาษาธรรมชาติเปรียบเทียบระหว่างผลการพยากรณ์ป้ายกำกับภาษาธรรมชาติกับผลการประเมินความหมายของป้ายกำกับจากผู้เชี่ยวชาญ พบว่า ผลการพยากรณ์ป้ายกำกับภาษาธรรมชาติภายใต้ 3 เงื่อนไข ได้แก่ 1) ข้อความบรรยายรูปภาพ ในลักษณะป้ายกำกับของรูปภาพ 2) ข้อความบรรยายรูปภาพเกี่ยวกับแนวคิดระดับสูงของภาพ และ 3) ข้อความบรรยายรูปภาพที่อธิบายความหมายเชิงคุณภาพของรูปภาพเท่ากับ 0.905, 0.830 และ 0.585 ตามลำดับ โดยผลการประเมินด้วยข้อความบรรยายรูปภาพที่อธิบายความหมายเชิงคุณภาพของรูปภาพอยู่ในระดับปานกลาง เนื่องจากข้อความที่อยู่ในรูปแบบของป้ายกำกับภาษาธรรมชาตินั้นถือว่าเป็นแนวคิดระดับสูง ดังนั้นประสบการณ์ของแต่ละบุคคลจึงส่งผลให้การประเมินนั้นแตกต่างกันตามหลักการรับรู้ของมนุษย์ อันจะเห็นได้จากผลการพยากรณ์ที่ใกล้เคียงกันของข้อความบรรยายรูปภาพที่มากกว่า 1 ตัวเลือก การเรียกค้นรูปภาพตามความหมายจึงควรให้ความสำคัญกับการลดช่องว่างความหมายของคำค้นหา และช่วยสนับสนุนผู้ใช้ด้วยการใช้คำค้นภายใต้รูปแบบของภาษาธรรมชาติที่ยึดโยงกับความหมายของรูปภาพแทนที่จะยึดตามหลักไวยากรณ์ของภาษา อันจะเป็นแนวทางการเรียกคืนสารสนเทศในอนาคต

คำสำคัญ: การจำแนกรูปภาพ รูปภาพตามความหมาย การเรียนรู้ภาษาธรรมชาติ การเรียกค้นรูปภาพ

[*]ผู้ประสานงานหลัก (Corresponding Author)

E-mail: chakkarin_san@vu.ac.th

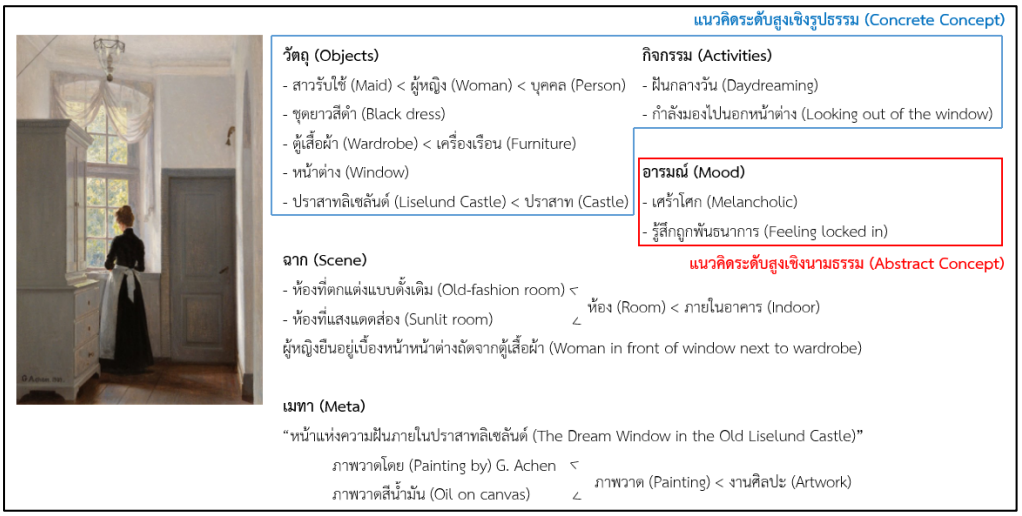
Abstract

The objective of this research is to develop a model for image classification based on semantic meaning using pre-trained deep learning models with text and image data, and iterative learning on custom datasets. Evaluation results of the trained models for image retrieval using natural language label were compared against expert-assessed label meanings, revealing that the prediction performance for natural language label under three conditions, namely 1) image descriptive text resembling image label, 2) high-level conceptual text related to the image content, and 3) text describing the qualitative meaning of the image, yielded scores of 0.905, 0.830, and 0.585, respectively. The evaluation results for text describing the qualitative meaning of the image were found to be at a moderate level, as the text in the form of natural language label is considered a high-level concept. Consequently, individual perceptual experiences influenced the evaluation differently based on human cognition principles, as evidenced by the closely aligned prediction results for similar descriptive text for more than one option. Therefore, meaningful image retrieval should emphasize reducing the semantic gap in search queries and assist users by utilizing query terms aligned with image meaning rather than adhering strictly to grammatical language rules. This approach is suggested as a future direction for information retrieval.

Keywords: image classification, semantic-based image, natural language supervision, image retrieval

บทนำ

ช่องว่างความหมาย (semantic-gap) เป็นข้อจำกัดสำคัญของการเรียกค้นรูปภาพจากเนื้อหา (content-based image retrieval: CBIR) [1] เนื่องจากคุณลักษณะระดับต่ำ (low-level features) เช่น สี รูปทรง พื้นผิว หรือตำแหน่งนั้นไม่สามารถสื่อความหมายของรูปภาพได้อย่างถูกต้อง แม้ปัจจุบันจะสามารถทำงานได้อย่างมีประสิทธิภาพ แต่ในความเป็นจริงพฤติกรรมผู้ใช้โดยมากจะเรียกค้นรูปภาพโดยใช้ความสัมพันธ์ระหว่างคุณลักษณะระดับต่ำ และแนวคิดระดับสูง (high-level semantics) ที่เป็นรูปธรรม (concrete concept) อย่างวัตถุ กิจกรรม หรือเหตุการณ์ และแนวคิดที่เป็นนามธรรม (abstract concept) อย่างอารมณ์และความรู้สึก ดังภาพที่ 1 รวมเข้าด้วยกันเป็นข้อความค้นหาที่เป็นแนวคิดระดับสูง ในทางกลับกัน คุณลักษณะที่ถูกแยกโดยอัตโนมัติโดยใช้คอมพิวเตอร์มักเป็นคุณลักษณะระดับต่ำ ซึ่งโดยทั่วไปไม่มีการเชื่อมโยงโดยตรงระหว่างแนวคิดระดับสูงและคุณลักษณะระดับต่ำทำให้ผลลัพธ์จากการเรียกค้นมีประสิทธิภาพไม่สูงเท่าที่ควร เป็นที่มาของแนวคิดการเรียกค้นรูปภาพตามความหมาย (semantic-based image retrieval: SBIR) [2] ที่มุ่งเชื่อมโยงคุณลักษณะระดับต่ำของรูปภาพให้กลายเป็นแนวคิดระดับสูง เพื่อระบุถึงความหมายของรูปภาพในระดับสูงที่มนุษย์สามารถเข้าใจได้



ภาพที่ 1 ความสัมพันธ์ระหว่างแนวคิดระดับสูงที่เป็นรูปธรรม และที่เป็นนามธรรม (ดัดแปลงจาก [2])

ปัจจุบันมีการนำเสนอแนวคิดเกี่ยวกับโครงข่ายประสาทเทียม (neural network) ในการจำแนกประเภทรูปภาพ (image classification) [3] โดยมีจุดเริ่มต้นมาจากใช้แนวคิดของคอนโวลูชันในการทำงานกับข้อมูล 2 มิติเกิดเป็นโครงข่ายประสาทเทียมคอนโวลูชัน (convolutional neural network: CNN) ในการสกัดคุณลักษณะ (feature extraction) ของรูปภาพออกมาวิเคราะห์ และนำไปจำแนกประเภทรูปภาพเป็นที่มาของสถาปัตยกรรม LeNet-5 [4]

เมื่อคอมพิวเตอร์ถูกพัฒนาให้มีประสิทธิภาพสูงขึ้นสูงขึ้น จึงเกิดการพัฒนาอัลกอริทึมเพื่อจำแนกวัตถุบนชุดข้อมูลขนาดใหญ่ ที่ได้รับความนิยมมาจนถึงปัจจุบัน เช่น AlexNet และ VGG ซึ่งใช้แนวคิดการเพิ่มความลึกของจำนวนชั้นให้ตัวแบบมีความซับซ้อนมากขึ้น ถือว่าเข้าสู่ยุคการเรียนรู้เชิงลึก (deep learning) [5] ที่มุ่งเพิ่มความลึกของชั้นซ่อนเร้นให้มากยิ่งขึ้น เนื่องจากส่งผลต่อความแม่นยำอย่างไรก็ดี เมื่อถึงจุดหนึ่ง ๆ ความแม่นยำจะลดลง เนื่องจากปัญหาค่าเกรเดียนต์สูญหาย (gradient vanishing) เป็นที่มาของสถาปัตยกรรม ResNet [6] ที่แก้ปัญหาค่าเกรเดียนต์สูญหายด้วยเทคนิค skip connection ร่วมกับ batch normalization เกิดเป็นตัวแบบที่ให้ความแม่นยำสูงแม้จะมีความลึกมากก็ตาม

พัฒนาการของการเรียนรู้เชิงลึกในปัจจุบันส่วนสำคัญเริ่มต้นมาจากงานด้านคอมพิวเตอร์วิทัศน์ ในการพัฒนาตัวแบบที่ถูกฝึกฝนใหม่ตั้งแต่ต้น (pre-training) จากค่าน้ำหนักที่เริ่มต้นจากการสุ่มและเริ่มฝึกฝนโดยไม่มีความรู้อะไรเลย ดังนั้นในการประยุกต์ใช้จึงสามารถปรับแต่ง (fine-tuning) ตัวแบบเพิ่มเติมด้วยชุดข้อมูลที่เฉพาะเจาะจงในงานที่ต้องการ เนื่องจากตัวแบบได้เรียนรู้การเป็นตัวแทน (representation) ของคุณลักษณะที่สำคัญ ๆ ในรูปภาพเอาไว้ อีกทั้งสามารถถ่ายโอนความรู้ (transfer learning) ไปใช้ได้เลยโดยไม่จำเป็นต้องฝึกฝนตัวแบบใหม่ตั้งแต่ต้น ดังนั้นการประมวลผลภาษาธรรมชาติ (natural language processing: NLP) จึงได้นำแนวคิดดังกล่าวมาปรับใช้ในการฝึกฝนตัวแบบทางภาษา (language model) บนข้อความดิบขนาดใหญ่ด้วยการเรียนรู้ด้วยตนเอง (self-supervised) อย่างสถาปัตยกรรม Transformer [7] ต่อมาทีมผู้พัฒนาจากกูเกิลเสนอสถาปัตยกรรม Vision in Transformer (ViT) [8] ในการจำแนกรูปภาพโดยใช้แนวคิดเช่นเดียวกับการทำงานกับข้อความแต่เปลี่ยนมาใช้กับรูปภาพ แบบไม่มีการใช้โครงข่ายประสาทเทียมคอนโวลูชันแม้แต่กระบวนการเดียว (end-to-end transformer)

ตัวแบบ CLIP (Contrastive Language–Image Pre-training) [9] ที่ถูกฝึกฝนล่วงหน้าเป็นหนึ่งในตัวแบบที่นำแนวคิดของ Transformer มาใช้ในการจำแนกรูปภาพจากการเรียนรู้หลายรูปแบบ (multimodal learning) ร่วมกันระหว่างรูปภาพและข้อความ เพื่อใช้เรียนรู้ความหมายรูปภาพจากข้อความภาษาธรรมชาติ ประกอบด้วยตัวเข้ารหัสข้อความ (text encoder) ด้วย Transformer สำหรับแปลงข้อความให้อยู่ในรูปแบบเวกเตอร์ (text embedding) และตัวเข้ารหัสรูปภาพ (image encoder) ด้วย Vision in Transformer สำหรับแปลงรูปภาพให้อยู่ในรูปแบบเวกเตอร์ (image embedding) เพื่อฝึกฝนให้ตัวแบบเรียนรู้จากความสัมพันธ์ระหว่างคำบรรยายภาพและรูปภาพ

งานวิจัยนี้มุ่งฝึกฝนตัวแบบจำแนกรูปภาพตามความหมายจากการเรียนรู้ความสัมพันธ์ระหว่างรูปภาพกับข้อความที่อยู่ในลักษณะข้อความภาษาธรรมชาติ แทนที่จะเรียนรู้ภายใต้คลาสที่จำกัดอย่างสุนัขหรือแมว เมื่อตัวแบบได้รับการฝึกฝนจากข้อความทั้งประโยค จะส่งผลให้ตัวแบบสามารถเรียนรู้สิ่งต่าง ๆ ได้มากขึ้น และสามารถค้นหารูปแบบความสัมพันธ์บางอย่างระหว่างรูปภาพกับข้อความด้วยการเรียนรู้ด้วยตนเอง เพื่อจำแนกรูปภาพตามความหมายคำค้นหาในรูปแบบภาษาธรรมชาติที่ยึดโยงกับความหมายของรูปภาพแทนที่จะยึดตามหลักไวยากรณ์ของภาษา อันจะเป็นแนวทางการเรียกค้นสารสนเทศในอนาคต

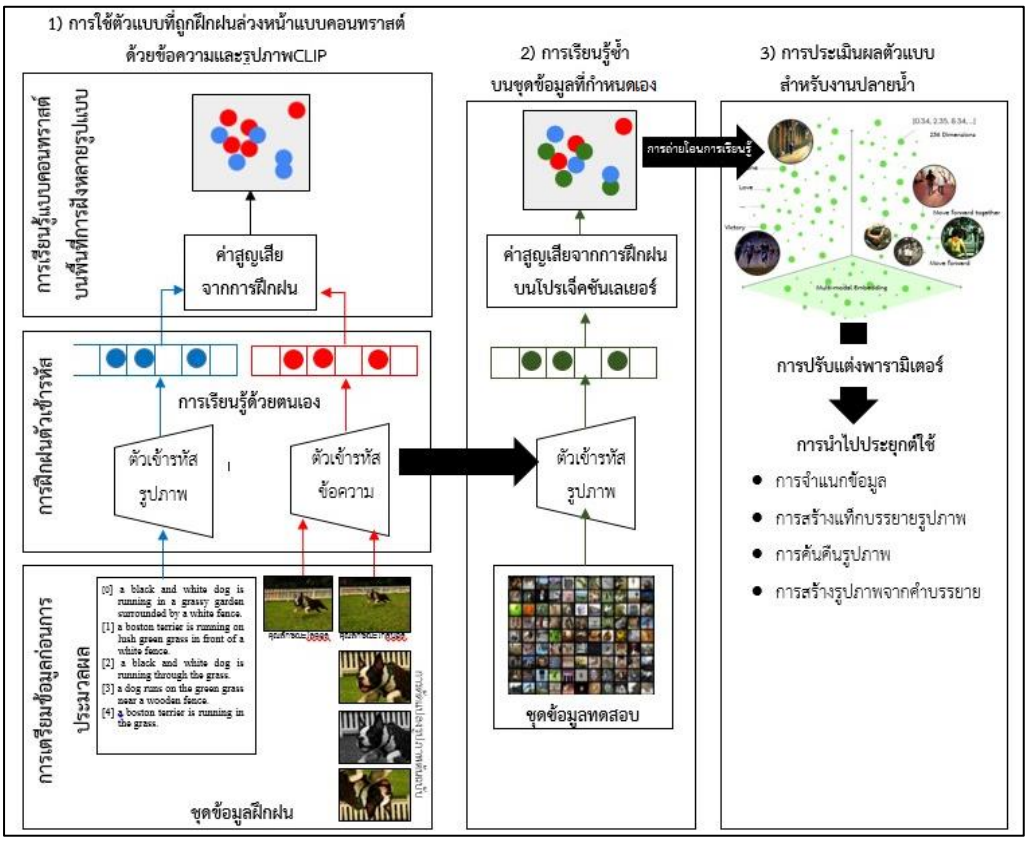
วัตถุประสงค์

1. เพื่อพัฒนาแบบจำลองการจำแนกรูปภาพตามความหมายสำหรับการเรียกค้นรูปภาพโดยใช้ภาษาธรรมชาติ

2. เพื่อประเมินประสิทธิภาพการจำแนกรูปภาพด้วยป้ายกำกับภาษาธรรมชาติ

วิธีดำเนินการวิจัย

งานวิจัยนี้ได้ศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้องเพื่อพัฒนาแบบจำลองการจำแนกรูปภาพตามความหมายได้รับการฝึกฝนสำหรับการเรียกค้นรูปภาพโดยใช้ภาษาธรรมชาติ ดังภาพที่ 2



ภาพที่ 2 แบบจำลองการจำแนกรูปภาพตามความหมายได้รับการฝึกฝนสำหรับการเรียกค้นรูปภาพ

1) การใช้ตัวแบบที่ถูกฝึกฝนล่วงหน้าแบบคอนทราสต์ด้วยข้อความและรูปภาพ ประกอบด้วย

1.1) การเตรียมข้อมูลก่อนการประมวลผล (dataset pre-processing) งานวิจัยนี้ใช้ชุดข้อมูล Flickr30k ประกอบด้วยรูปภาพจำนวน 31,783 รูปภาพและข้อความบรรยายรูปภาพ กำหนด 1 รูปภาพต่อ 5 ข้อความบรรยายรูปภาพ ดังตารางที่ 1 มีรายละเอียดดังนี้

1.1.1) การดัดแปลงรูปภาพต้นฉบับ (image augmentation) [10] มาดัดแปลงเป็นรูปภาพใหม่เพื่อเป็นชุดข้อมูลฝึกฝน ประกอบด้วยคุณลักษณะโกลบอล (global features) เช่น สี พื้นผิว

รูปทรง และตำแหน่ง ร่วมกับคุณลักษณะแบบโลคอล (local features) ที่เป็นคุณลักษณะเฉพาะของแต่ละส่วนในรูปภาพ เพื่อให้ตัวแบบเรียนรู้ และจดจำคุณลักษณะสำคัญของรูปภาพได้ดียิ่งขึ้น

1.1.2) การทำความสะอาดและกำหนดมาตรฐานข้อความ (text cleansing and normalization) ด้วยการกำจัดเครื่องหมายต่าง ๆ และเปลี่ยนเป็นตัวอักษรพิมพ์เล็กทั้งหมด ตลอดจนปรับข้อความให้อยู่ในเซลล์เดียวกันทั้งประโยค เพื่อให้อยู่ในรูปแบบที่เหมาะสมต่อการใช้งาน

ตารางที่ 1 ตัวอย่างรูปภาพและข้อความบรรยายรูปภาพจากชุดข้อมูล Flickr30k

ลำดับ	ข้อความบรรยายรูปภาพ	รูปภาพ
1	a black and white dog is running in a grassy garden surrounded by a white fence.	
2	a boston terrier is running on lush green grass in front of a white fence.	
3	a black and white dog is running through the grass.	
4	a dog runs on the green grass near a wooden fence.	
5	a boston terrier is running in the grass.	

1.2) การฝึกฝนตัวเข้ารหัส (encoder pre-training) กำหนดค่าไฮเปอร์พารามิเตอร์ (hyperparameters) ได้แก่ ความละเอียดอินพุต (input resolution) ความยาวบริบท (context length) และขนาดคำศัพท์ (vocab size) เท่ากับ 224, 77 และ 49,408 ตามลำดับ มีรายละเอียดดังนี้

1.2.1) ตัวเข้ารหัสข้อความ (text encoder) โดยใช้ตัวแบบภาษา DistilBERT [11] บนพื้นฐานสถาปัตยกรรม Transformer เริ่มจากการแยกข้อความบรรยายรูปภาพเป็นโทเค็น จากนั้นป้อนรหัสโทเค็น และสร้างมาสก์ระบุความสนใจ (attention masks) ที่เป็นเทนเซอร์ขนาดเท่ากับรหัสโทเค็น ประกอบด้วยเลข 1 ที่แปลว่าโทเค็นในตำแหน่งนั้น ๆ จำเป็นต้องพิจารณา ตรงกันข้ามกับเลข 0 ที่จะไม่ถูกพิจารณาบนชั้นความสนใจ (attention layers) จากนั้นเพิ่มโทเค็น CLS เพื่อใช้แทนตำแหน่งเริ่มต้นของข้อมูล และโทเค็น SEP สำหรับกำหนดขอบเขตระหว่างประโยค ซึ่งเป็นจุดเริ่มต้นและจุดสิ้นสุดของประโยค เพื่อเรียนรู้ความหมายโดยรวมของข้อความบรรยายรูปภาพแต่ละประโยค ผลลัพธ์จากการฝังข้อความ (text embedding) แปลงเป็นเวกเตอร์คุณลักษณะข้อความ (text feature vector) ขนาด 768

1.2.2) ตัวเข้ารหัสรูปภาพ (image encoder) โดยฝึกฝนตัวเข้ารหัสรูปภาพตามสถาปัตยกรรม Vision Transformer แบบ ViT-B/32 จากการสร้างคอนโวลูชันฟิลเตอร์ขนาด 32*32*3 มาจากแพทช์ (patches) ขนาด 32*32 หากอินพุตเป็นรูปภาพสีต้องคูณ 3 เข้าไป เพราะแต่ละพิกเซลจะมีข้อมูลความสว่าง 3 สี คือสีแดง เขียว และน้ำเงิน และเข้ารหัสเป็นอาร์เรย์ 7*7*3 จากการเปรียบเทียบความสัมพันธ์ระหว่างแพทช์เพื่อสกัดคุณลักษณะรูปภาพ โดยผลลัพธ์จากการฝังรูปภาพ (image embedding) จะแปลงเป็นเวกเตอร์คุณลักษณะรูปภาพ (image feature vector) ขนาด 2,048

1.3) การเรียนรู้แบบคอนทราสต์บนพื้นที่การฝังหลายรูปแบบ (contrastive learning on multi-modal embedding space) โดยใช้ projection head ในการเปลี่ยนขนาดเวกเตอร์คุณลักษณะข้อความและเวกเตอร์คุณลักษณะรูปภาพบนพื้นที่การฝังที่มีมิติเท่ากับ 256 เพื่อฝึกฝนร่วมกันระหว่างตัวเข้ารหัสรูปภาพและตัวเข้ารหัสข้อความ โดยเอาต์พุตจากโครงข่ายประสาทเทียมจะอยู่ในรูปแบบ logits score จึงต้องแปลงค่าให้เป็นมาตรฐานสำหรับการแจกแจงความน่าจะเป็นตามสัดส่วนของค่าเอาต์พุตแต่ละคลาสด้วยฟังก์ชันซอฟต์แม็กซ์ [12] ดังสมการที่ 1

$$f(x_i) = \frac{\exp(x_i)}{\sum_j^n \exp(x_j)} \quad (1)$$

กำหนดให้

x_i คือ เวกเตอร์อินพุตของฟังก์ชันซอฟต์แม็กซ์

$\exp(x_i)$ คือ ฟังก์ชันเอกซ์โพเนนเชียลมาตรฐาน มีค่าคงที่เท่ากับ 2.71828 ยกกำลังด้วยเวกเตอร์อินพุต

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

จากตารางที่ 1 สมมติค่า logits score จากค่าความคล้ายคลึงระหว่างเวกเตอร์คุณลักษณะรูปภาพกับเวกเตอร์คุณลักษณะข้อความบรรยายรูปภาพลำดับที่ 1 ถึงลำดับ 5 เท่ากับ 0.5, 3.9, 0.2, 0.3 และ 5.2 ตามลำดับ เมื่อแทนค่าอินพุตลำดับที่ 1 ในสมการที่ 1 มีรายละเอียดดังนี้

$$f(0.5) = \frac{\exp(0.5)}{\exp(0.5) + \exp(3.9) + \exp(0.2) + \exp(0.3) + \exp(5.2)}$$

$$= \frac{1.6487}{1.6487 + 49.4023 + 1.2214 + 1.3499 + 181.2716}$$

$$= 0.0070$$

ผลลัพธ์จากฟังก์ชันซอฟต์แม็กซ์ของคำบรรยายรูปภาพลำดับที่ 1 ถึง 5 เท่ากับ 0.0070, 0.2103, 0.0052, 0.0057 และ 0.7717 ตามลำดับ ซึ่งจะมีผลรวมเท่ากับ 1 จากนั้นจึงวัดค่าความน่าจะเป็นของป้ายกำกับภาษาธรรมชาติว่ามีผลการพยากรณ์ใกล้เคียงกับผลการประเมินความหมายของป้ายกำกับจากผู้เชี่ยวชาญเพียงใดด้วยการหาค่าครอสเอนโทรปี (cross-entropy) [13] ดังสมการที่ 2

$$H(p, q) = -\sum_{x\epsilon\ classes} p(x) \log q\ (x) \quad (2)$$

กำหนดให้

$p(x)$ คือ ป้ายกำกับจริง (true label) แบบ one-hot โดยผลการประเมินความหมายของป้ายกำกับที่ผู้เชี่ยวชาญเลือกจะเท่ากับ 1 ตรงกันข้ามกับป้ายกำกับที่ไม่ถูกเลือกจะเท่ากับ 0 ทั้งหมด

$q(x)$ คือ ผลการพยากรณ์ความน่าจะเป็นของป้ายกำกับภาษาธรรมชาติจากตัวแบบ

จากตารางที่ 1 ป้ายกำกับที่ผู้เชี่ยวชาญเลือกคือลำดับที่ 5 ถือว่าเป็นป้ายกำกับจริง ดังนั้นตัวเลือกลำดับที่ 1 ถึง 4 จะเท่ากับ 0 ทั้งหมด แทนค่า p[0, 0, 0, 0, 1] ในทางกลับกัน ผลการพยากรณ์ความน่าจะเป็นของป้ายกำกับภาษาธรรมชาติจากฟังก์ชันซอฟต์แม็กซ์ในตัวเลือกที่ 1 ถึง 5 เมื่อแทนค่าใน q[0.0070, 0.2103, 0.0052, 0.005, 0.7717] ดังสมการที่ 2 มีรายละเอียดดังนี้

$$H(p,q) = -[0 * log_2(0.0070) + 0 * log_2(0.2103) + 0 * log_2(0.0052) + 0 * log_2(0.0057) + 1 * log_2(0.7717)]$$

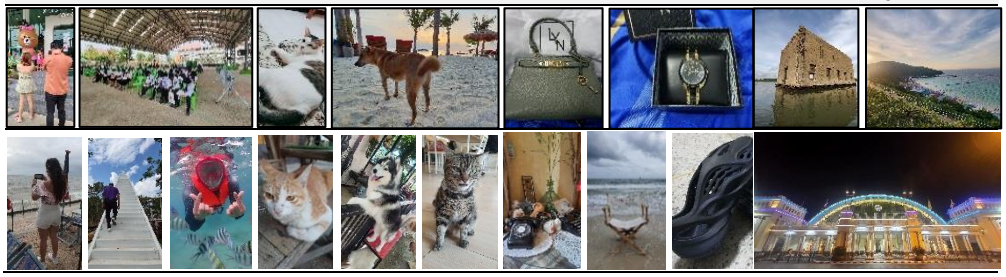
$$H(p,q) = 0.3739$$

ตัวอย่างรูปภาพจากตารางที่ 1 ผลการเปรียบเทียบระหว่างผลการประเมินความหมายของรูปภาพจากผู้เชี่ยวชาญคือตัวเลือกที่ 5 ตรงกับผลการพยากรณ์ความน่าจะเป็นของป้ายกำกับภาษาธรรมชาติที่สูงที่สุดในตัวเลือกที่ 5 ส่งผลให้ค่าครอสเอนโทรปี เท่ากับ 0.3739 โดยค่าครอสเอนโทรปีจะอยู่ระหว่าง 0 ถึงไม่จำกัด (infinity) โดย 0 คือไม่มีค่าสูญเสียเลย ซึ่งสื่อถึงความมั่นใจในผลการพยากรณ์ของตัวแบบ หากค่าครอสเอนโทรปีมีค่าน้อย แสดงถึงตัวแบบพยากรณ์ความน่าจะเป็นของป้ายกำกับถูกต้องตรงกับผลประเมินจากผู้เชี่ยวชาญได้อย่างมั่นใจ ตรงกันข้าม แม้ผลพยากรณ์ความน่าจะเป็นของป้ายกำกับถูกต้องแต่มีค่าความน่าจะเป็นน้อยค่าครอสเอนโทรปีก็จะมีค่ามาก เช่นเดียวกับผลพยากรณ์ความน่าจะเป็นของป้ายกำกับที่ไม่ถูกต้อง แม้จะมีค่าความน่าจะเป็นสูงก็จะทำให้ค่าครอสเอนโทรปีมากตามไปด้วย เมื่อเสร็จสิ้นจะนำคุณลักษณะเข้าสู่การฝังรูปภาพและผสานเป็นคุณลักษณะใหม่ของรูปภาพแล้วสร้างเป็นดัชนีคุณลักษณะ (index object) จัดเก็บไว้ที่ชุดคำอธิบายรูปภาพต่อไป

2) การเรียนรู้ซ้ำบนชุดข้อมูลที่กำหนดเอง โดยชุดข้อมูลรูปภาพที่ผู้วิจัยรวบรวมเอง กำหนดเป็น 4 โดเมน ได้แก่ 1) ภาพบุคคล 2) ภาพสัตว์เลี้ยง 3) ภาพสิ่งของ และ 4) ภาพสิ่งปลูกสร้าง รวมทั้งสิ้น 92,168 รูปภาพ ดังตารางที่ 2

ตารางที่ 2 ตัวอย่างรูปภาพจากชุดข้อมูลที่ผู้วิจัยรวบรวมเอง

ภาพบุคคล	ภาพสัตว์เลี้ยง	ภาพสิ่งของ	ภาพสิ่งปลูกสร้าง



กระบวนการเรียนรู้ด้วยตนเอง [14] เพื่อฝึกฝนแบบจำลองให้สามารถจัดหมวดหมู่วัตถุที่ไม่เคยเห็นมาก่อน (zero-shot prediction) คล้ายกับมนุษย์ที่สามารถแยกประเภทสิ่งของที่แตกต่างกันได้โดยที่ไม่มีการสอนหรือเรียนรู้มาก่อน เริ่มจากแบ่งข้อมูลออกเป็นชุดย่อย ๆ (batch size) เท่ากับ 256 สำหรับประมวลผลในแต่ละครั้ง และปรับรอบของการฝึกฝน (epochs) เท่ากับ 32 รอบ โดยผลลัพธ์จากการฝึกฝนตัวแบบจะเป็นตัวแบบที่มีค่าการสูญเสียค่อย ๆ ลดลงจนค่าการสูญเสียนั้นน้อยที่สุดจนแทบไม่มีการเปลี่ยนแปลงแล้วบันทึกไว้เพื่อเรียกใช้ในขั้นตอนถัดไป

3) การประเมินผลตัวแบบสำหรับงานปลายน้ำ ภายใต้ข้อความค้นหา 3 เงื่อนไขในรูปแบบภาษาอังกฤษที่รวบรวมโดยผู้วิจัย ได้แก่ 1) ข้อความบรรยายรูปภาพ ในลักษณะป้ายกำกับของรูปภาพ 2) ข้อความบรรยายรูปภาพเกี่ยวกับแนวคิดระดับสูงของภาพ และ 3) ข้อความบรรยายรูปภาพที่อธิบายความหมายเชิงคุณภาพของรูปภาพ อย่างไรก็ดี ความหมายของรูปภาพที่อยู่ในลักษณะนามธรรมนั้นยากที่จะประเมินผลด้วยค่าความถูกต้องเพียงอย่างเดียว จึงให้ผู้เชี่ยวชาญ 3 ท่านร่วมเป็นส่วนหนึ่งในการประเมินความหมายของแต่ละรูปภาพ ตามหลักการทดสอบทัวริง (Turing test) [15] ที่มุ่งทดสอบการทำงานของตัวแบบปัญญาประดิษฐ์เปรียบเทียบกับการทำงานของมนุษย์ โดยเปรียบเทียบระหว่างผลการประเมินความหมายของรูปภาพจากผู้เชี่ยวชาญกับผลการพยากรณ์ความน่าจะเป็นของป้ายกำกับภาษาธรรมชาติจากตัวแบบว่าแปลความหมายของรูปภาพได้ใกล้เคียงกับมนุษย์มากน้อยเพียงใดกำหนดคุณสมบัติผู้เชี่ยวชาญ [16] ว่าเป็นผู้สำเร็จการศึกษาด้านการจัดเก็บและเรียกค้นสารสนเทศ (information storage and retrieval) และมีประสบการณ์การเรียกค้นรูปภาพ 10 ปีขึ้นไป

งานวิจัยนี้คำนึงถึงการป้องกันผลประโยชน์ทับซ้อน (conflict of interest) โดยคัดเลือกผู้เชี่ยวชาญที่ไม่มีส่วนได้ส่วนเสียกับงานวิจัย ตลอดจนให้ความสำคัญกับความคลาดเคลื่อนเชิงสถิติ (statistical error) [17] โดยใช้การวัดซ้ำหลาย ๆ ครั้งเพื่อให้ความคลาดเคลื่อนลดน้อยลงจนอยู่ในระดับที่ยอมรับได้ โดยสร้างชุดคำถามแบบเลือกคำตอบที่ถูกต้องเพียงข้อเดียว จากนั้นให้ผู้เชี่ยวชาญแต่ละท่านประเมินความหมายทีละรูปภาพแบบเป็นอิสระต่อกันแทนสัญลักษณ์ด้วยวงกลมสีน้ำเงิน เมื่อครบรอบจะทำการประเมินอีกจนครบ 3 รอบ แล้วเก็บคำตอบไว้ในชุดคำตอบดังตารางที่ 3

ตารางที่ 3 ตัวอย่างผลการประเมินความหมายของรูปภาพแบบเลือกคำตอบที่ถูกต้องเพียงข้อเดียว

รูปภาพ	การประเมินความหมายของรูปภาพ		
	ข้อความบรรยายรูปภาพในลักษณะป้ายกำกับของรูปภาพ	ข้อความบรรยายรูปภาพเกี่ยวกับแนวคิดระดับสูงของรูปภาพ	ข้อความบรรยายรูปภาพที่อธิบายความหมายเชิงคุณภาพของรูปภาพ
1) 	(A) a photo of a dog (B) a photo of a cat (C) a photo of a cow (D) a photo of a pig (E) a photo of a horse	(A) a black and white dog is running in a grassy garden surrounded by a white fence. (B) a boston terrier is running on lush green grass in front of a white fence. (C) a black and white dog is running through the grass. (D) a dog runs on the green grass near a wooden fence. (E) a boston terrier is running in the grass.	(A) the dog is moving forward. (B) the dog was having fun playing on the lawn. (C) lonely dog running alone. (D) the dog is running towards something. (E) the dog ran aimlessly forward.

ผลการวิจัย

งานวิจัยนี้ถือว่ารูปภาพจากชุดข้อมูล flickr30k และชุดข้อมูลที่ผู้วิจัยรวบรวมเอง รวมทั้งสิ้น 123,951 รูปภาพที่นำมาใช้คือประชากร จึงเลือกกลุ่มตัวอย่างที่เป็นไปตามโอกาสทางสถิติ (probability sampling) ด้วยการสุ่มอย่างมีระบบ (systematic random sampling) [18] ได้กลุ่มตัวอย่างเป็นรูปภาพตามที่กำหนดในงานวิจัย 4 โดเมน โดเมนละ 100 รูปภาพ รวมทั้งสิ้น 400 รูปภาพ จากชุดข้อมูล Flickr30k จำนวน 100 รูปภาพ และอีก 300 รูปภาพจากชุดข้อมูลที่ผู้วิจัยรวบรวมเองตามสัดส่วนของจำนวนรูปภาพที่ใช้ในงานวิจัย การเปรียบเทียบผลการประเมินความหมายของรูปภาพจากผู้เชี่ยวชาญกับผลการพยากรณ์ความน่าจะเป็นของป้ายกำกับภาษาธรรมชาติ หากรูปภาพใดมีผลการพยากรณ์ความ

น่าจะเป็นของป้ายกำกับที่สูงที่สุดตรงกับคำตอบของผู้เชี่ยวชาญในชุดข้อมูลที่แทนสัญลักษณ์ด้วยวงกลมสีน้ำเงินถือว่าผลลัพธ์นั้นถูกต้องเท่ากับ 1 คะแนน ในทางกลับกัน หากรูปภาพใดที่ผลการพยากรณ์ความน่าจะเป็นของป้ายกำกับที่สูงที่สุดไม่ตรงกับคำตอบของผู้เชี่ยวชาญ จะเท่ากับ 0 คะแนน ดังตารางที่ 4

ตารางที่ 4 การเปรียบเทียบผลการประเมินความหมายของรูปภาพจากผู้เชี่ยวชาญกับผลการพยากรณ์ความน่าจะเป็นของป้ายกำกับภาษาธรรมชาติ

รูปภาพ	ผลการประเมินความหมายของรูปภาพ จากผู้เชี่ยวชาญ	ผลการพยากรณ์ความน่าจะเป็น
1) 	**1) ข้อความบรรยายรูปภาพ ในลักษณะป้ายกำกับของรูปภาพ**	
	(A) a photo of a dog	**0.82834270**
	(B) a photo of a cat	0.00169742
	(C) a photo of a cow	0.01496507
	(D) a photo of a pig	0.14255421
	(E) a photo of a horse	0.01244060
	2) ข้อความบรรยายรูปภาพที่อธิบายความหมายเชิงคุณภาพของรูปภาพ	
	(A) a black and white dog is running in a grassy garden surrounded by a white fence.	0.00701898
	(B) a boston terrier is running on lush green grass in front of a white fence.	0.21031745
	(C) a black and white dog is running through the grass.	0.00519979
	(D) a dog runs on the green grass near a wooden fence.	0.00574666 **0.77171712**
	(E) a boston terrier is running in the grass.	
	3) ข้อความบรรยายรูปภาพที่อธิบายความหมายเชิงคุณภาพของรูปภาพ	
	(A) the dog is moving forward.	0.26929662
	(B) the dog was having fun playing on the lawn.	0.04025207
	(C) lonely dog running alone.	0.28437802
	(D) the dog is running towards something.	**0.35570347**
	(E) a dog ran aimless forward.	0.05036977

ผลการประเมินความหมายของรูปภาพจากผู้เชี่ยวชาญกับผลการพยากรณ์ความน่าจะเป็นของป้ายกำกับภาษาธรรมชาติภายใต้ข้อความค้นหาใน 3 เงื่อนไข ค่าเฉลี่ยเท่ากับ 0.905, 0.830 และ 0.585 ตามลำดับ ดังตารางที่ 5

ตารางที่ 5 ผลการประเมินประสิทธิภาพการจำแนกรูปภาพด้วยป้ายกำกับภาษาธรรมชาติ

ชุดข้อมูลรูปภาพ	ผลการพยากรณ์ความน่าจะเป็นของป้ายกำกับภาษาธรรมชาติ		
	ข้อความบรรยายรูปภาพในลักษณะป้ายกำกับของรูปภาพ	ข้อความบรรยายรูปภาพเกี่ยวกับแนวคิดระดับสูงของรูปภาพ	ข้อความบรรยายรูปภาพที่อธิบายความหมายเชิงคุณภาพของรูปภาพ
Flickr30k	0.950	0.860	0.630
Custom	0.890	0.820	0.570
ค่าเฉลี่ย	**0.905**	**0.830**	**0.585**

อภิปรายผลการวิจัย

เมื่อพิจารณาลงไปในรายละเอียดจากผลการประเมินประสิทธิภาพการจำแนกรูปภาพด้วยป้ายกำกับภาษาธรรมชาติ พบว่า

1) ผลการพยากรณ์ป้ายกำกับภาษาธรรมชาติด้วยข้อความบรรยายรูปภาพในลักษณะป้ายกำกับของรูปภาพอยู่ในระดับสูงมาก แสดงถึงการกำหนดข้อความบรรยายรูปภาพส่งผลอย่างมากต่อผลการพยากรณ์ เนื่องจากหากกำหนดเพียงชื่อคลาส เช่น “boston terrier” ผลการพยากรณ์จะไม่สูงมากนัก แต่หากเพิ่มบริบทของคำจำกัดความ เช่น “boston terrier dog” ผลลัพธ์จะดีขึ้นเรื่อย ๆ และหากกำหนดเป็น “photo of a boston terrier dog” จะให้ผลลัพธ์สูงที่สุด โดยผลการพยากรณ์ความน่าจะเป็นของป้ายกำกับภาษาธรรมชาติเท่ากับ 0.27951103, 0.30876088 และ 0.4117281 ตามลำดับ

2) ผลการพยากรณ์ป้ายกำกับภาษาธรรมชาติด้วยข้อความบรรยายรูปภาพเกี่ยวกับแนวคิดระดับสูงของภาพอยู่ในระดับสูง แสดงถึงแบบจำลองทำงานได้ดีกับการจำแนกวัตถุในรูปภาพ (recognizing common objects) แต่ประสิทธิภาพจะลดลงเมื่อต้องจำแนกรายละเอียดเชิงลึก (very fine-grained classification) [19] เพื่อจำแนกความแตกต่างระหว่างคลาสย่อย (subclass) ของหมวดหมู่คลาสเดียวกัน โดยรูปภาพอาจคล้ายกันมากและอาจแตกต่างกันในบางคุณลักษณะที่สามารถแยกแยะได้ เช่น การจำแนกสายพันธุ์สุนัข ยกตัวอย่างรูปภาพจากตารางที่ 1 หากกำหนดป้ายกำกับของรูปภาพใหม่ในตัวเลือก (A) photo of a french bulldog, (B) photo of a boston terrier, (C) photo of a pug, (D) photo of a bulldog, (E) photo of a bullmastiff และ (F) photo of a chinese shar-pei ที่เป็นสุนัขสายพันธุ์หน้าสั้น ซึ่งมีเอกลักษณ์เฉพาะตัวคล้ายคลึงกันมาก พบว่า ผลการพยากรณ์ป้ายกำกับตัวเลือก (A), (B), (C), (D), (E) และ (F) จะเท่ากับ 2.2035880e-02, 9.1603160e-01, 2.6491896e-04, 6.0518362e-02, 8.5088005e-04 และ 2.9830670e-04 แม้ว่าผลการพยากรณ์ป้ายกำกับในตัวเลือก (B) ที่เป็นคำตอบที่ถูกต้องจะมีค่าความน่าจะเป็นสูงที่สุด แต่เมื่อรวมค่าจากการพยากรณ์ทั้งหมดจะเกิน 1 ดังนั้นหากต้องการคำตอบแบบจัดอันดับสำหรับคลาสที่มีหลายระดับชั้นเกี่ยวข้องกัน (multiclass relevance) การใช้งานฟังก์ชันซิกมอยด์ (sigmoid function) [13] จะเหมาะสมกว่าฟังก์ชันซอฟต์แม็กซ์

3) ผลการพยากรณ์ป้ายกำกับภาษาธรรมชาติด้วยข้อความบรรยายรูปภาพที่อธิบายความหมายเชิงคุณภาพของรูปภาพอยู่ในระดับปานกลาง โดยภาษาธรรมชาตินั้นถือว่าเป็นแนวคิดระดับสูงที่มีความซับซ้อน และมีความเป็นนามธรรมเข้ามาเกี่ยวข้อง ดังนั้นการแปลความหมายตามหลักการรับรู้ของมนุษย์ (human perception) [20] ด้วยประสบการณ์หรือความรู้เดิมของแต่ละบุคคลจึงแตกต่างกัน เห็นได้จากการผลการพยากรณ์จากตารางที่ 4 ที่มีความใกล้เคียงกันในหลายตัวเลือก เช่น ตัวเลือกที่ (A), (C) และ (D) เท่ากับ 0.26929662, 0.28437802 และ 0.35570347 ตามลำดับ เมื่อเทียบกับข้อความบรรยายรูปภาพเกี่ยวกับแนวคิดระดับสูงของภาพที่จะมีผลการพยากรณ์ในตัวเลือกใดตัวเลือกหนึ่งอย่างชัดเจน

4) ผลการพยากรณ์ป้ายกำกับภาษาธรรมชาตินั้นยังทำงานได้ไม่ดีนักกับประโยคปฏิเสธ เช่น หากเปลี่ยนป้ายกำกับในตัวเลือก (B) จากตารางที่ 4 เปลี่ยนเป็น “a photo containing no cat” พบว่า ผลการพยากรณ์ป้ายกำกับตั้งแต่ตัวเลือก (A), (B), (C), (D) และ (E) จะเปลี่ยนไปเป็น 0.8556212, 0.00291535, 0.01234535, 0.11747061 และ 0.01164759 เห็นได้ว่าผลการพยากรณ์ป้ายกำกับของตัวเลือก (B) นั้นต่ำมาก แม้จะเป็นป้ายกำกับที่ถูกต้องเช่นเดียวกับป้ายกำกับ “photo of a dog” ที่ให้ผลลัพธ์สูงที่สุด อีกทั้งผลการพยากรณ์ยังต่ำกว่าป้ายกำกับอื่น ๆ ที่ไม่ถูกต้องอีกด้วย

5) ผลการพยากรณ์ป้ายกำกับภาษาธรรมชาติบนชุดข้อมูล Flickr30k เปรียบเทียบกับชุดข้อมูลที่ผู้วิจัยกำหนดเอง พบว่า ผลการประเมินจะลดลงเล็กน้อย แต่เป็นไปในทิศทางเดียวกัน และไม่เกิดปัญหาที่ตัวแบบจะประสิทธิภาพจะลดลงเมื่อทำงานกับข้อมูลที่ไม่เคยพบมาก่อน

สรุปผลการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองการจำแนกรูปภาพตามความหมายสำหรับการเรียกค้นรูปภาพโดยใช้ภาษาธรรมชาติโดยใช้ตัวแบบที่ถูกฝึกฝนล่วงหน้าจากความสัมพันธ์ระหว่างรูปภาพกับข้อความ เพื่อแยกความแตกต่างของแต่ละคลาสด้วยการวัดความคล้ายคลึงโคไซน์ของเวกเตอร์รูปภาพและข้อความบรรยายรูปภาพตามหลักการเรียนรู้แบบคอนทราสต์และเรียนรู้ซ้ำบนชุดข้อมูลที่กำหนดเอง ผลการประเมินประสิทธิภาพการจำแนกรูปภาพ เพื่อเปรียบเทียบผลการประเมินความหมายของรูปภาพจากผู้เชี่ยวชาญกับผลการพยากรณ์ความน่าจะเป็นของป้ายกำกับภาษาธรรมชาติจากตัวแบบว่าแปลความหมายของรูปภาพได้ใกล้เคียงกับมนุษย์มากน้อยเพียงใดตามหลักการทดสอบทัวริง พบว่า 1) ข้อความบรรยายรูปภาพในลักษณะป้ายกำกับของรูปภาพนั้นมีผลการประเมินอยู่ในระดับสูงมาก โดยลักษณะข้อความบรรยายรูปภาพส่งผลอย่างมากต่อผลการพยากรณ์ 2) ข้อความบรรยายรูปภาพเกี่ยวกับแนวคิดระดับสูงของภาพที่มีผลการประเมินระดับสูงในการจำแนกวัตถุภายในรูปภาพ แต่ยังมีประสิทธิภาพลดลงเมื่อมีการนับจำนวนวัตถุในภาพ ตลอดจนการจำแนกที่มีการลงรายละเอียดเชิงลึก และ 3) ข้อความบรรยายรูปภาพที่อธิบายความหมายเชิงคุณภาพของรูปภาพมีผลการประเมินอยู่ในระดับปานกลาง เนื่องจากข้อความที่อยู่ในรูปแบบของป้ายกำกับภาษาธรรมชาตินั้นถือว่าเป็นแนวคิดระดับสูง ดังนั้นประสบการณ์ของแต่ละบุคคลจึงส่งผลให้การประเมินนั้นแตกต่างกันตามหลักการรับรู้ของมนุษย์

การวิจัยในครั้งต่อไปนั้นควรให้ความสำคัญกับการปรับค่าไฮเปอร์พารามิเตอร์ก่อนทำการเรียนรู้ เนื่องจากจะส่งผลให้ตัวแบบมีความแม่นยำที่สูงขึ้น หรือลดค่าการสูญเสียให้ต่ำลงได้ ตลอดจนวัดผลการแปลความหมายจากรูปภาพเป็นข้อความ (image-to-text) ด้วยค่าคะแนน BLEU เนื่องจากวัดได้ทั้งความครบถ้วน (adequacy) และความไพเราะ (fluency) จากความคล้ายกันของคำตามจำนวนเอ็นแกรม (n-gram) ในประโยคระหว่างผลการแปลจากตัวแบบเปรียบเทียบกับประโยคที่แปลโดยมนุษย์ ยิ่งมีคำที่ตรงกันมากเท่าไหร่จะส่งผลให้คะแนนสูงขึ้น โดยค่าจะอยู่ระหว่าง 0 ถึง 100

เอกสารอ้างอิง

[1] Tyagi, V. (2017). *Content-Based Image Retrieval Ideas, Influences, and Current Trends*. Gateway East: Springer.

[2] Barz, B., & Denzler, J. (2020). Content-based Image Retrieval and the Semantic Gap in the Deep Learning Era (pp 2 - 19). In *International Workshop on Content-Based Image Retrieval: where have we been, and where are we going (CBIR 2020)*. Italy.

[3] Aggarwal, C. C. (2018). *Neural Networks and Deep Learning: A Textbook*. Cham: Springer.

[4] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998).Gradient-based learning applied to document recognition (pp 2278-2324). In *Proceedings of the IEEE*. USA.

[5] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Massachusetts: MIT Press.

[6] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition (pp 770-778). In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. USA.

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need (pp 6000-6010). In *31st Conference on Neural Information Processing Systems.* USA.

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

[9] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision (pp 8748-8763). In *38th International Conference on Machine Learning (ICML 2021)*.

[10] Xu, M., Yoon, S., Fuentes, A., & Park, D. S. (2023). A Comprehensive Survey of Image Augmentation Techniques for Deep Learning. *Pattern Recognition*, 137, 109347.

[11] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv: 1910.01108, 1-5.

[12] Nwankpa, C., Ijomah, W., Gachagan, A., & Marshall, S. (2018). Activation Functions: Comparison of trends in Practice and Research for Deep Learning. arXiv preprint arXiv: 1910.01108, 1-20.

[13] Rosebrock, A. (2017). *Deep Learning for Computer Vision with Python*. New York: PYIMAGESEARCH.

[14] Sawarka, K. (2022). *Deep Learning with PyTorch Lightning Swiftly build high-performance Artificial Intelligence (AI) models using Python*. Birmingham: Packt.

[15] Christian, B. (2011). *The Most Human Human: What Talking with Computers Teaches Us About What It Means to Be Alive*. New York: Doubleday.

[16] อรนุช ศรีสะอาด. (2561). การตรวจสอบความเที่ยงตรงของเครื่องมือวัดผลโดยผู้เชี่ยวชาญ. *วารสารการวัดผลการศึกษา มหาวิทยาลัยมหาสารคาม*, 1(1), 45-49.

[17] เรวัต แสงสุริยงค์. (2565). ความเสี่ยงของการเกิดความคลาดเคลื่อนในการวิจัยเชิงปริมาณด้านสังคมวิทยา. *วารสารวิชาการมนุษยศาสตร์และสังคมศาสตร์ มหาวิทยาลัยบูรพา*, 30(1), 158-185.

[18] Brase, C. H., & Brase, C. P. (2018). *Understanding Basic Statistics*. Boston: Cengage Learning.

[19] Benois-Pineau, J., & Zemmari, A. (2021). *Multi-faceted Deep Learning: Models and Data*. Cham: Springer.

[20] Dix, A., Finlay, J., Abowd, G. D., & Beale, R. (2004). *Human–Computer Interaction* (3rd edition). Harlow: Pearson.