การพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมาย โดยใช้โครงข่ายประสาทเทียมที่ถูกฝึกฝนล่วงหน้า
The Development of a Semantic-based Image Retrieval Model by Pre-training Neural Network

จักรินทร์ สันติรัตนภักดี[1*] และ ศุภกฤษฎิ์ นิวัฒนากูล[1]
Chakkarin Santirattanaphakdi[1] and Suphakit Niwattanakul[1]
[1]สำนักวิชาศาสตร์และศิลป์ดิจิทัล มหาวิทยาลัยเทคโนโลยีสุรนารี
[1]Institute of Digital Arts and Science, Suranaree University of Technology
วันที่ส่งบทความ : 23 เมษายน 2566 วันที่แก้ไขบทความ : 10 กรกฎาคม 2566 วันที่ตอบรับบทความ : 15 พฤศจิกายน 2566
Received: 23 April 2023, Revised: 10 July 2023, Accepted: 15 November 2023

บทคัดย่อ

วิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมาย โดยประยุกต์ใช้ตัวแบบ Contrastive Language-Image Pre-training (CLIP) ที่ถูกฝึกฝนล่วงหน้า ผลประเมินประสิทธิภาพการค้นคืนรูปภาพด้วยค่าความแม่นยำ ค่าความครบถ้วน และค่าประสิทธิภาพโดยรวม พบว่าผลการค้นคืนรูปภาพจากข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ในลักษณะป้ายกำกับ และข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงนั้นมีค่าความแม่นยำอยู่ในระดับสูงมาก แสดงถึงตัวแบบสามารถค้นคืนรูปภาพจากเนื้อหาได้อย่างมีประสิทธิภาพ อย่างไรก็ดี ผลการค้นคืนรูปภาพจากข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพที่แม้จะมีค่าความแม่นยำในระดับดี แต่ผลลัพธ์นั้นยังห่างไกลจากความคาดหวังของผู้ใช้ เนื่องจากรูปภาพจะถูกแปลความหมายด้วยประสบการณ์ของแต่ละบุคคลตามหลักการรับรู้ของมนุษย์ อีกทั้งความหมายของรูปภาพที่อยู่ในลักษณะนามธรรมนั้นยากที่จะประเมินผลว่าถูกต้องหรือไม่ถูกต้อง ผลผลิตจากงานวิจัยนี้สามารถแก้ไขปัญหาช่องว่างความหมาย และช่วยสนับสนุนผู้ใช้ด้วยข้อความค้นหาภาษาธรรมชาติที่ยึดโยงกับความหมายของรูปภาพแทนที่จะยึดตามหลักไวยากรณ์ของภาษา อันก่อให้เกิดผลลัพธ์ที่เป็นแนวทางการค้นคืนสารสนเทศเชิงความหมายต่อไปในอนาคต

คำสำคัญ : การค้นคืนรูปภาพ คุณลักษณะเชิงความหมาย โครงข่ายประสาทเทียมที่ถูกฝึกฝนล่วงหน้า

Abstract

This research aims to develop a semantic-based image retrieval model applying the Contrastive Language-Image Pre-training (CLIP) model. Evaluation of image retrieval

[*]ที่อยู่ติดต่อ E-mail address: d6200701@g.sut.ac.th

performance with precision, recall and f-measure, it was found that image search results with the query by global labels condition and the query by high level concepts of the images condition had a very good level of precision, the model can efficiently retrieve images from the content. However, image retrieval results with the query by qualitative semantic concepts of the image condition, despite having a good level of precision. But the results are far from the user's expectations because the semantic of image is interpreted by experience on human perception principles. In addition, also, the semantic of image is difficult to evaluate whether they are correct or not. The output from this research can resolve the semantic gap problem and support users by query within a natural language that attaches to the semantic of the image rather than the grammar of the language. This impact of results in a guideline for semantic information retrieval in the future.

Keywords: Image retrieval, Semantic feature, Pre-training neural network

1. บทนำ

การค้นคืนสารสนเทศในปัจจุบันนั้นมิได้จำกัดเพียงแค่ตัวอักษรหรือข้อความเท่านั้น เนื่องจากข้อมูลบนอินเทอร์เน็ตอยู่ในลักษณะมัลติมีเดีย (Multimedia) ซึ่งประกอบด้วยข้อมูลที่เป็นรูปภาพ เสียง และวิดีโอมากพอ ๆ กับข้อมูลที่เป็นตัวอักษร แต่โปรแกรมค้นคืนสารสนเทศหรือเสิร์ชเอนจิน (Search engine) ที่มีอยู่ในปัจจุบันนั้นมีประสิทธิภาพสูงสำหรับค้นคืนข้อมูลที่เป็นตัวอักษรเท่านั้น แต่สำหรับข้อมูลประเภทอื่น ๆ ถือว่าประสิทธิภาพยังห่างไกลจากการค้นคืนข้อมูลที่เป็นตัวอักษรอยู่มาก เนื่องจากความแตกต่างของลักษณะข้อมูล ดังนั้นระบบค้นคืนสารสนเทศสำหรับข้อมูลเหล่านี้จึงมีการทำงานที่ซับซ้อนมากกว่าระบบค้นคืนตัวอักษร โดยเฉพาะอย่างยิ่งข้อมูลประเภทรูปภาพที่มีจำนวนมากในปัจจุบันตามข้อมูลของเว็บไซต์ Phototutorial [1] รายงานว่าทั่วโลกมีรูปภาพถึง 1.2 ล้านล้านภาพในปี ค.ศ. 2021 และเพิ่มเป็น 1.72 ล้านล้านในปี ค.ศ. 2022 จำนวนนี้จะเพิ่มขึ้นเป็น 1.81 ล้านล้านในปี ค.ศ. 2023 เมื่อถึงปี ค.ศ. 2025 จะมีรูปภาพมากกว่า 2 ล้านล้านภาพ และมีแนวโน้มที่จะเพิ่มมากขึ้นอีกในอนาคต โดยมีปัจจัยหลักอยู่ที่การพัฒนาจากกล้องฟิล์มเป็นอุปกรณ์ถ่ายภาพดิจิทัล ประกอบกับอุปกรณ์ถ่ายภาพที่มีราคาถูกลงและสะดวกต่อการใช้งานในลักษณะของสมาร์ทโฟน อีกทั้งพฤติกรรมของผู้คนที่นิยมใช้สื่อสังคมออนไลน์ในการแบ่งปันเรื่องราวและบันทึกความทรงจำ ก่อให้เกิดรูปภาพจำนวนมหาศาลบนอินเทอร์เน็ต อีกทั้งยากต่อการนำไปใช้ประโยชน์

การค้นคืนรูปภาพเป็นหนึ่งในศาสตร์คอมพิวเตอร์วิทัศน์ที่มุ่งเปลี่ยนรูปภาพให้กลายเป็นข้อมูลที่มีความหมาย เพื่อให้คอมพิวเตอร์สามารถเข้าใจรูปภาพได้ใกล้เคียงมนุษย์มากที่สุด โดยในยุคแรกจะใช้การค้นคืนรูปภาพจากการเปรียบเทียบคำสำคัญ (Keyword) กับชื่อไฟล์รูปภาพหรือคำอธิบายภาพ เมื่อพบข้อความที่ตรงกัน (String matching) [2] รูปภาพนั้นก็จะถูกดึงมาเป็นผลลัพธ์ แต่ปัญหาที่พบคือบางครั้งการตั้งชื่อไฟล์ไม่ตรงกับเนื้อหาของรูปภาพ หรือคำอธิบายภาพที่ใส่โดยมนุษย์ อาจไม่สามารถอธิบายได้ครอบคลุมเนื้อหาของรูปภาพได้หมดส่งผลให้ผลลัพธ์จากการค้นคืนมีประสิทธิภาพต่ำ อย่างไรก็ดี การค้นคืน

รูปภาพด้วยข้อความ (Text Based Image Retrieval: TBIR) [3] ก็ยังเป็นวิธีการที่นิยมใช้อยู่ในปัจจุบัน เช่น การค้นคืนรูปภาพผ่านเสิร์ชเอนจินที่ผลลัพธ์จากการค้นคืนอาจเป็นรูปภาพที่มีข้อความค้นหานั้นฝังอยู่ในภาพ หรืออาจเป็นรูปภาพภายในเว็บเพจที่มีหัวข้อหรือเนื้อหาที่ตรงกับข้อความค้นหาดังกล่าว ตามแนวคิดการใช้ข้อความที่อยู่รอบ ๆ รูปภาพในการอธิบายความหมายของรูปภาพมานำเสนอแก่ผู้ใช้ ก่อนจะพัฒนาเป็นการค้นคืนรูปภาพเชิงเนื้อหา (Content Based Image Retrieval: CBIR) [3] จากเดิมที่ใช้คุณลักษณะระดับต่ำของรูปภาพ (Low-level features) [4] ซึ่งเป็นข้อมูลที่ไม่มีความหมายในตัวเอง และไม่สื่อถึงความหมายใด ๆ ที่อยู่ในรูปภาพ แบ่งเป็นคุณลักษณะแบบโกลบอล (Global features) เช่น สี (Color) รูปทรง (Sharp) พื้นผิว (Texture) รวมถึงตำแหน่งของพื้นที่ (Spatial) และคุณลักษณะแบบโลคอล (Local features) ซึ่งจะเป็นคุณลักษณะเฉพาะของแต่ละส่วนในรูปภาพไปใช้ค้นคืนโดยตรง แบบไม่มีการแปลงให้อยู่ในรูปแบบของข้อมูลที่มนุษย์สามารถเข้าใจได้ โดยใส่ข้อความค้นหาเข้าไปในระบบ และทำการค้นหาภาพในฐานข้อมูลที่มีคุณลักษณะความคล้ายคลึงกับข้อความค้นหามากที่สุดมาแสดงเป็นผลลัพธ์ ทำให้ระบบที่ใช้คุณลักษณะระดับต่ำเพื่อค้นคืนรูปภาพนั้น มีประสิทธิภาพไม่สูงเท่าที่ควร เนื่องจากปัญหาช่องว่างความหมาย (Semantic gap) [5] ที่คุณลักษณะระดับต่ำเหล่านี้ไม่สามารถสื่อความหมายที่อยู่ในรูปภาพได้อย่างถูกต้อง อันเป็นข้อจำกัดสำคัญของการค้นคืนรูปภาพเชิงเนื้อหา

อย่างไรก็ตาม “A picture is worth a thousand words” แสดงถึงเนื้อหารูปภาพมีความหมายหลากหลายมากกว่าข้อความ และพฤติกรรมการค้นคืนรูปภาพของผู้ใช้ที่โดยมากจะค้นคืนรูปภาพโดยใช้ความสัมพันธ์ระหว่างคุณลักษณะระดับต่ำ และแนวคิดระดับสูง (High-level concept) เช่น นามธรรม (Abstract) วัตถุ (Object) หรือเหตุการณ์ (Event) เข้าด้วยกันในลักษณะข้อความค้นหาภาษาธรรมชาติ ที่จัดว่าเป็นแนวคิดระดับสูง เนื่องจากมีความซับซ้อน และมีความเป็นนามธรรมเข้ามาเกี่ยวข้อง ดังนั้น การแปลความหมายจึงแตกต่างกันตามประสบการณ์หรือความรู้เดิมของแต่ละบุคคล ส่งผลให้ผลลัพธ์จากการค้นคืนรูปภาพจึงยังห่างไกลจากความคาดหวังของผู้ใช้ เนื่องจากคุณลักษณะเชิงเนื้อหาเพียงอย่างเดียวนั้นไม่สามารถสื่อถึงความหมายที่แท้จริงของรูปภาพได้อย่างสมบูรณ์ จึงเกิดเป็นแนวคิดการค้นคืนรูปภาพเชิงความหมาย (Semantic Based Image Retrieval: SBIR) [6] ที่มุ่งเชื่อมโยงคุณลักษณะระดับต่ำและแนวคิดระดับสูงของรูปภาพให้อยู่ในรูปแบบที่มนุษย์สามารถเข้าใจได้ ปัจจุบันมีการนำเสนอแนวคิดเกี่ยวกับการค้นคืนรูปภาพเชิงความหมายมากมาย โดยประยุกต์หลายศาสตร์มาบูรณาการร่วมกัน เพื่อแก้ไขปัญหาช่องว่างความหมาย โดยหนึ่งในวิธีที่ได้รับความนิยมอย่างมากนั่นคือแนวคิดการเรียนรู้ของเครื่อง (Machine learning) ที่พัฒนามาเป็นการเรียนรู้เชิงลึก (Deep learning) [7] ซึ่งมีจุดเด่นในการวิเคราะห์และจำแนกคุณลักษณะรูปภาพได้อย่างหลากหลาย และมีประสิทธิภาพการทำนายผลได้ดีกว่าการเรียนรู้ของเครื่องแบบเดิมเป็นอย่างมาก [8]

การปริทัศน์งานวิจัยที่มุ่งค้นคืนรูปภาพเชิงความหมายในปัจจุบัน โดย Caicedo และคณะ [9] นำเสนอวิธีการค้นคืนรูปภาพเชิงเนื้อหาด้วยการแยกคุณสมบัติระดับสูงจากการวิเคราะห์ความหมายของรูปภาพค้นหา (Image query) พบว่า การค้นคืนด้วยคุณลักษณะระดับต่ำเพียงอย่างเดียว มีค่าความแม่นยำสูงสุดเท่ากับ 0.67 เมื่อผลลัพธ์จากการค้นคืนมีเพียง 1 รูปภาพ แต่เมื่อใช้ร่วมกับคุณลักษณะเชิงความหมาย (Semantic features) ค่าความแม่นยำเพิ่มเป็น 0.80 อย่างไรก็ดี ค่าความแม่นยำจะลดลงเรื่อย ๆ แปรผันตามจำนวนผลลัพธ์จากการค้นคืนที่เพิ่มมากขึ้น ในขณะที่ Khodaskar และ Ladhake [10] นำเสนอการ

วิเคราะห์เนื้อหารูปภาพค้นหาจากการแทนคุณลักษณะเนื้อหาเชิงความหมาย (Semantic content representation) ในฐานความรู้ ร่วมกับการดึงข้อมูลและการจัดทำดัชนีสำหรับการค้นคืนรูปภาพ พบว่าค่าความแม่นยำอยู่ระหว่าง 0.79 ถึง 0.89 จากนั้น Thanh และคณะ [11] นำเสนอระบบค้นคืนรูปภาพเชิงความหมายด้วยรูปภาพค้นหา พบว่า การใช้เทคนิคเคมีน (K-mean) เพียงอย่างเดียวมีค่าความแม่นยำเท่ากับ 0.65 แต่เมื่อผสมผสานเทคนิคเพื่อนบ้านใกล้ที่สุด (K-nearest neighbor) เข้าด้วยกัน ค่าความแม่นยำเพิ่มเป็น 0.70 ในขณะที่ Nhi และคณะ [12] นำเสนอตัวแบบการค้นคืนรูปภาพเชิงเนื้อหาจากการวิเคราะห์รูปภาพค้นหาด้วยกราฟ C-tree และเทคนิคเพื่อนบ้านใกล้ที่สุด พบว่า ค่าความแม่นยำบนชุดข้อมูล COREL เท่ากับ 0.89 เมื่อพิจารณาในรายละเอียด พบว่า งานวิจัยส่วนใหญ่มุ่งใช้รูปภาพค้นหาในการค้นคืนรูปภาพเชิงความหมาย เนื่องจากระบบจะวิเคราะห์คุณลักษณะรูปภาพเพื่อค้นคืนรูปภาพตามความเหมือนหรือคล้ายคลึงกับรูปภาพค้นหา จึงสามารถลดปัญหาช่องว่างความหมายได้อย่างชัดเจน แต่ในความเป็นจริงพฤติกรรมผู้ใช้โดยมากจะค้นคืนรูปภาพโดยใช้ความสัมพันธ์ระหว่างคุณลักษณะระดับต่ำ ผสมผสานแนวคิดระดับสูงที่เป็นรูปธรรม (Concrete concept) และแนวคิดระดับสูงที่เป็นนามธรรม (Abstract concept) รวมเข้าด้วยกันเป็นข้อความค้นหาภาษาธรรมชาติที่อยู่ในลักษณะแนวคิดระดับสูง [13] แต่ในทางกลับกันคุณลักษณะที่ถูกแยกอัตโนมัติโดยใช้คอมพิวเตอร์มักเป็นคุณลักษณะระดับต่ำเท่านั้น ซึ่งโดยทั่วไปไม่มีการเชื่อมโยงระหว่างคุณลักษณะระดับต่ำและแนวคิดระดับสูง ส่งผลให้ผลลัพธ์จากการค้นคืนมีประสิทธิภาพไม่สูงเท่าที่ควร อีกทั้งการค้นคืนโดยใช้รูปภาพค้นหายังไม่สะดวก และเป็นการผลักภาระให้กับผู้ใช้ในการหารูปภาพต้นฉบับอีกด้วย

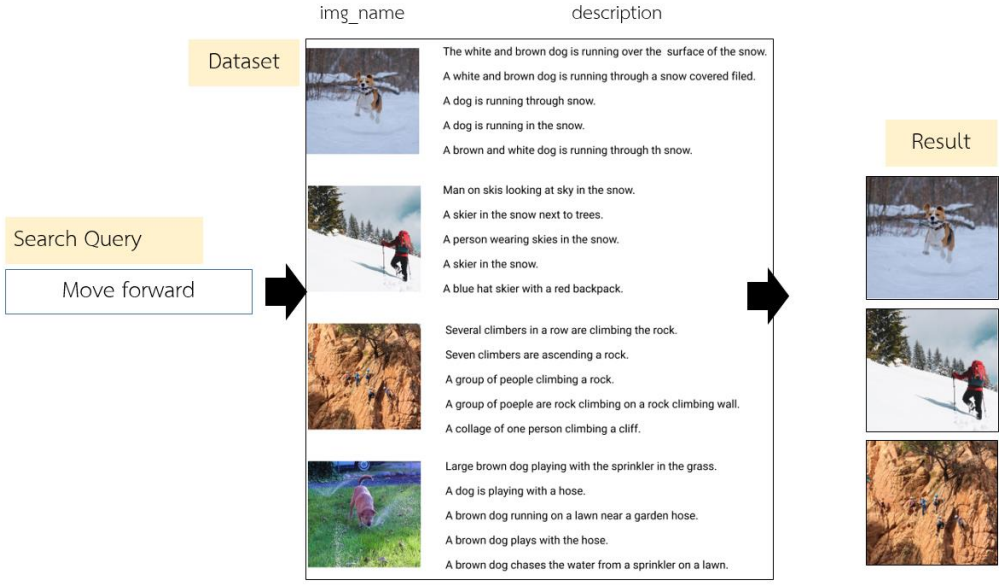
งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายโดยประยุกต์ใช้ตัวแบบ Contrastive Language Image Pre-training (CLIP) [14] ที่ถูกฝึกฝนล่วงหน้าในการเรียนรู้คุณลักษณะของรูปภาพ เพื่อสร้างดัชนีรองรับการค้นคืนรูปภาพเชิงความหมาย และประเมินประสิทธิภาพการค้นคืนรูปภาพบนชุดข้อมูล Flickr30k และ Unsplash เปรียบเทียบกับชุดข้อมูลรูปภาพที่ผู้วิจัยรวบรวมเอง (Custom dataset) เพื่อป้องกันปัญหา Overfitting ที่แบบจำลองสามารถจำแนกข้อมูลได้อย่างมีประสิทธิภาพกับชุดข้อมูลทดสอบ แต่ประสิทธิภาพจะลดลงเมื่อทำงานกับข้อมูลที่ไม่เคยพบมาก่อน และช่วยสนับสนุนผู้ใช้ที่ไม่สามารถกำหนดคำหลักหรือคำที่เฉพาะเจาะจงสำหรับการค้นหาที่เหมาะสมได้ ด้วยการใช้ข้อความค้นหาภายใต้รูปแบบของภาษาธรรมชาติที่ยึดโยงกับความหมายของรูปภาพแทนที่จะยึดตามหลักไวยากรณ์ของภาษา อันจะเป็นแนวทางการค้นคืนสารสนเทศเชิงความหมายในอนาคต

2. วิธีการทดลอง

งานวิจัยนี้ประยุกต์ใช้แนวคิดแบบจำลองการพัฒนาแอปพลิเคชันแบบเร่งรัด (Rapid Application Development: RAD) [15] มาเป็นกรอบในการดำเนินงานด้วยการพัฒนาตัวแบบให้มีขั้นตอนการดำเนินงานที่รวบรัดมากขึ้นจากการใช้เครื่องมือและเทคนิคต่าง ๆ ก่อนจะพัฒนาเป็นระบบที่สมบูรณ์ต่อไป ประกอบด้วยการวางแผนความต้องการ (Requirement planning) การออกแบบ (Design) การสร้าง (Construction) และการทดสอบและตรวจสอบซ้ำ (Testing and turnover) มีรายละเอียดดังนี้

2.1 การวางแผนความต้องการ

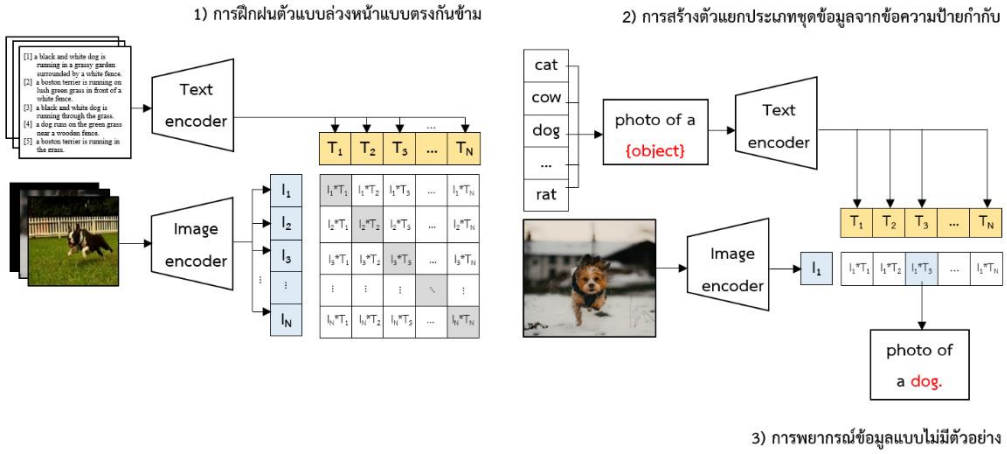
กระบวนการค้นคืนรูปภาพเชิงความหมาย [6] คือ การเรียนรู้คุณลักษณะของรูปภาพ (Visual image extraction) เพื่อเชื่อมโยงไปยังคุณลักษณะเชิงความหมายที่อยู่ในรูปภาพด้วยการแปลความหมายรูปภาพระดับสูง (High level interpretation) ก่อนจะสร้างเป็นดัชนีคุณลักษณะของรูปภาพในการค้นคืนรูปภาพเชิงความหมาย ดังรูปที่ 1 ตัวอย่างข้อความค้นหาที่อาจทำให้เกิดปัญหาช่องว่างความหมาย เช่น “move forward” เมื่อเปรียบเทียบโดยใช้อักขระกับคำบรรยายภาพ หรือเปรียบเทียบกับคุณลักษณะเชิงเนื้อหาของรูปภาพทั้งหมดแล้ว อาจไม่พบผลลัพธ์ที่เกี่ยวข้องกับข้อความค้นหาเลย แต่เมื่อนำมาเปรียบเทียบกับคุณลักษณะเชิงความหมาย จะพบว่ามีผลลัพธ์จำนวน 3 รูปภาพ ได้แก่ รูปภาพสุนัขที่กำลังวิ่งไปข้างหน้า รูปภาพนักสกีที่กำลังมุ่งหน้าไปบนยอดเขาหิมะ หรือรูปภาพนักไต่เขาที่กำลังปีนมุ่งหน้าไปยังยอดเขาที่ทั้ง 3 รูปภาพไม่ปรากฏคำว่า “move forward” ภายในรูปภาพหรือคำบรรยายภาพเลย แต่อาจมีความหมายแฝงว่า “move forward” ที่แปลว่ามุ่งไปข้างหน้า อันเป็นคุณลักษณะที่เกี่ยวข้องกับข้อความค้นหาจากผู้ใช้



รูปที่ 1. กระบวนการค้นคืนรูปภาพเชิงความหมาย

การวิจัยนี้ประยุกต์ใช้ตัวแบบ CLIP ที่ถูกฝึกฝนล่วงหน้า [14] ในการเรียนรู้คุณลักษณะรูปภาพเพื่อแปลเป็นความหมายรูปภาพระดับสูงด้วยการเรียนรู้แบบตรงกันข้าม (Contrastive learning) [16] จากรูปภาพและคำบรรยายภาพ เพื่อฝึกฝนตัวเข้ารหัสรูปภาพ (Image encoder) และตัวเข้ารหัสข้อความ (Text encoder) ด้วยการพยายามขยับเวกเตอร์ตัวแทน (Vector representation) ระหว่างรูปภาพและข้อความบรรยายภาพคู่เดียวกัน (Positive point) ให้เข้าใกล้กัน และขยับตัวแทนของรูปภาพและข้อความบรรยายภาพคู่ที่แตกต่างกัน (Negative point) ให้ออกห่างกันมากขึ้นด้วยค่าความคล้ายคลึงเชิงมุมโคไซน์

(Cosine similarity) ดังนั้นรูปภาพและข้อความบรรยายภาพคู่ที่ถูกต้องจะมีค่าความคล้ายคลึงเข้าใกล้กับ 1 มากขึ้น ตรงกันข้ามกับค่าความคล้ายคลึงของคู่ที่ไม่ถูกต้องจะลดลงเรื่อย ๆ จนมีค่าเข้าใกล้กับ 0 ก่อนจะสร้างตัวแยกประเภทชุดข้อมูลจากข้อความป้ายกำกับ (Create dataset classifier from label text) โดยฝัง อ็อบเจกต์คลาสลงในข้อความบรรยายรูปภาพ (Object classes into captions) ดังรูปที่ 2 อ็อบเจกต์คลาสที่ฝังเป็นข้อความบรรยายรูปภาพ คือ “a photo of a dog” เพื่อนำไปใช้ในการค้นคืนรูปภาพต่อไป



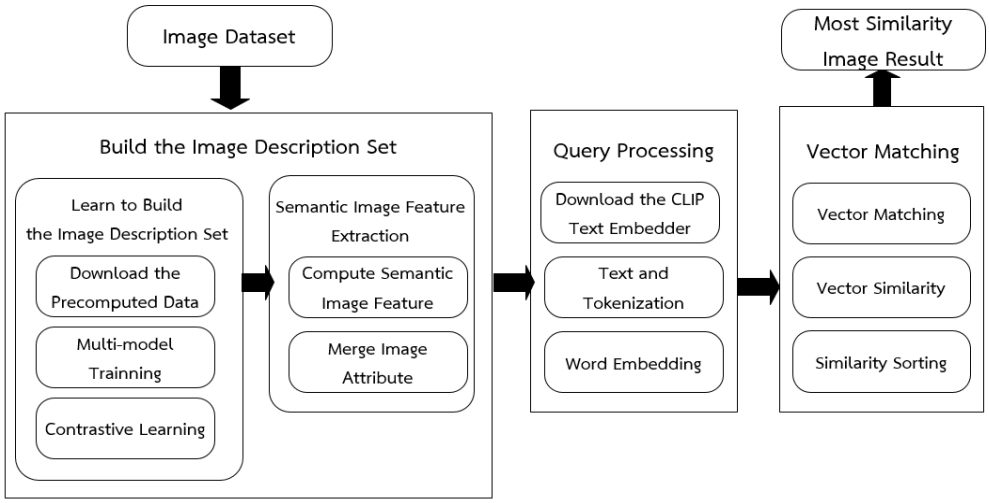
รูปที่ 2. กระบวนการทำงานของตัวแบบ CLIP ที่ถูกฝึกฝนล่วงหน้า [14]

ปัจจุบันมีหลายงานวิจัยที่นำตัวแบบ CLIP ที่ถูกฝึกฝนล่วงหน้ามาเป็นโมเดลฐาน (Baseline model) ในการเรียนรู้ระหว่างรูปภาพและข้อความ ได้แก่ Lu และคณะ [17] นำเสนอตัวแบบ Vision and Language BERT (ViLBERT) สำหรับการเรียนรู้จากรูปภาพและข้อความภาษาธรรมชาติร่วมกัน จากการนำสถาปัตยกรรม BERT มาปรับปรุงเป็นตัวแบบสองทางหลายรูปแบบ (Multimodal two stream model) ประมวลผลทั้งรูปภาพและข้อความในสตรีมแยกกันผ่านเลเยอร์ Transformer เช่นเดียวกับ Li และคณะ [18] นำเสนอตัวแบบ Unicoder-VL ซึ่งเป็นตัวเข้ารหัสเพื่อเรียนรู้ความหมายรูปภาพและภาษาธรรมชาติโดยผสานแนวคิดจากโมเดลที่ฝึกฝนล่วงหน้าข้ามภาษา เช่น Extensible Markup Language (XML) และ Unicoder จากการใช้รูปภาพและข้อความเป็นอินพุตเข้าสู่โมเดล Transformer แบบหลายเลเยอร์สำหรับการฝึกฝนตัวแบบล่วงหน้า ประกอบด้วยการสร้างแบบจำลองภาษามาสก์ (Masked Language Modeling: MLM) การจำแนกวัตถุมาสก์ (Masked Object Classification: MOC) เพื่อเรียนรู้ความหมายรูปภาพ และการจับคู่ระหว่างรูปภาพและภาษาศาสตร์ (Visual Linguistic Matching: VLM) สำหรับคาดการณ์ว่าคู่รูปภาพและข้อความนั้นอธิบายความหมายซึ่งกันและกันหรือไม่ จากนั้นจะถ่ายโอนความรู้ไปยังการค้นคืนรูปภาพตามข้อความบรรยายรูปภาพ (Caption-based image-text retrieval) และการให้เหตุผลภาพ (Visual commonsense reasoning) และ Chen และคณะ [19] นำเสนอตัวแบบ Universal Image Text Representation (UNITER) ที่ถูกฝึกฝนล่วงหน้าบนชุดข้อมูลรูปภาพขนาดใหญ่ที่มีการฝังต่อเนื่องหลายรูปแบบร่วมกัน ประกอบด้วยการสร้างแบบจำลองภาษามาสก์ และการจำแนกวัตถุมาสก์ในการเรียนรู้

รูปภาพและข้อความ การจับคู่รูปภาพและข้อความ (Image Text Matching: ITM) และการจัดรูปแบบข้อความ (Word Region Alignment: WRA) เพื่อสนับสนุนการจัดตำแหน่งแบบละเอียดระหว่างข้อความและขอบเขตรูปภาพระหว่างการฝึกฝนตัวแบบ

2.2 การออกแบบ

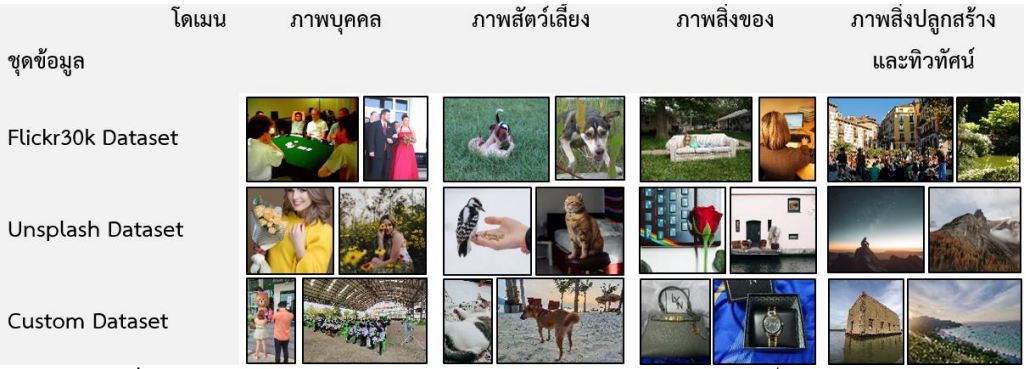
งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายโดยประยุกต์ใช้ตัวแบบ CLIP ที่ถูกฝึกฝนล่วงหน้าในการเรียนรู้คุณลักษณะเชิงความหมายจากความสัมพันธ์ระหว่างรูปภาพกับข้อความบรรยายรูปภาพที่อยู่ในลักษณะข้อความภาษาธรรมชาติ แทนที่จะเรียนรู้ภายใต้คลาสที่จำกัดเช่นเดิมอย่างสุนัขหรือแมว ดังนั้นเมื่อตัวแบบได้รับการฝึกฝนทั้งประโยค จะส่งผลให้ตัวแบบสามารถเรียนรู้สิ่งต่าง ๆ ได้มากขึ้น และสามารถค้นหารูปแบบความสัมพันธ์บางอย่างระหว่างรูปภาพกับข้อความด้วยการเรียนรู้ด้วยตนเอง ก่อนจะผสานเป็นคุณลักษณะรูปภาพใหม่จัดเก็บในชุดคำอธิบายรูปภาพ จากนั้นจึงแปลงเป็นเวกเตอร์เพื่อเปรียบเทียบความคล้ายคลึงระหว่างเวกเตอร์คุณลักษณะรูปภาพกับเวกเตอร์คุณลักษณะข้อความค้นหาจากผู้ใช้ด้วยการวัดค่าความคล้ายคลึงเวกเตอร์ ก่อนจะเรียงลำดับตามความเกี่ยวข้อง และนำเสนอรูปภาพมาแสดงเป็นผลลัพธ์แก่ผู้ใช้ การออกแบบจึงนำทฤษฎีการวิเคราะห์และออกแบบระบบเชิงวัตถุ โดยใช้แบบจำลอง Unified Modeling Language (UML) [20] ที่เป็นประเภทของไดอะแกรมโครงสร้างสแตติกที่อธิบายโครงสร้างของระบบโดยแสดงคลาสของระบบ คุณลักษณะ การดำเนินการ และความสัมพันธ์ระหว่างอ็อบเจ็กต์ ดังรูปที่ 3



รูปที่ 3. การอธิบายโครงสร้างของแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมาย
โดยใช้โครงข่ายประสาทเทียมที่ถูกฝึกฝนล่วงหน้าโดยใช้แบบจำลอง UML

2.3 การสร้าง

แบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายพัฒนาขึ้นด้วยการเขียนโปรแกรมภาษาไพธอน (Python) บน Google Colaboratory เรียกใช้ตัวแบบ CLIP ที่ถูกฝึกฝนล่วงหน้าเรียนรู้แบบตรงกันข้ามบนชุดข้อมูลรูปภาพ Flickr30k และชุดข้อมูลรูปภาพ Unsplash เพื่อประเมินประสิทธิภาพการค้นคืนรูปภาพเปรียบเทียบกับชุดข้อมูลรูปภาพที่ผู้วิจัยรวบรวมเองกำหนดเป็น 4 โดเมน ได้แก่ 1) ภาพบุคคล 2) ภาพสัตว์เลี้ยง 3) ภาพสิ่งของ และ 4) ภาพสิ่งปลูกสร้างและทิวทัศน์ ดังรูปที่ 4



รูปที่ 4. ตัวอย่างรูปภาพในชุดข้อมูล Flickr30k, Unsplash และชุดข้อมูลรูปภาพที่ผู้วิจัยรวบรวมเอง

2.4 การทดสอบและการกลับมาตรวจสอบซ้ำ

การวัดประสิทธิภาพการค้นคืนรูปภาพดิจิทัลแบบไม่ทราบจำนวนผลลัพธ์ที่ถูกต้อง (Order-unaware metrics) [21] ซึ่งถูกใช้วัดประสิทธิภาพการค้นคืนสารสนเทศบนเว็บไซต์ ด้วยการประเมินค่าความแม่นยำที่ k อันดับ ($precision@k$) ค่าความครบถ้วนที่ k อันดับ ($recall@k$) และค่าประสิทธิภาพโดยรวมที่ k อันดับ ($f-measure@k$) ดังสมการที่ (1) - (3)

$$precision@k = \frac{true\ positive@k}{(true\ positive@k)+(false\ positive@k)} \quad (1)$$

$$recall@k = \frac{true\ positive@k}{(true\ positive@k)+(false\ negative@k)} \quad (2)$$

$$f-measure@k = \frac{2*(precision@k)*(recall@k)}{(precision@k)+(recall@k)} \quad (3)$$

กำหนดข้อความค้นหาใน 3 เงื่อนไข [22] ได้แก่ 1) ข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ในลักษณะป้ายกำกับ เช่น the dog, the cat หรือ the animal เป็นต้น 2) ข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูง เช่น the dog running on a grass เป็นต้น และ 3) ข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพ เช่น the dog feeling lonely เป็นต้น เงื่อนไขละ 100 รายการ รวมทั้งสิ้น 300 รายการ กำหนดการแสดงผลเป็น 3 รูปภาพ 5 รูปภาพ และ 10 รูปภาพ ตามลำดับ ดังรูปที่ 5 ตัวอย่างผลลัพธ์จากข้อความค้นหา “the dog running on a grass” กำหนดจำนวนรูปภาพที่แสดงผลคือ 5 รูปภาพ แทนด้วย k ที่ผ่านการประเมินผลจากผู้เชี่ยวชาญท่านที่ 1

รูปที่ 5. ตัวอย่างผลลัพธ์ 5 รายการจากการค้นคืนรูปภาพดิจิทัลที่ผ่านการประเมินความถูกต้อง

จากรูปที่ 5 กำหนดคะแนนเมื่อผลลัพธ์ถูกต้องตามข้อความค้นหาเท่ากับ 1 คะแนน ตรงกันข้ามกับเมื่อผลลัพธ์ไม่ถูกต้องหรือไม่ครบถ้วนตามข้อความค้นหาเท่ากับ 0 คะแนน ตัวอย่างผลการประเมินแบบไม่ทราบจำนวนผลลัพธ์ที่ถูกต้อง เมื่อมีรูปภาพที่ถูกต้องจำนวน 4 รูปภาพจาก 5 รูปภาพ โดยรูปภาพที่ 1, 2, 3 และ 5 ถูกต้อง และรูปภาพที่ 4 ไม่ถูกต้อง ดังตารางที่ 1

ตารางที่ 1. ตัวอย่างผลการประเมินประสิทธิภาพการค้นคืนรูปภาพดิจิทัลแบบไม่ทราบจำนวนผลลัพธ์ที่ถูกต้อง

ลำดับ	ค่าความแม่นยำ	ค่าความครบถ้วน	ค่าประสิทธิภาพโดยรวม
รูปภาพที่ 1 ถูกต้อง	1/1 = 1.00	1/4 = 0.25	$\frac{2 * (1/1) * (1/4)}{(1/1) + (1/4)} = 0.40$
รูปภาพที่ 2 ถูกต้อง	2/2 = 1.00	2/4 = 0.50	$\frac{2 * (2/2) * (2/4)}{(2/2) + (2/4)} = 0.67$
รูปภาพที่ 3 ถูกต้อง	3/3 = 1.00	3/4 = 0.75	$\frac{2 * (3/3) * (3/4)}{(3/3) + (3/4)} = 0.86$
รูปภาพที่ 4 ไม่ถูกต้อง	3/4 = 0.75	3/4 = 0.75	$\frac{2 * (3/4) * (3/4)}{(3/4) + (3/4)} = 0.75$
รูปภาพที่ 5 ถูกต้อง	4/5 = 0.80	4/4 = 1.00	$\frac{2 * (4/5) * (4/4)}{(4/5) + (4/4)} = 0.89$
ค่าเฉลี่ย	**0.91**	**0.65**	**0.71**

จากตารางที่ 1 ผลการประเมินประสิทธิภาพการค้นคืนรูปภาพดิจิทัลแบบไม่ทราบจำนวนผลลัพธ์ที่ถูกต้องจากข้อความค้นหา “the dog running on a grass” กำหนดการแสดงผลคือ 5 รูปภาพ พบว่า

มีค่าความแม่นยำ ค่าความครบถ้วน และค่าประสิทธิภาพโดยรวม เท่ากับ 0.91, 0.65 และ 0.71 ตามลำดับ จากนั้นดำเนินการทดสอบจนครบทั้ง 300 รายการก่อนจะคำนวณเป็นประสิทธิภาพการค้นคืนรูปภาพของตัวแบบ ประกอบด้วยค่าความแม่นยำเฉลี่ย ค่าความครบถ้วนเฉลี่ย และค่าประสิทธิภาพโดยรวมเฉลี่ย

งานวิจัยนี้ให้ความสำคัญความคลาดเคลื่อน (Error) คือ ผลต่างระหว่างค่าที่วัดได้กับค่าที่แท้จริง หากค่าที่วัดได้ใกล้เคียงกับค่าจริงมากแสดงว่าการวัดนั้นมีความถูกต้องสูง [23] โดยการวัดทุกครั้งมักมีค่าความคลาดเคลื่อนเกิดขึ้นเสมอ แบ่งออกเป็น 3 ประเภท ได้แก่ 1) ความคลาดเคลื่อนที่เกิดจากผู้วัด (Human error) อันเกิดจากความประมาทเลินเล่อในการวัด 2) ความคลาดเคลื่อนเชิงระบบ (Systematic error) มักเกิดจากเครื่องมือวัดไม่ได้ประสิทธิภาพ และ 3) ความคลาดเคลื่อนเชิงสถิติ (Statistical error) ซึ่งความคลาดเคลื่อนประเภทที่ 1 และ 2 สามารถจัดการได้ด้วยความรอบคอบ ระมัดระวัง และการพัฒนาเครื่องมือวัดให้มีประสิทธิภาพ แต่ความคลาดเคลื่อนประเภทที่ 3 นั้นไม่มีทางกำจัดให้หมดไปได้ เนื่องจากอยู่นอกเหนือการควบคุม จึงมักใช้การวัดซ้ำหลาย ๆ ครั้งเพื่อให้ความคลาดเคลื่อนลดน้อยลง ดังนั้นเมื่อครบรอบจะทำการประเมินอีกครั้งจนครบ 3 รอบด้วยข้อความค้นหาเดิมแบบผลลัพธ์เป็นอิสระต่อกัน แล้วนำผลการประเมินมาหาค่าเฉลี่ยเพื่อให้ความคลาดเคลื่อนลดน้อยลงจนถึงในระดับที่ยอมรับได้

3. ผลการทดลองและวิจารณ์

การพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมาย โดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า แบ่งผลการทดลองและวิจารณ์ออกเป็น 2 ส่วน ได้แก่ 1) การพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมาย และ 2) การประเมินประสิทธิภาพการค้นคืนรูปภาพดิจิทัล

3.1 ผลการพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมาย

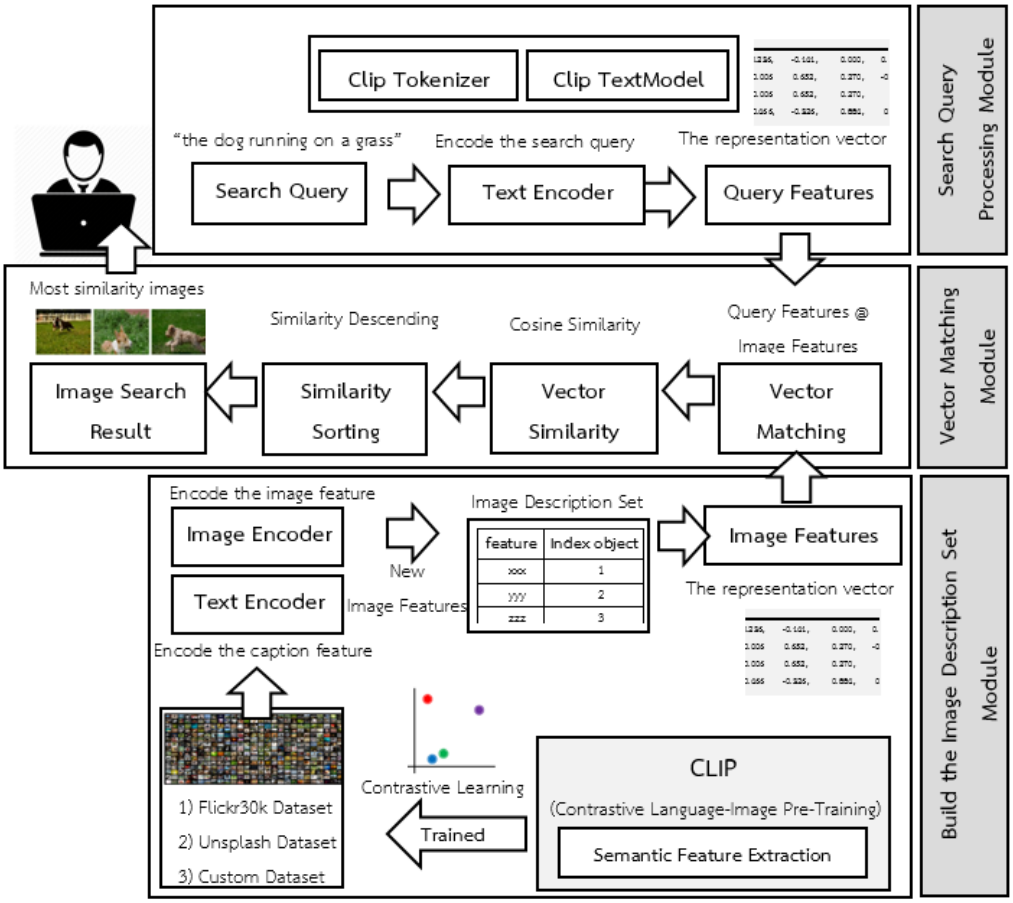
ภาพรวมกระบวนการทำงานของแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมาย โดยใช้โครงข่ายประสาทเทียมที่ถูกฝึกฝนล่วงหน้า ดังรูปที่ 6 ประกอบด้วย 3 มอดูล ได้แก่ มอดูลการสร้างชุดคำอธิบายรูปภาพ มอดูลการประมวลผลข้อความค้นหา และมอดูลการจับคู่เวกเตอร์ มีรายละเอียดดังนี้

1) มอดูลการสร้างชุดคำอธิบายรูปภาพ (Build the image description set module) จากการใช้ตัวแบบ CLIP ที่ถูกฝึกฝนล่วงหน้ามาเรียนรู้คุณลักษณะเชิงความหมายของรูปภาพด้วยกระบวนการเรียนรู้ด้วยตนเองแบบตรงกันข้าม ระหว่างรูปภาพและข้อความบรรยายรูปภาพจากค่าความคล้ายคลึงโคไซน์บนชุดข้อมูล Flickr30k, Unsplash และชุดข้อมูลที่ผู้วิจัยรวบรวมเองก่อนจะผสานเป็นคุณลักษณะใหม่ของรูปภาพแล้วจัดเก็บไว้ที่ชุดคำอธิบายรูปภาพเพื่อสร้างเป็นเวกเตอร์คุณลักษณะรูปภาพ

2) มอดูลการประมวลผลข้อความค้นหา (Search query processing module) จากผู้ใช้ในรูปแบบภาษาธรรมชาติ โดยเรียกใช้ CLIPTokenizer สำหรับเข้ารหัสข้อความ เพื่อสร้างเป็นเวกเตอร์คุณลักษณะข้อความ และ CLIPTextModel สำหรับฝังข้อความบนพื้นที่การฝังหลายรูปแบบ (Multi-modal embedding space) เพื่อสร้างเป็นเวกเตอร์คุณลักษณะข้อความค้นหา

3) มอดูลการจับคู่เวกเตอร์ (Vector matching module) ระหว่างเวกเตอร์คุณลักษณะรูปภาพและเวกเตอร์คุณลักษณะข้อความค้นหา เพื่อนำมาเปรียบเทียบค่าความคล้ายคลึงของเวกเตอร์ ก่อนจะเรียงลำดับตามความเกี่ยวข้อง และแสดงเป็นผลลัพธ์การค้นคืนรูปภาพทั้งในบริบทของเนื้อหาและบริบท

เชิงความหมายแก่ผู้ใช้ กำหนดจำนวนผลลัพธ์การค้นคืนรูปภาพที่เกี่ยวข้องกับข้อความค้นหาแต่ละครั้ง เท่ากับ 3 รูปภาพ 5 รูปภาพ และ 10 รูปภาพ ตามลำดับ



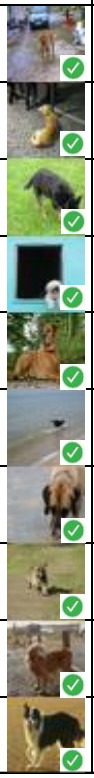
รูปที่ 6. ภาพรวมกระบวนการทำงานของแบบจำลองการค้นคืนรูปภาพดิจิทัล

3.2 ผลการประเมินประสิทธิภาพการค้นคืนรูปภาพดิจิทัล

งานวิจัยนี้ประเมินประสิทธิภาพการค้นคืนรูปภาพแบบไม่ทราบจำนวนผลลัพธ์ที่ถูกต้องด้วยค่าความแม่นยำที่ k อันดับ ค่าความครบถ้วนที่ k อันดับ และค่าประสิทธิภาพโดยรวมที่ k อันดับ โดยผู้วิจัยกำหนดข้อความค้นหาภาษาอังกฤษในรูปแบบภาษาธรรมชาติจากการสุ่มคำบรรยายรูปภาพใน 3 เงื่อนไข [22] ได้แก่ 1) ข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ ในลักษณะป้ายกำกับ 2) ข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูง และ 3) ข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพ อย่างละ 100 รายการ รวมทั้งสิ้น 300 รายการ กำหนดรูปภาพที่แสดงผลในแต่ละชุดข้อมูลเป็น 3 รูปภาพ 5 รูปภาพและ 10 รูปภาพต่อ 1

ข้อความค้นหาผ่าน Google Colaboratory ดังตารางที่ 2 ตัวอย่างผลลัพธ์จากการค้นคืนรูปภาพในแต่ละเงื่อนไขที่ผ่านการประเมินความถูกต้องโดยผู้เชี่ยวชาญ

ตารางที่ 2. ผลลัพธ์จากการค้นคืนรูปภาพที่ผ่านการประเมินความถูกต้องโดยผู้เชี่ยวชาญ

ตัวอย่างข้อความค้นหา		**เงื่อนไขที่ 1** "the dog"			**เงื่อนไขที่ 2** "the dog running on a grass"			**เงื่อนไขที่ 3** "the dog feeling lonely"		
ชุดข้อมูล (Dataset)		Flickr30k	Unsplash	Custom	Flickr30k	Unsplash	Custom	Flickr30k	Unsplash	Custom
จำนวนรูปภาพจากการค้นคืน (*k*)	*k* @3									
	k @5									
	k @10									

จากตารางที่ 2 คุณลักษณะเชิงความหมายนั้นยากที่จะประเมินผลว่าถูกต้องหรือไม่ถูกต้องเท่านั้น เนื่องจากคุณลักษณะเชิงความหมายของรูปภาพจะอยู่ในลักษณะนามธรรม จึงเป็นที่มาของการให้ผู้เชี่ยวชาญ จำนวน 3 ท่าน [24] เข้ามาร่วมเป็นส่วนหนึ่งในการประเมินผลความหมายของแต่ละรูปภาพให้สอดคล้องกับหลักคอมพิวเตอร์วิทัศน์ (Computer vision) ประยุกต์แนวคิดจากอรนุช ศรีสะอาด [24] มากำหนดคุณสมบัติผู้เชี่ยวชาญว่าเป็นผู้สำเร็จการศึกษาด้านการจัดเก็บและค้นคืนสารสนเทศ (Information storage and retrieval) หรือสาขาที่เกี่ยวข้อง ตลอดจนมีประสบการณ์การค้นคืนรูปภาพอย่างน้อย 10 ปีขึ้นไป จากนั้นให้ผู้เชี่ยวชาญแต่ละท่านประเมินความถูกต้องทีละรูปภาพแบบเป็นอิสระ

ต่อกันเพื่อลดความคลาดเคลื่อนให้อยู่ในระดับที่ยอมรับได้ หากรูปภาพใดมีผลการประเมินจากผู้เชี่ยวชาญทั้ง 3 ท่านเป็นถูกต้องทั้งหมด จะถือว่ารูปภาพนั้นถูกต้องตามข้อความค้นหาเท่ากับ 1 คะแนน ในทางกลับกัน หากรูปภาพใดถูกผู้เชี่ยวชาญประเมินว่าไม่ถูกต้องแม้เพียงท่านเดียว จะถือว่ารูปภาพนั้นไม่ถูกต้องเท่ากับ 0 คะแนน ก่อนจะคำนวณเป็นค่าความแม่นยำเฉลี่ยที่ k อันดับ ค่าความครบถ้วนเฉลี่ยที่ k อันดับ และค่าประสิทธิภาพโดยรวมเฉลี่ยที่ k อันดับ ดังตารางที่ 3

ตารางที่ 3. ผลการประเมินประสิทธิภาพการค้นคืนรูปภาพแบบไม่ทราบจำนวนผลลัพธ์ที่ถูกต้อง

เงื่อนไขการค้นหา		1) ข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ ในลักษณะป้ายกำกับ			2) ข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูง			3) ข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพ		
ผลการประเมินประสิทธิภาพการค้นคืนรูปภาพ		Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure
k @3	Flickr30k & Unsplash	**0.950**	0.560	0.705	**0.903**	0.521	0.661	**0.760**	0.420	0.541
	Custom	**0.929**	0.506	0.655	**0.900**	0.505	0.647	**0.749**	0.415	0.534
k @5	Flickr30k & Unsplash	0.941	0.609	0.739	0.885	0.572	0.695	0.734	0.475	0.577
	Custom	0.901	0.569	0.698	0.882	0.569	0.692	0.722	0.466	0.566
k @10	Flickr30k & Unsplash	0.927	**0.668**	**0.776**	0.862	**0.635**	**0.731**	0.695	**0.534**	**0.604**
	Custom	0.854	**0.639**	**0.731**	0.837	**0.632**	**0.720**	0.689	**0.525**	**0.596**

จากตารางที่ 3 ผลการประเมินประสิทธิภาพการค้นคืนรูปภาพดิจิทัลแบบไม่ทราบจำนวนผลลัพธ์ที่ถูกต้อง มีรายละเอียดดังนี้

1) ผลการประเมินประสิทธิภาพการค้นคืนรูปภาพดิจิทัลจากข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ในลักษณะป้ายกำกับ และข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงนั้นมีค่าความแม่นยำอยู่ในระดับสูงมาก โดยเฉพาะผลลัพธ์การค้นคืนจำนวน 3 ลำดับแรก ค่าเฉลี่ย 0.950 และ 0.903 ตามลำดับ เช่นเดียวกับผลการค้นคืน 5 อันดับที่มีค่าเฉลี่ย 0.941 และ 0.885 ตามลำดับ ซึ่งค่าความแม่นยำนั้นลดลงเรื่อย ๆ แปรผันตามจำนวนผลลัพธ์การค้นคืนรูปภาพที่เพิ่มมากขึ้น

2) ผลการประเมินประสิทธิภาพการค้นคืนรูปภาพดิจิทัลด้วยข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพ โดยผลลัพธ์การค้นคืนจำนวน 3 ลำดับแรกนั้นมีค่าความแม่นยำเพียง 0.760 แม้จะอยู่ในระดับ

ดี แต่ผลลัพธ์นั้นยังห่างไกลจากความคาดหวังของผู้ใช้ อันจะเห็นได้จากค่าความแม่นยำที่ลดลงอย่างเห็นได้ชัดเมื่อเปรียบเทียบกับข้อความค้นหาในเงื่อนไขที่ 1 และ 2 คิดเป็นลดลงร้อยละ 25 และ 19 ตามลำดับ เนื่องจากความหมายของรูปภาพที่อยู่ในลักษณะนามธรรมนั้นจะถูกแปลความหมายตามหลักการรับรู้ของมนุษย์ ดังนั้นประสบการณ์ของแต่ละบุคคลจึงส่งผลต่อการประเมินความถูกต้องที่แตกต่างกัน

3) ผลการประเมินประสิทธิภาพการค้นคืนรูปภาพดิจิทัลบนชุดข้อมูล Flickr30k และ Unsplash เปรียบเทียบกับชุดข้อมูลที่ผู้วิจัยรวบรวมเองนั้น พบว่า ค่าความแม่นยำของข้อความค้นหาเงื่อนไขที่ 1 มีค่าเฉลี่ยความแม่นยำของผลลัพธ์การค้นคืนรูปภาพจำนวน 3 รูปภาพ 5 รูปภาพและ 10 รูปภาพ เท่ากับ 0.929, 0.901 และ 0.854 ตามลำดับ อยู่ในระดับดีมาก เช่นเดียวกับเงื่อนไขที่ 2 ที่มีค่าเฉลี่ยความแม่นยำเท่ากับ 0.900, 0.882 และ 0.837 ตามลำดับ อยู่ในระดับดีมาก และ เงื่อนไขที่ 3 มีค่าเฉลี่ยความแม่นยำเท่ากับ 0.749, 0.722 และ 0.689 ตามลำดับ อยู่ในระดับดี ผลลัพธ์ดังกล่าวเป็นไปในทิศทางเดียวกันตามความซับซ้อนของข้อความค้นหา และไม่เกิดปัญหาที่ตัวแบบจะประสิทธิภาพลดลงเมื่อทำงานกับข้อมูลที่ไม่เคยพบมาก่อน (Poor real-world performance)

4. สรุปผลการทดลอง

การพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมาย โดยใช้โครงข่ายประสาทเทียมที่ถูกฝึกฝนล่วงหน้าประกอบด้วย 3 มอดูล ได้แก่ 1) มอดูลการสร้างชุดคำอธิบายรูปภาพ โดยประยุกต์ใช้ตัวแบบ CLIP ที่ถูกฝึกฝนล่วงหน้าในการเรียนรู้คุณลักษณะเชิงความหมายของรูปภาพและผสานเป็นคุณลักษณะใหม่ของรูปภาพ ก่อนจะแปลงเป็นเวกเตอร์คุณลักษณะรูปภาพ 2) มอดูลการประมวลผลข้อความค้นหา สำหรับเข้ารหัสข้อความค้นหาจากผู้ใช้ ก่อนจะแปลงเป็นเวกเตอร์คุณลักษณะข้อความค้นหา และ 3) มอดูลการจับคู่เวกเตอร์คุณลักษณะรูปภาพและคุณลักษณะข้อความค้นหาด้วยการเปรียบเทียบค่าความคล้ายคลึงของเวกเตอร์ ก่อนจะเรียงลำดับตามความเกี่ยวข้อง และแสดงผลการค้นคืนรูปภาพแก่ผู้ใช้

การประเมินประสิทธิภาพการค้นคืนรูปภาพด้วยข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ในลักษณะป้ายกำกับของรูปภาพ และข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของภาพนั้นมีค่าความแม่นยำอยู่ในระดับสูงมาก แสดงถึงตัวแบบมีประสิทธิภาพสูงในการจำแนกวัตถุภายในรูปภาพ (Recognizing common objects) อย่างไรก็ดี ความแปรผันของรูปภาพถือเป็นอุปสรรคสำคัญต่อการจำแนกรูปภาพ [25] ได้แก่ 1) ความแปรผันภายในคลาส (Intra-class variation) ที่มีวัตถุหลายประเภทที่ถูกจัดอยู่ในคลาสเดียวกัน จึงจำเป็นต้องลดความแตกต่างภายในคลาส ขณะเดียวกันก็ขยายความแตกต่างระหว่างคลาส (Inter-class variation) เพื่อเพิ่มประสิทธิภาพการจำแนก 2) ความแปรผันของมาตราส่วน (Scale variation) เนื่องจากวัตถุเดียวกันนั้นมีหลายขนาด ดังนั้นความละเอียดของภาพ ขนาดของวัตถุและขนาดภาพ จึงมีผลต่อความถูกต้องของผลลัพธ์ 3) ความแปรผันของมุมมอง (View-point variation) โดยที่วัตถุเดียวกันแต่สามารถวางแนวหรือหมุนได้หลายมิติ 4) วัตถุบางส่วนถูกบดบัง (Occlusion) หรือซ่อนอยู่หลังวัตถุอื่น ๆ ส่งผลให้ประสิทธิภาพการจำแนกลดลง ตามขนาดของพื้นที่ที่บดบังวัตถุ 5) ความสว่าง (Illumination) โดยวัตถุในรูปภาพที่มีค่าแสงน้อย หรือวัตถุในที่มืดนั้นประสิทธิภาพการจำแนกจะต่ำกว่าวัตถุในแสงสว่างปกติ และ 6) พื้นหลังยุ่งเหยิง (Background clutter) จากการมีวัตถุจำนวนมากในภาพ ทำให้ยากต่อการจำแนกวัตถุ เนื่องจากมีสิ่งรบกวนมากเกินไป อย่างไรก็ดี ยังพบข้อจำกัดใน

สถานการณ์ที่มีการนับจำนวนวัตถุในภาพ ตลอดจนการจำแนกรายละเอียดเชิงลึก (Fine-grained classification) เช่น สายพันธุ์สุนัข สายพันธุ์ดอกไม้ หรือยี่ห้อรถยนต์ เป็นต้น

ผลการค้นคืนรูปภาพเชิงความหมายในงานวิจัยนี้มีค่าความแม่นยำบนชุดข้อมูล Flickr30k และ Unsplash เท่ากับ 0.760, 0.734 และ 0.695 ที่จำนวนรูปภาพเท่ากับ 3 รูปภาพ 5 รูปภาพ และ 10 รูปภาพตามลำดับ อยู่ในระดับดี และเป็นระดับที่ยอมรับได้ เนื่องจากความหมายของรูปภาพที่อยู่ในลักษณะนามธรรมนั้นยากที่จะประเมินผลว่าถูกต้องหรือไม่ถูกต้องเท่านั้น ดังนั้นการแปลความหมายตามหลักการรับรู้ของมนุษย์ (Human perception) [26] ด้วยประสบการณ์หรือความรู้เดิมจากการเรียนรู้ของแต่ละบุคคลจึงแตกต่างกัน เห็นได้จากผลลัพธ์การค้นคืนจากข้อความค้นหาในรูปแบบของภาษาธรรมชาติที่ยึดโยงกับความหมายของรูปภาพแทนที่จะยึดตามหลักไวยากรณ์ของภาษามีค่าความแม่นยำไม่แตกต่างจากการใช้รูปภาพค้นหามากนัก อย่างไรก็ดี ปัญหาช่องว่างความหมายของรูปภาพนั้นยากที่จะทำให้หมดไป ทำได้เพียงลดช่องว่างความหมายระหว่างคอมพิวเตอร์ ผู้ใช้ และรูปภาพลงด้วยเทคนิคและวิธีการต่าง ๆ เท่านั้น [27] งานวิจัยในครั้งต่อไปจึงควรให้ความสำคัญกับการลดช่องว่างความหมายของข้อความค้นหาจากผู้ใช้ควบคู่ไปกับการวิเคราะห์ความหมายของรูปภาพ เพื่อเพิ่มประสิทธิภาพในการค้นคืน ผลผลิตจากงานวิจัยนี้ได้นำเสนออีกหนึ่งแนวทางการค้นคืนสารสนเทศเชิงความหมายต่อไปในอนาคต

เอกสารอ้างอิง (References)

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

[2] AbdElrazek, E.E. 2017. A Comparative Study of Image Retrieval Algorithms for Enhancing a Content-based Image Retrieval System. *Global Journal of Computer Science and Technology: (F) Graphics & Vision*, 17(3), 1-9.

[3] Tyagi, V. 2017. Content-Based Image Retrieval Ideas, Influences, and Current Trends. 1st ed, Springer Nature, Springer Nature Singapore Pte Ltd.

[4] Liu, Y., Huang, Y., Zhang, S., Zhang, D. and Ling, N. 2017. Integrating object ontology and region semantic template for crime scene investigation image retrieval. Proceedings 12th IEEE Conference on Industrial Electronics and Applications (ICIEA), Siem Reap, 49-153.

[5] Alkhawlani, M., Elmogy, M. and Elbakry, H. 2015. Content-Based Image Retrieval using Local Features Descriptors and Bag-of-Visual Words. *International Journal of Advanced Computer Science and Applications*, 6(9), 212-219.

[6] Barz, B. 2020. Semantic and Interactive Content-based Image Retrieval. Ph.D. Thesis, University of Jena.

[7] Goodfellow, I., Bengio, Y. and Courville, A. 2016. Deep Learning. MIT Press, Cambridge, Massachusetts Institute of Technology.

[8] Aggarwal, C.C. 2018. Neural Networks and Deep Learning A Textbook. 1st ed, Springer Nature, Springer International Publishing.

[9] Caicedo, J.C., Gonzalez, F.A. and Romero, E. 2008. A Semantic Content-Based Retrieval Method for Histopathology Images. Proceedings 4th Asia Information Retrieval Symposium (AIRS 2008), Harbin, 51-60.

[10] Khodaskar, A. and Ladhake, S. 2015. Semantic Image Analysis for Intelligent Image Retrieval. *Procedia Computer Science*, 48(2015), 192-197.

[11] Thanh, L.M., Nhi, N.T.U., Thi, N.T.U., Han, P.T.A. and Thanh, V. T. 2020. A Semantic-Based Image Retrieval System Using A Hybrid Method K-Means And K-Nearest-Neighbor. *Annales Univ. Sci. Budapest., Sec. Comp*., 51, 253-274.

[12] Nhi, N.T.U., Le, T.M. and Van, T.T. 2022. A Model of Semantic-Based Image Retrieval Using C-Tree and Neighbor Graph. *International journal on Semantic Web and information systems*, 18(1), 1-23.

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

[14] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. Proceedings 38th International Conference on Machine Learning, Virtual Event, 8748-8763.

[15] McConnell, S. 1996. Rapid Development: Taming Wild Software Schedules. 1st ed, Microsoft Press, Microsoft.

[16] Mu, N., Kirillov, A., Wagner, D. and Xie, S. 2021. SLIP: Self-supervision meets Language-Image Pre-training. Available at: https://arxiv.org/pdf/2112.12750.pdf. Retrieved 16 September 2023.

[17] Lu, J., Batra, D., Parikh, D. and Lee, S. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Task. Available at: https://arxiv.org/pdf/1908.02265.pdf. Retrieved 16 September 2023.

[18] Li, G., Duan, N., Fang, Y., Gong, M., Jiang, D. and Zhou, M. 2019. Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training. Available at: https://arxiv.org/pdf/1908.06066.pdf. Retrieved 16 September 2023.

[19] Chen, Y., Li, L., Yu, L., Kholy, A.E. Ahmed, F., Gan, Z. Cheng, Y. and Liu, J. 2019. UNITER: UNiversal Image-TExt Representation Learning. Available at: https://arxiv.org/pdf/1909.11740.pdf. Retrieved 16 September 2023.

[20] Sommerville, L. 2015. Software Engineering. 10th ed. Pearson Education, Courier Westford.

[21] Chaudhary, A. 2022. Evaluation Metrics For Information Retrieval. Available at: https://amitness.com/2020/08/information-retrieval-evaluation/. Retrieved 5 April 2023.

[22] Sarwara, S., Qayyuma, Z.U. and Majeedb, S. 2013. Ontology Based Image Retrieval Framework using Qualitative Semantic Image Description. *Procedia Computer Science*, 22(2013), 285–294.

[23] เรวัต แสงสุริยงค์. 2565. ความเสี่ยงของการเกิดความคลาดเคลื่อนในการวิจัยเชิงปริมาณด้านสังคมวิทยา. *วารสารวิชาการมนุษยศาสตร์และสังคมศาสตร์ มหาวิทยาลัยบูรพา*, 30(1), 158-185. [Rewat Sangsuriyong. 2022. Risks of Error in the Quantitative Sociology Research. *Journal of Humanities and Social Sciences Burapha University*, 30(1), 158-185. (in Thai)]

[24] อรนุช ศรีสะอาด. 2561. การตรวจสอบความเที่ยงตรงของเครื่องมือวัดผลโดยผู้เชี่ยวชาญ. *วารสารการวัดผลการศึกษา มหาวิทยาลัยมหาสารคาม*, 1(1), 45-49. [Oranuch Srisa-ard. 2018. Validation of Measurement and Evaluation Tools by Expert. *Journal of Educational Measurement Mahasarakham University*, 1(1), 45-49. (in Thai)]

[25] Gilani, R. 2020. Main Challenges in Image Classification. Available at: http://towards datascience.com/mainchallengesinimageclassification-ba24dc78b558. Retrieved 29 March 2023.

[26] Dix, A., Finlay, J., Abowd, G.D. and Beale, R. 2004. Human–Computer Interaction. 3rd ed. Pearson Education, Scotprint Book Printers Ltd.

[27] Liu, C. and Song, G. 2011. A Method of Measuring the Semantic Gap in Image Retrieval: Using the Information Theory. Proceedings 2011 International Conference on Image Analysis and Signal Processing (IASP 2011), Hubei, 1-5.