บทที่ 1
บทนำ

1.1 ความสำคัญและที่มาของปัญหาการวิจัย

เทคโนโลยีการถ่ายภาพที่พัฒนาจากกล้องฟิล์มเป็นอุปกรณ์ถ่ายภาพดิจิทัล ประกอบกับมีราคาที่ถูกลง และสะดวกต่อการใช้งานในลักษณะอุปกรณ์พกพา ส่งผลให้เกิดรูปภาพจำนวนมาก ตามข้อมูลของเว็บไซต์ Phototutorial (Broz, 2024, www: 1) รายงานว่ามีจำนวนรูปภาพถึง 1.2 ล้านล้านภาพในปี 2021 และเพิ่มเป็น 1.72 ล้านล้านภาพในปี 2022 จำนวนนี้เพิ่มขึ้นเป็น 1.81 ล้านล้านภาพในปี 2023 และภายในปี 2025 จะมีรูปภาพมากกว่า 2 ล้านล้านภาพ อีกทั้งมีแนวโน้มที่จะเพิ่มมากขึ้นอีกในอนาคต โดยมีปัจจัยหลักอยู่ที่การพัฒนากล้องถ่ายภาพของสมาร์ทโฟน และพฤติกรรมการใช้เครือข่ายสังคมออนไลน์ในการแบ่งปันเรื่องราวและบันทึกความทรงจำ ส่งผลให้มีรูปภาพจำนวนมหาศาลบนอินเทอร์เน็ต ดังนั้นก่อนการนำไปใช้ประโยชน์จำเป็นต้องมีการจำแนกประเภทข้อมูลภาพ (Image classification) เพื่อแยกรูปภาพออกเป็นกลุ่มหรือคลาส (Class) ที่มีคุณลักษณะเป็นไปในทิศทางเดียวกัน (Canty, 2014)

หนึ่งในแนวคิดที่ถูกนำมาใช้อย่างแพร่หลายในการจำแนกประเภทข้อมูลภาพก็คือการเรียนรู้ของเครื่อง (Machine learning) (Aggarwal, 2015) เพื่อสร้างตัวจำแนกข้อมูล (Classifier) ด้วยโครงข่ายประสาทเทียม (Neural network classifier) (Wozniak, 2014) ที่เลียนแบบการทำงานของโครงข่ายประสาทของมนุษย์ (Vasilev, Slater, Spacagna, Roelants and Zocca, 2019) อย่างไรก็ดี โครงข่ายประสาทเทียมทั่วไปนั้นอาจทำงานได้ดีกับรูปภาพที่ไม่ซับซ้อนมากนัก เช่น รูปภาพที่มีวัตถุอยู่ตรงกลาง รูปภาพวัตถุในคลาสเดียวกัน หรือรูปภาพวัตถุที่มีคุณลักษณะเฉพาะตัวแตกต่างกันอย่างชัดเจน เนื่องจากมนุษย์เป็นผู้กำหนดคุณลักษณะในการเรียนรู้ ตรงกันข้ามหากรูปภาพมีความซับซ้อนจะยากต่อการกำหนดคุณลักษณะ และส่งผลให้ตัวจำแนกเรียนรู้ได้ไม่ดีนัก เป็นที่มาของการนำการเรียนรู้เชิงลึก (Deep learning) (Zhang, 2019) เข้ามาประยุกต์ใช้กับการจำแนกข้อมูลรูปภาพ ด้วยการใช้ตัวแบบที่ผ่านการฝึกฝนล่วงหน้า (Pre-trained model) โดยอาศัยเทคนิคการถ่ายโอนการเรียนรู้ (Transfer learning) บนชุดข้อมูลขนาดใหญ่ แล้วนำมาปรับแต่งพารามิเตอร์ (Fine-tuning) เพื่อให้สามารถเรียนรู้ซ้ำ (Retrain) บนชุดข้อมูลที่กำหนดเอง (Custom dataset) จึงสามารถใช้เพียงไม่กี่ร้อยภาพต่อคลาสในการฝึกตัวฝน ส่งผลให้ประหยัดเวลาในการฝึกฝนและประสิทธิภาพสูงกว่าตัวแบบที่เริ่มต้นการฝึกจากศูนย์ (Training from scratch) ทั้งนี้ กระบวนการเรียนรู้ของตัวแบบยังคงอาศัยกลไกการแพร่กระจายย้อนกลับ (Backpropagation) เพื่อนำค่าความสูญเสียกลับมาปรับปรุงค่าน้ำหนักของโครงข่ายประสาทเทียมในแต่ละชั้นอย่างเป็น

ลำดับจนกว่าจะได้ผลลัพธ์ที่เหมาะสมที่สุด โดยการวิจัยนี้ประยุกต์ใช้ตัวแบบ CLIP (Contrastive Language–Image Pre-training) ที่ถูกฝึกฝนล่วงหน้าจากคู่ข้อความและรูปภาพ (Radford et al., 2021) มาฝึกฝนตัวแบบให้เข้าใจความสัมพันธ์ระหว่างคำบรรยายภาพและรูปภาพให้ถูกต้องมากที่สุด ก่อนจะสร้างป้ายกำกับเชิงความหมายจากผลการพยากรณ์สำหรับการค้นคืนรูปภาพเชิงความหมาย เพื่อให้คอมพิวเตอร์สามารถเข้าใจรูปภาพได้ใกล้เคียงมนุษย์มากที่สุด (Tyagi, 2017)

แม้การค้นคืนรูปภาพเชิงเนื้อหาจะทำงานได้อย่างมีประสิทธิภาพด้วยการระบุแนวคิดระดับสูงที่เป็นรูปธรรม (Concrete concept) อย่างวัตถุ (Object) กิจกรรม (Activities) ฉาก (Scene) หรือเหตุการณ์ (Event) แต่ในความเป็นจริงพฤติกรรมผู้ใช้โดยมากจะค้นคืนรูปภาพโดยใช้ความสัมพันธ์ระหว่างแนวคิดระดับสูงที่เป็นรูปธรรม และแนวคิดที่เป็นนามธรรม (Abstract concept) อย่างอารมณ์และความรู้สึก (Mood and feelings) รวมเข้าด้วยกันเป็นข้อความค้นหาภาษาธรรมชาติ ในทางกลับกันคุณลักษณะที่ถูกแยกอัตโนมัติด้วยคอมพิวเตอร์มักเป็นแนวคิดระดับสูงที่เป็นรูปธรรมเท่านั้น ดังนั้นการค้นคืนรูปภาพจากเนื้อหาเชิงประจักษ์เพียงอย่างเดียวอาจส่งผลให้เกิดปัญหาช่องว่างความหมาย (Semantic-gap) (Liu and Song, 2011) ระหว่างความต้องการของผู้ใช้ที่อยู่ในลักษณะข้อความค้นหา ความหมายของรูปภาพ และคอมพิวเตอร์ ซึ่งโดยทั่วไปไม่มีการเชื่อมโยงโดยตรงระหว่างแนวคิดระดับสูงที่เป็นรูปธรรมและที่เป็นนามธรรมทำให้ผลลัพธ์จากการค้นคืนมีประสิทธิภาพไม่สูงเท่าที่ควร เป็นที่มาของงานวิจัยชิ้นนี้ที่มุ่งเชื่อมโยงระหว่างแนวคิดระดับสูงที่เป็นรูปธรรมและนามธรรมผ่านแบบจำลองการค้นคืนรูปภาพเชิงความหมาย (Semantic-based Image Retrieval: SBIR) โดยเปรียบเทียบระหว่างผลการพยากรณ์ความน่าจะเป็นของป้ายกำกับภาษาธรรมชาติกับผลการประเมินความหมายรูปภาพจากผู้เชี่ยวชาญ ก่อนจะนำไปปรับค่าน้ำหนักตามแนวคิดการแผ่กระจายย้อนหลัง เพื่อให้แบบจำลองเรียนรู้คุณลักษณะเชิงความหมายของรูปภาพให้ใกล้เคียงกับมนุษย์มากยิ่งขึ้น

ปัจจุบันงานวิจัยที่ประยุกต์ใช้เทคนิคต่าง ๆ เพื่อเพิ่มประสิทธิภาพการค้นคืนรูปภาพเชิงความหมายมีเป็นจำนวนมาก แต่ละงานวิจัยก็มีเทคนิคและวิธีการแตกต่างกันไป แต่เมื่อพิจารณาลงไปในรายละเอียดพบว่างานวิจัยส่วนใหญ่มุ่งใช้รูปภาพค้นหา (Image query) ในการค้นคืนรูปภาพ เนื่องจากระบบจะวิเคราะห์คุณลักษณะรูปภาพทั้งหมดเพื่อค้นคืนรูปภาพตามความเหมือนหรือคล้ายคลึงกับรูปภาพค้นหา จึงสามารถลดปัญหาช่องว่างความหมายลงได้ แต่ในความเป็นจริงพฤติกรรมผู้ใช้โดยมากจะค้นคืนรูปภาพโดยใช้ความสัมพันธ์ระหว่างคุณลักษณะระดับต่ำ ผสมผสานแนวคิดระดับสูงที่เป็นรูปธรรม และนามธรรมรวมเข้าด้วยกันเป็นข้อความค้นหาภาษาธรรมชาติ แต่ยังมีผู้ใช้จำนวนไม่น้อยที่ขาดความชำนาญในการใช้คำค้น ส่วนใหญ่มักใช้คำสำคัญ (Keyword) เพียง 1 หรือ 2 คำเป็นหลักในการค้นหาแต่ละครั้ง ประกอบกับบางอัลกอริทึมยังไม่เอื้อต่อการค้นหาด้วยข้อความในรูปแบบภาษาธรรมชาติ (Natural language) จึงอาจเกิดผลลัพธ์จำนวนมากทั้งที่ตรง

และไม่ตรงกับความต้องการของผู้ใช้ ส่งผลให้ผลลัพธ์จากการค้นคืนมีประสิทธิภาพไม่สูงเท่าที่ควร เนื่องจากไม่มีการแปลงเป็นแนวคิดเชิงความหมาย (High-level semantic) ในรูปแบบที่มนุษย์สามารถเข้าใจได้ อีกทั้งการค้นคืนโดยใช้รูปภาพค้นหายังไม่สะดวก และเป็นการผลักภาระให้กับผู้ใช้ในการหารูปภาพต้นฉบับอีกด้วย

งานวิจัยนี้มุ่งพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมาย เพื่อเชื่อมโยงระหว่างแนวคิดระดับสูงที่เป็นรูปธรรมและที่เป็นนามธรรมให้เป็นแนวคิดเชิงความหมายในรูปแบบที่มนุษย์สามารถเข้าใจได้โดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า อันจะสามารถลดช่องว่างความหมายระหว่างข้อความค้นหาภาษาธรรมชาติ ความหมายของรูปภาพ และการแปลความหมายของรูปภาพด้วยคอมพิวเตอร์ ตลอดจนช่วยสนับสนุนผู้ใช้ด้วยข้อความค้นหาภาษาธรรมชาติที่ยึดโยงกับความหมายของรูปภาพแทนที่จะยึดตามหลักไวยากรณ์ของภาษา

1.2 วัตถุประสงค์ของการวิจัย

1.2.1 เพื่อพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายโดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า

1.2.2 เพื่อประเมินประสิทธิภาพแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมาย

1.3 สมมติฐานการวิจัย

แบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมาย มีผลการประเมินประสิทธิภาพด้วยค่าความแม่นยำที่ k อันดับ (Precision@k) ค่าความครบถ้วนที่ k อันดับ (Recall@k) และค่าประสิทธิภาพโดยรวมที่ k อันดับ (F-measure@k) มากกว่าร้อยละ 80

1.4 ขอบเขตของการวิจัย

1.4.1 การพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายด้วยการเขียนโปรแกรมภาษาไพทอน โดยประยุกต์ใช้ตัวแบบ CLIP ที่ถูกฝึกฝนล่วงหน้ารุ่น ViT-B/32 เป็นโมเดลฐานบน Google Colaboratory

1.4.2 การประเมินประสิทธิภาพการค้นคืนรูปภาพแบบไบนารีด้วยมาตรวัดแบบไม่สนใจลำดับของผลลัพธ์ ประกอบด้วย ค่าความแม่นยำที่ k อันดับ ค่าความครบถ้วนที่ k อันดับ และค่าประสิทธิภาพโดยรวมที่ k อันดับ และมาตรวัดความเกี่ยวข้องแบบให้คะแนนด้วยค่า NDCG ที่ k อันดับ

1.4.3 การประเมินประสิทธิภาพการค้นคืนรูปภาพโดยผู้เชี่ยวชาญแบ่งออกเป็น 3 กลุ่ม ได้แก่ 1) กลุ่มผู้เชี่ยวชาญด้านคอมพิวเตอร์และเทคโนโลยี 2) กลุ่มผู้เชี่ยวชาญด้านการค้นคืนสารสนเทศ และ 3) กลุ่มผู้ใช้ทั่วไป กลุ่มละ 10 ท่าน รวมทั้งสิ้น 30 ท่าน

1.4.4 กำหนดคุณสมบัติผู้เชี่ยวชาญ ดังนี้ 1) จบการศึกษาระดับปริญญาโทขึ้นไป หรือมีความรู้ความเชี่ยวชาญเฉพาะด้านอันเป็นที่ประจักษ์ 2) มีประสบการณ์การใช้งานคอมพิวเตอร์และอินเทอร์เน็ตอย่างน้อย 10 ปีขึ้นไป 3) มีประสบการณ์การใช้งานคอมพิวเตอร์และสมาร์ทโฟนอย่างน้อย 10 ปีขึ้นไป 4) มีความสามารถในการให้ข้อความค้นหาในรูปแบบภาษาอังกฤษ และ 5) สามารถประเมินความถูกต้องของการค้นคืนรูปภาพจากข้อความค้นหาภาษาอังกฤษ

1.5 ข้อตกลงเบื้องต้น

1.5.1 กำหนดการสร้างข้อความบรรยายรูปภาพจากโมเดล LLaVA โดยใช้คำสั่ง (Prompt) เพื่อสร้างคำบรรยายรูปภาพในโดเมนอารมณ์พื้นฐานของมนุษย์ ประกอบด้วย ความรื่นเริง ความไว้วางใจ ความกลัว ความประหลาดใจ ความเศร้า ความรังเกียจ ความโกรธ และความคาดหวัง โดยในแต่ละโดเมนจะประกอบด้วยอารมณ์ย่อยโดเมนละ 5 อารมณ์

1.5.2 กำหนดคำบรรยายรูปภาพในรูปแบบภาษาอังกฤษจำนวน 5 รายการต่อ 1 รูปภาพ

1.5.3 กำหนดข้อความค้นหาในรูปแบบภาษาอังกฤษเท่านั้น โดยไม่มีข้อจำกัดความยาวของคำหรือประโยค ตลอดจนตัวอักษรพิมพ์เล็กพิมพ์ใหญ่ หรืออักขระพิเศษจะไม่มีผลในการค้นหา

1.5.4 กำหนดข้อความค้นหาภาษาอังกฤษภายใต้ 3 เงื่อนไข ได้แก่ 1) ข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ ในลักษณะป้ายกำกับของรูปภาพ 2) ข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพ และ 3) ข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ

15.5 กำหนดขอบเขตการค้นคืนรูปภาพบนชุดข้อมูล Flickr30k จำนวน 31,783 รูปภาพ และชุดข้อมูลที่ผู้วิจัยรวบรวมเองจำนวน 92,168 รูปภาพ รวมทั้งสิ้น 123,951 รูปภาพ

1.5.6 กำหนดขนาดของรูปภาพในการแสดงผลความกว้างและความยาวเท่ากับ 250 * 250 พิกเซล โดยเรียงลำดับผลลัพธ์ตามค่าความคล้ายคลึงจากมากไปน้อยจำนวน 20 รูปภาพ อย่างไรก็ดีจะไม่ปรากฏค่าคะแนนความคล้ายคลึงแก่ผู้ใช้ในขั้นตอนของการทำงาน

1.5.7 กำหนดการทดสอบแบบจำลองบน Google Colaboratory แบบมีค่าใช้จ่าย โดยใช้ GPU เป็นหน่วยประมวลผล และใช้หน่วยความจำขนาด 25 กิกะไบต์ในโหมด High-RAM

1.5.8 กำหนดการใช้งานรูปภาพอย่างคำนึงถึงสิทธิส่วนบุคคล โดยไม่ปรากฏใบหน้าของบุคคลหรือส่วนหนึ่งส่วนใดอย่างชัดเจนจนสามารถนำไปสู่การระบุอัตลักษณ์บุคคลได้

1.6 ประโยชน์ที่คาดว่าจะได้รับ

1.6.1 ได้แบบจำลองการค้นคืนรูปภาพดิจิทัลที่ลดข้อจำกัดช่องว่างความหมาย ภายใต้ข้อความค้นหาภาษาธรรมชาติที่ยึดโยงกับความหมายของรูปภาพแทนที่จะยึดความถูกต้องตามหลักไวยากรณ์ของภาษา

1.6.2 ได้ชุดข้อมูลคุณลักษณะเชิงความหมายของรูปภาพจากการเรียนรู้ของตัวแบบโครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้าจากการเรียนรู้ร่วมกันระหว่างรูปภาพและข้อความบรรยายรูปภาพ

1.6.3 ได้แนวทางการค้นคืนรูปภาพดิจิทัลเชิงความหมาย อันเป็นแนวทางในการพัฒนาการค้นคืนสารสนเทศในรูปแบบที่เกี่ยวข้องต่อไปในอนาคต

1.7 คำอธิบายศัพท์

1.7.1 รูปภาพดิจิทัล คือข้อมูลรูปภาพที่ถูกจัดเก็บในรูปแบบดิจิทัล โดยใช้เลขฐานสองในการกำหนดรายละเอียดแต่ละส่วนของภาพ ซึ่งสามารถเปลี่ยนแปลง แก้ไข ประมวลผล หรือถ่ายโอนได้โดยใช้ระบบคอมพิวเตอร์

1.7.2 โครงข่ายประสาทเทียมเชิงลึก คือพัฒนาการจากการเรียนรู้ของเครื่องในการสร้างแบบจำลองด้วยการประมวลผลข้อมูลหลายชั้นที่มีโครงสร้างที่ซับซ้อน

1.7.3 ตัวแบบที่ถูกฝึกฝนล่วงหน้า หมายถึง ตัวแบบที่ถูกฝึกฝนมาแล้วล่วงหน้า โดยใช้เทคนิคการถ่ายโอนการเรียนรู้ จากนั้นนำมาปรับปรุงพารามิเตอร์ แล้วจึงเรียนรู้ซ้ำบนชุดข้อมูลที่กำหนดเองเพื่อลดภาระการฝึกฝนแบบจำลองด้วยชุดข้อมูลจำนวนมาก

1.7.4 ตัวแบบ CLIP ที่ถูกฝึกฝนล่วงหน้า คือโครงข่ายประสาทเทียมเชิงลึกที่เรียนรู้แบบคอนทราสต์ระหว่างข้อความและรูปภาพ เพื่อเรียนรู้ความหมายของรูปภาพด้วยภาษาธรรมชาติ

1.7.5 การค้นคืนรูปภาพเชิงความหมาย คือการค้นคืนรูปภาพจากการแปลงคุณลักษณะของรูปภาพโดยเชื่อมโยงระหว่างแนวคิดระดับสูงที่เป็นรูปธรรมและนามธรรมเข้าด้วยกัน เพื่อค้นคืนรูปภาพด้วยเทคนิคต่าง ๆ ในการแปลความหมายที่เป็นนามธรรมที่ซ่อนอยู่ในรูปภาพ