บทที่ 5
สรุปและข้อเสนอแนะ

การพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมาย โดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายโดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า และเพื่อประเมินประสิทธิภาพการค้นคืนรูปภาพดิจิทัลเชิงความหมาย เนื้อหาในบทนี้ประกอบด้วยสรุปผลการวิจัย การประยุกต์ผลการวิจัย และข้อเสนอแนะในการวิจัยครั้งต่อไป มีรายละเอียดดังนี้

5.1 สรุปผลการวิจัย

งานวิจัยนี้แบ่งสรุปผลการวิจัยออกเป็น 2 ส่วน ตามวัตถุประสงค์ ได้แก่ 1) เพื่อออกแบบและพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายโดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า และ 2) เพื่อประเมินประสิทธิภาพการค้นคืนรูปภาพดิจิทัลเชิงความหมาย มีรายละเอียดดังนี้

5.1.1 สรุปผลการพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายโดยประยุกต์ใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า

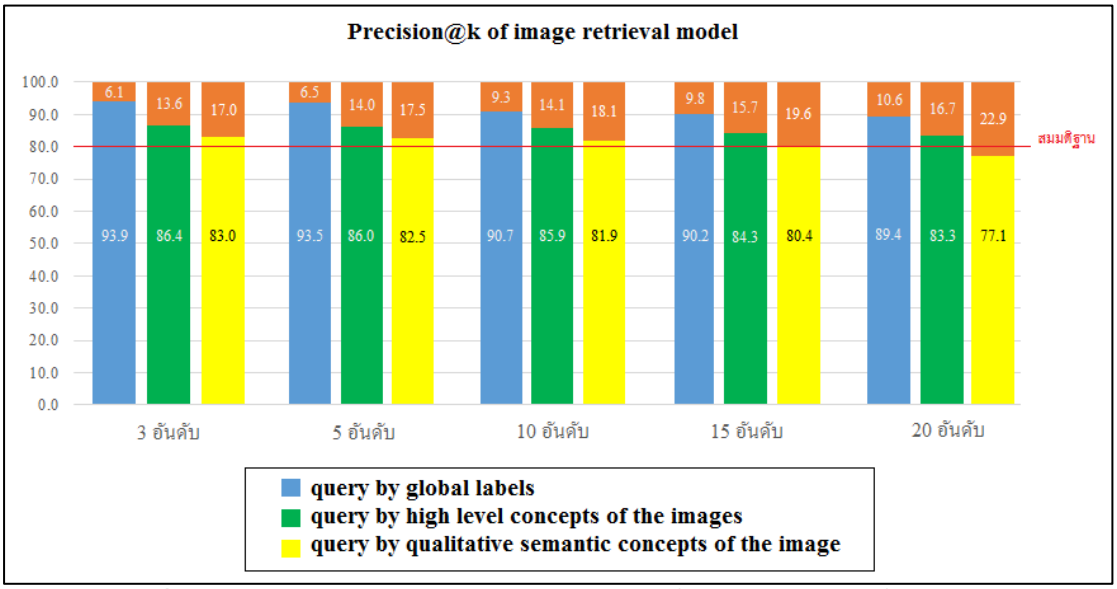
การพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายด้วยการเขียนโปรแกรมภาษาไพทอนบน Google Colaboratory ประกอบด้วย 3 มอดูล ได้แก่ 1) มอดูลการสร้างชุดคำอธิบายรูปภาพ โดยประยุกต์ใช้ตัวแบบ CLIP ที่ถูกฝึกฝนล่วงหน้ารุ่น ViT-B/32 สำหรับฝึกฝนตัวเข้ารหัสรูปภาพ และตัวเข้ารหัสข้อความบนพื้นที่การฝังหลายรูปแบบ จากการคำนวณหาค่าความผิดพลาดเปรียบเทียบระหว่างผลลัพธ์การพยากรณ์ป้ายกำกับจากตัวแบบกับผลลัพธ์ที่ถูกประเมินความหมายเชิงคุณภาพของรูปภาพโดยผู้เชี่ยวชาญจำนวน 10 ท่านกับรูปภาพจำนวน 400 รูปภาพที่สุ่มจากชุดข้อมูล จากนั้นนำค่าการสูญเสียมาปรับปรุงพารามิเตอร์ที่ใช้ในการเรียนรู้ของโมเดลด้วยการปรับแต่งด้วยตนเอง เพื่อแยกความแตกต่างของแต่ละคลาสด้วยค่าความคล้ายคลึงโคไซน์ของเวกเตอร์รูปภาพและข้อความบรรยายรูปภาพตามหลักการเรียนรู้แบบคอนทราสต์ จากการเปรียบเทียบความคล้ายคลึงเชิงความหมายให้ใกล้เคียงกับมนุษย์มากที่สุดก่อนจะเรียนรู้ซ้ำบนชุดข้อมูลที่กำหนดเอง เพื่อสกัดคุณลักษณะเชิงความหมายของรูปภาพบนชุดข้อมูล Flickr30k จำนวน 31,783 รูปภาพ และชุดข้อมูลที่ผู้วิจัยรวบรวมเองจำนวน 92,168 รูปภาพ รวมทั้งสิ้น 123,951 รูปภาพ เพื่อสร้างเป็นเวกเตอร์คุณลักษณะรูปภาพขนาด 2,048 แล้วจัดเก็บไว้ในชุดคำอธิบายรูปภาพ

2) มอดูลการประมวลผลข้อความค้นหาจากผู้ใช้ในรูปแบบภาษาธรรมชาติด้วยโมเดลภาษา DistilBERT ที่ถูกฝึกฝนล่วงหน้ามาปรับแต่งให้เหมาะสมสำหรับเข้ารหัสข้อความค้นหาภาษาธรรมชาติในรูปแบบภาษาอังกฤษจากผู้ใช้ เพื่อสร้างเป็นเวกเตอร์คุณลักษณะข้อความค้นหา ขนาด 768 ใช้ในการค้นคืนรูปภาพต่อไป 3) มอดูลการจับคู่เวกเตอร์คุณลักษณะรูปภาพและเวกเตอร์คุณลักษณะข้อความค้นหาบนพื้นที่การฝังหลายรูปแบบเพื่อแปลงเป็นเวกเตอร์ขนาด 256 และนำมาเปรียบเทียบค่าความคล้ายคลึงของเวกเตอร์ ก่อนจะเรียงลำดับตามความเกี่ยวข้องจากมากไปน้อย และแสดงเป็นผลลัพธ์การค้นคืนรูปภาพทั้งในบริบทของเนื้อหาและบริบทเชิงความหมายแก่ผู้ใช้

5.1.2 สรุปผลการประเมินประสิทธิภาพการค้นคืนรูปภาพดิจิทัลเชิงความหมายด้วยค่าความแม่นยำที่ k อันดับ ค่าความครบถ้วนที่ k อันดับ และค่าประสิทธิภาพโดยรวมที่ k อันดับ

การประเมินประสิทธิภาพการค้นคืนรูปภาพดิจิทัลเชิงความหมายภายใต้ข้อความค้นหา 3 เงื่อนไข ได้แก่ 1) ข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ในลักษณะป้ายกำกับของรูปภาพ 2) ข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพ และ 3) ข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ ค้นหาเงื่อนไขละ 100 รายการ รวมทั้งสิ้น 300 รายการ โดยผู้วิจัยระบุข้อความค้นหาผ่าน Google Colaboratory เมื่อการค้นคืนรูปภาพเสร็จสิ้นจะแสดงผลรูปภาพแต่ละครั้งเท่ากับ 20 รูปภาพ เมื่อสิ้นสุดการค้นคืนผลลัพธ์ใน แต่ละรายการผู้วิจัยประเมินความถูกต้องทีละรูปภาพ กำหนดให้ หากรูปภาพจากการค้นคืนถูกต้องตามข้อความค้นหา เท่ากับ 1 คะแนน ในทางกลับกันหากรูปภาพจากการค้นคืนไม่ถูกต้องหรือไม่ครบถ้วนตามข้อความค้นหา เท่ากับ 0 คะแนน มีรายละเอียดดังนี้

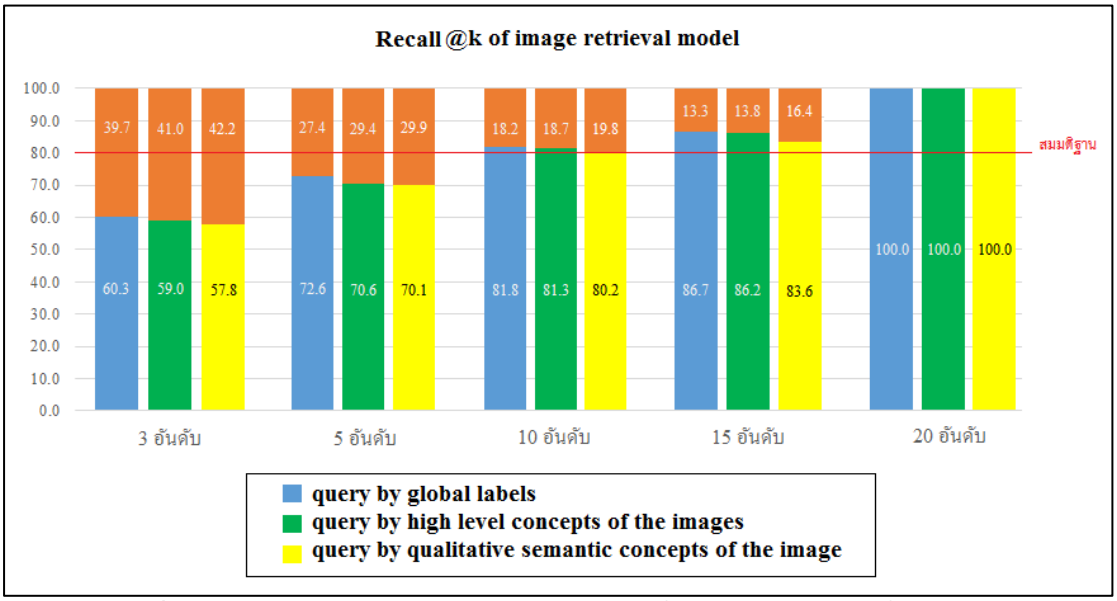
5.1.2.1 สรุปผลการประเมินประสิทธิภาพการค้นคืนรูปภาพดิจิทัลเชิงความหมายเมื่อพิจารณาค่าความแม่นยำที่ k อันดับ อันแสดงถึงประสิทธิภาพของระบบในการค้นคืนรูปภาพโดยดูจากอัตราส่วนของจำนวนรูปภาพที่ถูกต้องจากรูปภาพที่ถูกเลือกมาทั้งหมด แสดงถึงแบบจำลองมีความสามารถในการขจัดรูปภาพที่ไม่เกี่ยวข้อง พบว่า ผลลัพธ์การค้นคืนจำนวน 3, 5 และ 10 ลำดับแรกภายใต้เงื่อนไขข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ในลักษณะป้ายกำกับของรูปภาพคิดเป็นร้อยละ 93.9, 93.5 และ 90.7 ตามลำดับ อยู่ในระดับดีมาก โดยลักษณะข้อความบรรยายรูปภาพส่งผลอย่างมากต่อผลลัพธ์จากการค้นคืน เช่นเดียวกับข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพนั้นมีค่าความแม่นยำคิดเป็นร้อยละ 86.4, 86.0 และ 85.9 ตามลำดับ อยู่ในระดับดีมากแสดงถึงตัวแบบมีประสิทธิภาพสูงการจำแนกวัตถุภายในรูปภาพ แต่ยังมีประสิทธิภาพลดลงเมื่อมีการนับจำนวนวัตถุในภาพ ตลอดจนการจำแนกที่มีรายละเอียดเชิงลึก และข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ คิดเป็นร้อยละ 83.0, 82.5 และ 81.9 ตามลำดับ อยู่ในระดับดี



รูปที่ 5.1 ประสิทธิภาพการค้นคืนรูปภาพดิจิทัล เมื่อค่าความแม่นยำที่ k อันดับ เท่ากับ 3, 5, 10, 15 และ 20 ตามลำดับ

จากรูปที่ 5.1 ค่าความแม่นยำนั้นลดลงเรื่อย ๆ แปรผันตามจำนวนผลลัพธ์การค้นคืนรูปภาพที่เพิ่มมากขึ้น ดังรูปที่ 5.1 แสดงให้เห็นถึงข้อความที่อยู่ในรูปแบบของป้ายกำกับภาษาธรรมชาตินั้นถือว่าเป็นแนวคิดระดับสูง ดังนั้นประสบการณ์ของแต่ละบุคคลจึงส่งผลให้การประเมินนั้นแตกต่างกันตามหลักการรับรู้ของมนุษย์ อย่างไรก็ดี ผลการประเมินประสิทธิภาพการค้นคืนรูปภาพด้วยค่าความแม่นยำสูงกว่าสมมติฐานการวิจัยที่กำหนดไว้ที่ร้อยละ 80

5.1.2.2 สรุปผลการประเมินประสิทธิภาพการค้นคืนรูปภาพด้วยค่าความครบถ้วนที่ k อันดับ อันแสดงถึงประสิทธิภาพของการค้นคืนรูปภาพโดยดูจากอัตราส่วนจำนวนรูปภาพที่ถูกต้องที่เลือกมาต่อจำนวนรูปภาพที่ถูกต้องทั้งหมดที่อยู่ในชุดข้อมูล แสดงถึงแบบจำลองมีความสามารถในการดึงรูปภาพที่เกี่ยวข้อง พบว่า ผลการประเมินประสิทธิภาพการค้นคืนรูปภาพด้วยค่าความครบถ้วนจะค่อย ๆ เพิ่มขึ้นตามจำนวนผลลัพธ์ที่เพิ่มมากขึ้น และจะเข้าใกล้ 1 มากขึ้นเมื่อจำนวนผลลัพธ์เท่ากับ 10 และ 15 ลำดับ ภายใต้เงื่อนไขข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ในลักษณะป้ายกำกับของรูปภาพอยู่ในระดับดี คิดเป็นร้อยละ 81.8 และ 86.7 ตามลำดับ และภายใต้เงื่อนไขข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพนั้นมีค่าความครบถ้วนอยู่ในระดับดีเช่นกัน คิดเป็นร้อยละ 81.3 และ 86.2 ตามลำดับ เป็นไปในทิศทางเดียวกันกับข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพที่อยู่ในระดับดี คิดเป็นร้อยละ 80.2 และ 83.6 ตามลำดับ

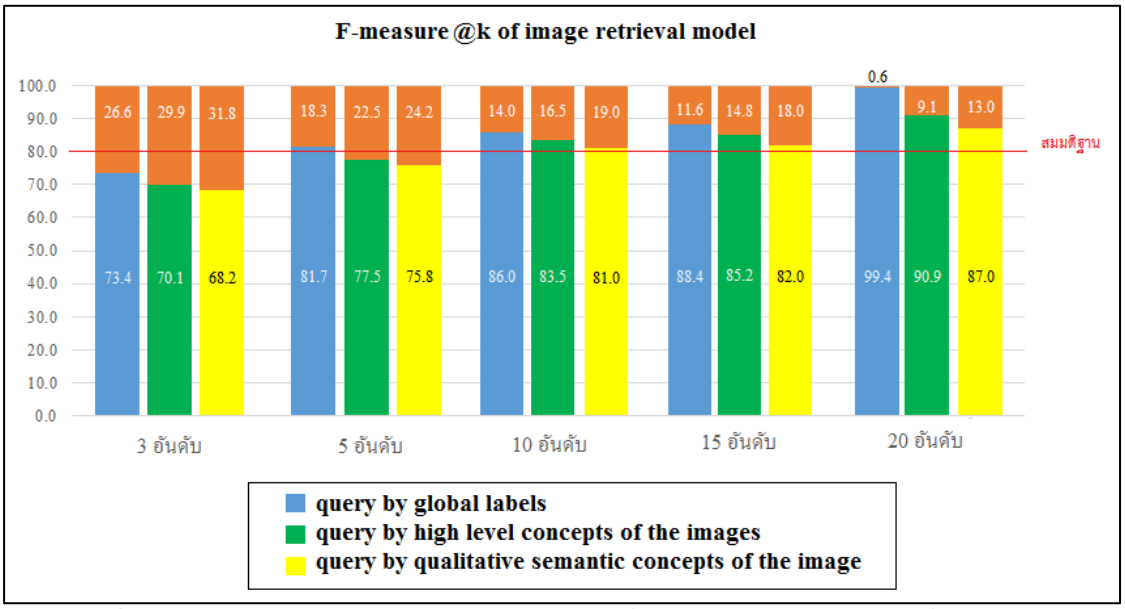


รูปที่ 5.2 ประสิทธิภาพการค้นคืนรูปภาพดิจิทัล เมื่อค่าความครบถ้วนที่ k อันดับ เท่ากับ 3, 5, 10, 15 และ 20 ตามลำดับ

จากรูปที่ 5.2 อย่างไรก็ดี ยังมีปัจจัยอื่นที่อาจส่งผลต่อค่าความถูกต้องและค่าความแม่นยำของแบบจำลองที่พัฒนาขึ้น ได้แก่ ปริมาณรูปภาพในฐานข้อมูล ซึ่งมีความเป็นได้น้อยมากที่จำนวนรูปภาพจะครอบคลุมความต้องการจากผู้ใช้ทั้งหมด อีกทั้งพฤติกรรมของผู้ใช้มักต้องการเพียงผลลัพธ์ที่ตนเองสนใจเท่านั้น ผู้ใช้จำนวนไม่น้อยที่ขาดความชำนาญในการใช้คำค้นหาหรือไม่สามารถกำหนดคำหลักหรือคำที่เฉพาะเจาะจงสำหรับการค้นหาที่เหมาะสมได้ ส่งผลให้เกิดเป็นผลลัพธ์รูปภาพจำนวนมากจากค้นคืนทั้งที่ตรงกับความต้องการและไม่ตรงกับความต้องการ อีกทั้งปริมาณผลลัพธ์ที่มากเกินไปอาจทำให้ผู้ใช้เสียเวลาในการคัดเลือกรูปภาพที่ตรงกับความต้องการอย่างแท้จริง อีกทั้งความต้องการของผู้ใช้ที่ไม่มีขีดจำกัดยังส่งผลต่อความคาดหวังต่อผลลัพธ์จากการค้นคืนอีกด้วย

5.1.2.3 สรุปผลการประเมินประสิทธิภาพการค้นคืนรูปภาพด้วยค่าประสิทธิภาพโดยรวมที่ k อันดับ เมื่อพิจารณาค่าความแม่นยำและค่าความครบถ้วนจะมีค่าอยู่ระหว่าง 0 กับ 1 หากการค้นคืนเอกสารนั้นได้ผลลัพธ์เป็นรูปภาพที่ตรงกับความต้องการทั้งหมดและไม่มีรูปภาพที่ไม่เกี่ยวข้องออกมาด้วยค่าความแม่นยำและค่าความครบถ้วนจะมีค่าเป็น 1 โดยทั่วไปค่าความแม่นยำและค่าความครบถ้วนจะมีความสัมพันธ์เป็นปฏิภาคผกผัน คือหากค่าความแม่นยำสูงขึ้นค่าความครบถ้วนก็จะมักลดลง และในทางตรงกันข้าม หากค่าความครบถ้วนสูงขึ้น ค่าความแม่นยำก็มักจะลดลง แต่ในบางกรณีค่าความแม่นยำและค่าความครบถ้วนอาจจะเพิ่มขึ้นหรือลดลงด้วยกันทั้งสองค่าก็ได้ ดังนั้นค่าประสิทธิภาพโดยรวมที่ k อันดับที่เกิดจากการเปรียบเทียบกันระหว่าง

ค่าความแม่นยำและค่าความครบถ้วน เพื่อเป็นมาตรวัดเดี่ยว ซึ่งเป็นสิ่งที่แสดงถึงประสิทธิภาพโดยรวมของแบบจำลองที่พัฒนาขึ้นที่สูงกว่าสมมติฐานการวิจัยที่กำหนดไว้ที่ร้อยละ 80 เมื่อจำนวนผลลัพธ์เท่ากับ 10 และ 15 ลำดับและจะสูงที่สุดเมื่อจำนวนผลลัพธ์เท่ากับ 20 อันดับ โดยเงื่อนไขข้อความค้นหาด้วย ชื่อรูปภาพและหมวดหมู่ในลักษณะป้ายกำกับของรูปภาพอยู่ในระดับดีมาก คิดเป็นร้อยละ 86.0, 88.4 และ 99.4 ตามลำดับ และภายใต้เงื่อนไขข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพนั้นมีค่าความครบถ้วนอยู่ในระดับดีเช่นกัน คิดเป็นร้อยละ 83.5, 85.2 และ 90.9 ตามลำดับ เป็นไปในทิศทางเดียวกันกับข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพที่อยู่ในระดับดี คิดเป็นร้อยละ 81.0, 82.0 และ 87.0 ตามลำดับ



รูปที่ 5.3 ประสิทธิภาพการค้นคืนรูปภาพดิจิทัล เมื่อค่าประสิทธิภาพโดยรวมที่ k อันดับเท่ากับ 3, 5, 10, 15 และ 20 ตามลำดับ

5.1.3 สรุปผลการประเมินประสิทธิภาพการค้นคืนรูปภาพดิจิทัลเชิงความหมายโดยจำแนกจากกลุ่มผู้เชี่ยวชาญ

งานวิจัยนี้แบ่งผู้เชี่ยวชาญออกเป็น 3 กลุ่ม ได้แก่ 1) กลุ่มผู้เชี่ยวชาญด้านคอมพิวเตอร์และเทคโนโลยี 2) กลุ่มผู้เชี่ยวชาญด้านการค้นคืนสารสนเทศ และ 3) กลุ่มผู้ใช้ทั่วไป เพื่อให้ผลการประเมินครอบคลุมทุกมิติ มีรายละเอียดดังนี้

5.1.3.1 สรุปผลการประเมินประสิทธิภาพการค้นคืนรูปภาพด้วยค่าความแม่นยำที่ k อันดับ ค่าความครบถ้วนที่ k อันดับ และค่าประสิทธิภาพโดยรวมที่ k อันดับภายใต้เงื่อนไขข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ในลักษณะป้ายกำกับของรูปภาพ พบว่า ค่าความแม่นยำที่ 3, 5

และ 10 อันดับจากกลุ่มผู้เชี่ยวชาญด้านคอมพิวเตอร์และเทคโนโลยี อยู่ในระดับดีมาก คิดเป็นร้อยละ 94.3, 93.0 และ 90.3 ตามลำดับ เช่นเดียวกับกลุ่มผู้เชี่ยวชาญด้านการค้นคืนสารสนเทศ คิดเป็นร้อยละ 92.6, 97.4 และ 94.3 ตามลำดับ และเป็นไปในทิศทางเดียวกันกับกลุ่มผู้ใช้ทั่วไปอยู่ในระดับดีมาก คิดเป็นร้อยละ 85.3, 90.0 และ 87.5 ตามลำดับ ซึ่งสูงกว่าสมมติฐานการวิจัยที่กำหนดไว้ที่ร้อยละ 80 แสดงถึงตัวแบบมีประสิทธิภาพการค้นคืนรูปภาพสูงเทียบเท่ากับการค้นคืนรูปภาพด้วยข้อความ จากการเปรียบเทียบคำสำคัญ (Keyword) กับคุณลักษณะของรูปภาพที่ถูกเก็บไว้ในลักษณะของเมทาดาตาของรูปภาพและใช้ดัชนีแบบหลายมิติในการค้นคืนรูปภาพ อย่างไรก็ดี การกำหนดข้อความค้นหาที่ส่งผลอย่างมากต่อความถูกต้องของผลการค้นคืนรูปภาพ นอกจากนี้ แบบจำลองยังมีข้อจำกัดในการค้นคืนรูปภาพที่มีความกำกวมเช่นเดียวกับมนุษย์ เช่น ข้อความค้นหาระหว่าง “photo of sunset” กับ “photo of sunrise” ซึ่งยากที่จะแยกความแตกต่างของรูปภาพได้ หากปราศจากองค์ประกอบอื่น ๆ มาเกี่ยวข้อง

5.1.3.2 สรุปผลการประเมินประสิทธิภาพการค้นคืนรูปภาพด้วยค่าความแม่นยำที่ k อันดับ ค่าความครบถ้วนที่ k อันดับ และค่าประสิทธิภาพโดยรวมที่ k อันดับภายใต้เงื่อนไขข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพ พบว่า ค่าความแม่นยำที่ 3, 5 และ 10 อันดับจากกลุ่มผู้เชี่ยวชาญด้านคอมพิวเตอร์และเทคโนโลยี อยู่ในระดับดีมาก คิดเป็นร้อยละ 82.7, 82.4 และ 82.4 ตามลำดับ เช่นเดียวกับกลุ่มผู้เชี่ยวชาญด้านการค้นคืนสารสนเทศ อยู่ในระดับดีมาก คิดเป็นร้อยละ 88.3, 87.2 และ 87.4 ตามลำดับ เป็นไปในทิศทางเดียวกับกลุ่มผู้ใช้ทั่วไปอยู่ในระดับดีมาก คิดเป็นร้อยละ 88.3, 88.4 และ 87.8 ตามลำดับ ซึ่งสูงกว่าสมมติฐานการวิจัยที่กำหนดไว้ที่ร้อยละ 80 อย่างไรก็ดี ความแปรปรวนของรูปภาพเป็นอุปสรรคสำคัญที่ทำให้ตัวแบบจำแนกข้อมูลภาพได้ไม่ดีเท่าที่ควร โดยเฉพาะอย่างยิ่ง ความคล้ายคลึงกันระหว่างคลาส (Inter-class similarity) และความต่างภายในคลาสเดียวกัน (Intra-class variance) ซึ่งตัวแบบที่มีอยู่ในปัจจุบัน เช่น ResNet50 มีแนวโน้มที่จะเลือกรู้จำวัตถุภายในภาพจากคุณลักษณะสีมากกว่าคุณลักษณะอื่น ๆ

5.1.3.3 สรุปผลการประเมินประสิทธิภาพการค้นคืนรูปภาพด้วยค่าความแม่นยำที่ k อันดับ ค่าความครบถ้วนที่ k อันดับ และค่าประสิทธิภาพโดยรวมที่ k อันดับ ภายใต้เงื่อนไขข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ พบว่า ค่าความแม่นยำที่ 3, 5 และ 10 อันดับจากกลุ่มผู้เชี่ยวชาญด้านคอมพิวเตอร์และเทคโนโลยี อยู่ในระดับดีมาก คิดเป็นร้อยละ 83.3, 84.8 และ 81.8 ตามลำดับ เช่นเดียวกับกลุ่มผู้เชี่ยวชาญด้านการค้นคืนสารสนเทศ อยู่ในระดับดีมาก คิดเป็นร้อยละ 88.3, 86.4 และ 85.0 ตามลำดับ เป็นไปในทิศทางเดียวกับกลุ่มผู้ใช้ทั่วไปอยู่ในระดับดี คิดเป็นร้อยละ 77.3, 77.2 และ 78.8 ตามลำดับ ซึ่งสูงกว่าสมมติฐานการวิจัยที่กำหนดไว้ที่ร้อยละ 80 อย่างไรก็ดี ความหมายของรูปภาพที่อยู่ในลักษณะนามธรรมนั้นยากที่จะประเมินผลว่าถูกต้องหรือไม่ถูกต้องเท่านั้น ดังนั้นการแปลความหมายจึงเป็นไปตามหลักการรับรู้ของมนุษย์ (Human

perception) ด้วยประสบการณ์หรือความรู้เดิมจากการเรียนรู้ของแต่ละบุคคลจึงแตกต่างกัน จะเห็นได้ว่า ปัญหาช่องว่างความหมายของรูปภาพนั้นยากที่จะทำให้หมดไป ทำได้เพียงลดช่องว่างความหมายระหว่างคอมพิวเตอร์ ผู้ใช้ และรูปภาพลงด้วยเทคนิคและวิธีการต่าง ๆ เท่านั้น จึงควรให้ความสำคัญกับการลดช่องว่างความหมายของข้อความค้นหาจากผู้ใช้ควบคู่ไปกับการวิเคราะห์ความหมายของรูปภาพ เพื่อเพิ่มประสิทธิภาพในการค้นคืน

5.1.4 สรุปผลการประเมินประสิทธิภาพการค้นคืนรูปภาพดิจิทัลเชิงความหมายด้วยค่า NDCG เปรียบเทียบระหว่างการค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิมกับการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนัก

โดยผู้วิจัยคัดเลือกข้อความค้นหาภาษาอังกฤษที่มีค่าความแม่นยำที่ 5 ลำดับแรกสูงที่สุด ภายใต้ 3 เงื่อนไข ได้แก่ 1) ข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ ในลักษณะป้ายกำกับของรูปภาพ 2) ข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพ และ 3) ข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ เงื่อนไขละ 100 รายการ รวมทั้งสิ้น 300 รายการ บนชุดข้อมูล Flickr30k เมื่อพิจารณาค่า NDCG ที่ลำดับ 1, 3 และ 5 พบว่า ผลลัพธ์จากการค้นคืนรูปภาพด้วยข้อความบรรยายรูปภาพในลักษณะป้ายกำกับของรูปภาพ โดยการค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิมเปรียบเทียบกับการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนักนั้นมีค่า NDCG ที่ลำดับ 1, 3 และ 5 ไม่แตกต่างกันมากนัก เช่นเดียวกับผลลัพธ์จากการค้นคืนรูปภาพด้วยข้อความบรรยายรูปภาพเกี่ยวกับแนวคิดระดับสูงของรูปภาพ โดยการค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิม เปรียบเทียบกับการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนักนั้นมีค่า NDCG ที่ลำดับ 1, 3 และ 5 เพิ่มขึ้นจากเดิมเล็กน้อย ตรงกันข้ามกับผลลัพธ์จากการค้นคืนรูปภาพด้วยข้อความบรรยายรูปภาพที่อธิบายความหมายเชิงคุณภาพของรูปภาพ โดยการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนักนั้นมี ค่า NDCG ที่ลำดับ 1, 3 และ 5 เพิ่มขึ้นจากการค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิมถึงร้อยละ 25.77, 18.23 และ 31.33 ตามลำดับ

อย่างไรก็ดี งานวิจัยนี้ไม่ได้เพียงเสนอวิธีการปรับแต่งโมเดลเชิงเทคนิคเท่านั้น แต่ยังเปิดมุมมองใหม่ในการพัฒนาโมเดลที่เข้าใจภาษาธรรมชาติและรูปภาพเชิงความหมายในแบบที่ใกล้เคียงกับการรับรู้ของมนุษย์ ทั้งในด้านอารมณ์ ความรู้สึก และบริบทอย่างลึกซึ้ง ซึ่งสามารถจำแนกผลจากการวิจัยนี้ว่าก่อให้เกิดประโยชน์ (Contribution) สำคัญใน 4 มิติที่ครอบคลุมทั้งด้านวิชาการ ทฤษฎี ปฏิบัติ และเทคโนโลยี รายละเอียดดังตารางที่ 5.1

ตารางที่ 5.1 ประโยชน์จากการพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายโดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า

Contribution (ประโยชน์จากการวิจัย)	What (สิ่งที่ดำเนินการ)	Why (เหตุผลในการดำเนินการ)	How (แนวทางการดำเนินการ)	Result (ผลที่ได้รับจากการดำเนินการ)
1) ประโยชน์เชิงวิชาการ (Academic contribution)	การปรับแต่งน้ำหนักของตัวแบบ CLIP ที่ถูกฝึกล่วงหน้าด้วยการกำหนดค่าน้ำหนักโครงข่ายด้วยตนเอง (Manually set weights) โดยอิงจากผลการประเมินของผู้เชี่ยวชาญ	เพื่อเพิ่มความสามารถของโมเดลในการจับคู่เวกเตอร์ข้อความกับรูปภาพอย่างมีประสิทธิภาพ โดยเฉพาะในด้านความหมายเชิงบริบท หรือรูปแบบที่เป็นนามธรรม ซึ่งสอดคล้องกับการรับรู้ของมนุษย์ มากกว่าการจับคู่อย่างผิวเผินจากลักษณะทางกายภาพ	งานวิจัยนี้นำผลการประเมินจากผู้เชี่ยวชาญจำนวน 400 ค่า มาสร้างเป็นฐานน้ำหนัก (base_weights) จากนั้นแปลงเป็น list และเติมค่าที่เหลือจนครบ 393,216 ค่าด้วยค่าที่สุ่มจากการแจกแจงแบบปกติ จากฐานน้ำหนัก แล้วแปลงเป็น PyTorch tensor และปรับรูปทรงให้ตรงกับเลเยอร์ visual_projection ก่อนอัปเดตเข้าไปในโมเดล CLIP เพื่อให้ได้ชุดน้ำหนักใหม่ที่ผสานระหว่างข้อมูลจริงและการสุ่มแบบควบคุม	แบบจำลองสามารถเรียนรู้แนวคิดนามธรรม เช่น ความสงบ โดดเดี่ยว หรือความอบอุ่น ที่ไม่ปรากฏชัดในรูปภาพได้ดีขึ้น เพิ่มความแม่นยำของระบบค้นคืนภาพในแง่ความหมายเชิงลึก (Semantic relevance) และสามารถประเมินภาพในลักษณะเดียวกับมนุษย์ที่ตีความความหมายผ่านบริบทและบรรยากาศโดยรวม ซึ่งมิใช่แค่ลักษณะวัตถุ กิจกรรม ฉาก หรือเหตุการณ์เท่านั้น

ตารางที่ 5.1 ประโยชน์จากการพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายโดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า (ต่อ)

Contribution (ประโยชน์จากการวิจัย)	What (สิ่งที่ดำเนินการ)	Why (เหตุผลในการดำเนินการ)	How (แนวทางการดำเนินการ)	Result (ผลที่ได้รับจากการดำเนินการ)
2) ประโยชน์เชิงทฤษฎี (Theoretical contribution)	การนำเสนอแนวทางใหม่ในการผสานข้อมูลเชิงมนุษย์ (Human-labeled semantics) เข้ากับการเรียนรู้ของแบบจำลองเชิงคณิตศาสตร์ โดยมุ่งให้แบบจำลองสามารถแปลความหมายรูปภาพได้ใกล้เคียงกับการรับรู้ของมนุษย์	เพื่อยกระดับการเรียนรู้ของโมเดลจากระดับสถิติ (Statistical learning) ไปสู่การตีความที่มีความหมาย (Semantic-level understanding) ซึ่งเป็นการลดช่องว่างความหมายของการค้นคืนรูปภาพจากเนื้อหาทั่วไปที่ยังมีข้อจำกัดในการประมวลผลภาษาธรรมชาติหรือการวิเคราะห์ความหมายเชิงนามธรรมจากรูปภาพ	งานวิจัยนี้ผสมผสานผลการประเมินจากผู้เชี่ยวชาญมาสร้างเป็นชุดฐานน้ำหนัก และออกแบบวิธีการสุ่มเติมน้ำหนักให้สอดคล้องทางสถิติกับค่าจริง โดยใช้ค่าที่สุ่มตามการแจกแจงแบบปกติและ tensor transformation พร้อมการปรับแต่งเลเยอร์เฉพาะใน CLIP เพื่อไม่กระทบโครงสร้างโดยรวม	สามารถเพิ่มขีดความสามารถของแบบจำลองในการประมวลผลรูปภาพเชิงความหมายที่มีบริบทหลากหลาย และสามารถใช้แนวคิดนี้ต่อยอดไปยังโมเดลอื่นด้านมัลติโมดัล (Multimodal) หรือ การรู้จำอารมณ์จากรูปภาพ (Emotional recognition) ได้

ตารางที่ 5.1 ประโยชน์จากการพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายโดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า (ต่อ)

Contribution (ประโยชน์จากการวิจัย)	What (สิ่งที่ดำเนินการ)	Why (เหตุผลในการดำเนินการ)	How (แนวทางการดำเนินการ)	Result (ผลที่ได้รับจากการดำเนินการ)
3) ประโยชน์เชิงปฏิบัติ/สังคม (Practical/ Social contribution)	การสนับสนุนผู้ใช้ด้วยข้อความค้นหาภาษาธรรมชาติที่ยึดโยงกับความหมายของรูปภาพ แม้จะอยู่ในรูปแบบที่ไม่เป็นทางการ เช่น “ภาพที่รู้สึกเหงา” หรือ “บรรยากาศอบอุ่น”	เนื่องจากผู้ใช้จำนวนไม่น้อยที่ขาดความชำนาญในการใช้คำค้นหาหรือไม่สามารถกำหนดคำหลักหรือคำที่เฉพาะเจาะจงสำหรับการค้นหาที่เหมาะสมได้ ส่งผลให้เกิดเป็นผลลัพธ์รูปภาพจำนวนมากจากค้นคืนทั้งที่ตรงกับความต้องการและไม่ตรงกับความต้องการ จึงต้องมีระบบที่สามารถวิเคราะห์ข้อความค้นหาภาษาธรรมชาติยึดโยงกับความหมายของรูปภาพแทนที่จะยึดตามหลักไวยากรณ์ของภาษาเพื่อง่ายต่อการใช้งาน	งานวิจัยนี้ปรับค่าน้ำหนักโครงข่ายด้วยตนเอง เพื่อให้แบบจำลองสามารถเรียนรู้คุณลักษณะเชิงความหมายของรูปภาพ และสามารถค้นคืนผลลัพธ์ได้ใกล้เคียงกับการรับรู้ของมนุษย์	แบบจำลองสามารถค้นคืนรูปภาพทั้งเชิงเนื้อหาและเชิงความหมายที่ตรงกับความต้องการของผู้ใช้ได้ แม้ไม่ใช้คำศัพท์ทางเทคนิค ทำให้ผู้ใช้ทั่วไปสามารถใช้งานได้โดยไม่จำเป็นต้องมีความรู้เฉพาะทาง ส่งเสริมการเข้าถึงเทคโนโลยีอย่างเสมอภาค และสามารถนำไปใช้ในสื่อดิจิทัล การศึกษา การดูแลสุขภาพจิต หรือแอปพลิเคชันแนะนำเนื้อหาเชิงความหมายได้ในอนาคต

ตารางที่ 5.1 ประโยชน์จากการพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายโดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า (ต่อ)

Contribution (ประโยชน์จากการวิจัย)	What (สิ่งที่ดำเนินการ)	Why (เหตุผลในการดำเนินการ)	How (แนวทางการดำเนินการ)	Result (ผลที่ได้รับจากการดำเนินการ)
4) ประโยชน์เชิงเทคโนโลยี (Technological contribution)	การผสมผสานข้อมูลแบบสุ่มและข้อมูลจริงในการปรับค่าน้ำหนักโครงข่ายของโมเดล ที่ปรับเปลี่ยนแบบเจาะจงเฉพาะบางชั้น (Layer-specific) ของโครงข่าย แทนที่จะทำการปรับเปลี่ยนทั้งโมเดล	เพื่อให้เกิดการควบคุมความเปลี่ยนแปลงของพารามิเตอร์ในลักษณะที่ไม่ส่งผลเสียต่อคุณลักษณะเชิงโครงสร้างของโมเดล ตลอดจนควบคุมความแม่นยำของการปรับค่าพารามิเตอร์ต่อการเรียนรู้ความหมายเชิงนามธรรมของโมเดล ซึ่งยากที่จะทำได้ด้วยกระบวนการเรียนรู้อัตโนมัติทั่วไป	งานวิจัยนี้ใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของฐานน้ำหนัก (base_weights) มาเป็นแนวทางกำหนดขอบเขตของค่าที่สุ่มลงในพารามิเตอร์ของเลเยอร์ visual_projection แบบควบคุม จากนั้นปรับรูปทรงและอัปเดตน้ำหนักโครงข่ายโดยตรงในโมเดล CLIP	ได้ชุดน้ำหนักใหม่ที่กลมกลืนกับโมเดลเดิม และสามารถเก็บรักษาโครงสร้างความรู้จากการฝึกฝนเดิมไว้ในขณะเดียวกันก็เพิ่มขีดความสามารถในการประมวลผลข้อมูลเชิงความหมายที่ใกล้เคียงกับการรับรู้ของมนุษย์

จากตารางที่ 5.1 งานวิจัยนี้นำเสนอกรอบแนวคิดใหม่เชิงบูรณาการที่รวมความรู้ทางเทคนิคเชิงลึกเข้ากับองค์ความรู้จากมนุษย์ โดยนำเสนอแนวทางใหม่ในการผสานข้อมูลเชิงมนุษย์เข้าสู่กระบวนการเรียนรู้ของแบบจำลองเชิงคณิตศาสตร์ที่ถูกฝึกไว้ล่วงหน้า เพื่อยกระดับความสามารถของการค้นคืนรูปภาพให้เข้าใจความหมายเชิงนามธรรมในลักษณะเดียวกับการรับรู้ของมนุษย์ ด้วยการปรับแต่งน้ำหนักของโครงข่ายเฉพาะบางชั้น โดยอิงข้อมูลจากผลการประเมินของผู้เชี่ยวชาญ ซึ่งถูกนำมาใช้เป็นฐานน้ำหนักสำหรับสุ่มค่าน้ำหนักในลักษณะที่ควบคุมทางสถิติ และไม่กระทบโครงสร้างโดยรวมของโมเดลเดิม ดังนั้นแบบจำลองจึงสามารถวิเคราะห์ข้อความค้นหาในรูปแบบภาษาธรรมชาติที่ไม่เป็นทางการ เช่น “อบอุ่น”, “เหงา” หรือ “สันโดษ” ที่สะท้อนคุณลักษณะเชิงนามธรรมได้อย่างมีประสิทธิภาพ ส่งผลให้ผู้ใช้ทั่วไปสามารถค้นคืนภาพผ่านภาษาธรรมชาติได้อย่างสะดวก โดยไม่ต้องพึ่งพาคำค้นหาเชิงเทคนิคหรือเฉพาะทาง ซึ่งถือเป็นการลดข้อจำกัดในการเข้าถึงเทคโนโลยีด้านปัญญาประดิษฐ์ และสามารถต่อยอดไปยังการประยุกต์ด้านมัลติโมดัลการวิเคราะห์และรู้จำอารมณ์จากภาพ และการแนะนำเนื้อหาที่อิงความหมายที่ไม่เพียงแม่นยำเชิงสถิติ แต่ยังสามารถเข้าใจความหมายเชิงบริบทได้อย่างลึกซึ้ง และสอดคล้องกับการใช้งานในสถานการณ์จริงได้อย่างครอบคลุมและเป็นรูปธรรม

นอกจากนี้แนวทางดังกล่าว ยังเป็นการขยายองค์ความรู้ในระดับวิชาการเกี่ยวกับการเรียนรู้ของแบบจำลองเชิงลึก ตลอดจนเสนอกรอบแนวคิดเชิงทฤษฎีที่สะท้อนความเป็นไปได้ในการเชื่อมโยงระหว่างการเรียนรู้เชิงลึกกับข้อมูลที่สะท้อนคุณลักษณะเชิงความหมายของมนุษย์ (Human-centered semantics) อย่างมีนัยสำคัญ ซึ่งเป็นประเด็นสำคัญในการพัฒนาแบบจำลองที่สามารถตีความภาพในระดับความหมาย มากกว่าการรับรู้เพียงลักษณะทางกายภาพ

5.2 การประยุกต์ผลการวิจัย

5.2.1 การประยุกต์ใช้ผลการวิจัยทางตรง คือการนำตัวแบบโครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้าจากการเรียนรู้ร่วมกันระหว่างรูปภาพและข้อความบรรยายรูปภาพที่ผ่านการปรับค่าน้ำหนักให้ใกล้เคียงกับมนุษย์ไปใช้เป็นกรอบการพัฒนาระบบเพื่อค้นคืนรูปภาพดิจิทัลจากชุดข้อมูลคุณลักษณะเชิงความหมายของรูปภาพที่รวบรวมไว้ ซึ่งไม่จำกัดเพียงการค้นคืนรูปภาพด้วยเนื้อหาจากการระบุแนวคิดระดับสูงที่เป็นรูปธรรมอย่างวัตถุ กิจกรรม ฉาก หรือเหตุการณ์เท่านั้น แต่ยังรวมถึงแนวคิดระดับสูงที่เป็นนามธรรมอย่างอารมณ์ และความรู้สึก อีกทั้งช่วยสนับสนุนผู้ใช้ที่ไม่สามารถกำหนดคำหลักหรือคำที่เฉพาะเจาะจงสำหรับการค้นหาที่เหมาะสมได้ด้วยการใช้ข้อความค้นหาภายใต้รูปแบบของภาษาธรรมชาติที่ยึดโยงกับความหมายของรูปภาพแทนที่จะยึดตามหลักไวยากรณ์ของภาษา เนื่องจากพฤติกรรมของผู้ใช้นั้นต้องการเพียงผลลัพธ์ที่ตนเองสนใจเท่านั้น จึงมีผู้ใช้จำนวนไม่น้อยที่ขาดความชำนาญในการใช้คำค้นหา ไม่สามารถกำหนดคำหลักหรือคำที่

เฉพาะเจาะจงสำหรับการค้นหาที่เหมาะสมได้ ส่งผลให้เกิดเป็นผลลัพธ์รูปภาพจำนวนมากจากค้นคืนทั้งที่ตรงกับความต้องการและไม่ตรงกับความต้องการ จึงอาจเป็นการเสียโอกาสสำหรับผู้ใช้หากมีรูปภาพบางส่วนที่ไม่ถูกเรียกดู ตลอดจนอำนวยความสะดวกให้กับผู้ใช้ในการใช้ข้อความค้นหาในรูปแบบภาษาธรรมชาติแทนที่จะใช้ภาพค้นหาที่ยังไม่สะดวก และเป็นการผลักภาระให้กับผู้ใช้ในการหารูปภาพต้นฉบับอีกด้วย

5.2.2 การประยุกต์ใช้ผลการวิจัยทางอ้อม คือการนำตัวแบบโครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้าที่ผ่านการปรับค่าน้ำหนักให้ใกล้เคียงกับมนุษย์ไปใช้เรียนรู้แบบคอนทราสต์ (Contrastive learning) เพิ่มเติมกับสารสนเทศในรูปแบบวีดีโอ เพื่อเพิ่มประสิทธิภาพการค้นคืนสารสนเทศเชิงความหมายของมัลติมีเดียให้ครอบคลุมทุกมิติ และลดช่องว่างความหมายระหว่างคอมพิวเตอร์ ผู้ใช้ และสารสนเทศ และสามารถต่อยอดไปยังการประยุกต์ใช้ด้านมัลติโมดัลการวิเคราะห์และรู้จำอารมณ์จากภาพ และการแนะนำเนื้อหาที่อิงความหมายที่ไม่เพียงแม่นยำเชิงสถิติ อันเป็นแนวทางในการพัฒนาการค้นคืนสารสนเทศในรูปแบบที่เกี่ยวข้องต่อไปในอนาคต

5.3 ข้อเสนอแนะในการวิจัยครั้งต่อไป

5.3.1 ปัญหาช่องว่างความหมายของรูปภาพนั้นยากที่จะทำให้หมดไป ทำได้เพียงลดช่องว่างความหมายระหว่างคอมพิวเตอร์ ผู้ใช้ และรูปภาพลงด้วยเทคนิคและวิธีการต่าง ๆ เท่านั้น งานวิจัยในครั้งต่อไปจึงควรให้ความสำคัญกับการลดช่องว่างความหมายของข้อความค้นหาจากผู้ใช้ควบคู่ไปกับการวิเคราะห์ความหมายของรูปภาพ เพื่อเพิ่มประสิทธิภาพในการค้นคืน

5.3.2 การพัฒนาแบบจำลองการค้นคืนรูปภาพนั้นควรให้ความสำคัญที่การปรับค่าไฮเปอร์พารามิเตอร์ (Hyperparameter) ก่อนที่ตัวแบบจะทำการเรียนรู้ เนื่องจากค่าไฮเปอร์พารามิเตอร์ที่เหมาะสมจะส่งผลให้ตัวแบบมีความแม่นยำที่สูงขึ้น หรือลดค่าการสูญเสียให้ต่ำลงได้

5.3.3 การเพิ่มประสิทธิภาพจากการค้นคืนควรมุ่งที่ผลลัพธ์ไม่เกิน 5 ลำดับ เป็นหลักเพื่อตอบสนองต่อความต้องการของผู้ใช้ เนื่องจากพฤติกรรมของผู้ใช้ต้องการเพียงผลลัพธ์ที่ตนเองสนใจเท่านั้น ผลลัพธ์จำนวนมากจากค้นคืนอาจทำให้ผู้ใช้เสียเวลาคัดเลือกรูปภาพที่ตรงกับความต้องการอย่างแท้จริง ซึ่งนับเป็นความสูญเปล่าทางสารสนเทศ เนื่องจากเป็นได้น้อยมากที่ผู้ใช้จะเลือกดูรูปภาพจากผลการค้นคืนทั้งหมด

5.3.4 ความหมายเชิงนามธรรมนั้นหลากหลาย และมีลักษณะที่ขึ้นอยู่กับบริบททางวัฒนธรรม อารมณ์ หรือประสบการณ์ส่วนบุคคล จึงควรมีการพัฒนาแบบจำลองที่เปิดรับข้อมูลจากผู้ใช้ เช่น การให้คะแนนหรือข้อเสนอแนะเกี่ยวกับผลลัพธ์ของการค้นคืน เพื่อให้โมเดลสามารถเรียนรู้และปรับปรุงตนเองได้อย่างต่อเนื่อง (Aaptive learning) ซึ่งจะช่วยเพิ่มศักยภาพในการเรียนรู้ความหมายของรูปภาพได้อย่างครอบคลุมและสอดคล้องกับการรับรู้ของมนุษย์ในสถานการณ์จริง