จักรินทร์ สันติรัตนภักดี : การพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายโดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า (THE DEVELOPMENT OF A SEMANTIC-BASED IMAGE RETRIEVAL MODEL USINGก DEEP NEURAL NETWORK BY PRE-TRAINING APPROACH)
อาจารย์ที่ปรึกษา : ผู้ช่วยศาสตราจารย์. ดร.ศุภกฤษฎิ์ นิวัฒนากูล, 226 หน้า.

คำสำคัญ: การค้นคืนรูปภาพเชิงความหมาย/การเรียนรู้ความหมายรูปภาพจากภาษาธรรมชาติ/โครงข่ายประสาทเทียมเชิงลึก

การวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองการค้นคืนรูปภาพดิจิทัลเชิงความหมายโดยใช้โครงข่ายประสาทเทียมเชิงลึกที่ถูกฝึกฝนล่วงหน้า ประกอบด้วย 3 มอดูลหลัก ได้แก่ 1) มอดูลการสร้างชุดคำอธิบายรูปภาพ โดยประยุกต์ใช้ตัวแบบ CLIP ที่ถูกฝึกฝนล่วงหน้าสำหรับฝึกฝนตัวเข้ารหัสรูปภาพ และตัวเข้ารหัสข้อความอธิบายรูปภาพ เพื่อแยกความแตกต่างของแต่ละคลาสด้วยการวัดความคล้ายคลึงโคไซน์ตามหลักการเรียนรู้แบบคอนทราสต์ โดยคำนวณค่าความผิดพลาดจากการเปรียบเทียบระหว่างผลการพยากรณ์ป้ายกำกับกับผลการประเมินความหมายของรูปภาพโดยผู้เชี่ยวชาญ จากนั้นนำค่าการสูญเสียไปปรับปรุงพารามิเตอร์ด้วยตนเองเพื่อใช้ในการเรียนรู้ความคล้ายคลึงเชิงความหมายให้ใกล้เคียงกับการรับรู้มนุษย์มากที่สุด ก่อนจะเรียนรู้ซ้ำบนชุดข้อมูลที่กำหนดเอง และนำมาแปลงเป็นเวกเตอร์คุณลักษณะรูปภาพเพื่อจัดเก็บไว้ในชุดคำอธิบายรูปภาพ 2) มอดูลการประมวลผลข้อความค้นหาจากผู้ใช้ในรูปแบบภาษาธรรมชาติด้วยโมเดลภาษา DistilBERT ที่ถูกฝึกฝนล่วงหน้าสำหรับเข้ารหัสข้อความ เพื่อแปลงเป็นเวกเตอร์คุณลักษณะข้อความค้นหา 3) มอดูลการจับคู่เวกเตอร์คุณลักษณะรูปภาพกับเวกเตอร์คุณลักษณะข้อความค้นหาบนพื้นที่การฝังหลายรูปแบบ เพื่อเปรียบเทียบค่าความคล้ายคลึงของเวกเตอร์ ก่อนจะเรียงลำดับตามความเกี่ยวข้อง และแสดงเป็นผลลัพธ์การค้นคืนรูปภาพทั้งในบริบทของเนื้อหาและบริบทเชิงความหมายแก่ผู้ใช้

การประเมินประสิทธิภาพแบบจำลองการค้นคืนรูปภาพโดยผู้เชี่ยวชาญ 3 กลุ่ม ประกอบด้วยกลุ่มผู้เชี่ยวชาญด้านคอมพิวเตอร์และเทคโนโลยี กลุ่มผู้เชี่ยวชาญด้านการค้นคืนสารสนเทศ และกลุ่มผู้ใช้ทั่วไป กลุ่มละ 10 ท่าน รวมทั้งสิ้น 30 ท่าน ด้วยการวัดประสิทธิภาพการค้นคืนรูปภาพแบบไบนารี มีรายละเอียดดังนี้ ค่าความแม่นยำที่ k อันดับ พบว่า ผลลัพธ์จากการค้นคืนจำนวน 3 ลำดับแรกด้วยชื่อรูปภาพและหมวดหมู่ในลักษณะป้ายกำกับของรูปภาพ และข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพนั้นมีค่าความแม่นยำอยู่ในระดับดีมาก คิดเป็นร้อยละ 93.9 และ 86.4 ตามลำดับ อย่างไรก็ดี ความหมายเชิงคุณภาพนั้นขึ้นอยู่กับหลักการรับรู้ของมนุษย์ ดังนั้นประสบการณ์ของแต่ละบุคคลจึงส่งผลให้การประเมินนั้นแตกต่างกัน ส่งผลให้ค่าความแม่นยำ

จากข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพที่อยู่ในระดับดี คิดเป็นร้อยละ 83.0 เช่นเดียวกับ ค่าความครบถ้วนที่ k อันดับ พบว่า ผลลัพธ์จากการค้นคืนจำนวน 3 ลำดับแรก ภายใต้เงื่อนไขข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ในลักษณะป้ายกำกับของรูปภาพ และข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพนั้นมีค่าความครบถ้วนอยู่ในระดับดี ตรงกันข้ามกับข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพที่อยู่ในระดับปานกลาง อย่างไรก็ดี ค่าความครบถ้วนจะค่อย ๆ เพิ่มตามจำนวนผลลัพธ์ที่เพิ่มมากขึ้น เมื่อจำนวนผลลัพธ์เท่ากับ 10 คิดเป็นร้อยละ 81.8, 81.3 และ 80.2 ตามลำดับ นอกจากนั้น ค่าประสิทธิภาพโดยรวมที่ k อันดับ พบว่า ผลลัพธ์จากการค้นคืนจำนวน 3 ลำดับแรก ภายใต้เงื่อนไขข้อความค้นหาด้วยชื่อรูปภาพและหมวดหมู่ในลักษณะป้ายกำกับของรูปภาพ และข้อความค้นหาสั้น ๆ เกี่ยวกับแนวคิดระดับสูงของรูปภาพนั้นมีค่าความครบถ้วนอยู่ในระดับดี เช่นเดียวกับข้อความค้นหาที่อธิบายความหมายเชิงคุณภาพของรูปภาพ โดยที่ค่าประสิทธิภาพโดยรวมมีค่าสูงขึ้น เมื่อทั้งค่าความแม่นยำในการค้นคืนรูปภาพ และค่าความครบถ้วนจากการค้นคืนรูปภาพมีค่าสูงขึ้นไปในทิศทางเดียวกัน ซึ่งเป็นสิ่งที่แสดงถึงประสิทธิภาพโดยรวมของแบบจำลองที่พัฒนาขึ้น เมื่อจำนวนผลลัพธ์เท่ากับ 10 คิดเป็น 86.0, 83.5 และ 81.0 ตามลำดับ

การประเมินประสิทธิภาพการค้นคืนรูปภาพดิจิทัลเชิงความหมายด้วยค่า NDCG ที่ k อันดับ เปรียบเทียบระหว่างการค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิมกับการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนัก พบว่า ผลลัพธ์จากการค้นคืนรูปภาพด้วยข้อความบรรยายรูปภาพในลักษณะป้ายกำกับของรูปภาพ โดยการค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิมเปรียบเทียบกับการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนัก มีค่า NDCG ที่ลำดับ 1, 3 และ 5 ไม่แตกต่างกันมากนัก เช่นเดียวกับผลลัพธ์จากการค้นคืนรูปภาพด้วยข้อความบรรยายรูปภาพเกี่ยวกับแนวคิดระดับสูงของรูปภาพ โดยการค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิม เปรียบเทียบกับการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนักนั้นมีค่า NDCG ที่ลำดับ 1, 3 และ 5 เพิ่มขึ้นจากเดิมเล็กน้อย ตรงกันข้ามกับ ผลลัพธ์จากการค้นคืนรูปภาพด้วยข้อความบรรยายรูปภาพที่อธิบายความหมายเชิงคุณภาพของรูปภาพ โดยการค้นคืนรูปภาพด้วยตัวแบบ CLIP ที่ผ่านการปรับค่าน้ำหนักนั้นมี ค่า NDCG ที่ลำดับ 1, 3 และ 5 เพิ่มขึ้นจากการค้นคืนรูปภาพด้วยตัวแบบ CLIP ดั้งเดิมถึงร้อยละ 21.50, 19.90 และ 22.80 ตามลำดับ

สำนักวิชาศาสตร์และศิลป์ดิจิทัล ลายมือชื่อนักศึกษา _____

ปีการศึกษา 2567 ลายมือชื่ออาจารย์ที่ปรึกษา _____

CHAKKARIN SANTIRATTANAPHAKDI : THE DEVELOPMENT OF A SEMANTIC-BASED IMAGE RETRIEVAL MODEL USING DEEP NEURAL NETWORK BY PRE-TRAINING APPROACH. THESIS ADVISOR : Asst. Prof. SUPHAKIT NIWATTANAKUL, Ph.D., 226 PP.

KEYWORD: Semantic-based image retrieval/Natural language supervision/ Deep neural network

This research aims to develop a semantic digital image retrieval model using a pre-trained deep neural network, comprising three main modules: 1) Image description set generate module: This module applies a pre-trained CLIP model to train an image encoder and a text encoder for image captions. It distinguishes between classes by measuring cosine similarity based on contrastive learning. The error is calculated by comparing predicted labels with expert semantic evaluations of the images. The loss is then used to manually adjust the weight parameters involved in learning semantic similarity to closely match human perception. Afterwards, the model is fine-tuned on a custom dataset, and image feature vectors are generated and stored in the image caption dataset. 2) Query processing module: This module processes natural language search queries using the pre-trained DistilBERT language model, which encodes queries into vector representations and 3) Vector Matching Module: This module matches image feature vectors with search query vectors in a multi-modal embedding space by comparing their similarity scores. The results are ranked by relevance and presented as image retrieval outputs, reflecting both content and semantic contexts to the user.

The retrieval performance was evaluated by 30 experts divided into three groups: 10 computer and technology specialists, 10 information retrieval experts, and 10 general users. Binary image retrieval performance was measured with the following results: Precision@k: For the top 3 retrieved results, queries based on image names and categories as image labels, as well as short queries concerning high-level image concepts, yielded very high precision rates of 93.9% and 86.4%, respectively. However, since qualitative semantic meaning depends on human

perception, individual experience caused variability in evaluation, resulting in a precision of 83.0% for queries describing qualitative image semantics. Recall@k: For the top 3 results, queries based on image names, category labels, and short high-level concept queries demonstrated good recall levels, whereas queries describing qualitative semantics showed moderate recall. Recall improved as the number of retrieved results increased, reaching 81.8%, 81.3%, and 80.2% at 10 results, respectively. F-measure@k: For the top 3 results, under queries based on image names, category labels, and short high-level concepts, effectiveness was good. Similarly, queries describing qualitative semantics also showed good effectiveness. Overall effectiveness increased as precision and recall improved together, reflecting the robust performance of the developed model. At 10 results, the effectiveness scores were 86.0%, 83.5%, and 81.0%, respectively.

The semantic image retrieval effectiveness was further evaluated using NDCG@k (Normalized Discounted Cumulative Gain), comparing retrieval results between the original CLIP model and the fine-tuned CLIP model. For image caption-based retrieval serving as image labels, NDCG scores at ranks 1, 3, and 5 were similar between the original and fine-tuned models. For retrieval with captions describing high-level image concepts, slight improvements in NDCG were observed at ranks 1, 3, and 5 with the fine-tuned model. In contrast, retrieval with captions describing qualitative semantic content showed significant NDCG increases of 21.50%, 19.90%, and 22.80% at ranks 1, 3, and 5, respectively, when using the fine-tuned CLIP model compared to the original.

Institute of Digital Arts and Science
Academic Year 2024

Student's Signature _____
Advisor's Signature S. Nimit.