การสร้างคำบรรยายภาพด้วยแบบจำลอง CLIP Prefix Caption บนชุดข้อมูล Traffy Fondue

Image Captioning with CLIP Prefix Caption Model On Traffy Fondue Datasets

สหัสวรรษ ศรีพล[1], วรท กาญจนาคม[1], ชนกันต์ เวชทัพ[1], สรรพฤทธิ์ มฤคทัต[2], วสิศ ลิ้มประเสริฐ[1]*

Sahatwat Sriphon[1], Warot Kanchanakom[1], Chanakan Wechtap[1], Sapparit Marikathat[2], Wasit Limprasert[1]*

[1]สาขาวิชาวิทยาศาสตร์และนวัตกรรมข้อมูล วิทยาลัยสหวิทยาการ มหาวิทยาลัยธรรมศาสตร์

[1]Division of Data science and Innovation, College of Interdisciplinary Studies Thammasat University

[2]ทีมวิจัยการประมวลผลและเข้าใจภาพ กลุ่มวิจัยปัญญาประดิษฐ์ ศูนย์เทคโนโลยีคอมพิวเตอร์และอิเล็กทรอนิกส์แห่งชาติ

[2]Image Processing and Understanding, AINRG National Electronics and Computer Technology Center

บทคัดย่อ

Traffy Fondue เป็นระบบรับแจ้งเรื่องร้องเรียนที่กรุงเทพมหานครในการรับความคิดเห็นและข้อเสนอแนะที่ประชาชนมีต่อเมือง อย่างไรก็ตาม พบว่าจำนวนข้อมูลจากผู้ใช้งานจำนวนมากยังมีความไม่ชัดเจนในการแจ้งเรื่อง เช่น คำอธิบายและรูปภาพไม่สอดคล้องกัน ทำให้ยากต่อการทำงานของเจ้าหน้าที่ผู้รับเรื่องในการประสานงานเพื่อแก้ปัญหา ทีมวิจัยจึงเสนอวิธีการจัดกลุ่มข้อมูลเพื่อเพิ่มความสามารถในการจัดกลุ่มข้อมูลให้สะดวกขึ้นโดยใช้เทคนิคการประมวลผลข้อมูล โดยงานวิจัยนี้ เป็นการประยุกต์ใช้แบบจำลองโมเดล CLIP Prefix Caption สำหรับสร้างคำบรรยายรูปภาพและให้ระบบนำคำที่ได้ไปจัดกลุ่มหรือค้นหาปัญหาที่เกี่ยวข้องต่อไป โดยนำแบบจำลอง CLIP, CLIP Prefix Caption และ GPT-2 มาสร้างคำบรรยายภาพโดยใช้ภาพจาก Traffy Fondue ซึ่งผลการทดลองสรุปได้ว่า ค่า BLEU เท่ากับ 0.93% และ ROUGE-1 เท่ากับ 16.39% ซึ่งผลลัพธ์นี้ยังไม่ดีพอสำหรับการประยุกต์ใช้งานจริง ดังนั้น จึงทดลองเพิ่มโดยเสนอให้ใช้เป็นการจัดกลุ่มรูปภาพโดยใช้ค่าจาก Prefix Embeddings แทนการสร้างคำบรรยายจากภาพโดยตรง ซึ่งผลลัพธ์ชี้ให้เห็นว่าการศึกษานี้สามารถใช้เป็นแนวทางการพัฒนาต่อยอดได้

คำสำคัญ: CLIP, CLIP Prefix Caption, GPT-2, Prefix Embeddings, การสร้างคำบรรยายภาพ, Traffy Fondue

* *Corresponding author : wasit@tu.ac.th*

Received: 23 มี.ค. 66; Revised: 15 พ.ค. 66; Accepted: 13 มิ.ย. 66

Abstract

Traffy Fondue is a grievance system provided by the Bangkok Metropolitan Administration to receive opinions and suggestions that citizens have on the city. However, due to the amount of information from many users, there is still unclear reporting, such as descriptions and images are inconsistent, which makes it difficult for the receiving officer to coordinate and solve the problem. Therefore, our team proposed a data clustering method to increase the ability of clustering data to be more convenient by using data processing techniques. In this research, the presenter applied the CLIP Prefix Caption model for creating captions and allowing the system to group words or search for related problems. The CLIP, CLIP Prefix Caption and GPT-2 models were created captions using images from Traffy Fondue. The experimental results can be summarized as follows: BLEU 0.93% and ROUGE-1 16.39%. The result is not good enough for real applications. The experiment was further reorganized by proposing to group images using vectors from Prefix Embeddings instead of captioning directly from the image. The results indicate embedding can be applied for further development.

Keywords: CLIP, CLIP Prefix Caption, GPT-2, Prefix Embeddings, Image Captioning, Traffy Fondue

1. บทนำ

การสืบค้นข้อมูลในปัจจุบันนี้เป็นเรื่องพื้นฐานที่คนส่วนใหญ่จะต้องทำกันในชีวิตประจำวัน ซึ่งวิธีการที่ใช้สำหรับสืบค้นนั้นก็มีความหลากหลายมากขึ้นนอกจากการค้นหาด้วยข้อความแบบเดิมที่ทำกันเป็นประจำแล้วนั้น เรายังสามารถที่จะค้นหาด้วยรูปภาพหรือเสียงได้อีกด้วยแต่ทั้งนี้ไม่ว่าจะเป็นการค้นหารูปแบบใด ข้อมูลที่ถูกป้อนเข้ามาเพื่อค้นหานั้นก็จะต้องถูกแปลงให้อยู่ในรูปแบบและประเภทของข้อมูลที่เหมาะสมกับการนำไปค้นข้อมูลในระบบ

จากที่กล่าวไปข้างต้นนี้ ผู้ศึกษาต้องการที่จะทำระบบค้นหาปัญหาต่าง ๆ ที่ได้รับแจ้งจากผู้ใช้งานหรือผู้ที่ประสบปัญหา โดยผู้ศึกษาจะมีหน้าที่พัฒนาในส่วนการค้นหาด้วยรูปภาพ เช่น เมื่อผู้ใช้งานค้นหาด้วยการป้อนข้อมูลเข้ามาเป็นรูปภาพนั้นระบบจะนำภาพมาประมวลผลและทำคำบรรยายภาพ (Image Captioning) เพื่อนำเอาบริบทที่อยู่ในรูปภาพนั้น ๆ บรรยายออกมาเป็นประโยคที่เป็นข้อความเพื่อให้ระบบสามารถนำข้อมูลนี้ไปประมวลผลและค้นหาปัญหาที่เกี่ยวข้องต่อไปได้ และด้วยระบบค้นหาที่ดีขึ้นนี้จะช่วยให้ผู้ใช้งานทราบได้ว่ามีปัญหาอะไรบ้างตามเรื่องที่ผู้ใช้งานสนใจอีกทั้งยังช่วยลดการแจ้งปัญหาซ้ำซ้อนได้อีกด้วยหากระบบสามารถที่จะค้นหาและแสดงปัญหาที่มีอยู่แล้วได้ตรงกันกับปัญหาที่ถูกแจ้งเข้ามา

ภาพที่ 1 ตัวอย่างการสร้างคำบรรยายภาพด้วยแบบจำลอง CLIP Prefix Caption บนรูปภาพจากชุดข้อมูล Traffy Fondue

ในส่วนของการสร้างคำบรรยายจากรูปภาพผู้ศึกษาจะทดลองใช้แบบจำลองเครือข่ายประสาทเทียมที่มีความสามารถด้านการจำแนกรูปภาพ และแบบจำลองการเข้ารหัสและถอดรหัสรูปเพื่อสร้างคำบรรยายภาพที่เหมาะสม โดยแบบจำลองที่จะทดลองใช้งานคือ CLIP Prefix Captioning (Mokady et al., 2021)

แบบจำลอง Image Caption คือ แบบจำลองที่ใช้สำหรับสร้างคำอธิบายให้กับเนื้อหาของรูปภาพให้ออกมาเป็นภาษามนุษย์ ซึ่งจะใช้ความรู้ในส่วนของ Computer Vision และ NLP โดยไอเดียของ Image Caption แบบจำลองนั้น อยู่ในส่วนของกลุ่ม Encoder-Decoder ซึ่งจะประกอบไปด้วยแบบจำลองย่อยสองส่วน แบบจำลองแรก คือ “encoder” โดยจะทำหน้าที่แปลความหมายของรูปภาพให้อยู่ในรูปเวกเตอร์ทางคณิตศาสตร์ และแบบจำลองที่สอง คือ decoder” โดยจะทำหน้าที่ถอดความหมายที่ซ่อนอยู่ในเวกเตอร์ทางคณิตศาสตร์ให้ออกมาในรูปคำบรรยายบนตัวหนังสือในภาษาอังกฤษดังภาพที่ 1 โดยมีตัวอย่างการประยุกต์ใช้แบบจำลอง Image Caption เช่น ใช้ใน Virtual assistants หรือบอทต่าง ๆ เพื่อโต้ตอบกับรูปภาพผู้ใช้เพื่อระบุประเภทของรูปภาพสำหรับจัดหมวดหมู่ใช้สำหรับสร้างคำบรรยายและอ่านให้ผู้พิการทางสายตา เป็นต้น

2. วัตถุประสงค์การวิจัย

เพื่อนำเสนอแบบจำลอง CLIP Prefix Caption สำหรับใช้สร้างคำบรรยายรูปภาพและให้ระบบนำคำที่ได้ไปจัดกลุ่มหรือค้นหาปัญหาที่เกี่ยวข้อง

3. วิธีดำเนินการวิจัย

การสร้างคำบรรยายภาพในงานนี้ผู้ศึกษาใช้ชุดข้อมูลจาก Traffy Fondue โดยคัดข้อมูลในส่วนที่เป็นรูปภาพและความคิดเห็นที่ผู้ใช้งานแจ้งเข้ามา เพื่อนำมาใช้เป็นชุดข้อมูลสำหรับการสร้างคำบรรยายภาพสำหรับป้อนให้กับแบบจำลองแล้วสร้างคำบรรยายภาพขึ้นมาใหม่และความคิดเห็นใช้สำหรับเป็นตัวประเมินผลคุณภาพของคำบรรยายภาพที่ถูกสร้างขึ้น ซึ่งในงานนี้จะเป็นการนำเอาแบบจำลอง CLIP Prefix Caption มาใช้โดยไม่มีการฝึกสอนเพิ่มแต่อย่างใดและมีการเตรียมการเพื่อใช้งานและประเมินผลทั้งหมด ดังนี้

3.1 เตรียมข้อมูลรูปภาพและคำบรรยาย

ชุดข้อมูลที่ใช้ในที่นี้สามารถดาวน์โหลดได้จาก #Traffy x TeamChadChart โดย ณ วันที่ผู้ศึกษาเก็บรวบรวมข้อมูลมานี้ ข้อมูลมีทั้งหมด 100,000 แถว และมีคอลัมน์ ได้แก่ ticket_id, type, organization, comment, coordinate, photo, address, district, subdistrict, province, timestamp, และ state โดยทั้งหมดนี้จัดเก็บอยู่ในรูปแบบของไฟล์ CSV (Comma Separated Values) ซึ่งคอลัมน์ที่ผู้ศึกษาเลือกนำมาใช้งาน คือ photo ที่จัดเก็บเป็น URL ของรูปภาพแต่ละรูป และ comment ที่เป็นข้อความที่ผู้ใช้ส่งมาพร้อมคู่กันกับรูปภาพ เพื่อความสะดวกและความรวดเร็วในการทดลองผู้ศึกษาขอแนะนำให้ดาวน์โหลดรูปภาพจาก URL ดังกล่าวลงไว้บนเครื่องที่ใช้สำหรับทดลองสร้างคำบรรยายภาพเพื่อให้การทดลองอนุมานค่าสามารถทำได้เร็วขึ้น

3.2 การติดตั้งและเตรียมสภาพแวดล้อมสำหรับใช้แบบจำลอง CLIP Prefix Caption

ขั้นตอนการเตรียมการสามารถทำได้โดย git clone https://github.com/openai/CLIP.git (Radford et al., 2021) เพื่อคัดลอกเอาไลบรารี่และโมดูลต่าง ๆ ที่จำเป็นสำหรับแบบจำลอง CLIP จากนั้นสร้าง Python Virtual Environment โดยมีข้อกำหนดและไลบรารี่ที่ต้องติดตั้ง ดังนี้

- Python เวอร์ชันของรันไทม์ 3.7 ขึ้นไป
- Pytorch + CUDA (ทั้ง 2 จะใช้เวอร์ชันใดก็ได้ แต่ขอแนะนำเป็นการใช้เวอร์ชันที่เสถียรที่สุด)
- ไลบรารี่ Transformers
- ไลบรารี่ทั้งหมดตามใน requirements.txt ของ Git ที่คัดลอกมา

จากนั้นให้ดาวน์โหลดค่า Weights ของแบบจำลองที่ถูกฝึกสอนแล้วบนชุดข้อมูล Conceptual Captions เตรียมไว้สำหรับแบบจำลอง Mapping Network

3.3 หน้าที่ของแบบจำลอง CLIP, Mapping Network, และ GPT-2 มีดังนี้

Mapping Network ทำหน้าที่จับคู่ระหว่างคุณลักษณะที่ได้จากรูปภาพที่แบบจำลอง CLIP ทำการเข้ารหัสไว้และนำไปจับกับค่า Constant เพื่อส่งไปยังแบบจำลอง GPT-2 เพื่อสร้างเป็น Prefix Embeddings และป้อนเข้าแบบจำลอง GPT-2

แบบจำลอง CLIP สามารถโหลดโดยใช้โมดูล clip.py และโหลดแบบจำลองโดยกำหนดเป็น ViT-B/32 (Vision Transformers ตัวที่ฝึกสอนโดย Open AI) ซึ่งแบบจำลอง CLIP นี้มีหน้าที่หลักคือการเข้ารหัสรูปภาพ

โดยดึงเอาคุณลักษณะต่าง ๆ ของรูปภาพที่เป็นค่าพิกเซลเก็บไว้และแปลงเป็น Tensor โดยแบบจำลอง CLIP นี้มาพร้อมกับฟังก์ชัน Preprocess ที่ช่วยเตรียมรูปภาพก่อนนำเข้าแบบจำลองนี้

แบบจำลอง GPT-2 นั้นผู้ศึกษาต้องการในส่วนของ Tokenizer เพื่อใช้ในการเข้ารหัสฟีเจอร์ที่ได้จาก CLIP Prefix Caption (Mapping Network) และ GPT-2 จะนำเอาโทเคนที่ได้ไปสร้างเป็นคำบรรยายภาพขึ้นใหม่ โดยการสร้างและประกอบคำขึ้นใหม่นั้นสามารถเลือกได้ว่าจะใช้อัลกอริทึม Beam Search (Meister et al., 2020) ด้วยหรือไม่

3.4 ภาพรวมการใช้งานแบบจำลองเพื่อสร้างคำบรรยายภาพ

แบบจำลอง CLIP Prefix Caption นี้ โดยรวมแล้วคือการใช้แบบจำลอง 3 ตัว ทำงานร่วมกัน คือ CLIP เพื่อเข้ารหัสรูปภาพและดึงคุณลักษณะของรูป, GPT-2 เพื่อสร้างคำบรรยายภาพขึ้นใหม่โดยใช้แค่ส่วน Tokenizer, และ Mapping Network ที่แปลงคุณลักษณะที่ได้จาก CLIP ให้อยู่ในสเปซของ GPT-2 โดยที่แบบจำลอง CLIP Prefix Caption จะรับเอารูปภาพเข้ามาและให้ผลลัพธ์ออกไปเป็นข้อความคำบรรยายภาพ

4. ผลการวิจัย

ผลการทดลองของแบบจำลองสร้างคำบรรยายภาพ CLIP Prefix Caption นี้ผู้ศึกษาใช้รูปภาพจำนวน 6,001 รูป ในการวัดผลโดยได้แปลข้อความในชุดข้อมูล Traffy Fondue ที่มาคู่กับรูปภาพจากภาษาไทยเป็นภาษาอังกฤษจำนวน 15,000 คู่ ข้อความและรูปภาพ จากนั้นจึงเลือกใช้เฉพาะคู่ข้อความและรูปภาพที่จำนวนตัวอักษรของข้อความเฉลยมีไม่เกิน 70 ตัวอักษร เพราะว่าแบบจำลอง CLIP Prefix Caption สามารถที่จะสร้างได้สูงสุดไม่ค่อยเกิน 70-80 ตัวอักษร (ส่วนใหญ่มักจะไม่เกิน 40 ตัวอักษร) การที่เราเลือกความยาวของข้อความผลเฉลยและความยาวของข้อความที่แบบจำลองสร้างขึ้นให้ใกล้เคียงกันก็เพื่อให้ตัววัดประเมินผลสามารถทำงานได้เหมาะสม และในการสร้างคำบรรยายนั้นผู้ศึกษาจะมีการเปรียบเทียบระหว่างการนำอัลกอริทึม Beam Search มาใช้เพื่อค้นหาคำที่เหมาะสม กับแบบที่ไม่ใช้อัลกอริทึม Beam Search

ตารางที่ 1 แบบจำลอง CLIP Prefix Caption ผลเปรียบเทียบระหว่างคำบรรยายที่แบบจำลองสร้างขึ้นกับบรรยายของมนุษย์ความยาว 31,840 ตัวอักษร กับ 30,913 ตัวอักษร ตามลำดับ

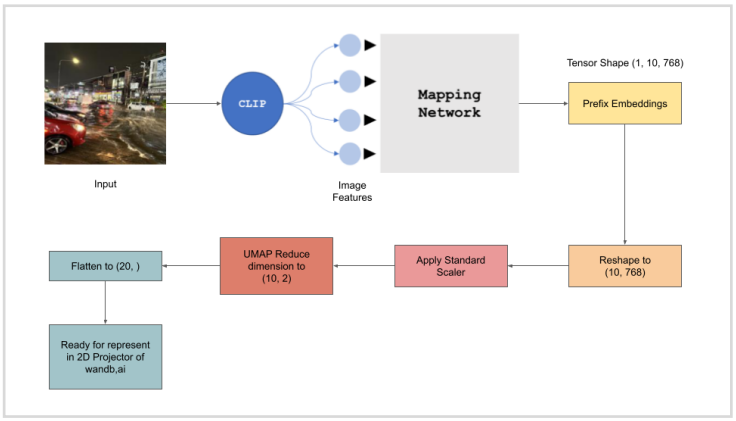
	สร้างคำบรรยายโดยไม่ใช้ Beam Search	**สร้างคำบรรยายโดยใช้ Beam Search**
BLEU	0.93%	0.78%
ROUGE-1	16.39%	14.9%
ROUGE-2	3.35%	2.58%
ROUGE-L	15.45%	13.24%

จากตารางที่ 1 จะเห็นได้ว่าการสร้างคำบรรยายโดยใช้เพียงแค่ความสามารถของแบบจำลองเพียงอย่างเดียวนั้นให้ผลลัพธ์ที่ดีกว่าการนำอัลกอริทึม Beam Search มาช่วยค้นหาคำ แต่ทั้งนี้คะแนนประเมินผลบนทุกตัววัดผลมีเปอร์เซนต์ที่ต่ำมาก อันเนื่องมาจากคำบรรยายของมนุษย์นั้นถูกแปลจากภาษาไทยเป็นภาษาอังกฤษด้วยเครื่องมืออีกทีหนึ่ง และข้อความเหล่านั้นไม่ใช่คำบรรยายภาพโดยตรงแต่เป็นข้อความร้องเรียน แจ้งปัญหา ติหรือชมเป็นส่วนใหญ่ จึงทำให้แตกต่างกับคำบรรยายภาพของแบบจำลองที่พยายามจะอธิบายถึงสิ่งต่าง ๆ ในรูปภาพออกมาเป็นข้อความที่สื่อถึงรูปนั้นเพียงอย่างเดียว

นอกจากนี้จากการสังเกตของผู้ศึกษา พบว่า แบบจำลองยังบรรยายภาพได้ไม่ค่อยตรงกับสิ่งที่อยู่ในรูปภาพนัก ดังภาพที่ 1 หากสังเกตดูให้ชัดจะเห็นว่าแบบจำลองบรรยายไปคนละอย่างกับสิ่งที่มนุษย์เห็นและตีความ แต่ก็ยังมีคำที่เกี่ยวข้องกับรูปนั้น ๆ อยู่ในคำบรรยายเช่นกัน ดังภาพที่ 2 ทำให้ผู้ศึกษาจึงคิดว่าหากต้องการใช้ประโยชน์จากแบบจำลอง CLIP Prefix Caption กับชุดข้อมูล Traffy Fondue ควรที่จะเปลี่ยนจากการนำไปสร้างคำบรรยายภาพโดยตรงเป็นการนำคุณลักษณะที่ได้จาก Prefix Embeddings ไปจัดกลุ่มรูปภาพเพื่อนำเสนอว่ารูปภาพนั้น ๆ อยู่ในกลุ่มรูปภาพที่คล้ายกันเองหรือไม่ เพื่อให้ระบบสามารถที่จะจัดกลุ่มข้อมูลที่ผู้ใช้ส่งเข้ามาและพอทราบได้ว่าเป็นปัญหาประเภทใด กระบวนการทดลองในส่วนของการนำ Prefix Embeddings มาแสดงผลแบบ Embedding Projector นี้ผู้ศึกษาเริ่มจากการเตรียมข้อมูลสำคัญทั้งหมด 3 อย่าง ได้แก่

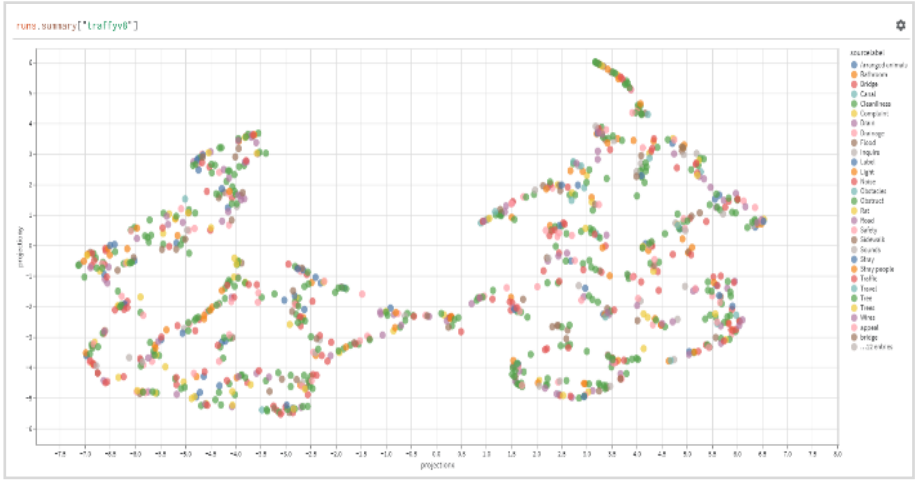
1. รูปภาพที่จัดเตรียมเป็นรูปแบบของ wandb.ai โดยแปลงจาก URL รูปภาพในชุดข้อมูล Traffy Fondue เป็นรูปแบบ PIL Image และแปลงเป็นวัตถุบน wandb.ai
2. หมวดหมู่ของรูปภาพ โดยผู้ศึกษาเตรียมจากคอลัมน์ Type ในชุดข้อมูล Traffy Fondue ซึ่งเป็นหมวดหมู่ของปัญหาที่เกี่ยวข้องกับรูปนั้น ๆ ซึ่งผู้ใช้งานเป็นคนระบุเข้ามา

ภาพที่ 2 ตัวอย่างการสร้างคำบรรยายภาพด้วยแบบจำลอง CLIP Prefix Caption บนรูปภาพจากชุดข้อมูล Traffy Fondue

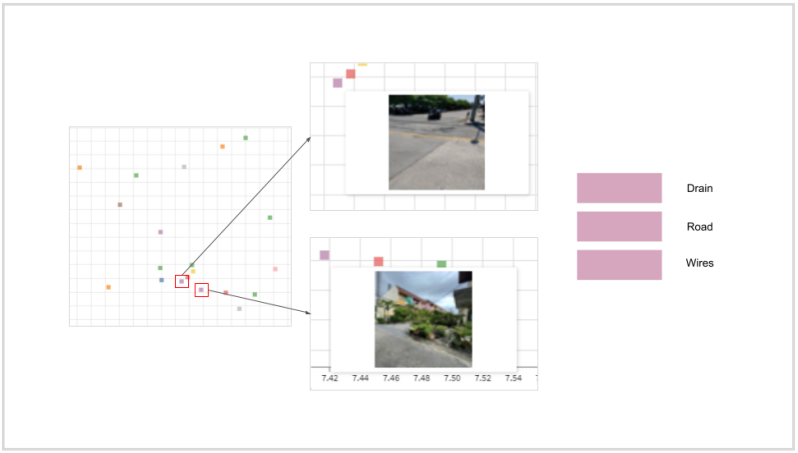


ภาพที่ 3 กระบวนการสร้าง Prefix Embeddings ให้พร้อมสำหรับนำไปทำ Embedding Projector ใน wandb.ai

3. ชุดค่าของตัวเลขในอาเรย์ (เวกเตอร์) ที่เป็นค่าของ Prefix Embeddings ได้มาจากการนำรูปภาพเข้าสู่แบบจำลอง CLIP และนำค่าที่ได้จาก CLIP เข้าแบบจำลอง CLIP Caption เพื่อให้ได้ค่า Prefix Embeddings โดยค่าที่ได้จะเป็น Tensor ขนาด (1, 10, 768) จากนั้นผู้ศึกษาใช้การ reshape เพื่อลดให้เหลือ 2 มิติเป็น (10, 768) ดังภาพที่ 3 แล้วจึงปรับช่วงของข้อมูลและใช้ UMAP (McInnes et al., 2018) สำหรับลดมิติของข้อมูลลง สุดท้ายคือนำไปเรียงต่อกันเป็น 1 มิติเพื่อให้สามารถนำเข้าสู่เครื่องมือ wandb.ai และแสดงผลลัพธ์แบบ Embedding Projector ได้ดังภาพที่ 4 ผลลัพธ์ของการทดลองในส่วนนี้ผู้ศึกษาจะใช้เครื่องมือทดลองอย่าง Wandb.ai (Weight and Bias AI) ผลลัพธ์ที่ได้มีดังต่อไปนี้สำหรับผู้สนใจผลการทดลองนี้สามารถเข้าไปดูได้ที่นี่ Traffy Image Prefix Embedding in 2D Projector



ภาพที่ 4 ผลลัพธ์จากการนำ Prefix Embeddings ที่ได้จากแบบจำลอง Clip Prefix Caption มาแสดงผลแบบ 2D Projection เพื่อสังเกตดูกลุ่มก้อนของข้อมูลโดย Prefix Embeddings ที่คล้ายกันควรจะอยู่ด้วยกัน ซึ่งจะหมายถึงรูปภาพที่คล้ายกัน



ภาพที่ 5 ตัวอย่างจากจุดสีม่วงสองจุดที่อยู่ใกล้กันมีรูปภาพเป็นรูปเกี่ยวกับทางถนนทั้งคู่ โดยที่สีม่วงในแผนภาพนี้จะหมายถึง 3 เรื่องได้แก่ Drain (ท่อระบายน้ำ), Road (ถนน), Wires (สายไฟ) และทั้ง 2 รูปนี้เกี่ยวข้องกับถนน

5. อภิปรายผลการวิจัยและข้อเสนอแนะ

จากการทดลองสร้างคำบรรยายภาพด้วยแบบจำลอง CLIP Prefix Caption บนชุดข้อมูล Traffy Fondue นี้สรุปได้ว่าผลลัพธ์จากคำบรรยายที่แบบจำลองสร้างขึ้นนั้นมีประสิทธิภาพที่ต่ำเกินกว่าจะใช้กับงานรับแจ้งปัญหาของ Traffy Fondue x TeamChadChart อันเนื่องมาจากคำบรรยายหรือคอมเมนต์ที่ผู้ใช้งานส่งเข้ามาพร้อมกับรูปภาพนั้นมีความหลากหลายและมีความหมายที่มากกว่าสิ่งที่ปรากฏอยู่ในรูปภาพ แต่ทั้งนี้คำบรรยายที่แบบจำลองสร้างขึ้นมานั้นยังคงคำสำคัญไว้ได้บ้าง ฉะนั้นหากนำคำบรรยายที่สร้างขึ้นได้นี้ไปสกัดคำสำคัญต่อและนำเข้าไปค้นหาในกราฟความรู้ก็อาจทำให้ได้คอนเซปต์หรือบ่งชี้ถึงเรื่องที่ผู้ใช้ต้องการจะแจ้งได้

ส่วนที่สำคัญและสามารถนำไปใช้งานได้คือ Prefix Embedding ดังภาพที่ 2 โดยการนำค่าที่ได้จาก CLIP เข้าสู่แบบจำลอง Mapping Network (แบบจำลอง Clip Prefix Caption) นั้นช่วยให้ผู้ศึกษาสามารถสร้างค่าบางอย่างที่พอจะอธิบายถึงรูปภาพและจัดกลุ่มพวกมันได้ ซึ่งความสามารถในส่วนนี้อาจช่วยให้การจัดกลุ่มของปัญหาที่แจ้งเข้ามาทำได้ง่ายขึ้น เพราะในบางครั้งหมวดหมู่ของปัญหาที่ผู้ใช้ส่งเข้ามาอาจไม่ตรงกับรูปภาพหรือมีหลายหมวดหมู่ ดังนั้นถ้าหากใช้คุณลักษณะที่สกัดจากรูปภาพเพื่อจัดกลุ่มปัญหาของเรื่องที่แจ้งเข้ามาเป็นรูปภาพได้เหมาะสมมากยิ่งขึ้น

การนำไปใช้งานจริงสามารถประยุกต์ใช้กับระบบที่มีความต้องการจัดกลุ่มรูปภาพแบบอัตโนมัติซึ่งเมื่อรูปภาพถูกจัดกลุ่มแล้วก็จะสะดวกต่อการค้นหาและเรียกใช้งานมากยิ่งขึ้น นอกจากนี้ยังสามารถปรับใช้กับเรื่องการค้นหารูปภาพจากข้อความได้อีกด้วยเช่น ผู้ใช้งานส่งข้อความเข้าสู่ระบบ ระบบเข้ารหัสข้อความและแปลงเป็น Prefix Embeddings จากนั้นนำค่าที่ได้นี้ไปเปรียบเทียบกับ Prefix Embeddings ของรูปภาพที่มีอยู่เพื่อหาค่าที่ใกล้เคียงกันมากที่สุด จากนั้นส่งผลลัพธ์กลับมาเป็นรูปภาพที่ใกล้เคียงกับคำค้นหามากที่สุดให้กับผู้ใช้งาน ดังภาพที่ 5 เป็นต้น แนวทางการต่อยอดในการพัฒนาและเปรียบการทดลองโดยใช้ image captioning

ที่หลากหลายเพื่อหาเทคนิคที่แม่นยำมากขึ้น โดยมีแนวทางได้แก่ mPLUG (Li et al., 2022) และ OFA (Wang et al., 17--23 Jul 2022) ส่วนการแสดงผล clustering จะมีการเปรียบเทียบการทำ dimension reduction โดยใช้ PCA และ t-SNE (Maaten & Hinton, 2008)

7. เอกสารอ้างอิง

Li, C., Xu, H., Tian, J., Wang, W., Yan, M., Bi, B., Ye, J., Chen, H., Xu, G., Cao, Z., Zhang, J., Huang, S., Huang, F., Zhou, J., & Si, L. (2022). mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections. *arXiv:2205.12005v2 [cs.CL]*. https://doi.org/10.48550/arXiv.2205.12005

Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research 9*(86), 2579–2605. http://www.jmlr.org/papers/v9/vandermaaten08a.html

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426v3 [stat.ML]*. https://doi.org/10.48550/arXiv.1802.03426

Meister, C., Vieira, T., & Cotterell, R. (2020). Best-first beam search. *Transactions of the Association for Computational Linguistics 8*, 795-809. https://doi.org/10.1162/tacl_a_00346/96473.

Mokady, R., Hertz, A., & Bermano, A. H. (2021). ClipCap: CLIP Prefix for Image Captioning. arXiv:2111.09734v1. https://doi.org/10.48550/arXiv.2111.09734

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. *arXiv:2103.00020v1.* https://doi.org/10.48550/arXiv.2103. 00020

Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., & Yang, H. (17-23 Jul 2022). OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, & S. Sabato (Eds.), *Proceedings of the 39th International Conference on Machine Learning Vol. 162* (pp. 23318–23340). https://proceedings.mlr.press/v162/wang22al/wang22al.pdf