การพัฒนาวิธีการสร้างคำบรรยายภาพภาษาไทยโดยใช้การเรียนรู้เชิงลึก

โดย
นายวิชญ์พล เทียนชอ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์ แผน ก แบบ ก 2 ระดับปริญญามหาบัณฑิต
ภาควิชาวิศวกรรมไฟฟ้า
มหาวิทยาลัยศิลปากร
ปีการศึกษา 2566

การพัฒนาวิธีการสร้างคำบรรยายภาพภาษาไทยโดยใช้การเรียนรู้เชิงลึก

โดย
นายวิชญ์พล เทียนชอ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์ แผน ก แบบ ก 2 ระดับปริญญามหาบัณฑิต
ภาควิชาวิศวกรรมไฟฟ้า
มหาวิทยาลัยศิลปากร
ปีการศึกษา 2566

DEVELOPMENT OF THAI IMAGE CAPTIONING METHOD USING DEEP LEARNING

By

MR. Witchaphon TIEANCHO

A Thesis Submitted in Partial Fulfillment of the Requirements

for Master of Engineering (ELECTRICAL AND COMPUTER ENGINEERING)

Department of ELECTRICAL ENGINEERING

Academic Year 2023

หัวข้อ	การพัฒนาวิธีการสร้างคำบรรยายภาพภาษาไทยโดยใช้การเรียนรู้เชิงลึก
โดย	นายวิชญ์พล เทียนชอ
สาขาวิชา	วิศวกรรมไฟฟ้าและคอมพิวเตอร์ แผน ก แบบ ก 2 ระดับปริญญามหาบัณฑิต
อาจารย์ที่ปรึกษาหลัก	อาจารย์ ดร. โสภณ ผู้มีจรรยา

คณะวิศวกรรมศาสตร์และเทคโนโลยีอุตสาหกรรม มหาวิทยาลัยศิลปากร ได้รับพิจารณาอนุมัติให้เป็นส่วนหนึ่งของการศึกษา ตามหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต

.....คณบดีคณะวิศวกรรมศาสตร์และ
(ผู้ช่วยศาสตราจารย์ ดร. อรุณศรี ลีจีรจำเนียร) เทคโนโลยีอุตสาหกรรม

พิจารณาเห็นชอบโดย

.....ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร. ระพีพันธ์ แก้วอ่อน)

.....อาจารย์ที่ปรึกษาหลัก
(อาจารย์ ดร. โสภณ ผู้มีจรรยา)

.....ผู้ทรงคุณวุฒิภายใน
(ผู้ช่วยศาสตราจารย์ ดร. ภมร ศิลาพันธ์)

.....ผู้ทรงคุณวุฒิภายนอก
(ผู้ช่วยศาสตราจารย์ ดร. วีรพล จิรจริต)

640920027 : วิศวกรรมไฟฟ้าและคอมพิวเตอร์ แผน ก แบบ ก 2 ระดับปริญญามหาบัณฑิต
คำสำคัญ : คำบรรยายภาพภาษาไทย, ชุดข้อมูลการจราจร, ชุดข้อมูล Flickr8k, โครงข่ายประสาทเทียมแบบ Convolutional, LSTM แบบสองทิศทาง, ตัวชี้วัด BLEU

นาย วิชญ์พล เทียนชอ: การพัฒนาวิธีการสร้างคำบรรยายภาพภาษาไทยโดยใช้การเรียนรู้เชิงลึก อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก : อาจารย์ ดร. โสภณ ผู้มีจรรยา

วิทยานิพนธ์เล่มนี้ได้ออกแบบและพัฒนาโมเดลการเรียนรู้เชิงลึกเพื่อสร้างคำบรรยายภาพภาษาไทยโดยใช้ Convolutional Neural Network (CNN) อย่างเช่น VGG16 และอื่นๆ เพื่อคัดแยกคุณลักษณะของรูปภาพและได้ใช้ Bidirectional LSTM ในการสร้างคำบรรยายภาพ โดยที่ CNN คือกระบวนการในการเข้ารหัส และ Bidirectional LSTM คือกระบวนการในการถอดรหัส ซึ่ง Bidirectional LSTM คือ LSTM อีกประเภทที่ช่วยให้โมเดลสามารถเรียนรู้ได้แบบสองทิศทางคือทิศทางไปข้างหน้าและทิศทางย้อนกลับทำให้โมเดลเรียนรู้และแยกแยะคำที่มีความคล้ายคลึงกันได้รวมถึงเพิ่มความสามารถของหน่วยความจำโมเดล และในส่วนของชุดข้อมูลที่ใช้สำหรับการฝึกสอนและทดสอบประกอบด้วย ฐานข้อมูลแรกคือ Flickr8k ซึ่งเป็นฐานข้อมูลสาธารณะที่ภายในฐานข้อมูลประกอบไปด้วยรูปภาพจำนวน 8091 รูป และคำบรรยายภาษาอังกฤษ 5 คำบรรยายซึ่งจะทำการแปลคำบรรยายเป็นภาษาไทยโดยใช้ Google Translate ก่อน โดยส่วนใหญ่ฐานข้อมูลชุดนี้จะเป็นรูปภาพและคำบรรยายที่เกี่ยวกับชีวิตประจำวันทั่วไป และฐานข้อมูลที่สองคือ ชุดข้อมูลการจราจรที่จัดทำขึ้นเองซึ่งภายในจะประกอบไปด้วยรูปภาพ 429 รูป และคำบรรยายภาษาไทย 5 คำบรรยาย โดยฐานข้อมูลชุดนี้คือรูปภาพและคำบรรยายที่เกี่ยวข้องกับการสัญจรบนท้องถนนอย่างเช่น เด็กผู้หญิงคนหนึ่งกำลังเดินข้ามถนน ไฟแดงเตือนให้รถยนต์และรถจักรยานยนต์ทุกคันต้องหยุด ซึ่งเหตุผลที่ได้จัดทำชุดข้อมูลนี้เพราะว่าวิทยานิพนธ์เล่มนี้หวังว่างานวิจัยชุดนี้ในอนาคตจะสามารถทำการสร้างระบบแจ้งเตือนให้กับผู้ขับขี่บนท้องถนนหรือแม้แต่ผู้ที่สัญจรอยู่ตามท้องถนนไม่ใช่กับผู้ขับขี่อย่างเดียวซึ่งระบบการแจ้งเตือนนั้นจะเป็นการแจ้งเตือนด้วยเสียงเมื่อโมเดลรับอินพุตภาพเข้ามาแล้วแต่วิทยานิพนธ์ฉบับนี้ไม่ได้ทำไปจนถึงระบบนั้น ดังนั้นการทดลองของวิทยานิพนธ์ฉบับนี้จะทำการรวมชุดข้อมูลทั้งสองเข้าด้วยกันเพราะไม่เพียงแต่ต้องการดูผลลัพธ์ที่เกี่ยวข้องกับการจราจรแต่ต้องการดูผลลัพธ์การบรรยายรูปภาพทั่วไปด้วยอีกทั้งการรวมชุดข้อมูลเข้าด้วยกันยังช่วยเสริมการเรียนรู้ให้กับโมเดลด้วย และสุดท้ายได้ทำการประเมินคำบรรยายที่โมเดลสร้างเทียบกับคำบรรยายอ้างอิงโดยการใช้ตัวชี้วัด BLEU

640920027 : Major (ELECTRICAL AND COMPUTER ENGINEERING)
Keyword : Thai Captions, Traffic Dataset, Flickr8k Dataset, Convolutional Neural Networks(CNN), Bidirectional LSTM, BLEU Metric

MR. Witchaphon TIEANCHO : Development of Thai Image Captioning Method Using Deep Learning Thesis advisor : SOPON PHUMEECHANYA, Ph.D.

This thesis designed and developed a deep learning model to create Thai image captions using Convolutional Neural Network (CNN) such as VGG16 and others to extract image features and use Bidirectional LSTM is used to create captions, where CNN is the encoding process and Bidirectional LSTM is the decoding process. Bidirectional LSTM is another type of LSTM that allows the model to learn in two directions. The forward and reverse directions allow the model to learn and distinguish similar words and improve the model's memory capacity. And the dataset used for training and testing includes: The first database is Flickr8k, which is a public database that contains 8091 images and 5 English subtitles, which will be translated into Thai using Google Translate first. Most of this database. It will be pictures and descriptions related to daily life. and the second database is A custom-made traffic dataset containing 429 images and 5 Thai language captions. This database contains images and captions related to road traffic such as A girl was walking across the road. A red light warns all cars and motorcycles to stop. The reason for creating this data set is because this thesis hopes that in the future this research will be able to create a warning system for drivers on the road or even people traveling on the road, not just drivers. The only notification system is an audio notification when the model receives image input, but this thesis does not go into that system. Therefore, the experiment of this thesis will combine the two datasets because we want to not only see traffic-related results but also to see general image description results. Moreover, combining the datasets also enhances learning for the model as well And finally, the subtitles generated by the model were evaluated against the reference subtitles using the BLEU metric.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้ได้รับการปรึกษาจากอาจารย์ ดร.โสภณ ผู้มีจรรยา เป็นที่ปรึกษาวิทยานิพนธ์ และคณะอาจารย์กรรมการสอบวิทยานิพนธ์ทุกท่าน ได้แก่ผู้ช่วยศาสตราจารย์ดร.ระพีพันธ์ แก้วอ่อน ประธานสอบวิทยานิพนธ์อาจารย์ดร.ภมร ศิลาพันธ์กรรมการผู้ทรงคุณวุฒิภายใน และผู้ช่วยศาสตราจารย์ดร.วีรพล จิรจริต กรรมการผู้ทรงคุณวุฒิภายนอก ที่ให้คำปรึกษา ปรับปรุง และแนวทางการแก้ไขให้วิทยานิพนธ์ฉบับนี้มีความสมบูรณ์ รวมถึงการนำเสนอผลงาน และกระบวนการวิจัย

ขอขอบพระคุณภาควิชาวิศวกรรมไฟฟ้า และคณะวิศวกรรมศาสตร์และเทคโนโลยีอุตสาหกรรมมหาวิทยาลัยศิลปากร คณาจารย์ประจำภาควิชาวิศวกรรมไฟฟ้า และเจ้าหน้าที่และบุคลากรภาควิชาวิศวกรรมไฟฟ้าทุกท่าน ที่ให้ความช่วยเหลือทั้งด้านการให้คำแนะนำ และให้ความช่วยเหลืออำนวยความสะดวกต่างๆต่อการวิทยานิพนธ์ฉบับนี้ หากมีข้อผิดผลาดประการใดทางผู้จัดทำขออภัยมา ณ ที่นี้

วิชญ์พล เทียนชอ

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

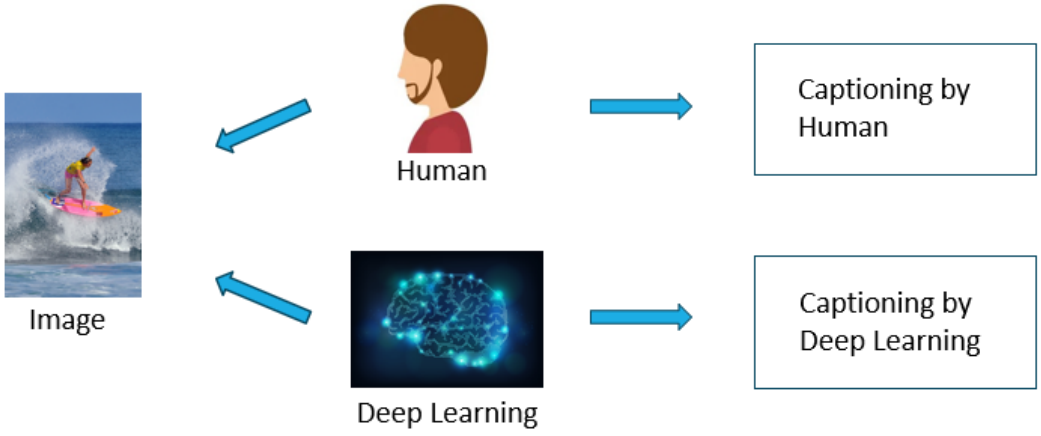
บทที่ 1
บทนำ

1.1 ความสำคัญและที่มาของปัญหา

ในปัจจุบันงานด้าน Computer Vision ได้รับความสนใจเป็นอย่างมากเพราะมีประโยชน์หลายอย่างเช่น การคัดแยกองค์ประกอบของภาพ การแบ่งกลุ่มภาพเป็นหมวดหมู่ การตราจจับวัตถุภายในภาพ การรู้จำใบหน้า รวมไปถึงการนำไปพัฒนาต่อยอดเป็นการบรรยายภาพ(Image Captioning) ที่ตัวโมเดลเมื่อได้รับข้อมูลอินพุตที่เป็นรูปภาพเข้ามาในระบบจากนั้นจะทำการบรรยายลักษณะขอรูปภาพออกมาในรูปแบบของข้อความหรือเสียงเพื่อให้ผู้อ่านหรือผู้ฟังสามารถเข้าใจรูปภาพนั้นได้โดยที่อาจจะเคยเห็นหรือไม่เคยเห็นรูปภาพนั้นมาก่อน โดยปกติแล้วชุดข้อมูลภาพและคำบรรยายที่เป็นอินพุตจะใช้จากชุดฐานข้อมูลที่มีอยู่ทั่วไปอย่าง Flickr8k, COCO เพื่อนำมาฝึกสอนและทดสอบโมเดลในการสร้างคำบรรยายภาพแต่ชุดข้อมูลทั้งสองคำบรรยายของรูปภาพนั้นเป็นภาษาอังกฤษแต่ในงานวิจัยฉบับนี้เห็นว่างานที่เกี่ยวข้องกับการบรรยายรูปภาพส่วนมากจะบรรยายมาเป็นรูปแบบภาษาอังกฤษอีกทั้งงานวิจัยนี้ยังมีความต้องการที่จะสื่อสารกับบุคคลที่เป็นคนไทยเป็นหลัก ดังนั้นหากต้องการที่จะบรรยายรูปภาพที่เฉพาะเจาะจงชุดฐานข้อมูลที่มีทั่วไปอาจไม่เพียงพอต่อการฝึกสอนและทดสอบ งานวิจัยฉบับนี้จึงเพิ่มฐานข้อมูลเฉพาะเจาะจงที่เกี่ยวข้องกับการจราจรทั้งรูปภาพและคำบรรยายภาษาไทยเข้าไปร่วมกับฐานข้อมูลที่มีอยู่เพราะเล็งเห็นว่าข้อมูลเหล่านี้สามารถนำไปใช้ประโยชน์ได้เป็นอย่างดีเกี่ยวกับการแจ้งเตือน เช่น การแจ้งเตือนให้กับผู้คนที่กำลังข้ามทางม้าลายว่ามีรถกำลังขับขี่อยู่ การแจ้งเตือนให้แก่ผู้ที่กำลังขับรถยนต์ว่าข้างหน้ามีไฟแดง เป็นต้น

การบรรยายภาพไม่เพียงแต่ต้องอาศัยชุดข้อมูลการฝึกสอนจำนวนมากแต่มันยังต้องการการออกแบบโมเดลที่ใช้ในการฝึกสอนที่ดีและมีประสิทธิภาพด้วย โดยใช้การเรียนรู้เชิงลึกในการออกแบบโมเดลซึ่งส่วนสำคัญเลยคือ การคัดแยกคุณลักษณะของรูปภาพ การสร้างคำบรรยาย ซึ่งจะอาศัยส่วนสำคัญทั้งสองอย่างนี้ในการคิดและออกแบบในการทดลองเพื่อให้ผลลัพธ์ในการสร้างคำบรรยายภาพมีความถูกต้องมากที่สุด เพราะมีการเช็คการบรรยายภาพที่โมเดลสร้างออกมาเทียบกับคำบรรยายต้นฉบับด้วยในการประเมินผลลัพธ์ซึ่งในการประเมินจะใช้ตัวชี้วัดสำหรับการประเมินการบรรยายซึ่งในวิทยานิพนธ์ฉบับนี้ได้เลือกใช้ BLEU (Bilingual Evaluation Understudy) สำหรับการประเมิน

และในงานการสร้างคำบรรยายการรับอินพุตรูปภาพเข้ามานั้นเพื่อทำการคัดแยกคุณลักษณะโดยใช้ Convolutional Neural Network (CNN) ซึ่ง CNN ก็จะถูกแบ่งออกเป็นการทำ convolution ในอีกหลากหลายรูปแบบและถูกตั้งชื่อตามวิธีการหรือจำนวนชั้นที่แตกต่างกันไป ต่อมาสิ่งสำคัญอีกอย่างสำหรับโมเดลการบรรยายภาพคือ Long Short Term Memory(LSTM) ซึ่งเป็นสถาปัตยกรรมที่อยู่ใน Recurrent Neural Network (RNN) โดย LSTM ที่นำมาใช้ในโมเดลการสร้างคำบรรยายภาพจะมีหน้าที่ในการสร้างคำภาษาไทยโดยจะทำการประมวลผลความน่าจะเป็นจากรูปภาพและคำก่อนหน้าทำกระบวนการเป็นลำดับต่อลำดับหรือเรียกวิธีการนี้ว่าการทำ decoder ส่วน CNN คือการทำ encoder แต่สำหรับวิทยานิพนธ์ฉบับนี้ได้เลือกใช้ LSTM อีกประเภทที่ชื่อว่า Bidirectional LSTM ซึ่งสามารถทำการประมวลผลแบบสองทิศทางได้ทำให้โมเดลทำการเรียนรู้ได้ดีกว่า LSTM แบบปกติ และสำหรับชุดข้อมูลในงานวิจัยนี้ได้เลือกใช้ Flickr8k ที่เป็นฐานข้อมูลสาธารณะที่ส่วนใหญ่รูปภาพและคำบรรยายจะเกี่ยวกับชีวิตประจำวันทั่วไปแต่คำบรรยายในข้อมูลชุดนี้เป็นภาษาอังกฤษดังนั้นจึงต้องทำการแปลเป็นภาษาไทยโดยใช้ Google Translate ก่อนเข้าโมเดลฝึกสอน และทำการรวมชุดข้อมูล Flickr8k กับชุดข้อมูลการจราจรที่จัดทำขึ้นเองโดยภายในชุดข้อมูลนี้จะประกอบไปด้วยรูปภาพที่เกี่ยวข้องกับการสัญจรบนท้องถนนซึ่งคำบรรยายในชุดข้อมูลนี้เป็นภาษาไทยอยู่แล้วเพราะได้จัดทำขึ้นเอง



รูปที่ 1-1 แสดงถึงหลักแนวคิดของการบรรยายภาพ

1.2 วัตถุประสงค์ของงานวิจัย

1.2.1 เพื่อศึกษาโมเดลการเรียนรู้เชิงลึกในการสร้างคำบรรยายภาพ

1.2.2 เพื่อออกแบบโมเดลในการสกัดคุณลักษณะสำคัญในภาพและการสร้างคำบรรยายภาพ

1.2.3 เพื่อปรับปรุงประสิทธิภาพของโมเดลในการสร้างคำบรรยายภาพภาษาไทย

1.3 ขอบเขตของงานวิจัย

1.3.1 สร้างคำบรรยายภาพภาษาไทย

1.3.2 ใช้ภาพจากฐานข้อมูลสาธารณะ Flickr8k ร่วมกับฐานข้อมูลรูปภาพและข้อความที่เกี่ยวกับการจราจรที่จัดทำขึ้นเอง

1.3.3 โมเดลในการสกัดคุณลักษณะสำคัญใช้ CNN

1.3.4 โมเดลในส่วนของการสร้างคำบรรยายใช้ Bidirectional LSTM

1.3.5 วัดประสิทธิภาพของโมเดลในการสร้างคำบรรยายภาพภาษาไทยโดยใช้ BLEU

1.4 ประโยชน์ที่คาดว่าจะได้รับ

ได้โมเดลใหม่ที่ถูกพัฒนาให้สามารถสร้างคำบรรยายภาพภาษาไทยได้มีประสิทธิภาพมากยิ่งขึ้น

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ทฤษฎีและงานวิจัยที่เกี่ยวข้องได้นำเสนอความรู้พื้นฐานที่สามารถนำไปสร้างการบรรยายภาพ (Image Captioning) ซึ่งประกอบด้วยวิธีการของ โครงข่ายประสาทแบบคอนโวลูชัน (Convolutional Neural Network) โครงข่ายประสาทเทียมแบบเกิดซ้ำ(Recurrent Neural Network) การตรวจจับวัตถุ(Object Detection) การประมวลผลภาษาธรรมชาติ(Natural Language Processing) และ งานวิจัยที่เกี่ยวข้องทางด้านการสร้างคำบรรยายภาพ

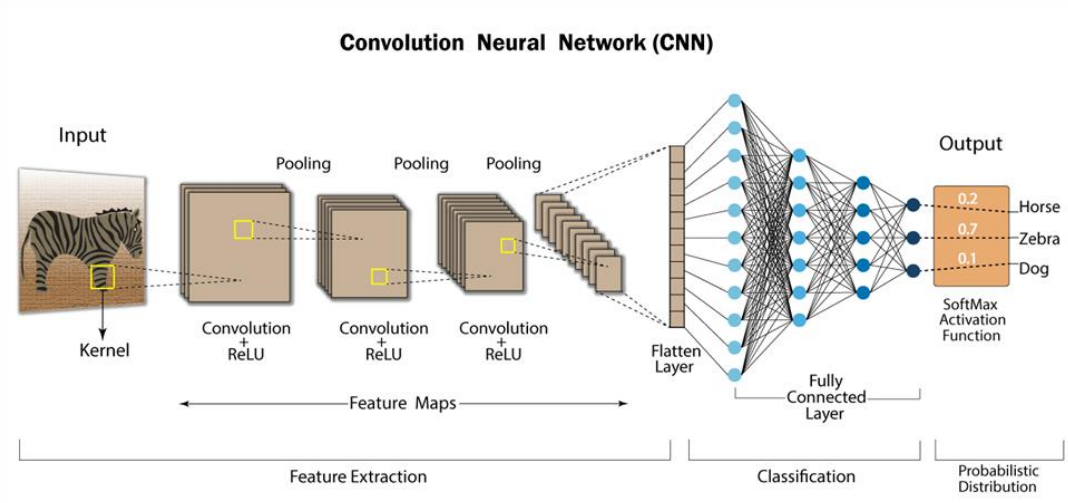
2.1 โครงข่ายประสาทแบบคอนโวลูชัน (Convolutional Neural Network)

โครงข่ายประสาทแบบคอนโวลูชัน(Convolutional Neural Network)[1] หรือ CNN เป็นการเลียนแบบระบบประสาทการมองเห็นของมนุษย์เพราะเมื่อมนุษย์จ้องมองไปที่วัตถุหรือเห็นภาพใดๆระบบประสาทจะทำการจำแนกลักษณะต่างๆ หรือคุณลักษณะต่างๆเอาไว้แล้วนำมาคิดวิเคราะห์เพื่อให้รู้ว่าสิ่งที่เห็นนั้นคือสิ่งใด ซึ่ง CNN จะเป็นการทำงานหลักของ 2 วิธีคือ Feature Extraction และ Neural Network โดยในขั้นตอนของ Feature Extraction จะมีการทำงานที่แบ่งย่อยออกไปอีกคือ คอนโวลูชัน(Convolution) พูลลิ่ง(Pooling) และขั้นตอนของ Neural Network ใช้สำหรับการจำแนกหรือแยกประเภทโดยมักจะถูกเรียกว่าการทำ Fully Connected Layer โดยหลักการโดยละเอียดมีดังนี้

1. คอนโวลูชัน (Convolution) เป็นการหาหรือคัดแยกคุณลักษณะของรูปภาพออกมาโดยเป็นการทำกับกระบวนการภาพที่รับเข้า (Input Image) กับ ภาพข้อมูลขนาดเล็กหรือ kernel ที่เป็นส่วนสำคัญที่ทำให้สามารถแยกลักษณะในรูปภาพได้อย่างเช่นต้องการตรวจจับเส้นขอบวัตถุในภาพ (Edge Detection) โดยทำการเลื่อน kernel ไป 1 ขั้นของภาพอินพุตทำจนครบทุกค่า pixel ในภาพอินพุตแล้วจะได้ผลลัพธ์เป็นภาพ Feature Map

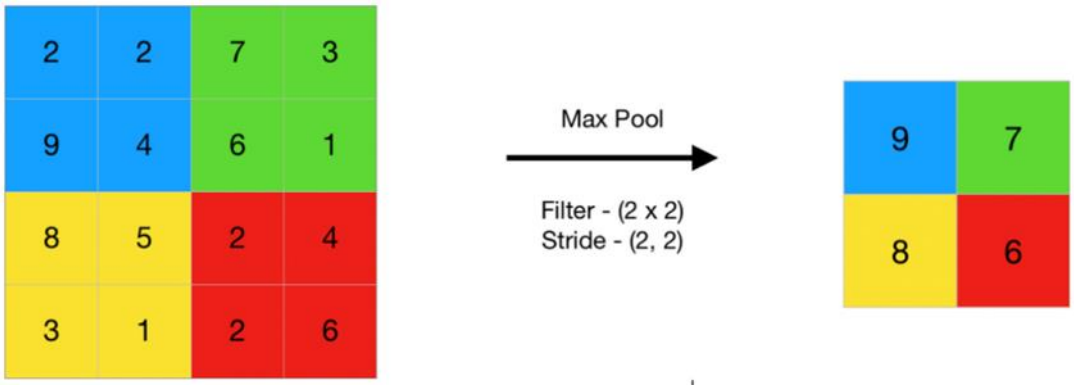
2. พูลลิ่ง (Pooling) เป็นการลดขนาดของรูปภาพลงจึงทำให้คุณลักษณะในรูปภาพนั้นมีขนาดเล็กลง แต่ไม่ได้ทำให้คุณลักณะนั้นหายไปยังสามารถเห็นและแยกคุณลักษณะนั้นออกมาได้อยู่ ซึ่งทำให้การทำงานในระบบไม่จำเป็นต้องใช้ข้อมูลที่มีขนาดใหญ่มากโดยการทำพูลลิ่งจะมีอยู่ด้วยกันคือ การหาพูลลิ่งสูงสุด (Max Pooling) ที่อาจใช้ kernel ขนาด 2x2หรือ Pooling2x2แล้วทำการทาบในภาพอินพุตเพื่อหาค่าสูงสุด การหาพูลลิ่งเฉลี่ย (Average Pooling) ที่อาจใช้ kernel นาด 2x2 หรือ Pooling2x2 ทำการทาบในภาพอินพุตเพื่อหาค่าเฉลี่ย

3. Neural Network ที่จริงแล้วการทำงานจะถูกเรียกว่า Fully Connected Layer ที่อยู่ในชั้น Neural Network โดยมีหน้าที่ในการจำแนกประเภทของข้อมูลที่ผ่านการทำ Feature Extraction หรือการแยกคุณลักษณะมาแล้ว



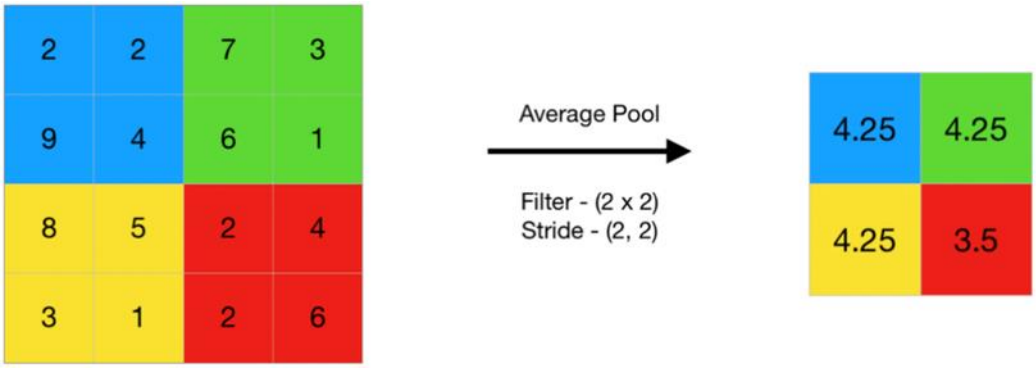
รูปที่ 2-1 โครงข่ายประสาทแบบคอนโวลูชัน(Convolutional Neural Network)

ที่มา https://i0.wp.com/developersbreach.com/wp-content/uploads/2020/08/cnn_banner.png?fit=1400%2C658&ssl=1



รูปที่ 2-2 พูลลิ่งสูงสุด (Max pooling)

ที่มา https://media.geeksforgeeks.org/wp-content/uploads/20190721025744/Screenshot-2019-07-21-at-2.57.13-AM.png

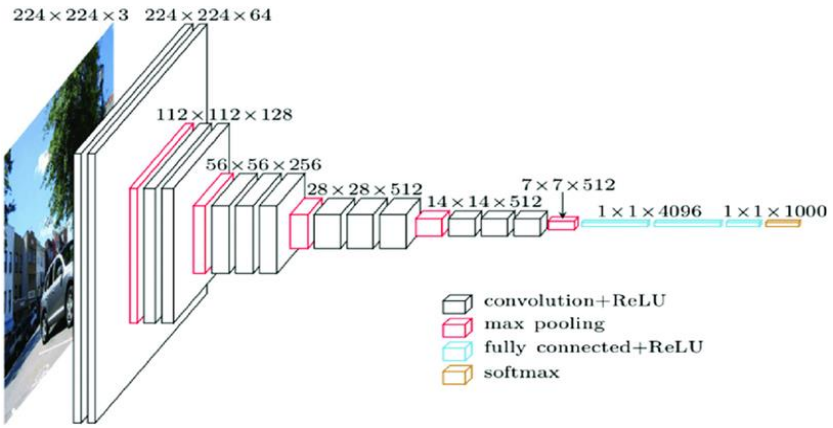


รูปที่ 2-3 พูลลิ่งเฉลี่ย (Average Pooling)

ที่มา https://media.geeksforgeeks.org/wp-content/uploads/20190721030705/Screenshot-2019-07-21-at-3.05.56-AM.png

2.1.1 VGG16

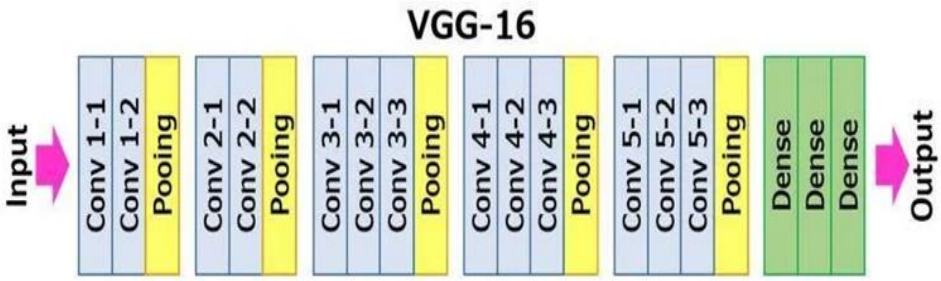
VGG16[2] คือโครงสร้างของ Convolutional Neural Network (CNN) ประเภทหนึ่งที่ถูกขนานนามว่าเป็นโมเดล Computer Vision ที่ดีที่สุดในยุคปัจจุบันซึ่ง VGG16 ถือเป็นโมเดลการเรียนรู้เชิงลึกที่ใช้สำหรับงานการจำแนกภาพเป็นหลักโดยสถาปัตยกรรมภายในประกอบไปด้วยเซลล์ประสาทเทียมทั้งหมด 16 ชั้น โดยในแต่ละชั้นจะทำงานการประมวลผลภาพข้อมูลอินพุตแบบค่อยเป็นค่อยไปและทำการเพิ่มประสิทธิภาพความแม่นยำในแต่ละชั้นไปด้วยซึ่ง VGG16 ได้ใช้ชั้น Convolutional แบบ 3"×" 3 และทำการ Stride สเต็ปละ1 และใช้ maxpool เป็น 2"×" 2 และทำ Stride สเต็ปละ 2 และจะถูกเชื่อมต่อด้วยชั้น Fully Connected 2 ชั้น และชั้นเอาต์พุตสุดท้ายเป็น Softmax



รูปที่ 2-4 สถาปัตยกรรมโมเดล VGG16

ที่มา

https://www.researchgate.net/publication/328966158/figure/fig2/AS:693278764720129@1542301946576/An-overview-of-the-VGG-16-model-architecture-this-model-uses-simple-convolutional-blocks.png



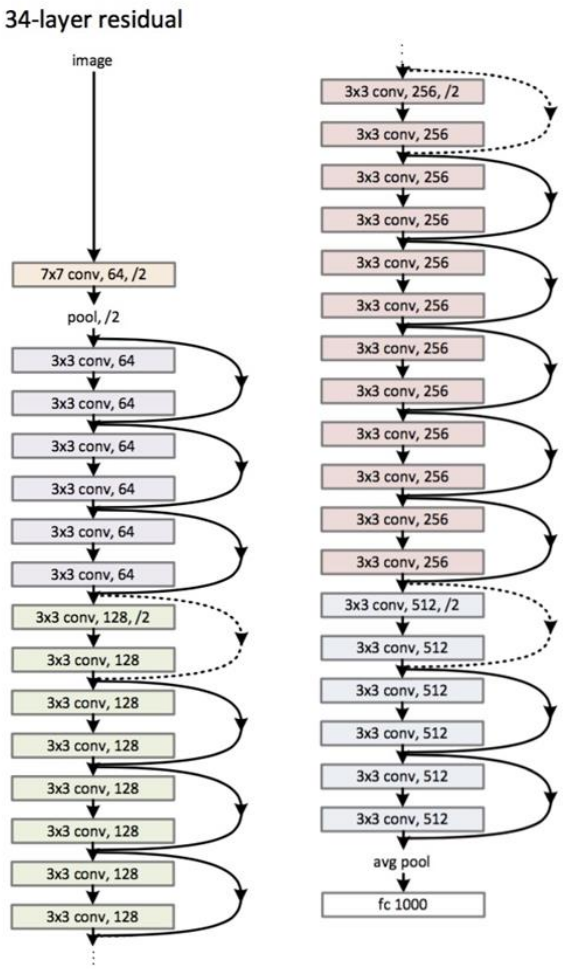
รูปที่ 2-5 การทำ Convolutional ของ VGG16

ที่มา

https://miro.medium.com/v2/resize:fit:828/format:webp/0*6VP81rFoLWp10FcG

2.1.2 Resnet50

Resnet50[3] คือ Convolutional Neural Network แบบ 50 ชั้น คือ ประกอบด้วยชั้น Convolutional 48 ชั้น และ ชั้น Maxpool 1 ชั้น รวมถึงชั้น average pool อีก 1 ชั้น รวมเป็น 50 ชั้น โดย Resdual Neural Network คือ โครงค่ายประสาทเทียมที่มีรูปแบบเครือข่ายเป็นการเรียงซ้อนบล็อกโดย Resnet50 จะประกอบด้วยการทำ Convolutional 7"x" 7 และทำ Stride สเต็ปละ 2 ร่วมถึง มีการทำ max pooling Stride สเต็ปละ 2 ร่วมถึงชั้นอื่นๆอีก รวมแล้วทั้งหมดจะมี 50 ชั้น

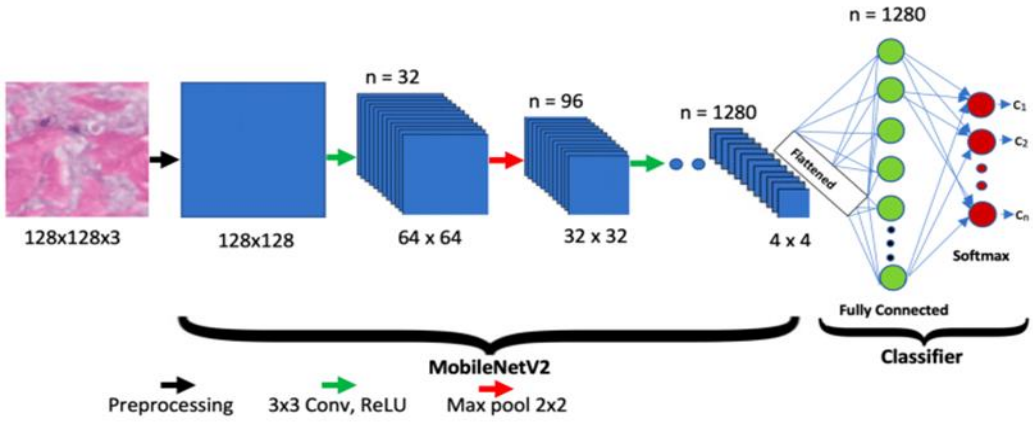


รูปที่ 2-6 สถาปัตยกรรม Resnet50

ที่มา https://i.stack.imgur.com/XTo6Q.png

2.1.3 MobileNetV2

MobileNetV2 [4] คือสถาปัตยกรรม Convolutional Neural Network(CNN) เป็นสถาปัตยกรรมที่ถูกออกแบบมาสำหรับแอพพลิเคชั่นการมองเห็นแบบมือถือหรือโทรศัพท์เคลื่อนที่ ซึ่งถูกพัฒนาโดยนักวิจัยของ Google ที่พัฒนา MobileNetV2 ให้ดีกว่ารุ่นก่อนหน้าซึ่งความพิเศษของ MobileNetV2 คือความสมดุลระหว่างขนาดของโมเดลและความแม่นยำทำให้มันเหมาะกับอุปกรณ์พกพาหรืออุปกรณ์ที่มีขนาดเล็กซึ่งในตัวของ MobileNetV2 มีองค์ประกอบที่สำคัญคือ 1. Depthwise Separable Convolution 2. Inverted Residuals 3. BottleNeck 4.Linear Bottlenecks 5.Squeeze-Excitation โดยองค์ประกอบเหล่านี้ช่วยลดความซับซ้อนการคำนวณของแบบจำลองทำให้อุปกรณ์ไม่จำเป็นต้องใช้ฮาร์ดแวร์ที่มีประสิทธิภาพสูงแต่ยังสามารถรักษาการคำนวณที่มีประสิทธิภาพไว้



รูปที่ 2-7 สถาปัตยกรรม MobileNetV2

ที่มา

https://www.researchgate.net/publication/350152088/figure/fig1/AS:1002717703045121@1616077938892/The-proposed-MobileNetV2-network-architecture.png

2.2 โครงข่ายประสาทเทียมแบบเกิดซ้ำ (Recurrent Neural Network)

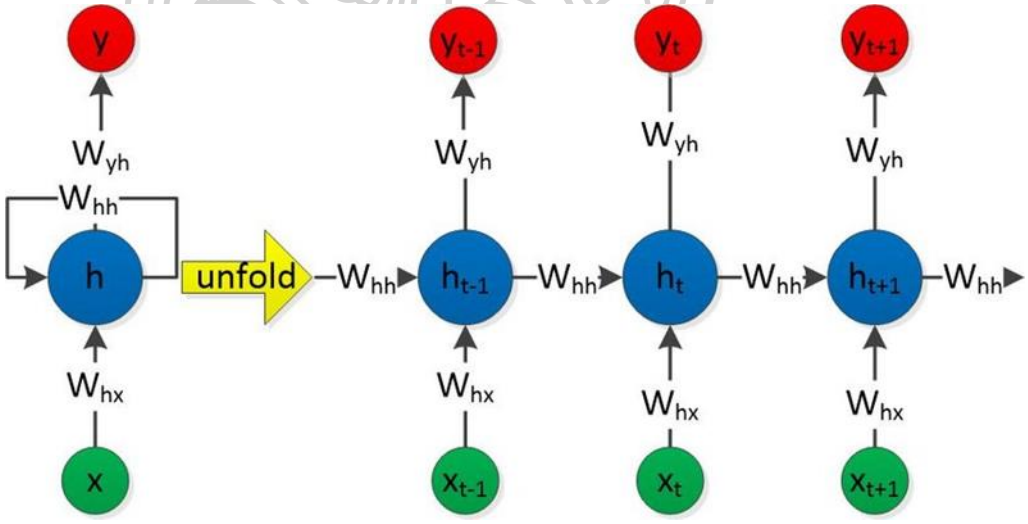
เป็นการทำงานเก็บค่าสถานะก่อนหน้าแล้วนำไปทำการวิเคราะห์ในส่วนถัดไปซึ่งมักนิยมใช้กับข้อมูลประเภทข้อความที่อาจเกี่ยวข้องกับการแปลภาษา การพยากรณ์อากาศ การวิเคราะห์หรือทำนายหุ้น ซึ่งหลักการของ RNN[1] จะอ่านข้อมูลทีละขั้นแล้วไว้เป็นสถานะ(state)หรือ สถานะซ่อน (hidden state) แล้วอาจทำการทำนายในขั้นตอนถัดไปโดยมีสมการในการคำนวณ State ดังนี้

$$h_t = f_h(w_i \cdot x_t + w_r \cdot h_{t-1} + b_n) \quad (1)$$

ค่า h_t คือค่า State หรือ Hidden State ณ ปัจจุบัน , f_h คือฟังก์ชันในการวิเคราะห์หรือประมวลผลซึ่งมีการเลือกใช้ Activation ให้เหมาะสมกับงาน , w_i คือค่า weight , x_t คือข้อมูลอินพุต(Input Data) , w_r คือค่า weight ของ hidden state , h_{t-1} คือ state ก่อนหน้านี้ , b_n คือ bias และการคำนวณค่า Output มีสมการดังนี้

$$y_t = f_y(w_y \cdot h_t + b_y) \quad (2)$$

ค่า y_tคือค่า Output , f_y คือฟังก์ชัน Activation โดยเลือกให้เหมาะสมกับงาน , w_y คือค่า weight , h_tคือ state , b_y คือ bias



รูปที่ 2-8 การวนรอบทำซ้ำของ RNN

ที่มา https://www.researchgate.net/profile/Jian-Zheng-17/publication/312593525/figure/fig3/AS:667699034210310@1536203263467/The-architecture-of-RNN.jpg

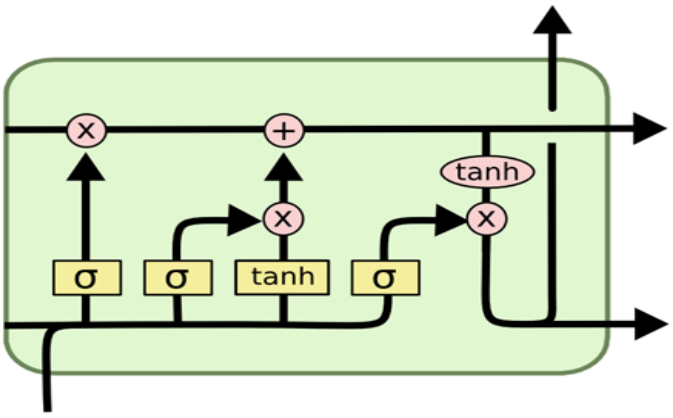
2.3 หน่วยความจำแบบสั่น-ยาว (Long Short-Term Memory)

หน่วยความจำแบบสั่น-ยาว (Long Short-Term Memory)[1] มีโครงสร้างหรือสถาปัตยกรรมคล้ายกับ RNN โดยพัฒนาขึ้นเพื่อมาแก้ปัญหาในกระบวนการของ RNN เนื่องจากRNN ไม่สามารถจัดการกับข้อมูลที่มีความยาวมากเกินไปได้ดีมากนักเพราะ RNN จะลืมข้อมูลในส่วนอันดับแรกๆที่เข้ามาก่อนหน้า ดังนั้น LSTM จึงมีหน้าที่ในการปรับปรุงข้อเสียของ RNN ด้วยวิธีการคือ LSTM จะมีสถานะของเซลล์(Cell state) ที่สามารถเลือกได้ว่าจะเก็บหรือจะลืมข้อมูลก่อนหน้าได้โดย ส่วนประกอบการทำงานที่สำคัญของ LSTM มีดังนี้

1. Forget Gate ทำการเลือกว่าจะเก็บข้อมูลหรือลืมข้อมูลของ state ก่อนหน้าแล้วนำไปวิเคราะห์ในกระบวนการถัดไปหรือไม่โดยส่วนใหญ่แล้วหากข้อมูลไหนที่เห็นว่าไม่สำคัญแล้วจะทำการลืมค่านั้นไป

2. Input Gate การที่จะทำกระบวนการนี้จะต้องผ่านกระบวนการ Forget Gate มาแล้วเพราะขั้นตอนในส่วนนี้จะทำการเพิ่มข้อมูลที่สำคัญเข้าไป

3. Output Gate คือตัวประเมินข้อมูลสุดท้ายที่สามารถเลือกได้ว่า Cell state ไปประมวลผลต่อไปซึ่งมักนิยมใช้ในการคาดเดารูปแบบของประโยคว่าประโยคต่อไปควรเป็นประโยคอะไร



รูปที่ 2-9 โครงสร้างสถาปัตยกรรม LSTM

ที่มา https://miro.medium.com/max/1400/0*D23ahAAuVce22goJ.png

2.4 Bidirectional LSTM (BiLSTM)

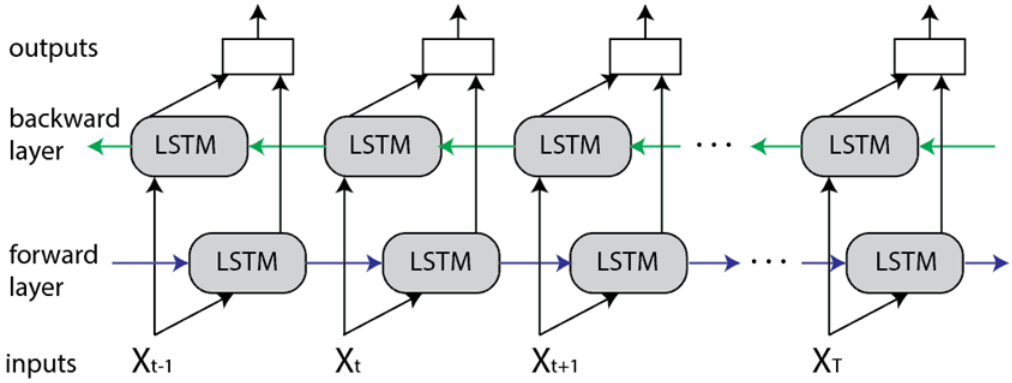
Bidirectional LSTM [5],[6] หรือ LSTM ที่ทำงานแบบสองทิศทางคือโมเดลแบบเป็นลำดับที่ประกอบด้วย LSTM 2 ชั้น ชั้นที่หนึ่งใช้สำหรับการประมวลผลข้อมูลอินพุตแบบไปข้างหน้า Forward Direction และอีกชั้นสำหรับการประมวลผลทิศทางย้อนกลับ Backward Direction ซึ่งกระบวนการทำงานไปกลับแบบสองทิศทางทำให้โมเดลเข้าใจความสัมพันธ์ระหว่างลำดับได้ดียิ่งขึ้น ดังเช่นตัวอย่างว่า “แกะตัวหนึ่งกำลังวิ่งอยู่ในสนามหญ้า” กับ “เด็กผู้หญิงกำลังแกะห่อของขวัญ” โดยให้สังเกตคำว่า แกะ ในประโยคทั้งสองที่เขียนเหมือนกันแต่ความหมายแตกต่างกัน ซึ่งความสัมพันธ์แบบนี้ขึ้นอยู่กับการเรียนรู้จากคำก่อนหน้ามา ซึ่ง Bidirectional LSTM ทำให้โมเดลเข้าใจความสัมพันธ์ดังที่ยกตัวอย่างไปได้ดีกว่า LSTM แบบปกติ ซึ่งสถาปัตยกรรมของ Bidirectional LSTM จะมี LSTM แบบทิศทางเดียวสองตัว ตัวหนึ่งทำการประมวลผลไปข้างหน้าและอีกตัวทำการประมวลผลย้อนกลับดังนั้นจะทำให้เกิดการประมวลผลแบบ 2 เครือข่ายโดยเครือข่ายแรกจะได้รับ token ตามที่เป็นอยู่ และอีกเครือข่ายได้รับคำสั่งย้อนกลับ ซึ่งเครือข่ายทั้งสองนี้จะคือค่า vector ความน่าจะเป็นออกมาและเอาต์พุตสุดท้ายจะรวมความน่าจะเป็นจากทั้งสองเครือข่ายมาแสดงผล

$$p_t = p_t^f + p_t^b \quad (3)$$

p_t คือ เวกเตอร์ความน่าจะเป็นอันสุดท้ายของเครือข่าย

p_t^f คือ เวกเตอร์ความน่าจะเป็นจากเครือข่ายการประมวลผลแบบทิศทางไปข้างหน้าของ *LSTM*

p_t^f คือ เวกเตอร์ความน่าจะเป็นจากเครือข่ายการประมวลผลแบบทิศทางย้อนกลับของ *LSTM*



รูปที่ 2-10 สถาปัตยกรรมของ Bidirectional LSTM

ที่มา https://www.baeldung.com/wp-content/uploads/sites/4/2022/01/bilstm-1.png

2.5 การตรวจจับวัตถุ (Object Detection)

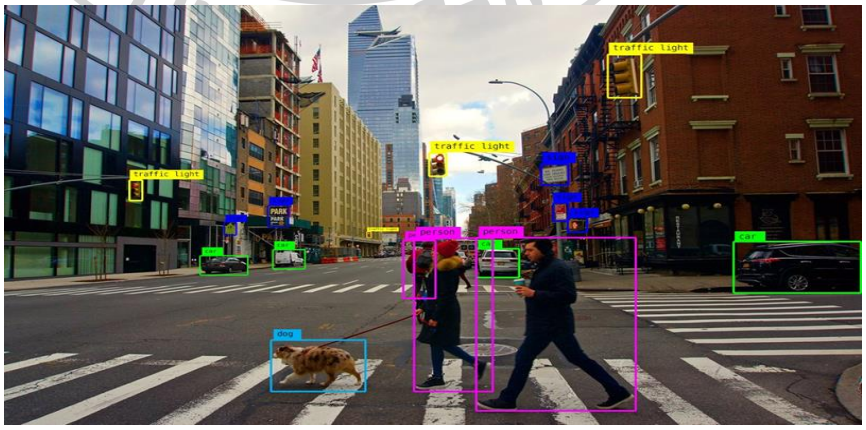
การตรวจจับวัตถุ (Object Detection)[1] เป็นงานที่สำคัญมากในงานด้านภาพเพราะเมื่อจะทำการประเมินภาพจะต้องตรวจสอบองค์ประกอบที่อยู่ในรูปภาพว่ามีกี่องค์ประกอบหรือมีวัตถุใดอยู่ในรูปภาพบางอย่างเช่น รถ หมา หมี รถไฟ เครื่องบิน สนามหญ้า และอื่นๆ และต้องทำการตีกรอบ (Bounding Box) ล้อมรอบวัตถุนั้นไว้หรือเรียกว่าต้องแบ่งแยกประเภทของวัตถุให้ชัดเจนนั้นเองโดยหลักการในการตรวจจับวัตถุจะมีหลักการที่สำคัญอยู่ 2 วิธีคือ

1. การเรียนรู้ของเครื่อง (Machine Learning) ทำการหาคุณลักษณะเด่นที่อยู่ในภาพโดยอาจจะหาจากขอบแนวดิ่ง แนวนอน แนวเอียง เป็นต้น

2. การเรียนรู้เชิงลึก (Deep Learning) จะใช้ CNN ในการเข้ามาทำการวิเคราะห์และยังแบ่งออกเป็นวิธีการย่อยอีก 2 อย่างคือ

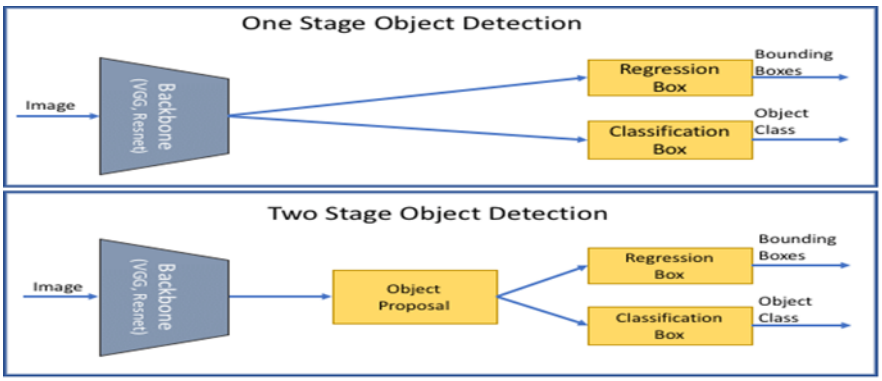
การตรวจจับ 2 ครั้ง (Two-stage detector) โดยจะมีองค์ประกอบหลักคือส่วนที่เรียกว่า Backbone ที่ใช้ CNN คัดแยกลักษณะเด่นในภาพหรือทำการแยกวัตถุที่สนใจ และการตรวจจับครั้งที่ 2 คือนำที่สิ่งที่คัดแยกมานั้นไปประมวลผลจำแนกออกมาว่าเป็นวัตถุใดแต่วิธีนี้ข้อเสียคืออาจใช้เวลาในการวิเคราะห์หรือประมวลผลนาน

การตรวจจับ 1 ครั้ง (Single-stage detector) โดยองค์ประกอบอย่าง Backbone จะทำการวิเคราะห์หรือประมวลผลเพียง 1 ครั้งเท่านั้นซึ่งจะทำการตีกรอบล้อมรอบวัตถุนั้นไว้เลยข้อดีของวิธีนี้คือมีความรวดเร็วแต่ข้อเสียคืออาจไม่แม่นยำมากนัก



รูปที่ 2-11 Object Detection

ที่มา https://learn.alwaysai.co/hubfs/object-dectection-4.jpg



รูปที่ 2-12 การตรวจจับ2ครั้ง และ การตรวจจับ 1ครั้ง

ที่มา

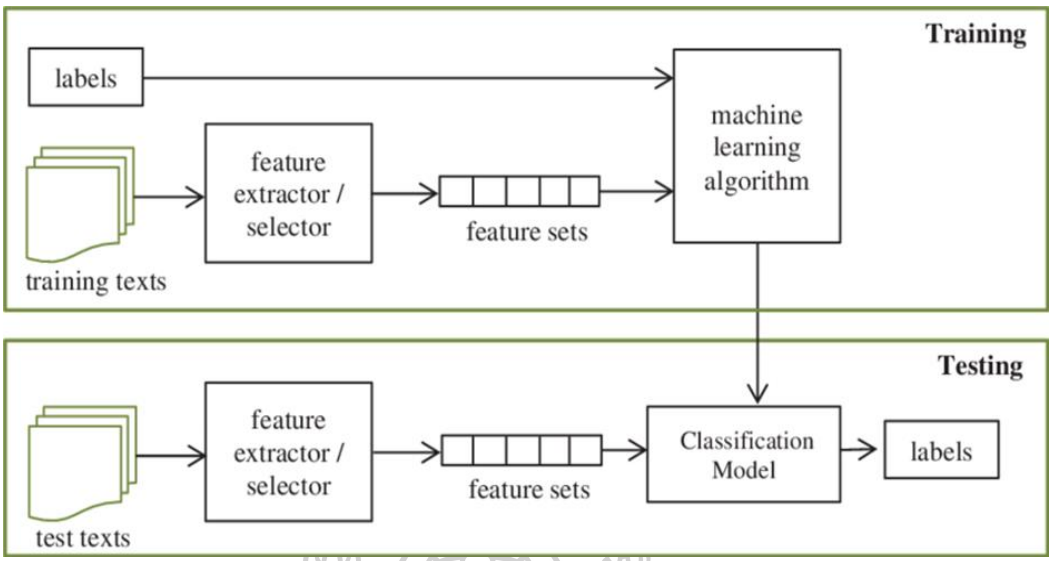
https://www.researchgate.net/publication/353284602/figure/fig3/AS:1046072046673927@1626414419841/Two-stage-vs-one-stage-object-detection-models.ppm

2.6 การประมวลผลภาษาธรรมชาติ(Natural Language Processing)

การประมวลผลภาษาธรรมชาติ(Natural Language Processing)[1] หรือ NLP เป็นการใช้งานในด้าน การแปลภาษาจากภาษาหรือแปลข้อความต่างๆที่เป็นลายลักษณ์อักษรโดยจะรับอินพุตเข้ามาเป็นข้อความแล้วทำการแปลงข้อความนั้นเป็นตัวเลขแล้วทำการวิเคราะห์ซึ่งหลักการทำงานจะประกอบไปด้วยส่วนที่สำคัญดังนี้

1. Text Classification คือการแยกประเภทของข้อความโดยทำการหาคุณลักษณะในข้อความนั้นหรือข้อความที่เป็นส่วนสำคัญแล้วนำมาทำการประมวลผลหรือเรียกว่าการฝึกสอน(Train) เพื่อที่ว่าพอมีข้อความชุดใหม่เข้ามาจะได้สามารถจำแนกประเภทของข้อมูลได้

2. การหาคุณลักษณะในข้อความ ต้องทำการแยกคุณลักษณะในข้อความที่เป็นตัวอักษรต้องแปลงข้อความที่เป็นตัวอักษรนั้นให้เป็นตัวเลขก่อน



รูปที่ 2-13 ไดอะแกรมของ Text Classification

ที่มา https://www.researchgate.net/profile/Mingyu-Wan/publication/335234564/figure/fig2/AS:807733112610820@1569589989596/A-diagram-of-machine-learning-for-automatic-text-classification.png

2.7 การประเมินคุณภาพของการเรียนรู้คำบรรยายเชิงลึกด้วยเมทริกซ์ (Assessing the quality of deep subtitle learning using matrices)

ในงานด้านการประมวลผลภาษาธรรมชาติหรือ NLP มีความสำคัญเป็นอย่างมากและเพื่อวัดประสิทธิภาพของแบบจำลอง NLP จึงต้องมีการนำตัวชี้วัดหรือเมทริกซ์มาประเมินซึ่งตัวชี้วัดที่นิยมใช้ได้แก่ [8] มีดังนี้

2.7.1 BLEU (Bilingual Evaluation Understudy)

ตัวชี้วัด BLEU คือหน่วยวัดที่นิยมใช้กันอย่างมากสำหรับงานที่เกี่ยวข้องทางด้าน Machine Translation ซึ่งตัวของ BLEU คือวิธีการประเมินคุณภาพของคำบรรยายที่ถูกสร้างขึ้นจากโมเดลเปรียบเทียบกับชุดคำบรรยายอ้างอิงที่ทำจากมนุษย์ โดย BLEU จะใช้ n-grams ที่เป็นลำดับคำของความต่อเนื่องซึ่งส่วนใหญ่ระดับคำสูงสุดต่อเนื่องที่นิยมใช้ในการวัดจะอยู่ที่ 4 คำต่อเนื่องซึ่งมีคำเรียกลำดับคำอีกอย่างว่า unigram, bigram, trigram และอื่นๆ ดังนั้น BLEU จะคำนวณความเที่ยงตรงของจำนวน n-grams ที่ใช้วัดรวมถึงยังมีการปรับความสั้นในคำบรรยายที่โมเดลสร้างนั้นสั้นกว่าคำบรรยายอ้างอิง ซึ่งช่วงคะแนนของ BLEU จะอยู่ที่ 0-1 ที่ 0 คือไม่มีคำบรรยายใดเลยที่โมเดลสร้างตรงกับคำบรรยายอ้างอิง ส่วน 1 คือคำบรรยายที่โมเดลสร้างตรงกับคำบรรยายอ้างอิงทั้งหมด

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad (4)$$

$$BP = \begin{cases} 1 & if\ c > r \\ e^{(1-\frac{r}{c})} & if\ c \leq r \end{cases} \quad (5)$$

2.7.2 ROUGE (Recall-Oriented Understudy for Gisting Evaluation) Score

ตัวชี้วัด ROUGE คือเมทริกซ์ที่ถูกออกแบบเพื่อการวัดข้อมูลคำบรรยายสรุปที่สร้างโดยโมเดลเทียบกับข้อมูลคำบรรยายจากมนุษย์ ซึ่ง ROUGE จะใช้ลำดับของคำ n-grams ที่ทับซ้อนกันซึ่งค่าคะแนนของ ROUGE จะคำนวณจากการเรียกคืน n-grams และช่วงคะแนนของ ROUGE จะอยู่ที่ 0-1 เช่นเดียวกับ BLEU ที่ 0 คือไม่มีคำบรรยายใดเลยที่โมเดลสร้างตรงกับคำบรรยายอ้างอิง ส่วน 1 คือคำบรรยายที่โมเดลสร้างตรงกับคำบรรยายอ้างอิงทั้งหมด แต่ ROUGE ยังถูกแบ่งออกเป็น ROUGE-N ที่ใช้วัดการทับซ้อนของ n-grams ROUGE-L วัดลำดับร่วมที่ยาวที่สุดของคำบรรยาย ROUGE-S วัดการทับซ้อนของคำที่ถูกละเว้นระหว่างคำบรรยายของโมเดลและคำบรรยายอ้างอิง

2.7.3 BERT Score

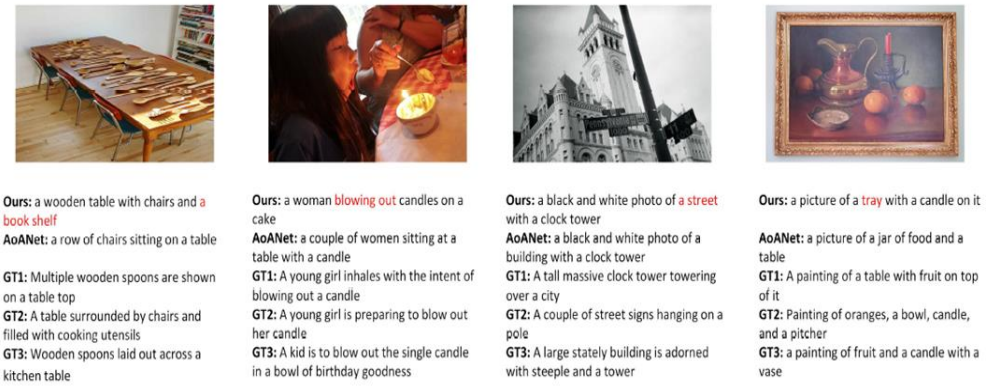
ตัวชี้วัด BERT คือเมทริกซ์สำหรับการประเมินโมเดลทางภาษาธรรมชาติซึ่งทำการประเมินโดยการวัดคำบรรยายที่โมเดลสร้างกับคำบรรยายอ้างอิงต้นฉบับว่ามีความคล้ายคลึงกันอย่างไร เหมือนกับ BLEU และ ROUGE ซึ่งสถาปัตยกรรมของ BERT หรือโครงสร้างการทำงานหลักจะประกอบไปด้วย 1.Contextual Embeddings 2.Cosine Similarity 3. Token Matching for Precision and Recall 4. Importance Weighting 5. Baseline Rescaling

2.8 งานวิจัยที่เกี่ยวข้อง

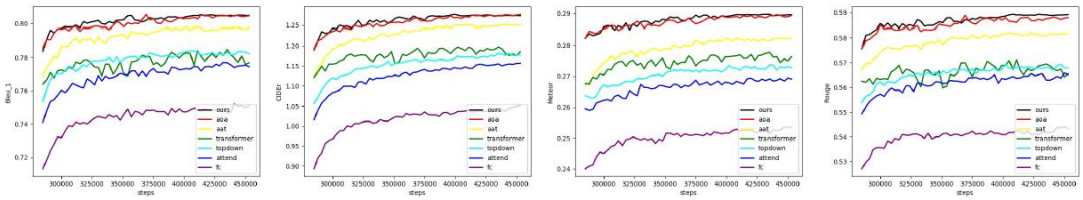
2.8.1 A Sparse Transformer-Based Approach for Image Captioning(CONGCONG ZHOU, ZHOU LEI, SHENGBO CHEN, YIYONG HUANG, XIANRUI LIU)[9]

ในบทความวิจัยนี้ได้นำเสนอโมเดลเชิงลึกการเข้ารหัสและถอดรหัสในรูปแบบใหม่ที่สามารถกระจายความสำคัญของรูปภาพได้ซึ่งการเข้ารหัสจะทำการแยกคุณลักษณะของรูปภาพในหลายระดับหรือเลือกการกระจายความสำคัญได้แต่การเข้ารหัสนี้จะต้องมีข้อมูลที่มากเพียงพอที่จะสามารถกระจายความสำคัญในรูปภาพได้ และในส่วนของการถอดรหัสจะเป็นส่วนที่ทำการปรับปรุงส่วนที่เกี่ยวข้องในรูปภาพที่อยู่ในรูปแบบแถวของMatrix ซึ่งการทำแบบนี้ในรูปภาพช่วยให้สามารถเลือก

ส่วนที่เกี่ยวข้องมากที่สุดได้และจะทำให้การสร้างคำบรรยายนั้นมีความแม่นยำมากขึ้น ซึ่งในงานด้าน Image Captioning การตรวจจับวัตุในภาพและความสัมพันธ์ในรูปภาพนั้นถือว่าเป็นสิ่งสำคัญซึ่งในกระบวนการทำงานของ Image Captioning รูปภาพอินพุตจะถูกทำการเข้ารหัสผ่านเครือข่าย CNN เพื่อทำการแยกคุณลักษณะในรูปภาพจากนั้นจะเป็นขั้นตอนของกระบวนการ RNN ที่เป็นส่วนของการถอดรหัสออกมาเป็นคำหลายๆคำ จากนั้นวิธีการของ LSTM จะทำการจัดเรียงตำแหน่งของคำที่ได้จากขั้นตอนก่อนหน้าให้สัมพันธ์กับรูปภาพอินพุต แต่ในบทความวิจัยนี้ได้ทำการศึกษาว่าตัวแปลงรุ่นก่อนมีการให้ความสำคัญกับองค์ประกอบในภาพทั้งหมดหรือมากเกินไปทำให้องค์ประกอบที่มีความสำคัญที่แท้จริงนั้นกลายเป็นมีความสำคัญเท่ากันกับองค์ประกอบอื่นทำให้อาจเกิดข้อผิดพลาดในการบรรยายที่คลาดเคลื่อนได้ดังนั้นจึงควรกระจายความสำคัญว่าสิ่งใดในภาพที่มีความสำคัญมากและสิ่งใดในภาพมีความสำคัญน้อยดังนั้นจึงมีการคิดค้น “Local Adaptive Threshold” เป็นโมเดลตัวใหม่ที่ช่วยในการให้ความสำคัญกับองค์ประกอบในภาพด้วยการกระจายความสำคัญไปยัง Matrix หรือเรียกว่า Attention Matrix ที่จะให้ความสนใจน้อยกับองค์ประกอบที่มีส่วนร่วมน้อย และ สนใจองค์ประกอบที่หลักหรือเข้มข้นในภาพได้เพื่อที่จะได้เน้นไปส่วนที่สำคัญมากที่สุดในภาพโดยได้ทำการทดลองบนชุดข้อมูล MSCOCO ที่ประกอบด้วยข้อมูลรูปภาพทั้งหมด 30,000 รูปมี 80 ประเภท และมี 5 คำบรรยายต่อ 1 รูปภาพร่วมถึงใช้ตัวชี้วัดในการประเมินประสิทธิภาพในการบรรยายซึ่งประกอบด้วย BLEU, METEOR, CIDEr และ Rouge โดย BLEU จะให้ค่าคะแนนที่ดีหากการบรรยายนั้นสั่น METEOR จะนำคำที่มีความหมายคล้ายหรือใกล้เคียงกันกับต้นกำเนิดประโยคมาปรับเทียบให้ดีขึ้น CIDEr จะคำนวณความเกิดบ่อยของคำ ROUGE วัดประสิทธิภาพของข้อความสุดท้าย



รูปที่ 2-14 ตัวอย่างคำบรรยายที่ถูกสร้างขึ้น

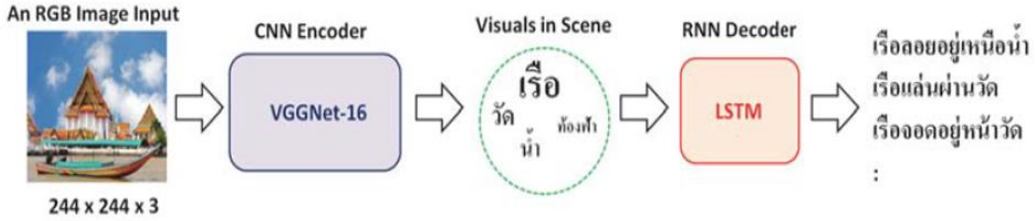


(a) Bleu (b) CIDEr (c) Meteor (d) Rouge

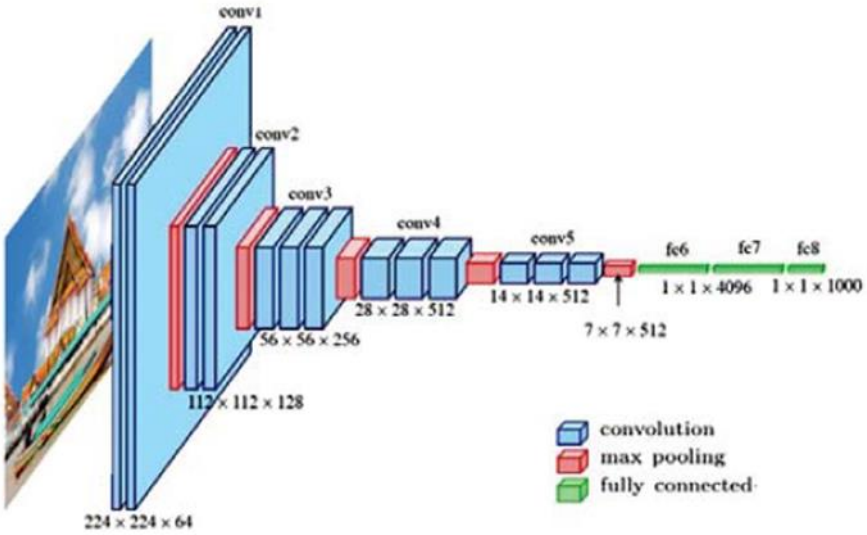
รูปที่ 2-15 ผลการประเมินการบรรยายรูปภาพที่แบ่งการตรวจสอบความถูกต้องบนชุดข้อมูล MSCOCO และเปรียบเทียบกับวิธีพื้นฐาน

2.8.2 Thai IC: Thai Image Captioning based on CNN-RNN Architecture(Pakpoom Mookdarsanit, Lawankorn Mookdarsanit)[10]

ในบทความวิจัยนี้ได้นำเสนอโมเดลการเรียนรู้เชิงลึกในการสร้างคำบรรยายภาษาไทยโดยใช้วิธีการต่างๆของการเรียนรู้เชิงลึก (deep learning) อย่างการเข้ารหัสด้วย CNN เพื่อแยกคุณลักษณะในรูปภาพโดยในที่นี้ใช้ VGGNet-16 เพื่อดูการจัดกลุ่มของวัตถุที่อยู่ในภาพแล้วทำการแยกชิ้นส่วนต่างๆในภาพโดยกระบวนการคือรับภาพอินพุตที่เป็นภาพสี RGB แล้วทำการลดสัดส่วนของภาพเป็น 224x224x3 หมายถึงภาพมีขนาดความกว้าง224pixelและยาว224pixelและมีสีองค์ประกอบเป็น แดง เขียว น้ำเงิน เมื่อเข้ากระบวนการของVGGNet-16จะต้องทำการลดสเกลตามนี้เพราะเป็นรูปแบบการกำหนดโมเดล รวมถึงมีการใช้ RNN เป็นตัวถอดรหัสหรือเรียกว่า RNN-decoder ที่มีหน่วยความจำสั่น-ยาวอยู่ภายใน (Long Short Term Memory) เพื่อทำการเรียบเรียงประโยคภาษาไทยให้ตรงตามความหมายของรูปภาพ ซึ่งการบรรยายออกมาเป็นภาษาไทยนั้นส่วนใหญ่จะทำกับฐานข้อมูลรูปภาพที่เกี่ยวข้องกับความเป็นไทยอย่างเช่น ประเพณีไทย แม่น้ำที่อยู่ในประเทศไทย แล้วสุดท้ายทำการประเมินด้วยตัวชี้วัด Bilingual Evaluation Understudy(BLEU) ซึ่งค่าสเกลของ BLEU จะมีค่าอยู่ในช่วง 0-1 หากแสดงค่าสูงสุดนั้นบ่งบอกถึงประสิทธิภาพดีที่สุด โดยการทดลองทั้งหมดนี้จะทำกับชุดข้อมูลรูปภาพ Flickr8k แล้วประกอบไปด้วยฐานข้อมูลข้อความที่เป็นภาษาอังกฤษที่มีอยู่แล้ว รวมถึงมีฐานข้อมูลคำบรรยายภาษาไทยที่สร้างขึ้นมาเองโดยเน้นไปที่วันสำคัญของไทย



รูปที่ 2-16 สถาปัตยกรรม CNN-RNN ในการสร้างคำบรรยายภาพ



รูปที่ 2-17 สถาปัตยกรรม VGGNet-16 ที่เป็นการเข้ารหัสCNN

รูปที่ 2-18 การบรรยายภาพเป็นภาษาไทยที่เป็นรูปแม่น้ำเจ้าพระยาที่เป็นสถานที่สำคัญในไทย

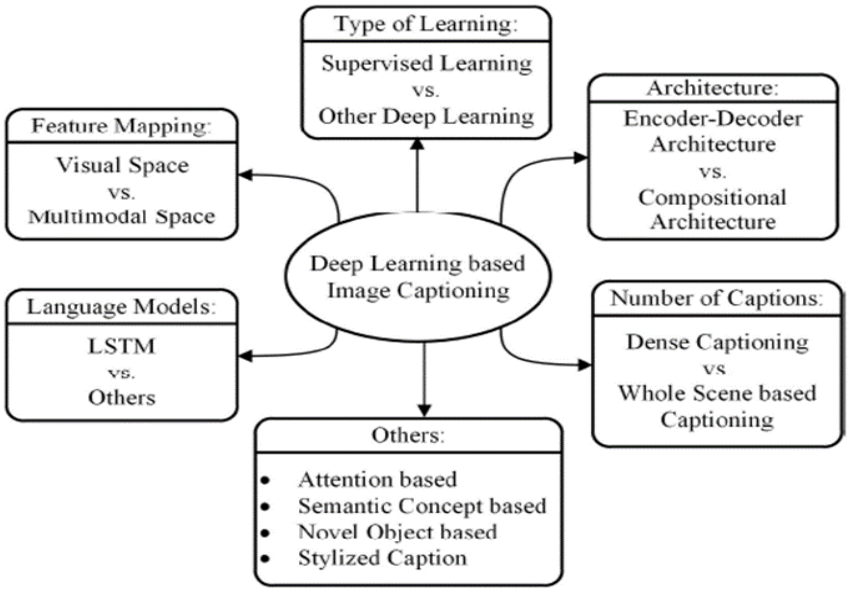
2.8.3 A Comprehensive Survey of Deep Learning for Image Captioning(MD. ZAKIR HOSSAIN, FERDOUS SOHEL, MOHD FAIRUZ SHIRATUDDIN, HAMID LAGA)[11]

ในบทความวิจัยนี้ได้นำเสนอการศึกษาที่ครอบคลุมเทคนิคพื้นฐานในการสร้างคำบรรยายภาพ(Image Captioning) ซึ่งได้ทำการศึกษาเกี่ยวกับการเรียงประโยคแบบอัตโนมัติคือหมายถึงลำดับการเกิดก่อนหน้าสิ่งที่อยู่ในภาพเพื่อการเรียงตามลำดับนัยสำคัญ เพราะหลักสำคัญระบบการสร้างคำบรรยายจะต้องเข้าใจองค์ประกอบในวัตถุแล้วพอถึงขั้นตอนการสร้างคำบรรยายจะสามารถบรรยายออกมาได้ตามหลักนัยสำคัญ ซึ่งการทำความเข้าใจในรูปภาพนั้นสามารถแบ่งเทคนิคได้เป็น 2 ประเภทหลักคือ

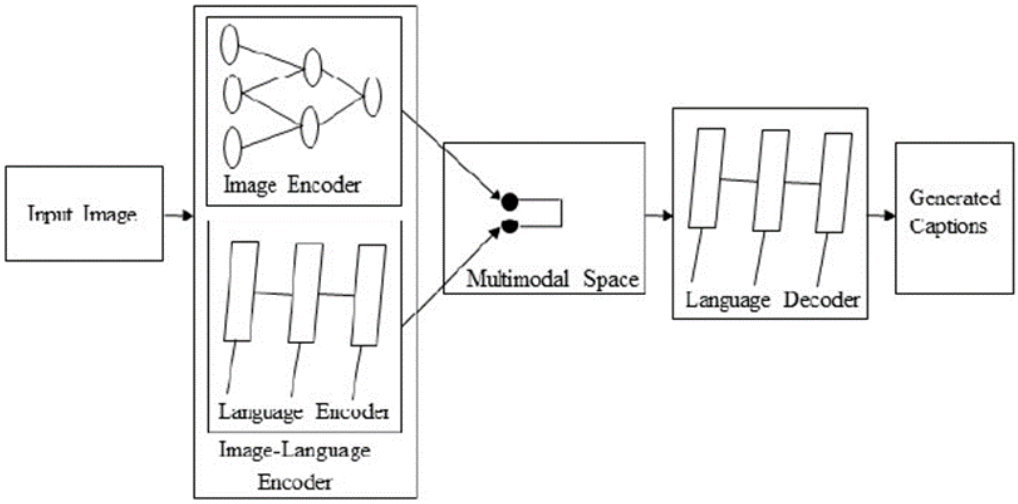
1. เทคนิคการเรียนรู้ของเครื่องแบบดั้งเดิม เป็นการแยกคุณลักษณะแบบดั้งเดิมกล่าวคือการแยกคุณลักษณะนั้นเป็นการแยกด้วยมือมนุษย์ซึ่งมีข้อเสียหากข้อมูลที่เข้ามามีขนาดใหญ่และจำนวนมากจะทำให้เกิดความล่าช้า

2. เทคนิคการเรียนรู้ด้วยเครื่องเชิงลึก คุณลักษณะในภาพจะถูกแยกและถูกฝึกสอน(Train) โดยอัตโนมัติและสามารถจัดการกับชุดข้อมูลที่ขนาดใหญ่และมีจำนวนมากได้

และเทคนิคสำคัญที่งานวิจัยนี้ได้ทำการศึกษาคือการสร้างคำบรรยายภาพแบบใหม่หรือ Novel Caption ที่หลักการสำคัญคือการสร้างคำบรรยายจากพื้นที่การมองเห็นและพื้นที่แบบหลายมิติ



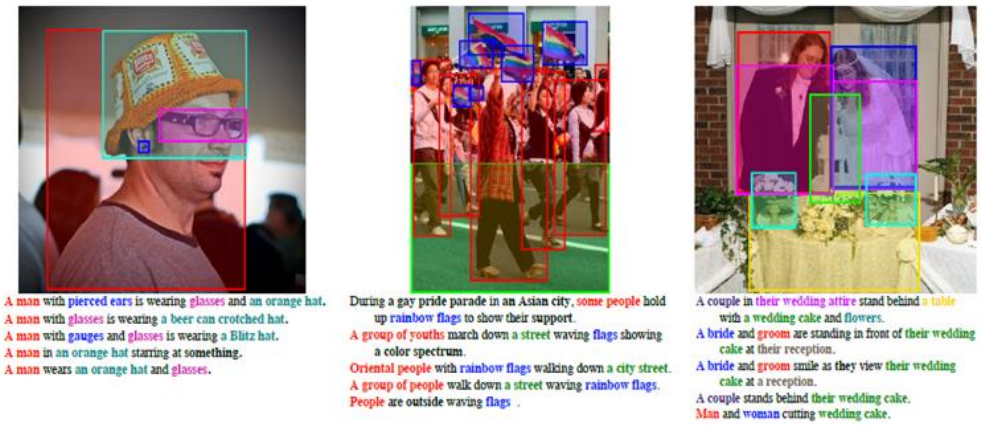
รูปที่ 2-19 ภาพรวมการศึกษาความครอบคลุมการเรียนรู้เชิงลึกในการสร้างคำบรรยายภาพ



รูปที่ 2-20 ไดอะแกรมของคำบรรยายภาพพื้นที่หลายมิติ

2.8.4 Flickr30k Entities: Collecting RegionRegion-to-Phrase Correspondences for Richer Image-to-Sentence Models(Bryan A. Plummer , Liwei Wang , Chris M. Cervantes , Juan C. Caicedo _ Julia Hockenmaier , Svetlana Lazebnik)[12]

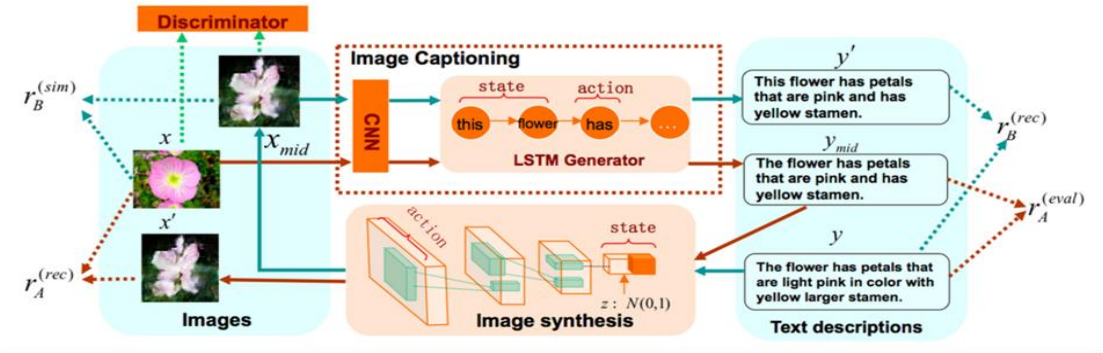
ในบทความวิจัยนี้เป็นการบรรยายเกี่ยวกับชุดข้อมูลที่นิยมนำมาใช้ใน Image Captioning นั้นก็คือชุดข้อมูล Flickr30k ที่เป็นมาตรฐานสำหรับการบรรยายภาพที่เป็นประโยคโดยในงานวิจัยนี้มีการเพิ่มเสริมคำบรรยายเข้าไปอีก 158k คำบรรยายโดยที่จากเดิมชุดข้อมูล Flickr30k มีคำบรรยายทั้งหมด 244k คำบรรยาย โดยจุดประสงค์ของงานวิจัยฉบับนี้คือการให้ชุดข้อมูลที่ครอบคลุมและมีขนาดใหญ่ที่ครอบคลุมความหมายในการบรรยายภาพจึงได้มีการเพิ่มคำบรรยายเข้าไปอีกทั้งในงานวิจัยนี้ต้องการสร้างวลีหรือประโยคที่เป็นมาตรฐานแบบใหม่ที่มีพื้นฐานมาจาก bounding box เพราะเป้าหมายในการสร้างประโยคคือการคาดเดาคำที่ได้มาจาก bounding box ซึ่ง bounding box เป็นการใช้หลักการในการตรวจจับวัตถุแบบดั้งเดิมคือจะดูรายการที่กำหนดไว้ล่วงหน้าของความหมายรูปภาพหรือองค์ประกอบที่ชัดเจน



รูปที่ 2-21 ตัวอย่างคำบรรยายจากชุดข้อมูล

2.8.5 Multitask Learning for Cross-domain Image Captioning(Min Yang, Wei Zhao, Wei Xu, Yabing Feng, Zhou Zhao, Xiaojun Chen, Kai Lei)[13]

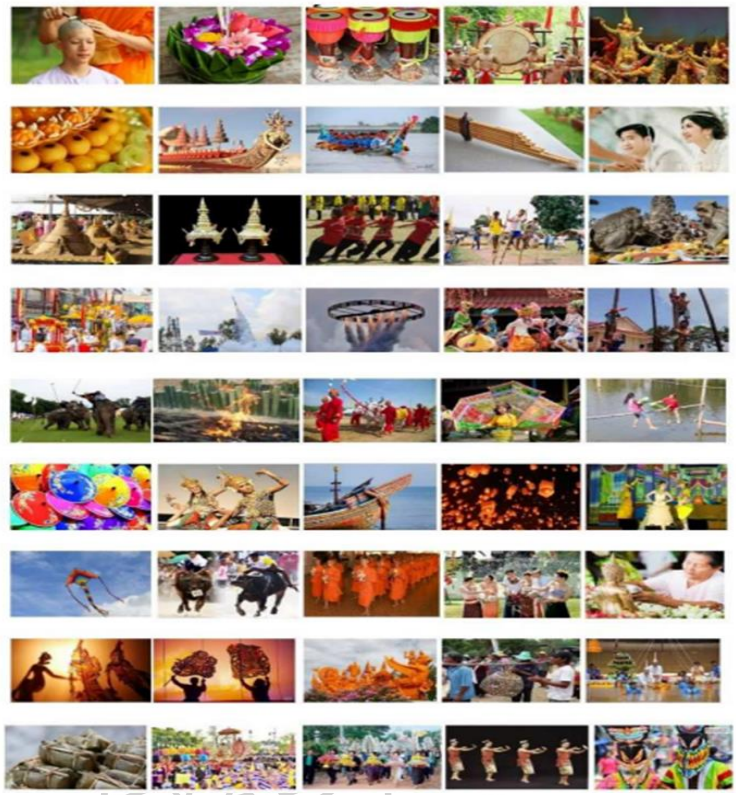
ในบทความวิจัยนี้ได้บรรยายถึงการศึกษาทางด้านปัญญาประดิษฐ์ล่าสุดที่เกี่ยวกับการสร้างคำบรรยายภาพแบบอัตโนมัติที่เรียกว่า “Image Captioning” แต่ในการสร้างคำบรรยายภาพนั้นจะต้องอาศัยข้อมูลที่มีจำนวนมากและใช้เวลานานซึ่งเป็นอุปสรรคในการสร้างคำบรรยายภาพดังนั้นในงานวิจัยนี้จึงได้นำเสนอ “MLADIC” เป็นอัลกอรีทึมแบบใหม่ที่ชื่อว่า “Multitask Learning Algorithm for cross-Domain Image Captioning” เพราะฉะนั้น MLADIC ก็คือระบบ Multitask ที่ทำการปรับวัตถุคู่กันสองสิ่งในเวลาเดียวกันผ่านการเรียนรู้ของเครื่องแบบคู่ขนานซึ่งแต่อย่างไรก็ตามงานด้าน Image Captioning จะถูกฝึกสอนด้วยโมเดลการเข้ารหัสและถอดรหัสอย่างเช่น CNN , LSTM เพื่อสร้างการบรรยายภาพจากภาพ input ที่เข้ามา และในส่วนสุดท้ายได้ทำการทดลองประสิทธิภาพของ MLADIC โดยใช้ MSCOCO เป็นข้อมูล domain ต้นทาง และใช้ Flickr30k และ Oxford-102 เป็นข้อมูล domain ปลายทางซึ่งผลลัพธ์ MLADIC ได้ประสบความสำเร็จเป็นอย่างมากและมีประสิทธิภาพที่ดีกว่าคู่แข่งในงานด้าน Image Captioning



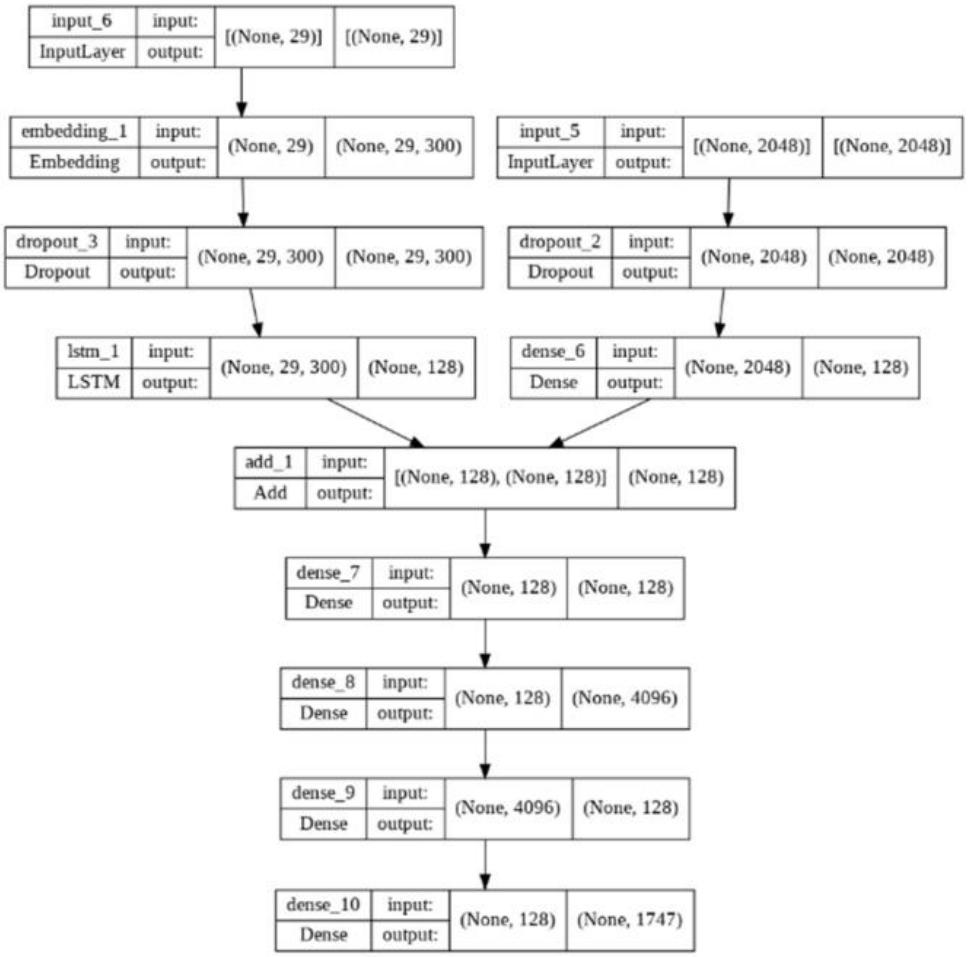
รูปที่ 2-22 กระบวนการทำงานข้อมูลกลไกการเรียนรู้แบบคู่

2.8.6 Image Captioning for Thai Cultures(Sarin Watcharabutsarakham, Sanparith Marukatat, Kantip Kiratiratanapruk, Pitchayagan Temniranrat)[14]

งานวิจัยนี้ได้ทำการบรรยายภาพภาษาไทยที่เกี่ยวข้องกับสถานที่ท่องเที่ยวโดยได้ทำฐานข้อมูลที่เกี่ยวกับสถานที่ท่องเที่ยวในประเทศไทย ซึ่งมีการนำ InceptionV3 มาใช้คัดแยกคุณลักษณะ 2048 คุณลักษณะจากรูปภาพและส่งคุณลักษณะเหล่านั้นเข้าสู่ LSTM ที่ถูกฝึกสอนด้วยภาษาไทย และในการทดลองมีการใช้ BLEU สำหรับประเมินการบรรยายของโมเดล โดยความยากหรือความท้าทายของงานวิจัยนี้คือรูปภาพของวัฒนธรรมไทยเนื่องจากมีความคล้ายคลึงกันมากในแต่ละรูป อย่างเช่นขบวนพาเหรดที่สามารถไปตรงกับวัฒนธรรมของประเทศอื่นๆได้ดังนั้นจึงมีการกำหนดข้อมูลประโยคบางประโยคและชื่อของวัฒนธรรมเพียง 1 class เท่านั้น และในส่วนของ BLEU ใช้วัดเปรียบเทียบค่าคะแนนระหว่างประโยคอ้างอิงและประโยคที่โมเดลสร้างออกมาซึ่งงานวิจัยนี้ได้ผลลัพธ์ที่แสดงถึงความแม่นยำด้วยเทคนิคการค้นหาแบบ Greedy Search ที่สูงกว่า BLEU โดย Greedy Search จะส่งคำที่มีความน่าจะเป็นสูงกลับออกมาทีละตำแหน่งซึ่งช่วยให้ output ที่ออกมามีความถูกต้องมาก



รูปที่ 2-23 ตัวอย่างรูปภาพวัฒนธรรมไทย



รูปที่ 2-24 สถาปัตยกรรมการฝึกสอนของโมเดล

Greedy	BLEU	
	n-gram-3	n-gram-10
55.23%	53.40%	43.12%

รูปที่ 2-25 ผลลัพธ์ของการทำนายที่ถูกเปรียบเทียบ

2.8.7 LEARNING DISCRIMINATIVE ACTION AND CONTEXT REPRESENTATIONS FOR ACTION RECOGNITION IN STILL IMAGES (Miao Xin, Hong Zhang, Ding Yuan, Mingui Sun)[15]

ในงานวิจัยนี้ได้อธิบายว่างานด้าน Computer นั้นเป็นงานที่ท้าทายเพราะยังมีปัญหาการรู้จำที่ผิดพลาดอย่างเช่นเมื่อมีรูปภาพ 2 รูปภาพที่มีบริบทร่วมกันหรือคล้ายคลึงกันมากซึ่งปัญหานี้ได้เกิดขึ้นมาเป็นเวลานาน ในงานนี้จึงมีการใช้การเรียนรู้แบบเมตริกซ์เพื่อระบุ class ภายใน โดยผู้แต่งของงานวิจัยฉบับนี้ได้นำเสนอฟังก์ชั่นการสูญเสียที่ชื่อว่า composite-triplet โดยเรียนรู้ฟังก์ชั่นที่คล้ายคลึงกันโดยตรงจากข้อมูล ซึ่งวิธีการจะถูกประเมินบนชุดข้อมูล PASCALVOC ซึ่งชุดข้อมูลนี้ประกอบไปด้วยรูปการกระทำ 10 รูปแบบ และคลาสอื่นๆ อีก 1 คลาส โดยวิธีการของงานวิจัยฉบับนี้มีประสิทธิภาพที่ดีกว่าวิธีที่นำมาเปรียบเทียบโดยประสบความสำเร็จด้วยค่า AP เฉลี่ย 90.6% ซึ่งมีการอธิบายฟังก์ชั่นการสูญเสีย composite-triplet ดังนี้

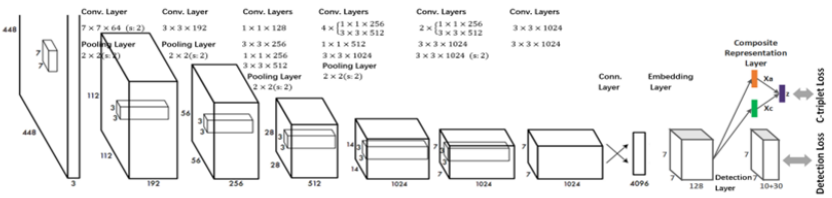
Case 1. $P(s|v1) \neq P(s|v2), P(c|v1) = P(c|v2)$

Case 2. $P(s|v1) = P(s|v2), P(c|v1) \neq P(c|v2)$

Case 3. $P(s|v1) \neq P(s|v2), P(c|v1) \neq P(c|v2)$

Case 4. $P(s|v1) = P(s|v2), P(c|v1) = P(c|v2)$

โดยที่ v1,v2 แสดงถึงรูปภาพการกระทำ 2 รูป คือ v1 และ v2 ตามลำดับ ต่อมา s และ c เป็นตัวแปรฝังของท่าทางบริบทของมนุษย์โดยทั้ง 4 กรณีนี้ครอบคลุมทุกสถานการณ์ในการเปรียบเทียบ และเพื่อให้ประสบความสำเร็จในการหาบริบทเหล่านี้จึงต้องหาขอบเขตการกระทำรวมถึงขอบเขตของภาพและองค์ประกอบที่ถูกแยกออกมาโดยใช้ AC-YOLO หรือชื่อเต็มคือ Action Context Yolo ที่ดำเนินงานพร้อมกับการตรวจจับขอบเขตบริบทโดยค่าสูญเสียการตรวจจับ AC-YOLO ถูกแยกออกเป็น 2 ค่าคือ compositetriplet loss และค่าสูญเสียการตรวจจับอื่น



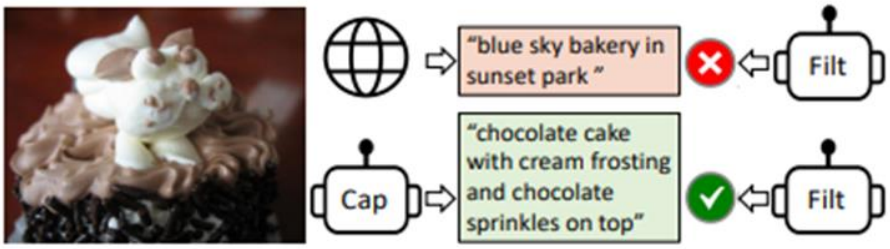
รูปที่ 2-26 สถาปัตยกรรมของ AC-YOLO

AP(%)	Jumping	Phoning	Playing Instrument	Reading	Riding Bike	Riding Horse	Running	Taking Photo	Using Computer	Walking	mAP
Maji [10]	59.3	32.4	45.4	27.5	84.5	88.3	77.2	31.2	47.4	58.2	55.1
Khosla [21]	69.1	75.7	44.8	66.6	44.4	93.2	94.2	87.6	38.4	70.6	75.6
Oquab [7]	74.8	46.0	75.6	45.3	93.5	95.0	86.5	49.3	66.7	69.5	70.2
Hoai [22]	82.3	52.9	84.3	53.6	95.6	96.1	89.7	60.4	76.0	72.9	76.3
Gkioxari [5]	84.7	67.8	91.0	66.6	96.6	97.2	90.2	76.0	83.4	71.6	82.6
Simonyan [6]	89.3	71.3	**94.7**	71.3	**97.1**	98.2	90.2	73.3	88.5	66.4	84.0
Gkioxari [1]	91.5	84.4	93.6	83.2	96.9	98.4	**93.8**	85.9	92.6	81.8	90.2
Ours	**91.8**	**84.7**	94.2	**83.6**	**97.1**	**98.5**	92.6	**87.2**	**94.1**	**82.2**	**90.6**

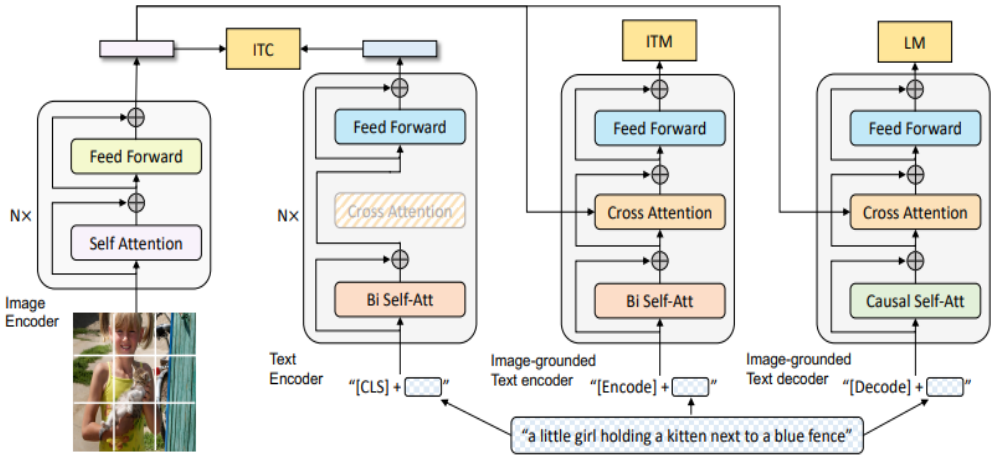
รูปที่ 2-27 การเปรียบเทียบผลลัพธ์กับวิธีการอื่น

2.8.8 BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation (Junnan Li, Dongxu Li, Caiming Xiong, Steven Hoi, Salesforce Research)[16]

ในงานวิจัยฉบับนี้ได้มีการนำเสนอ BLIP ซึ่งเป็นเฟรมเวิร์ก VLP รูปแบบใหม่ที่สามารถเรียนรู้คู่ของรูปภาพและข้อความที่มีสัญญาณรบกวน ซึ่ง BLIP ถือได้ว่ามีประสิทธิภาพที่ทันสมัยในงานด้าน downstream vision-language โดย BLIP จะฝึกสอนตัวเข้ารหัสและถอดรหัสแบบหลายรูปแบบล่วงหน้าโดยใช้ชุดข้อมูล bootstrapped จากรูปภาพและข้อความที่มีสัญญาณรบกวนจำนวนมาก และมีการคาดการณ์วิธีการที่จะสามารถเพิ่มประสิทธิภาพ BLIP ไว้ด้วยกัน 3 ข้อคือ 1. ทำการ Bootstrapping บนชุดข้อมูลที่ทำเป็นจำนวนหลายรอบ 2. สร้างคำบรรยายเทียมหลายรายการต่อรูปภาพเพื่อเพิ่มคลังข้อมูลก่อนการฝึกสอน 3. โมเดลที่ใช้ฝึกสอนคำบรรยายและตัวกรองทั้งหมดจะต้องถูกร่วมกันใน CapFilt ซึ่ง CapFilt คือการ bootstrapping ชุดข้อมูล ซึ่งผลสรุปแล้ว BLIP ได้ประสบความสำเร็จในการบรรยายภาพ การตอบคำถามด้วยภาพ การให้เหตุผลด้วยภาพ รวมไปถึงการโต้ตอบด้วยภาพ และได้ทำการเปรียบเทียบ BLIP กับ VLP



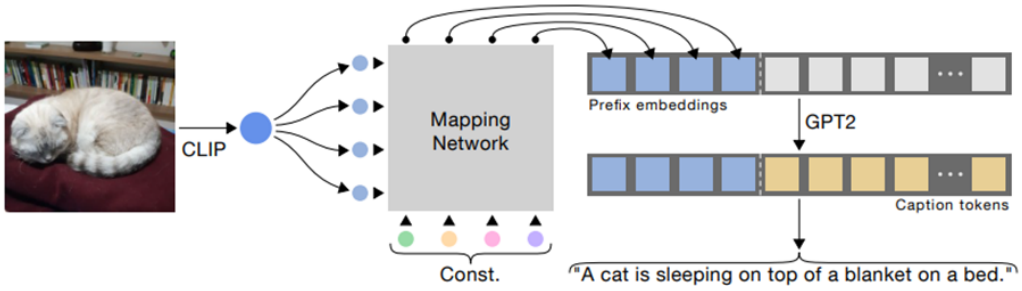
รูปที่ 2-28 การสร้างคำบรรยายเทียมและการใช้ตัวกรองเพื่อกรองคำบรรยายที่มีสัญญาณรบกวน



รูปที่ 2-29 สถาปัตยกรรมโมเดล pretraining และวัตถุประสงค์ของ BLIP

2.8.9 ClipCap: CLIP Prefix for Image Captioning (Ron Mokady, Amir Hertz, Amit H. Bermano)[17]

ในงานวิจัยนี้ได้มีการใช้ CLIP เป็นคำนำหน้าคำบรรยายภาพรวมถึงใช้วิธีการ mapping network และทำการปรับแต่งโมเดลทางภาษาเพื่อสร้างคำบรรยายภาพซึ่ง CLIP ถือว่าเป็นตัวเข้ารหัสรูปแบบหนึ่งโดยชื่อเต็มของ CLIP คือ Contrastive Language Image Pre-training ซึ่ง CLIP ถูกออกแบบมาเพื่อกำหนดการบรรยายร่วมกันทั้งรูปภาพและ text prompts โดยจะใช้การฝึกสอนเกี่ยวกับรูปภาพและคำบรรยายที่มีจำนวนมากที่ใช้ contrastive loss จึงทำให้ภาพที่แสดงออกกับข้อความที่ออกมามีความเชื่อมโยงกันเป็นอย่างมาก โดยวิธีการในการวิจัยนี้ได้ผลดีโดยการฝึกสอน mapping network ที่จะแปลการฝัง CLIP เป็นพื้นที่ GPT2 แต่ยังคง CLIP และโมเดลทางภาษาให้คงเดิมไว้อยู่ ส่งผลให้มีสถาปัตยกรรมที่ไม่ซับซ้อนเพราะมีจำนวนพารามิเตอร์น้อยรวมถึงสามารถฝึกสอนได้รวดเร็วและใช้เวลาไม่นาน



รูปที่ 2-30 ภาพรวมสถาปัตยกรรมโมเดลที่ฝึกสอน Mapping Network ขณะที่ยังคง CLIP และ GPT-2

(*A*) **Conceptual Captions**

Model	ROUGE-L ↑	CIDEr ↑	SPICE ↑	#Params (M) ↓	Training Time ↓
VLP	24.35	77.57	16.59	115	1200h (V100)
Ours; MLP + GPT2 tuning	**26.71**	**87.26**	**18.5**	156	80h (GTX1080)
Ours; Transformer	25.12	71.82	16.07	**43**	**72h** (GTX1080)

(*B*) **nocaps**

Model	in-domain		near-domain		out-of-domain		Overall		Params↓	Time↓
	CIDEr↑	SPICE ↑	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE		
BUTD [4]	74.3	11.5	56.9	10.3	30.1	8.1	54.3	10.1	52	960h
Oscar [19]	79.6	**12.3**	66.1	**11.5**	45.3	**9.7**	63.8	**11.2**	135	74h
Ours; MLP + GPT2 tuning	79.73	12.2	**67.69**	11.26	**49.35**	**9.7**	65.7	11.1	156	7h
Ours; Transformer	**84.85**	12.14	66.82	10.92	49.14	9.57	**65.83**	10.86	**43**	**6h**

(*C*) **COCO**

Model	B@4 ↑	METEOR ↑	CIDEr ↑	SPICE ↑	#Params (M) ↓	Training Time ↓
BUTD [4]	36.2	27.0	113.5	20.3	52	960h (M40)
VLP [47]	36.5	28.4	117.7	21.3	115	48h (V100)
Oscar [19]	**36.58**	**30.4**	**124.12**	**23.17**	135	74h (V100)
Ours; Transformer	33.53	27.45	113.08	21.05	**43**	**6h** (GTX1080)
Ours; MLP + GPT2 tuning	32.15	27.1	108.35	20.12	156	7h (GTX1080)

(*D*) **Ablation**

Ours; Transformer + GPT2 tuning	32.22	27.79	109.83	20.63	167	7h (GTX1080)
Ours; MLP	27.39	24.4	92.38	18.04	32	6h (GTX1080)

รูปที่ 2-31 เปรียบเทียบผลลัพธ์ของโมเดลกับวิธีการอื่น

2.8.10 Encoder-Decoder Model for Automatic Video Captioning Using Yolo Algorithm (Hanan Nasser Alkolouti, Dr. Mayada Ahmed AL Masre)[18]

จุดประสงค์ของงานวิจัยนี้คือพัฒนาการบรรยายวิดีโอแบบอัตโนมัติโดยใช้วิธีการเข้ารหัสและถอดรหัสตามการเรียนรู้เชิงลึกโดยมีขั้นตอนที่สำคัญสองอย่างคือ 1.ใช้โมเดลที่เรียกว่า KANTA ในการเลือกเฟรมสำคัญจากวิดีโอและลบส่วนที่ไม่สำคัญออกไป 2. เป็นการทำงานร่วมกันระหว่าง YOLO และ LSTM โดย YOLO หรือชื่อเต็มคือ You Only Look Once ถูกใช้ในการรับรู้องค์ประกอบที่อยู่ในเฟรมวิดีโอ และ LSTM หรือชื่อเต็มคือ Long Short-Term Memory ที่ใช้ในการสร้างคำบรรยายซึ่งได้ประยุกต์ใช้ YOLO บนชุดข้อมูล MSVD และในขั้นตอนสุดท้ายมีการประเมินโมเดลด้วยเมทริกซ์อย่าง METEOR

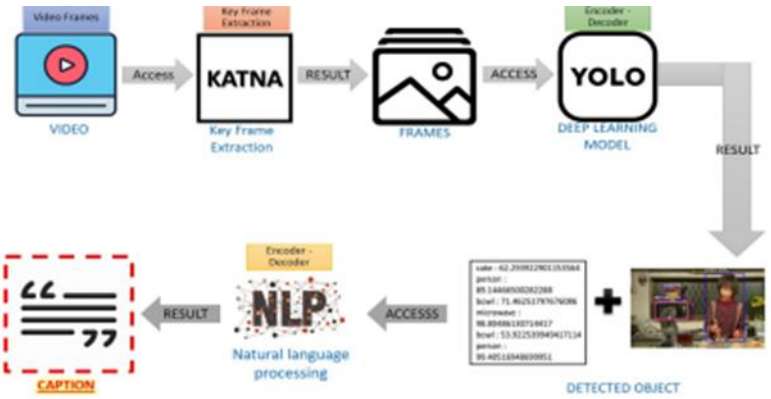


Figure 2: Architecture of Proposed Model.

รูปที่ 2-32 สถาปัตยกรรมภาพรวมของโมเดล

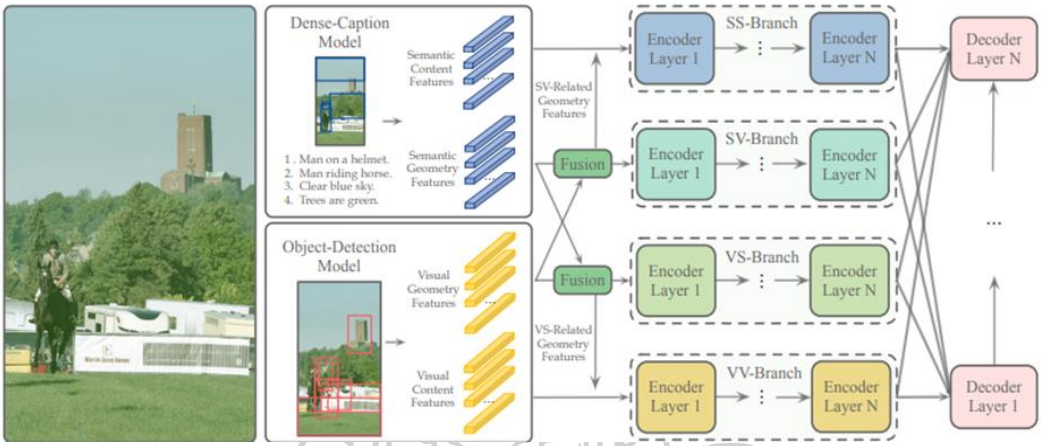
Paper	Percentage in METEOR on MSVD dataset
Less Is More: Picking Informative Frames for Video Captioning [7].	0.33
M3: Multimodal Memory Modelling for Video Captioning [6].	0.2658
STAT: Spatial-Temporal Attention Mechanism for Video Captioning [2].	0.33
Temporal Deformable Convolutional Encoder-Decoder Networks for Video Captioning [3].	0.308
Proposed model	0.35

รูปที่ 2-33 การเปรียบเทียบผลลัพธ์กับวิธีการอื่น

2.8.11 Geometry-Entangled Visual Semantic Transformer for Image Captioning (Ling Cheng, Wei Wei, Feida Zhu, Yong Liu, Member, Chunyan Miao, Senior Member)[19]

บทความวิจัยนี้ได้นำเสนอเครือข่ายระบบใหม่ที่ชื่อว่า Geometry-Entangled Visual Semantic Transformer(GEVST) และได้คิดหาวิธีในการเพิ่มประสิทธิภาพในการเชื่อมคำกริยากลางข้อมูลที่ส่งต่อความหมาย ซึ่งได้สร้างเครื่องเข้ารหัสการแปลแบบขนานสี่ตัว คือ VV(Pure Visual), VS(Semantic fused to Visual), SV(Visual fused to Semantic), SS(Pure Semantic) ในการสร้างคำบรรยายภาพสุดท้ายรวมถึงมีการใช้ประโยชน์จากความหมายที่มีคำบรรยายอยู่หลากหลาย

และมีจำนวนมากรวมไปถึงเนื้อหาที่เกี่ยวข้องกับเรขาคณิต ซึ่งการดำเนินงานจะทำบนชุดข้อมูล MSCOCO และสุดท้ายพบว่า GEVST ได้ประสบความสำเร็จและมีความสามารถแบบ real time ในการสร้างคำบรรยายภาพ



รูปที่ 2-34 ภาพรวมของโมเดล GEVST

Model	B-1		B-2		B-3		B-4		M		R		C	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
SCST CVPR2017	78.1	93.7	61.9	86.0	47.0	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.7
Stack-Cap AAAI2018	77.8	93.2	61.6	86.1	46.8	76.0	34.9	64.6	27.0	35.6	56.2	70.6	114.8	118.3
Up-Down CVPR2018	80.2	**95.2**	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
CVAP MM2018	80.1	94.9	64.7	88.8	50.0	79.7	37.9	69.0	28.1	37.0	**58.2**	73.1	121.6	123.8
HAN AAAI2019	80.4	94.5	63.8	87.7	48.8	78.0	36.5	66.8	27.4	36.1	57.3	71.9	115.2	118.2
SGAE CVPR2019	80.6	95.0	65.0	88.9	50.1	79.6	37.8	68.7	28.1	37.0	**58.2**	73.1	122.7	125.5
VSUA [52] MM2019	79.9	94.7	64.3	88.6	49.5	79.3	37.4	68.3	28.2	37.1	57.9	72.8	123.1	125.5
GEVST$_{ours}$	**80.8**	95.1	**65.1**	**89.2**	**50.4**	**80.5**	**38.2**	**70.1**	**28.7**	**37.9**	**58.2**	**73.3**	**125.1**	**127.8**

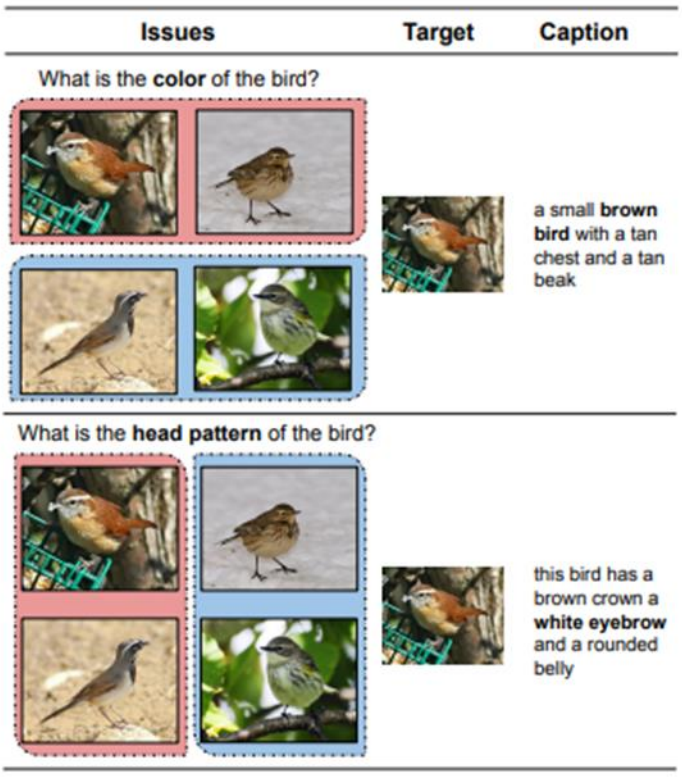
รูปที่ 2-35 รูปตารางแสดงการเปรียบเทียบ GEVST กับวิธีการอื่น

2.8.12 NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE(Dzmitry Bahdanau, KyungHyun Cho, Yoshua Bengio)[20]

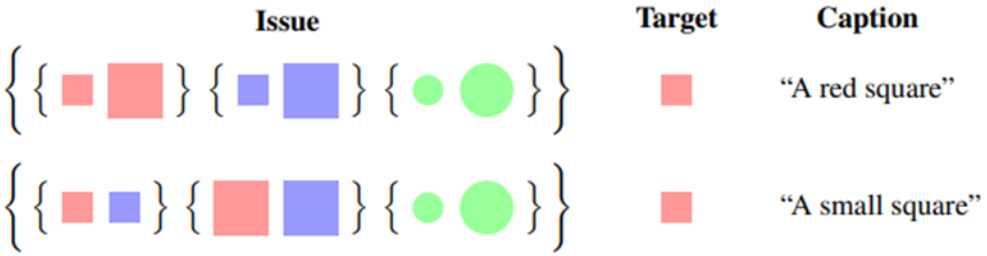
ในบทความวิจัยนี้ได้นำเสนอโมเดลจำลองที่ชื่อว่า RNNsearch ในการแปลจากภาษาอังกฤษไปเป็นฝรั่งเศสซึ่งในงานวิจัยนี้ได้พบปัญหาว่าการใช้ vector ที่มีความยาวคงที่จะเป็นคอขวดในการปรับปรุงประสิทธิภาพของการเข้ารหัสและถอดรหัส จึงได้ทำการหาวิธีขยายคอขวดโดยจะให้ตัวโมเดลค้นหาส่วนต่างๆของประโยคต้นทางหรือประโยคที่เป็นอินพุตก่อนแบบอัตโนมัติ และจากการทดลองพบว่า RNNsearch ได้ประสิทธิภาพที่ดีกว่าตัวเข้ารหัสและถอดรหัสแบบปกติอย่างเห็นได้ชัด

2.8.13 Pragmatic Issue-Sensitive Image Captioning (Allen Nie, Reuben Cohn-Gordon, Christopher Potts)[21]

ในบทความวิจัยนี้ได้กล่าวถึงปัญหาที่ว่าระบบการบรรยายภาพบางระบบยังประสบปัญหาที่ค่อนข้างจะมีความละเอียดอ่อนอย่างเช่นการบรรยายรูปภาพบางรูปยังไม่สามารถบรรยายได้ละเอียดมากพอ และเพื่อที่จะจัดการกับปัญหาจึงนำเสนอวิธีแก้ปัญหา ISIC(Issue-Sensitive Image Captioning) โดยการแก้ปัญหาจะเน้นไปที่การควบคุมข้อมูลให้มีความกระชับมากขึ้น ซึ่งใน ISIC อินพุตของคำบรรยายภาพทั้งหมดจะมีทั้ง ภาพและปัญหา แบบคู่กันซึ่งปัญหาคือชุดของรูปภาพที่ถูกแบ่งส่วนออกมาแบบเฉพาะเจาะจงเป็นอย่างมากว่ามีข้อมูลใดบ้างที่ต้องสนใจหรือเกี่ยวข้องเป็นพิเศษ โดยการทดลองของงานวิจัยนี้ได้ใช้ชุดข้อมูล Caltech-UC San Diego-Bird ซึ่งประกอบไปด้วยคำบรรยายที่มีเหตุผลที่ครอบคลุมทำให้สามารถศึกษาผลกระทบต่อโมเดลได้ครอบคลุมมากขึ้น



รูปที่ 2-36 ภาพการแบ่งกลุ่มของปัญหาที่ถูกแยกออกมาอย่างชัดเจน



รูปที่ 2-37 ตัวอย่างการกำหนดปัญหาของรูปภาพชุดเดียวกันแต่แบ่งแยกย่อยปัญหาออกมา

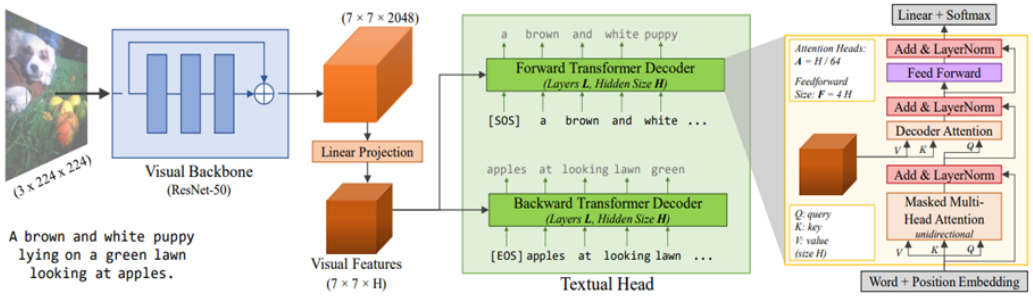
2.8.14 VinVL: Revisiting Visual Representations in Vision-Language Models (Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, Jianfeng Gao)[22]

ในบทความวิจัยฉบับนี้ได้นำเสนอการฝึกสอน OD หรือ โมเดลการตรวจจับวัตถุโดยเป็นการฝึกสอนแบบล่วงหน้า ซึ่งโมเดล OD จะแสดงภาพที่มีวัตถุเป็นศูนย์กลาง โดยในงานวิจัยนี้จะทำการปรับปรุง OD สำหรับงาน vision language กับชุดข้อมูลการตรวจจับวัตถุที่เปิดเผยแบบสาธารณะ 4 ชุด และทำการปรับแต่งโมเดลอย่างละเอียดบน Visual Genome ทำให้ตรวจจับวัตถุและคุณลักษณะได้ โดยผลลัพธ์แสดงให้เห็นว่าโมเดลสามารถพัฒนา SoTA ได้อย่างมากสำหรับงาน Vision Language

2.8.15 VirTex: Learning Visual Representations from Textual Annotations (Karan Desai, Justin Johnson)[23]

งานวิจัยฉบับนี้ได้นำเสนอ Virtex ที่ใช้สำหรับการฝึกสอนล่วงหน้าของคำบรรยายภาพจำนวนมากเพื่อเรียนรู้การแสดงและการบรรยายภาพ โดยทำการ convolutional ตั้งแต่เริ่มต้นจากนั้นจะส่งไปยังการจดจำแบบ downstream ซึ่ง Virtex สามารถให้ผลลัพธ์ที่ดีและแม่นยำกว่าที่เรียนรู้จากน้ำหนักข้อมูลจาก ImageNet ซึ่งในการทดลองได้ทำการฝึกสอน Virtex จาก scratch กับชุดข้อมูล COCO และทำการประเมินการเรียนรู้ด้วย visual backbone ที่ถูก freez เอาไว้บางส่วน อีกทั้งยังมีการตั้งสมมุติฐานก่อนที่จะทำการทดลองนั่นก็คือ จะต้องใช้คำบรรยายภาพที่มีความหมายหนักแน่นและชัดเจนซึ่งจะสามารถช่วยให้เรียนรู้คุณลักษณะการมองเห็นได้อย่างดีและใช้ข้อมูลสำหรับการฝึกสอนโมเดลน้อยลง อีกทั้งรายละเอียดที่สำคัญของโมเดล Virtex จะประกอบไปด้วย Visual Backbone, Textual Head, Model Size, Tokenization และ Training Details ซึ่งสิ่งสำคัญจะอยู่ที่ Visual Backbone ที่เป็นเครือข่าย Convolutional ที่ใช้ในการคัดแยกคุณลักษณะของรูปภาพ

กับ Textual Head ที่เป็นโมเดลภาษาที่จะคาดเดาคำบรรยายแบบสองทิศทางทั้งไปข้างหน้าและย้อนกลับตามลำดับ



รูปที่ 2-38 โมเดลการฝึกสอนล่วงหน้า Virtex

Method	Pretrain Images	COCO Instance Segmentation						LVIS Instance Segmentation			PASCAL VOC Detection			iNat 18
		AP^{bbox}_{all}	AP^{bbox}_{50}	AP^{bbox}_{75}	AP^{mask}_{all}	AP^{mask}_{50}	AP^{mask}_{75}	AP^{mask}_{all}	AP^{mask}_{50}	AP^{mask}_{75}	AP^{bbox}_{all}	AP^{bbox}_{50}	AP^{bbox}_{75}	Top-1
1) Random Init		36.7	56.7	40.0	33.7	53.8	35.9	17.4	27.8	18.4	33.8	60.2	33.1	61.4
2) IN-sup	1.28M	41.1	62.0	44.9	37.2	59.1	40.0	22.6	35.1	23.7	54.3	81.6	59.7	65.2
3) IN-sup-50%	640K	$40.3_{-0.8}$	$61.0_{-1.0}$	$44.0_{-0.9}$	$36.6_{-0.6}$	$58.0_{-1.1}$	$39.3_{-0.7}$	$21.2_{-1.4}$	$33.3_{-1.8}$	$22.3_{-1.4}$	$52.1_{-2.2}$	$80.4_{-1.2}$	$57.0_{-2.7}$	$63.2_{-2.0}$
4) IN-sup-10%	128K	$37.9_{-3.2}$	$58.2_{-3.8}$	$41.1_{-3.8}$	$34.7_{-2.5}$	$55.2_{-3.9}$	$37.1_{-2.9}$	$17.5_{-5.1}$	$28.0_{-7.1}$	$18.4_{-5.3}$	$42.6_{-11.7}$	$72.0_{-9.6}$	$43.8_{-15.9}$	$60.2_{-4.7}$
5) MoCo-IN	1.28M	$40.8_{-0.3}$	$61.6_{-0.4}$	$44.7_{-0.2}$	$36.9_{-0.3}$	$58.4_{-0.7}$	$39.7_{-0.3}$	$22.8_{+0.2}$	$35.4_{+0.3}$	$24.2_{+0.5}$	$56.1_{+1.8}$	$81.5_{-0.1}$	$62.4_{+0.7}$	$63.2_{-1.7}$
6) MoCo-COCO	118K	$38.5_{-0.6}$	$58.5_{-3.5}$	$42.0_{-2.9}$	$35.0_{-2.2}$	$55.6_{-3.5}$	$37.5_{-2.5}$	$20.7_{-1.9}$	$32.3_{-2.8}$	$21.9_{-1.8}$	$47.6_{-6.7}$	$75.4_{-6.2}$	$51.0_{-8.7}$	$60.5_{-4.4}$
7) VirTex	118K	$40.9_{-0.2}$	$61.7_{-0.3}$	$44.8_{-0.1}$	$36.9_{-0.3}$	$58.4_{-0.7}$	$39.7_{-0.3}$	$23.0_{+0.4}$	$35.4_{+0.4}$	$24.3_{+0.6}$	$55.3_{+1.0}$	$81.3_{-0.3}$	$61.0_{+1.3}$	$63.4_{-1.4}$

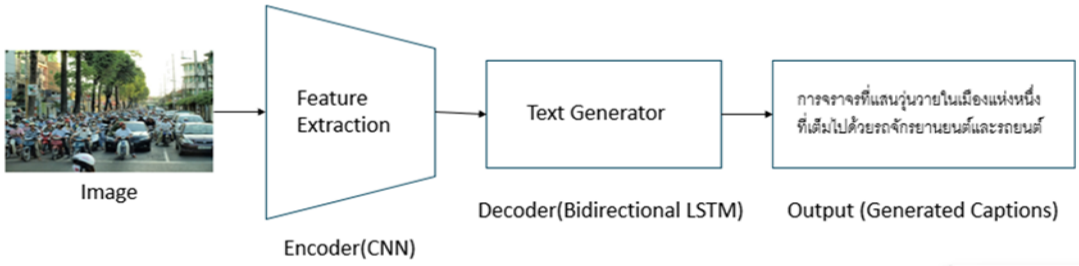
รูปที่ 2-39 เปรียบเทียบ Virtex กับงานอื่น

บทที่ 3
วิธีดำเนินการวิจัย

การดำเนินการวิจัยของวิทยานิพนธ์เล่มนี้ได้พัฒนาและต่อยอดจากโมเดลที่มีอยู่เพื่อเพิ่มประสิทธิภาพและทำการทดลองให้ได้ผลลัพธ์ที่เหมาะสมกับการบรรยายภาพภาษาไทย โดยได้มีการศึกษาถึงวิธีการสร้างคำบรรยายภาพทั้งจากวิธีที่บรรยายภาพเป็นภาษาอังกฤษและวิธีการบรรยายภาพที่เป็นภาษาไทยก่อนที่จะทำการทดลองและพัฒนาออกแบบโมเดล ซึ่งค้นพบว่ามีหลายวิธีในการพัฒนาโมเดลสร้างคำบรรยายภาพอย่าง การใช้โมเดล VGG16 เป็นที่นิยมใช้ในงานวิจัยหลายงานวิจัยเพื่อคัดแยกคุณลักษณะของภาพ รวมไปถึงการสร้างคำบรรยายภาพที่งานวิจัยส่วนใหญ่เลือกใช้ LSTM แบบปกติ ในวิทยานิพนธ์เล่มนี้จะมีการใช้ทั้ง LSTM แบบปกติและ LSTM ในรูปแบบประเภทอื่นอีกเพราะต้องการที่จะคิดค้นหาวิธีใหม่ในการพัฒนาจึงมีการทดลองด้วยการใช้ Bidirectional LSTM เข้ามาทำในการทดลองด้วย เนื่องจาก Bidirectional LSTM เป็น LSTM แบบ 2 ชั้น อีกทั้งในหนึ่งกระบวนการมันสามารถประมวลผลได้พร้อมกัน 2 ทิศทาง คือไปทิศทางไปข้างหน้าและทิศทางย้อนกลับและแต่ละชั้นการเรียนรู้สามารถเก็บความจำที่ซ่อนอยู่ของตัวเองได้ ถึง Bidirectional LSTM ในงานที่เกี่ยวข้องกับการบรรยายภาพจะมีส่วนน้อยมากที่นำมาใช้ในการสร้างคำบรรยายแต่ยังมีงานวิจัยอีกหลายงานวิจัยที่ได้กล่าวถึงประสิทธิภาพและข้อดีของ Bidirectional LSTM

ในบทนี้ได้กล่าวถึงขั้นตอนการดำเนินการวิจัย ร่วมถึงบรรยายเกี่ยงกับโครงสร้างของโมเดลในแต่ละขั้นตอนโดยลำดับแรกจะอธิบายถึงภาพรวมของโมเดล รวมถึงยังมีการอธิบายเกี่ยวกับขั้นตอนในการเตรียมข้อมูลก่อนการฝึกสอนโมเดล

3.1 ภาพรวมของวิธีการ



รูปที่ 3-1 ภาพรวมขั้นตอนการสร้างคำบรรยายภาพภาษาไทย

ในขั้นตอนแรกจะมีการรับอินพุตภาพเข้ามาแล้วทำการเข้ารหัสหรือเรียกขั้นตอนนี้ว่าการ Encoder คือกระบวนการคัดแยกคุณลักษณะหรือองค์ประกอบของรูปภาพด้วยวิธีการ CNN (Convolutional Neural Network) และในวิทยานิพนธ์ฉบับนี้ได้เลือกรูปแบบของ CNN ในการทำการทดลองเป็นทั้ง VGG16, Resnet50, Xception, MobilenetV2 และ InceptionV3 ในการทดลองเพื่อดูความแตกต่างของผลลัพธ์ในการประเมินการบรรยายภาพด้วย BLEU (Billingual Evaluation understudy) และสิ่งสำคัญของการสร้างคำบรรยายภาพภาษาไทยที่แตกต่างจากภาษาอังกฤษคือภาษาไทยจะเป็นการเขียนแบบตัวอักษรทุกตัวเรียงชิดติดกันเกือบทั้งหมดในประโยค ทำให้เมื่อนำเข้าโมเดลการฝึกสอนอาจนำมาซึ่งปัญหาได้เนื่องจากโมเดลการเรียนรู้จะต้องมีการเข้ารหัสคำเป็นคำทีละคำ ดังนั้นจึงต้องมีวิธีในการแยกประโยคเหล่านั้นให้ออกมาเป็นคำเดี่ยวและยังคงความหมายในคำนั้นไว้

3.2 การตัดคำหรือการแยกคำ

การตัดคำและแยกคำไทยคือสิ่งที่สำคัญมากก่อนนำข้อมูลเข้าสู่โมเดลการเรียนรู้ดังนั้นจึงมีการนำ library ที่ใช้ในการตัดคำภาษาไทยเข้ามาใช้โดยได้แสดงตัวอย่างดังรูปที่ 3.2

```
การจราจรที่แสนวุ่นวายในเมืองแห่งหนึ่งที่เต็มไปด้วยรถจักรยานยนต์และรถยนต์
['การ', 'จราจร', 'ที่', 'แสน', 'วุ่นวาย', 'ใน', 'เมือง', 'แห่ง', 'หนึ่ง', 'ที่', 'เต็ม', 'ไป', 'ด้วย', 'รถ', 'จักรย
าน', 'ยนต์', 'และ', 'รถ', 'ยนต์']
>>>
```

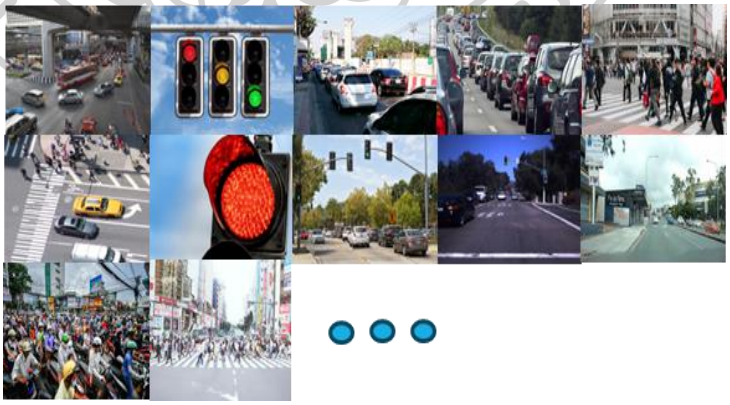
รูปที่ 3-2 ตัวอย่าง library ตัดคำไทย

จากรูปที่ 3-2 เห็นได้ว่าจากประโยค “การจราจรที่แสนวุ่นวายในเมืองแห่งหนึ่งที่เต็มไปด้วยรถจักรยานยนต์และรถยนต์” ถูกตัดออกมาเป็น ['การ', 'จราจร', 'ที่', 'แสน', 'วุ่นวาย', 'ใน', 'เมือง', 'แห่ง', 'หนึ่ง', 'ที่', 'เต็ม', 'ไป', 'ด้วย', 'รถ', 'จักรยาน', 'ยนต์', 'และ', 'รถ', 'ยนต์'] ซึ่งเห็นได้ว่าเมื่อประโยคถูกแยกคำออกมาจะทำให้มีความคล้ายคลึงกับการเขียนภาษาอังกฤษ และยังทำให้โมเดลสามารถที่จะแยกการเข้ารหัสคำทีละคำเพื่อการเรียนรู้ประเภทคำที่คล้ายกันหรือแตกต่างกันได้โดยในวิทยานิพนธ์เล่มนี้ได้โดยใช้ไลบรารี่การตัดคำที่ชื่อว่า Deepcut

3.3 การรวมชุดข้อมูลการจราจรกับ Flickr8k

เนื่องจากวิทยานิพนธ์ฉบับนี้ต้องการรวมทั้งชุดข้อมูลที่มีอยู่แล้วแบบสาธารณะ Flick8k ที่เป็นชุดข้อมูลรูปภาพและคำบรรยายที่เกี่ยวกับสถานการณ์ทั่วไปในชีวิตประจำวันอาทิเช่น “ผู้หญิงในเสื้อกั๊กสีน้ำเงินและหมวกกันน็อคสีฟ้ายืนอยู่กับจักรยานของเธอท่ามกลางการจราจร” กับชุดข้อมูลการจราจรที่ได้จัดทำขึ้นเองอาทิเช่น “รถจักรยานยนต์และรถยนต์จำนวนมากไม่สามารถเคลื่อนตัวได้เพราะติดไฟแดง” จากที่ได้กล่าวในบทนำที่หวังว่าในอนาคตจะถูกนำไปต่อยอดใช้เป็นการแจ้งเตือนให้แก่ผู้ขับขี่บนท้องถนนในรูปแบบการบรรยายออกมาเป็นเสียง อีกทั้งในชุดข้อมูล Flick8k ยังมีรูปภาพที่เกี่ยวกับรายละเอียดของการจราจรดังนั้นการเพิ่มฐานข้อมูลที่ได้จัดทำขึ้นเองยังเป็นการเพิ่มรายละเอียดองค์ประกอบให้กับกันและกันของชุดข้อมูล

รูปที่ 3-3 ภาพรวมชุดข้อมูล Flickr8k



รูปที่ 3-4 ภาพรวมชุดข้อมูลการจราจรที่จัดทำขึ้นเอง

3.4 การเตรียมชุดข้อมูล

3.4.1 ตัวอย่างชุดข้อมูล Flickr8k

รูปที่ 3-5 ตัวอย่างรูปภาพจากชุดข้อมูล Flickr8k

1.A lady wearing a helmet holding a bike.
2.A woman in a blue vest and a sky blue helmet stands with her bicycle in traffic.
3.A woman in a helmet rides her bike behind a car.
4.A woman with a helmet and a backpack walks next to her bike.
5.Women with bike and a helmet wait for traffic.

รูปที่ 3-6 ตัวอย่างคำบรรยายจากชุดข้อมูล Flickr8k

ในการเตรียมชุดข้อมูลของ Flickr8k เนื่องจากคำบรรยายทุกรูปภาพในฐานข้อมูลชุดนี้เป็นภาษาอังกฤษทั้งหมดดังนั้นจึงทำการแปลทุกข้อความจากภาษาอังกฤษเป็นภาษาไทยโดยได้นำเข้าไลบลารี่ที่มีอยู่เข้ามาในโปรแกรมไพธอน

1.ผู้หญิงสวมหมวกกันน็อคถือจักรยาน
2.ผู้หญิงในเสื้อกั๊กสีน้ำเงินและหมวกกันน็อคสีฟ้ายืนอยู่กับจักรยานของเธอท่ามกลางการจราจร
3.ผู้หญิงสวมหมวกกันน็อคขี่จักรยานอยู่หลังรถยนต์
4.ผู้หญิงสวมหมวกกันน็อคและเป้สะพายหลังเดินข้างจักรยาน
5.ผู้หญิงที่มีจักรยานและหมวกกันน็อครอการจราจร

รูปที่ 3-7 ตัวอย่างคำบรรยายภาษาอังกฤษที่ถูกแปลเป็นภาษาไทยด้วย Google Translate

3.4.2 ตัวอย่างชุดข้อมูลการจราจรที่จัดทำขึ้นเอง

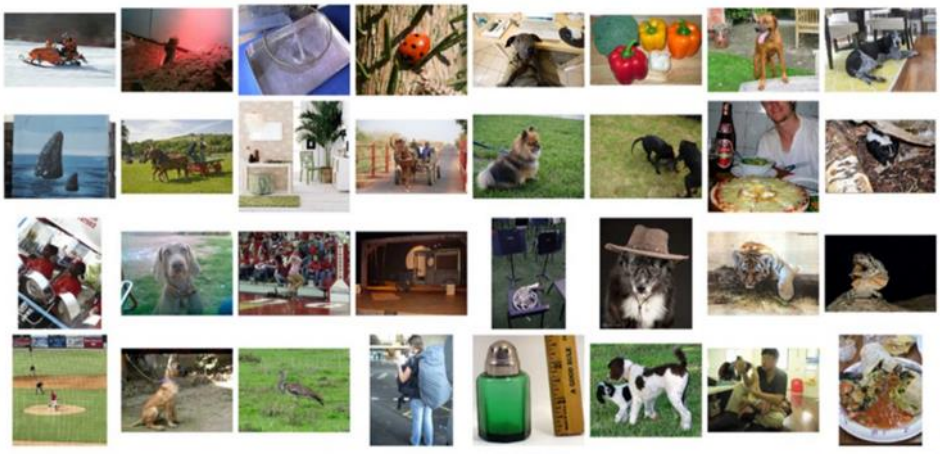
รูปที่ 3-8 ตัวอย่างรูปภาพการจราจรที่จัดทำขึ้นเอง

1.รถจักรยานยนต์จำนวนมากจอดติดไฟแดงบนถนน
2.รถจักรยานยนต์หลายคันและรถยนต์จำนวนหนึ่งกำลังจอดรอสัญญาณไฟบนถนน
3.รถจำนวนมากจอดรอสัญญาณไฟบนท้องถนน
4.รถจักรยานยนต์และรถยนต์จำนวนมากไม่สามารถเคลื่อนตัวได้เพราะติดไฟแดง
5. รถจักรยานยนต์และรถยนต์จำนวนมากบนท้องถนน

รูปที่ 3-9 ตัวอย่างคำบรรยายจากชุดข้อมูลการจราจรที่จัดทำขึ้นเอง

โดยจำนวนภาพในชุดข้อมูล Flickr8k มีทั้งหมด 8091 รูปและแต่ละรูปมีคำบรรยายภาษาอังกฤษ 5 คำบรรยาย ส่วนชุดข้อมูลการจราจรที่จัดทำขึ้นเองมีจำนวนรูปทั้งหมด 429 รูปมีคำบรรยายภาษาไทย 5 คำบรรยาย ดังนั้นเมื่อนำชุดข้อมูลทั้งสองมารวมกันจะได้รูปภาพทั้งหมด 8520 รูป และคำบรรยาย 42600 คำบรรยายโดยคำบรรยายที่รวมทั้งหมดจะเป็นคำบรรยายภาษาไทย แต่จำนวนรูปภาพทั้งหมดนั้นยังไม่เพียงพอต่อการฝึกสอนโมเดลเนื่องจากชุดข้อมูลยังมีความหลากหลายขององค์ประกอบหรือคุณลักษณะที่มากเกินไปทำให้เมื่อฝึกสอนโมเดลออกมาจะขาดการคาดเดาคำบรรยายที่หลากหลายดังนั้นจึงต้องมีการนำค่าน้ำหนักจาก Imagenet เข้ามาช่วยเสริมและเติมเต็มความหลากหลายของข้อมูล

3.4.3 ImageNet



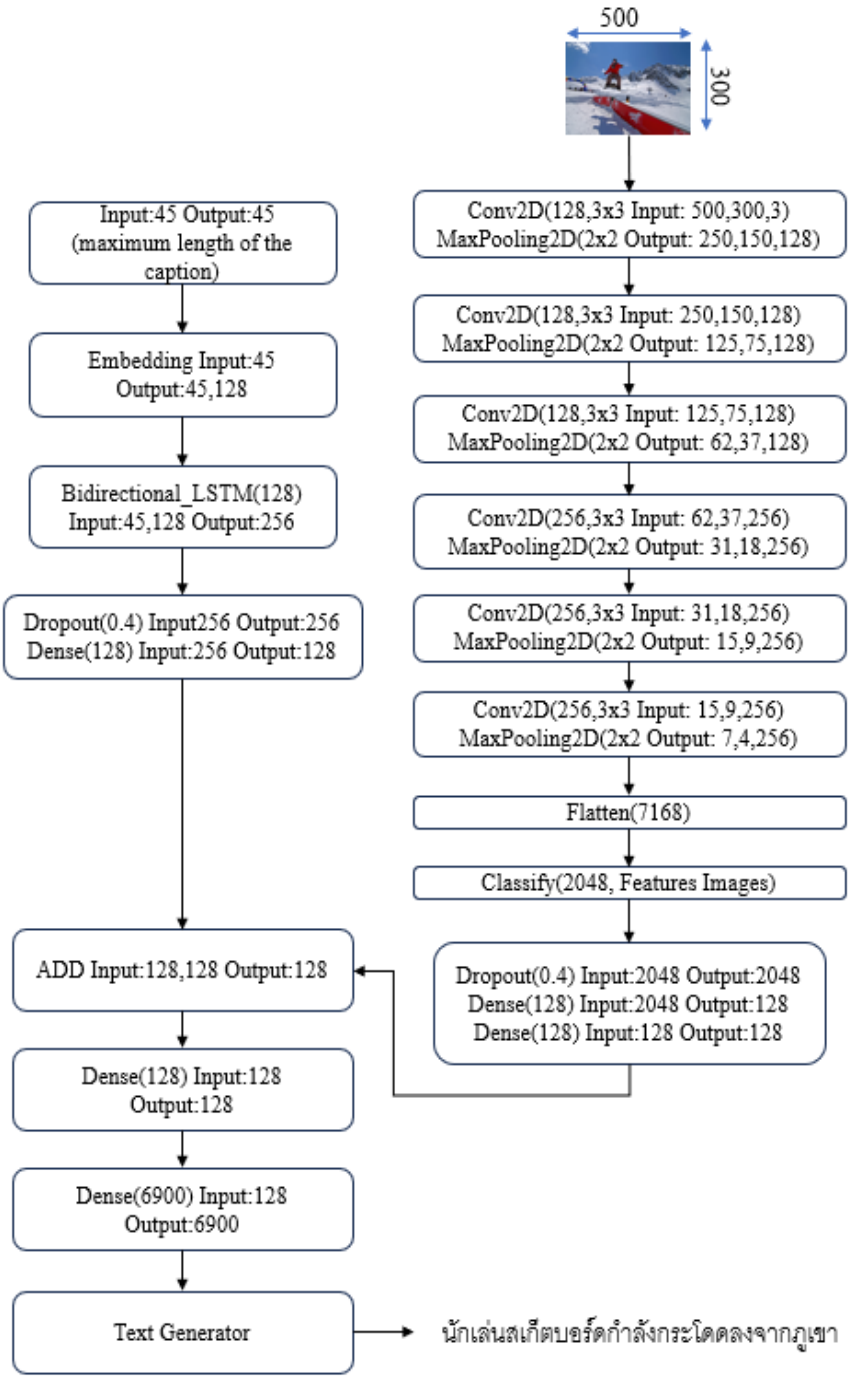
รูปที่ 3-10 ตัวอย่างรูปภาพจาก ImageNet

ที่มา https://thaikeras.com/wp-content/uploads/2018/06/2D0D1C81-15D3-4622-9749-8C7F68596B08.jpeg

ImageNet คือฐานชุดข้อมูลขนาดใหญ่ที่ถูกรวบรวมมาเพื่อทำงานในด้าน Visual Object Recognition หรืองานด้านการจำแนกคุณลักษณะของรูปภาพ ซึ่งใน ImageNet มีจำนวนรูปภาพมากกว่า 14 ล้านภาพ และประกอบไปด้วย 20000 หมวดหมู่และในแต่ละหมวดจะประกอบไปด้วยรูปภาพอีกหลายร้อยภาพ ทำให้ ImageNet คือแหล่งเก็บรวบรวมองค์ประกอบที่ยอดเยี่ยมอีกทั้งยังสามารถนำค่าน้ำหนักที่โมเดล CNN อย่าง VGG16, Resnet50 และ อื่นๆ ที่ทำการคัดแยกคุณลักษณะกับ ImageNet เข้ามาใช้กับโมเดลที่พัฒนาได้

3.5 ออกแบบ CNN และนำโมเดลการสร้างคำบรรยายมาพัฒนา

มีการออกแบบชั้นของ CNN และได้พัฒนาโมเดลในการสร้างคำบรรยายจากงานที่ [24] ซึ่งโมเดล CNN นี้ได้ทำการออกแบบตามรูปที่



รูปที่ 3-11 โมเดล CNN ที่ออกแบบเอง และ โมเดลการสร้างคำบรรยายภาพที่นำมาพัฒนา

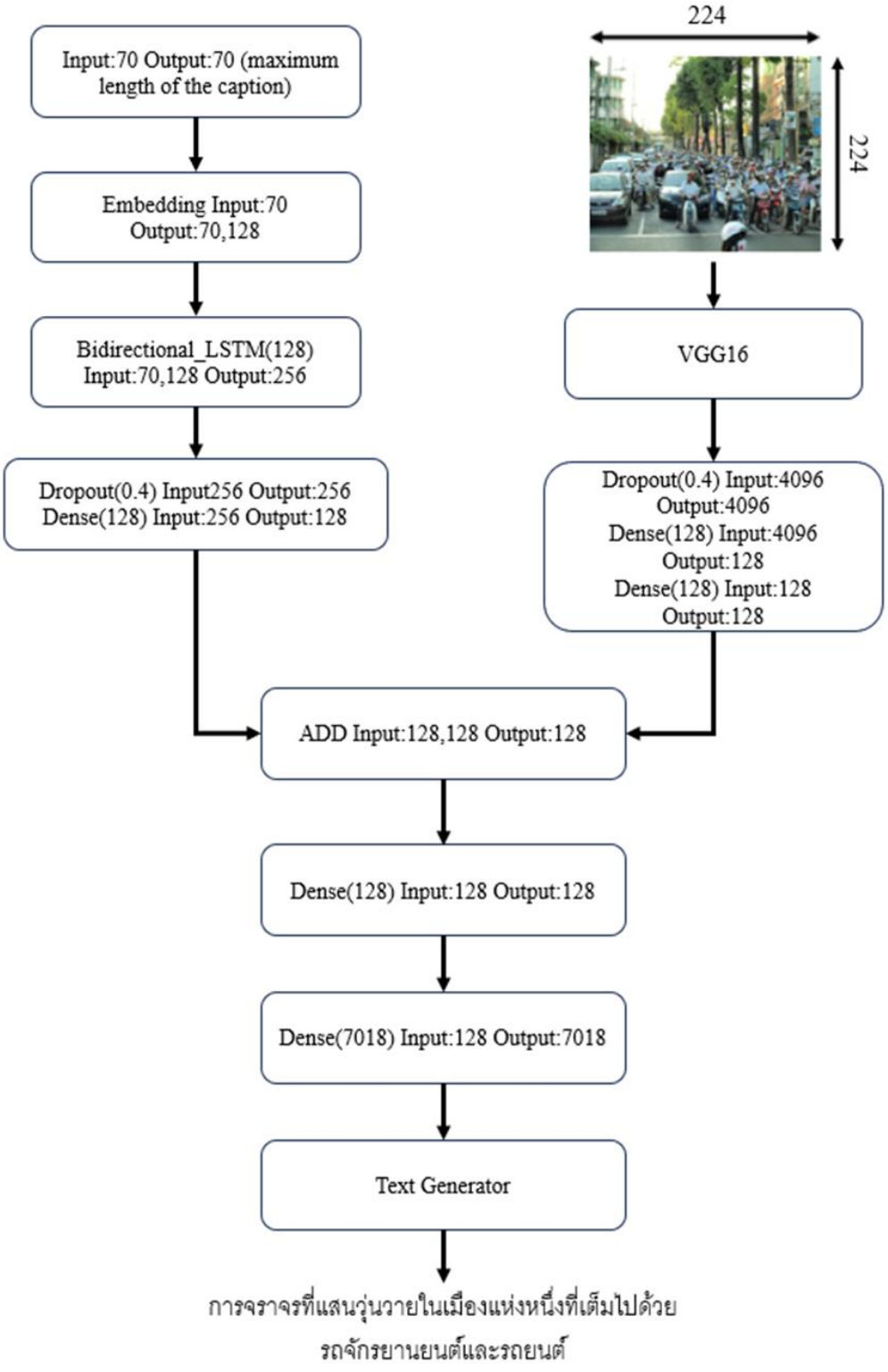
โดยจากรูปที่ 3-11 การทำ convolution ของทุกชั้นได้เลือกใช้ Activation Function เป็น relu และเมื่อทำ convolution เสร็จจะทำการ MaxPooling ให้กับทุกชั้น และในชั้นสุดท้ายหรือชั้นของการทำ Fully Connected ได้ใช้ 2048 และได้เลือกใช้ Activation Function เป็น Softmax

```
Model: "model_1"
__________________________________________________________________________________________________
 Layer (type)                Output Shape              Param #   Connected to
==================================================================================================
 input_2 (InputLayer)        [(None, 45)]              0         []

 input_1 (InputLayer)        [(None, 2048)]            0         []

 embedding (Embedding)       (None, 45, 128)           883200    ['input_2[0][0]']

 dropout (Dropout)           (None, 2048)              0         ['input_1[0][0]']

 bidirectional (Bidirection  (None, 256)               263168    ['embedding[0][0]']
 al)

 dense_1 (Dense)             (None, 128)               262272    ['dropout[0][0]']

 dropout_1 (Dropout)         (None, 256)               0         ['bidirectional[0][0]']

 dense_2 (Dense)             (None, 128)               16512     ['dense_1[0][0]']

 dense_3 (Dense)             (None, 128)               32896     ['dropout_1[0][0]']

 add (Add)                   (None, 128)               0         ['dense_2[0][0]',
                                                                  'dense_3[0][0]']

 dense_4 (Dense)             (None, 128)               16512     ['add[0][0]']

 dense_5 (Dense)             (None, 6900)              890100    ['dense_4[0][0]']

==================================================================================================
Total params: 2364660 (9.02 MB)
Trainable params: 2364660 (9.02 MB)
Non-trainable params: 0 (0.00 Byte)
```

รูปที่ 3-12 ตารางสรุปค่าพารามิเตอร์ของ CNN ที่ออกแบบเองและโมเดลการสร้างคำบรรยายภาพที่นำมาพัฒนา

3.6 ออกแบบและพัฒนาโมเดลการสร้างคำบรรยายภาพภาษาไทย

วิทยานิพนธ์เฉบับนี้ได้นำโมเดลที่ได้จาก [24] มาพัฒนาและต่อยอดซึ่งจากปกติที่มีการใช้ LSTM แบบปกติแต่สำหรับโมเดลที่พัฒนาได้มีการนำ Bidirectional LSTM เข้ามาทำในกระบวนการนี้รวมถึงมีการออกแบบชั้น Dense ใหม่โดยได้เพิ่มชั้น Dense เข้าไปอีก 1 ชั้น ในฝังของการคัดแยกคุณลักษณะของรูปภาพหรือการทำ Convolutional และกำหนดจำนวน Kernel ใหม่ให้กับทุกชั้น โดยสถาปัตยกรรมโมเดลได้แสดงดังรูปที่ 3-13

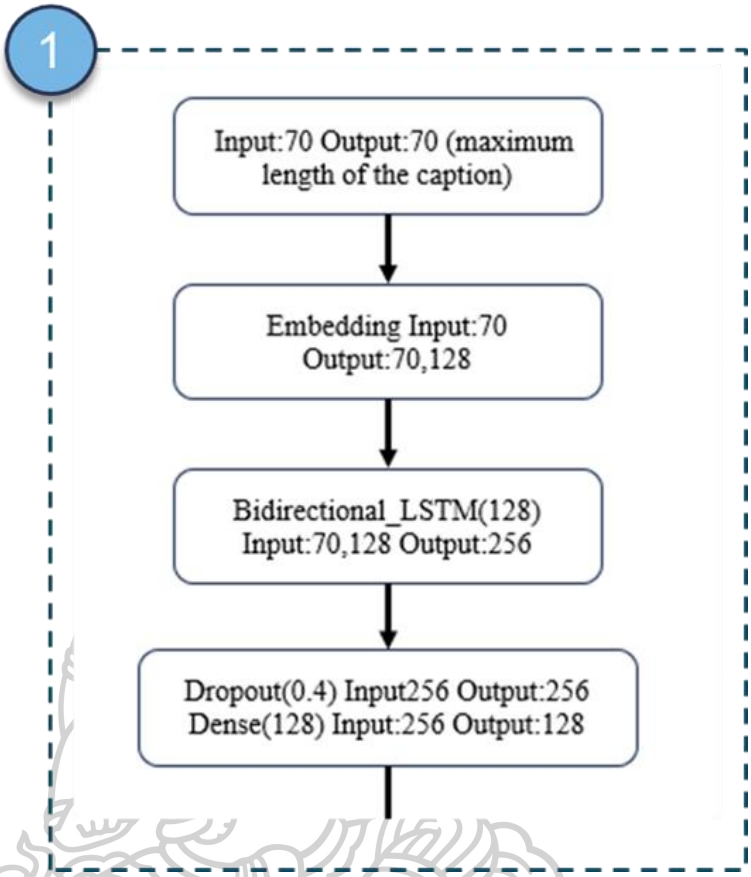


รูปที่ 3-13 สถาปัตยกรรมแบบจำลองที่ออกแบบและพัฒนา

```
Model: "model_1"
_________________________________________________________________________________________________
 Layer (type)                Output Shape              Param #   Connected to
=================================================================================================
 input_3 (InputLayer)        [(None, 70)]              0         []

 input_2 (InputLayer)        [(None, 4096)]            0         []

 embedding (Embedding)       (None, 70, 128)           898304    ['input_3[0][0]']

 dropout (Dropout)           (None, 4096)              0         ['input_2[0][0]']

 bidirectional (Bidirection  (None, 256)               263168    ['embedding[0][0]']
 al)

 dense (Dense)               (None, 128)               524416    ['dropout[0][0]']

 dropout_1 (Dropout)         (None, 256)               0         ['bidirectional[0][0]']

 dense_1 (Dense)             (None, 128)               16512     ['dense[0][0]']

 dense_2 (Dense)             (None, 128)               32896     ['dropout_1[0][0]']

 add (Add)                   (None, 128)               0         ['dense_1[0][0]',
                                                                  'dense_2[0][0]']

 dense_3 (Dense)             (None, 128)               16512     ['add[0][0]']

 dense_4 (Dense)             (None, 7018)              905322    ['dense_3[0][0]']

=================================================================================================
Total params: 2657130 (10.14 MB)
Trainable params: 2657130 (10.14 MB)
Non-trainable params: 0 (0.00 Byte)
```

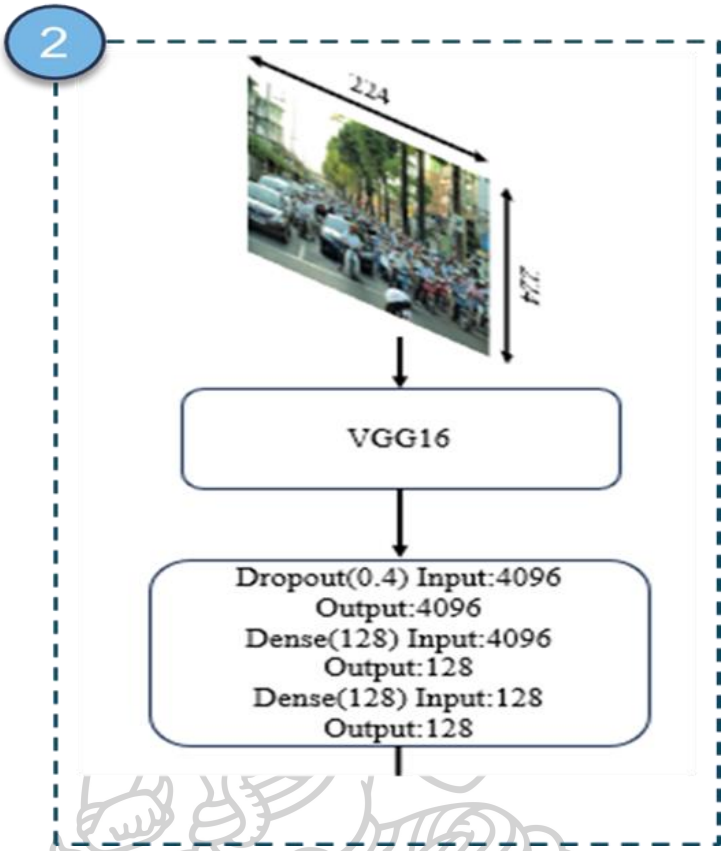
รูปที่ 3-14 ตารางสรุปค่าพารามิเตอร์ของโมเดลที่ออกแบบด้วย VGG16 และ Bidirectional LSTM

จากรูปที่ 3-13 จะเห็นได้โมเดลหลักในการสร้างคำบรรยายภาพจะประกอบด้วย VGG16 ที่รับภาพอินพุตเข้ามาแล้วทำการคัดแยกคุณลักษณะและ Bidirectional LSTM ทำหน้าที่ในการสร้างคำบรรยายภาพซึ่งข้อดีของ Bidirectional LSTM คือจะมี cell block เพิ่มขึ้นมาจากเดิมอีกหนึ่งแถวเพื่อทำการโมเดลข้อมูลในทิศทางย้อนกลับด้วยหรือที่เรียกว่า backward motion เพราะการที่มีการเรียนรู้แบบย้อนกลับรวมกับการเรียนรู้แบบไปข้างหน้าเป็นการเพิ่มประสิทธิภาพของความจำได้เป็นอย่างดี และจากรูปที่ 3-12 สามารถสังเกตเห็นได้ว่าชั้นที่เป็น Bidirectional LSTM ที่กำหนดจำนวน kernel ในตอนแรกเป็น 128 แต่ในตารางสรุปค่าพารามิเตอร์โมเดลจะเพิ่มขึ้นมาเป็น 256 เพราะต้องมีการแบ่งทิศทางของสถานะไปข้างหน้าและย้อนกลับ ดังนั้นจากรูปมีการอธิบายโครงสร้างโดยละเอียดตามรูปที่ 3-15 และ 3-16



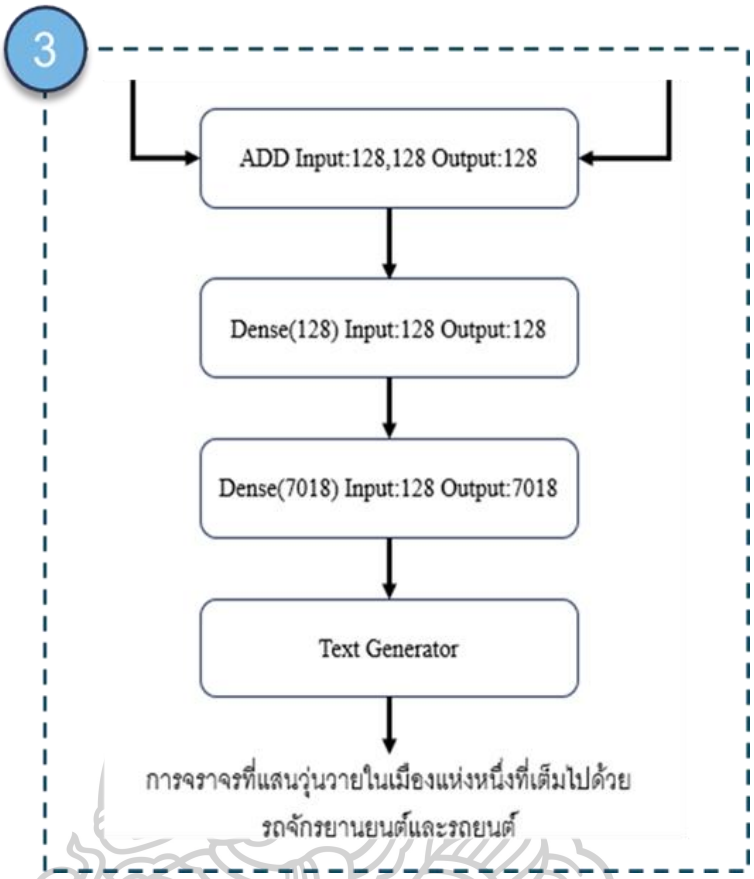
รูปที่ 3-15 การทำงานส่วนแรกของโมเดล

ในส่วนแรกจะมีการรับ Input ที่เป็นค่าความยาวสูงสุดของคำบรรยายที่มีอยู่เข้ามาซึ่งในชุดข้อมูลนี้จะได้ค่า 70 จากนั้น จะถูกนำเข้าฟังก์ชั่น Embedding เพื่อที่จะเปลี่ยนชุดข้อมูลให้ไปอยู่ในรูปแบบของเวกเตอร์ ก่อนนำเข้าสู่ชั้นสำคัญอย่าง Bidirectional LSTM ที่จะเป็นชั้นในการจำลองการขึ้นต่อกันตามลำดับระหว่างคำและวลีในทั้งสองทิศทาง แล้วก็จะมีการทำ drop out เพื่อลดการ overfitting จากนั้นส่งไปยัง Dense หรือ Fully connected ซึ่ง activation function ของชั้นนี้เลือกใช้เป็น relu



รูปที่ 3-16 การทำงานส่วนที่สองของโมเดล

ในส่วนที่สองคือการรับข้อมูล input ที่เป็นรูปภาพเพื่อทำการคัดแยกคุณลักษณะหรือองค์ประกอบในรูป ซึ่งเราได้เลือกใช้ CNN (Convolutonal Neural Network) รูปแบบ VGG16 ในการทำในส่วนนี้ ร่วมถึงมีการนำค่า weight ของ Imagenet ที่ผ่านกระบวนการคัดแยกคุณลักษณะของ VGG16 เข้ามาช่วยเสริมการเรียนรู้ให้กับโมเดลของเราอีกด้วยจากนั้นก็มีการทำ Dropout 1 ชั้น และ Dense 2 ชั้น ซึ่งการทำ Dense เพิ่มติดกันเข้ามาอีก 2 ชั้นนี้เป็นตัวช่วยอย่างมากในโมเดลของเราเพราะมันเรียนรู้ได้ดีส่วน activation function ของ dense 2 ชั้นนี้ใช้เป็น relu เช่นเดียวกัน



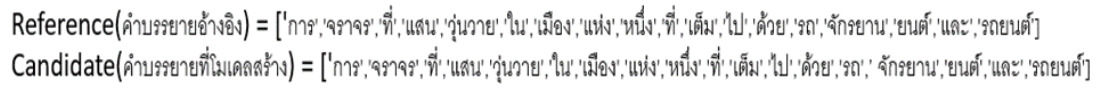
รูปที่ 3-17 การทำงานส่วนที่สามของโมเดล

ในส่วนที่3หรือส่วนสุดท้ายจะเป็นส่วนที่รวมค่าที่เป็น Feature จากชั้น Dense สุดท้ายของกระบวนการก่อนหน้าจากนั้นทำการกำหนด Dense อีก 2 ชั้น โดย Dense ก่อน Dense สุดท้าย ได้เลือกใช้ activation function เป็น relu และ Dense ชั้นสุดท้ายเลือกใช้เป็น softmax และเข้าสู่ Text Generator หรือฟังก์ชั่นสร้างคำบรรยาย

3.7 ประเมินผลลัพธ์คำบรรยายด้วย BLEU(Bilingual Evaluation Understudy)

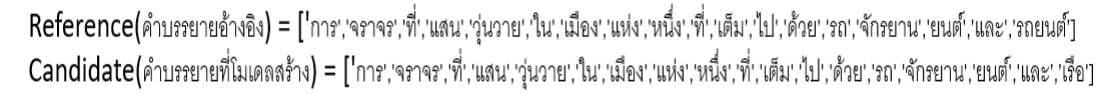
เมื่อโมเดลได้สร้างคำบรรยายภาพออกมาแล้วขั้นตอนต่อมาคือการประเมินผลลัพธ์ในการบรรยายเทียบกับคำบรรยายต้นฉบับ วิทยานิพนธ์ฉบับนี้จึงนำ matrix ที่ใช้ชี้วัดที่ชื่อว่า BLEU เข้ามาใช้ในการประเมินการบรรยายโดยผลลัพธ์ของ BLEU คือช่วงคะแนนที่อยู่ระหว่างช่วง 0-1 โดยที่ค่า 0 หมายถึงคำบรรยายที่โมเดลสร้างหรือคาดเดาออกมาไม่ตรงกับคำบรรยายอ้างอิงเลย และค่า 1 หมายถึงคำบรรยายที่โมเดลสร้างหรือคาดเดาออกมาตรงกับคำบรรยายอ้างอิงทั้งหมด โดยช่วงคะแนนของ matrix นี้รวมไปถึงทศนิยมด้วย ซึ่ง BLEU ต้องกำหนดค่าถ่วงน้ำหนักที่ขึ้นตามจำนวนของ n-gram ที่ซึ่งปกติจะวัดอยู่ที่ 4-gram คือการวัดคำบรรยายในประโยคจะเทียบเคียงความถูกต้องสูงสุดที่

ทุก 4 คำบรรยายจนกระทั่งจบประโยค โดยจะกำหนดค่าถ่วงเป็น (1, 0, 0, 0), (0.5, 0.5 , 0, 0), (0.33, 0.33, 0.33, 0.33) และ (0.25, 0.25, 0.25, 0.25) และตามปกติการทำงานของ BLEU ตามสูตรที่ (4) เพื่อเป็นการป้องกันความผิดพลาดในกรณีที่หากคำบรรยายที่เทียบระหว่างกันมีคำบรรยายใดคำบรรยายหนึ่งที่สั่นเกินไปอาจเกิดข้อผิดพลาดที่ทำให้ค่าคะแนนออกมาเป็น 1 ทั้งที่ความเป็นจริงอาจยังมีคำบรรยายที่ยาวกว่านั้น



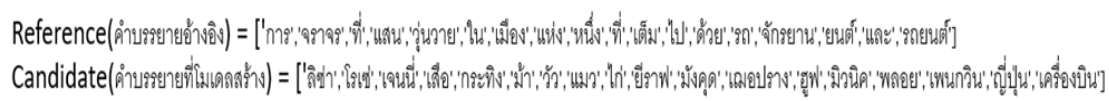
รูปที่ 3-18 ตัวอย่างคำบรรยายอ้างอิงเทียบกับคำบรรยายที่โมเดลสร้าง

จากรูปที่ 3-18 คำบรรยายอ้างอิงและคำบรรยายที่โมเดลสร้างตรงกันทุกข้อความแสดงว่าค่าคะแนน BLEU ของกรณีนี้จะมีค่าเป็น 1 เพราะไม่มีคำใดเลยที่ไม่ตรงกัน



รูปที่ 3-19 ตัวอย่างคำบรรยายอ้างอิงเทียบกับคำบรรยายที่โมเดลสร้างผิดหนึ่งคำ

จากรูปที่ 3-19 คำบรรยายอ้างอิงและคำบรรยายที่โมเดลสร้างที่ผิด 1 คำ เมื่อเทียบกันแล้ว BLEU พบข้อผิดพลาดในประโยคที่โมเดลสร้างเมื่อเทียบกับคำบรรยายอ้างอิงแล้วทำให้ได้ค่าคะแนน BLEU เป็น 0.9391 เพราะมีคำผิดจำนวนหนึ่งคำโดยการคิดค่า BLEU โดยละเอียดจะคำนวณได้จากสูตร (4)



รูปที่ 3-20 ตัวอย่างคำบรรยายอ้างอิงเทียบกับคำบรรยายที่โมเดลสร้างผิดทั้งประโยค

จากรูปที่ 3-20 คำบรรยายอ้างอิงและคำบรรยายที่โมเดลสร้างที่ผิดทั้งประโยค เมื่อเทียบกันแล้ว BLEU พบข้อผิดผลาดในประโยคที่โมเดลสร้างเมื่อเทียบกับคำบรรยายอ้างอิงแล้วทำให้ได้ค่าคะแนน BLEU เป็น 0 เพราะมีคำผิดทั้งประโยคในคำบรรยายที่โมเดลสร้าง

3.7.1 ค่าเฉลี่ย BLEU

ในวิทยานิพนธ์เล่มได้ทำการหาค่าเฉลี่ยของ BLEU ทั้งชุดข้อมูลฝึกสอนและทดสอบ โดยได้ทดสอบกับโมเดลการบรรยายภาพที่สร้างด้วย CNN(Convolutional Neural Network) ร่วมกับ Bidirectional LSTM ซึ่งการหาค่าเฉลี่ยของ BLEU ในตอนแรกจะนำคำบรรยายที่โมเดลได้สร้างขึ้นเทียบกับกับบรรยายต้นฉบับหรือคำบรรยายอ้างอิงที่เดิมที่ถูกเขียนบรรยายไว้ในชุดข้อมูลของ Flick8k และ ชุดข้อมูลการจราจรที่จัดทำขึ้นเองโดยผลลัพธ์ที่ได้จะแตกต่างกันไปในแต่ละโมเดลซึ่งโมเดลของการทดลองได้ค่า BLEU ที่มากกว่าโมเดลที่นำมาเปรียบเทียบและจะมีการแสดงโมเดลอื่นๆร่วมด้วย

```
References = [['สุนัข', 'สี', 'น้ำตาล', 'วิ่ง', 'อยู่', 'บน', 'พื้น','หญ้า'],
    ['สุนัข', 'สี', 'น้ำตาล', 'ตัว', 'หนึ่ง', 'กำลัง', 'วิ่ง', 'หายใจ', 'หอบ', 'อยู่','บน','พื้น','หญ้า'],
    ['สุนัข', 'สี', 'น้ำตาล', 'กำลัง', 'เดิน', 'อยู่', 'บน', 'พื้น', 'หญ้า'],
    ['สุนัข', 'สี', 'น้ำตาล', 'อุ้งเท้า', 'สี', 'ขาว', 'กำลัง', 'วิ่ง', 'เหยาะ', 'ผ่าน','ท่งหญ้า','สี','เขียว'],
    ['สุนัข', 'สี', 'น้ำตาล', 'หอบ', 'เดิน', 'อยู่', 'บน', 'พื้น', 'หญ้า']]
Candidates = ['สุนัข', 'สี', 'น้ำตาล', 'วิ่ง', 'อยู่', 'บน', 'พื้น','หญ้า']
```

รูปที่ 3-21 ตัวอย่างคำบรรยายที่โมเดลสร้างขึ้นเทียบกับคำบรรยายอ้างอิง 5 คำบรรยาย

จากรูปที่ 3-21 คือการนำคำบรรยายที่โมเดลสร้างขึ้นเทียบกับคำบรรยายต้นฉบับ 5 คำบรรยายโดยได้ใช้ค่าถ่วงน้ำหนักเป็น (1, 0, 0, 0), (0.25 ,0.25 ,0 ,0), (0.33 ,0.33 ,0.33 ,0.33) และ (0.25, 0.25, 0.25, 0.25) ตามลำดับหรือคือต้องการดูทั้ง 4-gram ซึ่งได้ค่า BLEU ออกมาเป็น 1-gram เป็น 1.0 2-gram เป็น 1.0 3-gram เป็น 0.894907 และ 4-gram เป็น 0.773055 ตามลำดับ โดยตัวอย่างคำบรรยายจากรูปที่ 3-21 มาจากชุดข้อมูลฝึกสอน Flickr8k ดังนั้นค่าเฉลี่ยจึงทำกระบวนการแบบนี้ไปกับทุกคำบรรยายที่โมเดลสร้างออกมาทำการบวกทุกค่าของ BLEU แล้วหารด้วยจำนวนที่นำมาบวกกันทั้งหมด

บทที่ 4
ผลการทดลองของงานวิจัย

ในบทนี้จะแสดงถึงผลลัพธ์การบรรยายภาพภาษาไทยที่ได้ทำบนชุดข้อมูล Flickr8k และ ชุดข้อมูลการจราจรที่ได้จัดทำขึ้นเองโดยจะแสดงรูปภาพและคำบรรยายภาษาไทยจากโมเดลที่สร้างขึ้นโดย CNN และ Bidirectional LSTM และแสดงผลการประเมินคุณภาพการเรียนรู้ของคำบรรยายที่โมเดลสร้างด้วยตัวชี้วัด BLEU ซึ่งการทดลองจะมีทั้งทำการทดลองกับชุดข้อมูล Flickr8k อย่างเดียวที่มีรูป 8091 รูปมีคำบรรยายทั้งหมด 40455 คำบรรยาย และการทดลองกับชุดข้อมูลรวม Flickr8k กับชุดข้อมูลจราจรที่จัดทำขึ้นเอง 429 รูป 2145 คำบรรยายเมื่อรวมทั้งสองข้อมูลจะได้รูปภาพ 8520 รูป และคำบรรยาย 42600 คำบรรยายและทุกโมเดลที่ทำการทดลองแบ่งเป็น การฝึกสอน 80% และทดสอบอีก 20% รวมถึงมีการแสดงตารางเปรียบเทียบโมเดลของงานวิจัยนี้เทียบกับงานวิจัยที่นำมาเปรียบเทียบซึ่งโมเดลของงานวิจัยนี้ที่นำโมเดลอื่นมาพัฒนาได้ผลลัพธ์ที่ดีกว่า

4.1 รูปและคำบรรยายภาพภาษาไทยที่สร้างจากโมเดล CNN ที่ทำขึ้นเอง + Bidirectional LSTM ทำการฝึกสอน 50 รอบ

เราได้ทำการทดลองโดยได้ทดลองโดยการออกแบบ CNN ด้วยวิธีการของเราเองแต่วิธีการนี้ยังมีข้อเสียในด้านขนาดของข้อมูลที่ทำการฝึกสอนเนื่องจากไม่สามารถนำ ImageNet เข้ามาใช้ในโมเดลได้ทำให้โมเดลขาดคุณลักษณะจำนวนมากและทำให้ค่าคะแนน BLEU ได้ผลลัพธ์ที่ไม่ค่อยดีนัก

4.1.1 ชุดข้อมูลทดสอบ Flickr8k

“นักสเก็ตบอร์ดกำลังกระโดดลงจากภูเขา”

รูปที่ 4-1 ตัวอย่างรูปภาพที่หนึ่งรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“สุนัขสีดำและสีขาวกำลังวิ่งอยู่บนชายหาด”

รูปที่ 4-2 ตัวอย่างรูปภาพที่สองรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“สุนัขสีน้ำตาลและสีขาวกำลังวิ่งอยู่บนพื้นหญ้า”

รูปที่ 4-3 ตัวอย่างรูปภาพที่สามรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

ตารางที่ 4-1 ค่าผลคะแนน BLEU ของโมเดล CNN ที่ทำขึ้นเอง + Bidirectional LSTM เฉพาะข้อมูล Flickr8k ฝึกสอน 50 รอบ

BLEU	Test
BLEU-1	0.513043
BLEU-2	0.283934
BLEU-3	0.177309
BLEU-4	0.112540

4.2 รูปและคำบรรยายภาพภาษาไทยที่สร้างจากโมเดล VGG16 + Bidirectional LSTM เฉพาะข้อมูล Flickr8k ทำการฝึกสอน 50 รอบ

4.2.1 ชุดข้อมูลทดสอบ Flickr8k

“สุนัขสีขาวตัวเล็กกำลังวิ่งอยู่บนพื้นหญ้า”

รูปที่ 4-4 ตัวอย่างรูปภาพที่สี่รวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“นักสโนว์บอร์ดกำลังกระโดดลงจากภูเขาที่เต็มไปด้วยหิมะ”

รูปที่ 4-5 ตัวอย่างรูปภาพที่ห้ารวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“นักสโนว์โมบิลคนหนึ่งกำลังขี่สโนว์โมบิล”

รูปที่ 4-6 ตัวอย่างรูปภาพที่หกรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“เด็กผู้หญิงสองคนกำลังเล่นอยู่ในทุ่งหญ้า”

รูปที่ 4-7 ตัวอย่างรูปภาพที่เจ็ดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“นักสเก็ตบอร์ดไถลลงมาตามทางลาด”

รูปที่ 4-8 ตัวอย่างรูปภาพที่แปดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“ชายคนหนึ่งโต้คลื่นในมหาสมุทร”

รูปที่ 4-9 ตัวอย่างรูปภาพที่เก้ารวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“สุนัขสีน้ำตาลกำลังวิ่งผ่านน้ำ”

รูปที่ 4-10 ตัวอย่างรูปภาพที่สิบรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“เด็กสาวคนหนึ่งกำลังเตะลูกฟุตบอล”

รูปที่ 4-11 ตัวอย่างรูปภาพที่สิบเอ็ดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“นักปั่นจักรยานกำลังกระโดดขึ้นไปในอากาศ”

รูปที่ 4-12 ตัวอย่างรูปภาพที่สิบสองรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“นกสีขาวบินอยู่เหนือน้ำ”

รูปที่ 4-13 ตัวอย่างรูปภาพที่สิบสามรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

ตารางที่ 4-2 ค่าผลคะแนน BLEU ของโมเดล VGG16 + Bidirectional LSTM เฉพาะข้อมูล Flickr8k ฝึกสอน 50 รอบ

BLEU	Test
BLEU-1	0.626148
BLEU-2	0.438592
BLEU-3	0.322355
BLEU-4	0.231805

4.3 รูปและคำบรรยายภาพภาษาไทยที่สร้างจากโมเดล ResNet50 + Bidirectional LSTM เฉพาะข้อมูล Flickr8k ทำการฝึกสอน 50 รอบ

4.3.1 ชุดข้อมูลทดสอบ Flickr8k

“สุนัขสีขาวและสีน้ำตาลกำลังวิ่งอยู่บนพื้นหญ้า”

รูปที่ 4-14 ตัวอย่างรูปภาพที่สิบสี่รวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“สุนัขสีน้ำตาลและสีขาวกำลังกระโดดข้ามหญ้า”

รูปที่ 4-15 ตัวอย่างรูปภาพที่สิบห้ารวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“นักสโนว์บอร์ดกำลังเล่นสโนว์บอร์ดลงจากภูเขา”

รูปที่ 4-16 ตัวอย่างรูปภาพที่สิบหกรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“ชายคนหนึ่งกำลังเล่นสเก็ตบอร์ดบนทางลาด”

รูปที่ 4-17 ตัวอย่างรูปภาพที่สิบเจ็ดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“ชายคนหนึ่งกำลังโต้คลื่นในมหาสมุทร”

รูปที่ 4-18 ตัวอย่างรูปภาพที่สิบแปดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“สุนัขสีขาววิ่งผ่านน้ำ”

รูปที่ 4-19 ตัวอย่างรูปภาพที่สิบเก้ารวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“สุนัขสีดำและสีน้ำตาลกำลังวิ่งอยู่บนชายหาด”

รูปที่ 4-20 ตัวอย่างรูปภาพที่ยี่สิบรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“นักสเก็ตบอร์ดกำลังเล่นสเก็ตบอร์ดบนถนน”

รูปที่ 4-21 ตัวอย่างรูปภาพที่ยี่สิบเอ็ดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“ชายคนหนึ่งปีนขึ้นไปบนหินสูงชัน”

รูปที่ 4-22 ตัวอย่างรูปภาพที่ยี่สิบสองรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“นกสีขาวตัวใหญ่บินอยู่ในอากาศ”

รูปที่ 4-23 ตัวอย่างรูปภาพที่ยี่สิบสามรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“คนสี่คนนั่งอยู่บนโขดหินที่มีตะไคร่น้ำ”

รูปที่ 4-24 ตัวอย่างรูปภาพที่ยี่สิบสี่รวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

ตารางที่ 4-3 ค่าผลคะแนน BLEU ของโมเดล ResNet50 + Bidirectional LSTM เฉพาะข้อมูล Flickr8k ฝึกสอน 50 รอบ

BLEU	Test
BLEU-1	0.630896
BLEU-2	0.448144
BLEU-3	0.332352
BLEU-4	0.242685

4.4 รูปและคำบรรยายภาพภาษาไทยที่สร้างจากโมเดล MobileNetV2 + Bidirectional LSTM เฉพาะข้อมูล Flickr8k ทำการฝึกสอน 50 รอบ

4.4.1 ชุดข้อมูลทดสอบ Flickr8k

“สุนัขสีขาวและสีน้ำตาลกำลังวิ่งอยู่ในสนามแข่ง”

รูปที่ 4-25 ตัวอย่างรูปภาพที่ยี่สิบห้ารวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“สุนัขสีขาวและสีน้ำตาลกำลังวิ่งอยู่ในสนามหญ้า”

รูปที่ 4-26 ตัวอย่างรูปภาพที่ยี่สิบหกรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“เด็กผู้ชายคนหนึ่งกำลังเล่นสกีลงไปในน้ำ”

รูปที่ 4-27 ตัวอย่างรูปภาพที่ยี่สิบเจ็ดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“ชายคนหนึ่งกำลังเล่นสเก็ตบอร์ดบนทางลาด”

รูปที่ 4-28 ตัวอย่างรูปภาพที่ยี่สิบแปดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“เด็กผู้หญิงสองคนกำลังเล่นอยู่ในสนามหญ้า”

รูปที่ 4-29 ตัวอย่างรูปภาพที่ยี่สิบเก้ารวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“สุนัขสีขาวกำลังวิ่งอยู่ในน้ำ”

รูปที่ 4-30 ตัวอย่างรูปภาพที่สามสิบรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

ตารางที่ 4-4 ค่าผลคะแนน BLEU ของโมเดล MobileNetV2 + Bidirectional LSTM เฉพาะข้อมูล Flickr8k ฝึกสอน 50 รอบ

BLEU	Test
BLEU-1	0.532048
BLEU-2	0.324067
BLEU-3	0.220295
BLEU-4	0.149306

4.5 รูปและคำบรรยายภาพภาษาไทยที่สร้างจากโมเดล VGG16 + Bidirectional LSTM เฉพาะข้อมูล Flickr8k ทำการฝึกสอน 100 รอบ

4.5.1 ชุดข้อมูลทดสอบ Flickr8k

“นักสโนว์บอร์ดกำลังกระโดดขึ้นไปในอากาศ”

รูปที่ 4-31 ตัวอย่างรูปภาพที่สามสิบเอ็ดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“นักสโนว์โมบิลกำลังกระโดด”

รูปที่ 4-32 ตัวอย่างรูปภาพที่สามสิบสองรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“เด็กผู้หญิงสองคนกำลังเล่นอยู่ในทุ่งหญ้า”

รูปที่ 4-33 ตัวอย่างรูปภาพที่สามสิบสามรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“สุนัขสีน้ำตาลกำลังวิ่งผ่านน้ำ”

รูปที่ 4-34 ตัวอย่างรูปภาพที่สามสิบสี่รวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“สุนัขสีขาวดำกำลังวิ่งอยู่บนพื้นหญ้า”

รูปที่ 4-35 ตัวอย่างรูปภาพที่สามสิบห้ารวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“ชายคนหนึ่งกำลังโต้คลื่นในมหาสมุทร”

รูปที่ 4-36 ตัวอย่างรูปภาพที่สามสิบหกรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“ชายคนหนึ่งปีนขึ้นไปบนหน้าผาหิน”

รูปที่ 4-37 ตัวอย่างรูปภาพที่สามสิบเจ็ดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“ชายคนหนึ่งขี่จักรยานไปตามถนนในเมือง”

รูปที่ 4-38 ตัวอย่างรูปภาพที่สามสิบแปดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

ตารางที่ 4-5 ค่าผลคะแนน BLEU ของโมเดล VGG16 + Bidirectional LSTM เฉพาะข้อมูล Flickr8k ฝึกสอน 100 รอบ

BLEU	Test
BLEU-1	0.610055
BLEU-2	0.424358
BLEU-3	0.310001
BLEU-4	0.222750

4.6 รูปและคำบรรยายภาพภาษาไทยที่สร้างจากโมเดล ResNet50 + Bidirectional LSTM เฉพาะข้อมูล Flickr8k ทำการฝึกสอน 100 รอบ

4.6.1 ชุดข้อมูลทดสอบ Flickr8k

“สุนัขสีขาวและสีน้ำตาลวิ่งไปตามหญ้า”

รูปที่ 4-39 ตัวอย่างรูปภาพที่สามสิบเก้ารวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“สุนัขสีน้ำตาลและสีขาวกำลังกระโดดข้ามหญ้า”

รูปที่ 4-40 ตัวอย่างรูปภาพที่สี่สิบรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“นักสโนว์บอร์ดกำลังเล่นกลในหิมะ”

รูปที่ 4-41 ตัวอย่างรูปภาพที่สี่สิบเอ็ดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“ชายคนหนึ่งกำลังเล่นสเก็ตบอร์ดลงบันได”

รูปที่ 4-42 ตัวอย่างรูปภาพที่สี่สิบสองรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“เด็กผู้หญิงสองคนเดินไปตามสนามหญ้า”

รูปที่ 4-43 ตัวอย่างรูปภาพที่สี่สิบสามรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“สุนัขขาวดำและสีน้ำตาลกำลังวิ่งอยู่บนพื้นหญ้า”

รูปที่ 4-44 ตัวอย่างรูปภาพที่สี่สิบสี่รวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“ชายคนหนึ่งกำลังโต้คลื่นในมหาสมุทร”

รูปที่ 4-45 ตัวอย่างรูปภาพที่สี่สิบห้ารวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“สุนัขสีน้ำตาลกำลังวิ่งอยู่ในน้ำ”

รูปที่ 4-46 ตัวอย่างรูปภาพที่สี่สิบหกรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“สุนัขสีขาวกำลังวิ่งผ่านพื้นที่ป่า”

รูปที่ 4-47 ตัวอย่างรูปภาพที่สี่สิบเจ็ดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“สุนัขสีดำและสีน้ำตาลกำลังวิ่งผ่านพื้นที่โล่ง”

รูปที่ 4-48 ตัวอย่างรูปภาพที่สี่สิบแปดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“เด็กผู้ชายคนหนึ่งกำลังเล่นสเก็ตบอร์ดบนทางลาด”

รูปที่ 4-49 ตัวอย่างรูปภาพที่สี่สิบเก้ารวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“ชายคนหนึ่งปีนขึ้นไปบนหินสูง”

รูปที่ 4-50 ตัวอย่างรูปภาพที่ห้าสิบรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“ชายคนหนึ่งขี่จักรยานบนทางลาด”

รูปที่ 4-51 ตัวอย่างรูปภาพที่ห้าสิบเอ็ดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“นกสีขาวตัวใหญ่บินอยู่เหนือน้ำ”

รูปที่ 4-52 ตัวอย่างรูปภาพที่ห้าสิบสองรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

ตารางที่ 4-6 ค่าผลคะแนน BLEU ของโมเดล ResNet50 + Bidirectional LSTM เฉพาะข้อมูล Flickr8k ฝึกสอน 100 รอบ

BLEU	Test
BLEU-1	0.607963
BLEU-2	0.425732
BLEU-3	0.313769
BLEU-4	0.228899

4.7 รูปและคำบรรยายภาพภาษาไทยที่สร้างจากโมเดล MobileNetV2 + Bidirectional LSTM เฉพาะข้อมูล Flickr8k ทำการฝึกสอน 100 รอบ

4.7.1 ชุดข้อมูลทดสอบ Flickr8k

“สุนัขสีขาวและสีน้ำตาลวิ่งอยู่ในสนามหญ้า”

รูปที่ 4-53 ตัวอย่างรูปภาพที่ห้าสิบสามรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“นักโต้คลื่นกำลังขี่คลื่นอยู่ในมหาสมุทร”

รูปที่ 4-54 ตัวอย่างรูปภาพที่ห้าสิบสี่รวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“นักสเก็ตบอร์ดกำลังกระโดดลงจากทางลาด”

รูปที่ 4-55 ตัวอย่างรูปภาพที่ห้าสิบห้ารวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“ชายคนหนึ่งกำลังเล่นสกีน้ำ”

รูปที่ 4-56 ตัวอย่างรูปภาพที่ห้าสิบหกรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“สุนัขสีขาววิ่งอยู่บนชายหาด”

รูปที่ 4-57 ตัวอย่างรูปภาพที่ห้าสิบเจ็ดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

จากรูปที่ 4-56 และ 4-57 จะเห็นได้ว่าคำบรรยายที่โมเดลบรรยายออกมานั้นไม่ตรงกับในรูปภาพหรือมีความหมายที่ผิดจากในรูปดังรูปที่ 4-56 ที่บรรยายว่า “ชายคนหนึ่งกำลังเล่นสกีน้ำ” แต่ที่จริงเมื่อมองจากรูปควรจะบรรยายว่า “ชายคนหนึ่งกำลังเล่นเซิร์ฟบอร์ด” หรือ “ชายคนหนึ่งกำลังโต้คลื่นในมหาสมุทร” เป็นต้นเพราะในรูปไม่ใช่สกี และในรูปที่ 4-57 ที่บรรยายว่า “สุนัขสีขาววิ่งอยู่บนชายหาด” แต่ที่จริงเมื่อมองจากรูปควรจะบรรยายว่า “สุนัขสีขาวกำลังวิ่งอยู่บนน้ำ” หรือ “สุนัขสีขาวกำลังวิ่งผ่านน้ำ” เป็นต้นเพราะในรูปไม่ใช่ชายหาด ซึ่งสาเหตุที่โมเดล MobileNetV2 + Bidirectional LSTM ที่ทำการฝึกสอนทั้งหมด 100 รอบ เฉพาะข้อมูล Flickr8k ในบางรูปที่มีการบรรยายผิดนั้นสาเหตุอาจเกิดจาก 1. MobileNetV2 มีการคัดแยกคุณลักษณะของรูปภาพที่น้อยเกินไป คือ 1280 คุณลักษณะ ซึ่งเมื่อมีคุณลักษณะน้อยแบบนี้อาจจะต้องเพิ่มจำนวนการฝึกสอนหรือเพิ่มรอบในการเรียนรู้ของโมเดลให้มากกว่านี้ 2. Bidirectional LSTM อาจจะต้องปรับเปลี่ยนพารามิเตอร์ต่างๆสำหรับการใช้งานร่วมกับMobileNetV2 เพื่อให้มีประสิทธิภาพที่ดีขึ้น

ตารางที่ 4-7 ค่าผลคะแนน BLEU ของโมเดล MobileNetV2 + Bidirectional LSTM เฉพาะข้อมูล Flickr8k ฝึกสอน 100 รอบ

BLEU	Test
BLEU-1	0.525464
BLEU-2	0.313897
BLEU-3	0.210837
BLEU-4	0.142086

4.8 รูปและคำบรรยายภาพภาษาไทยที่สร้างจากโมเดล VGG16 + Bidirectional LSTM (ชุดข้อมูลรวม)

โมเดลนี้คือโมเดลหลักที่วิยานิพนธ์เล่มนี้ได้นำเสนอโดยรูปแบบโมเดลสามารถดูได้จากรูปที่ 3-13 ในบทที่ 3 โดย

4.8.1 ชุดข้อมูลทดสอบที่เกี่ยวข้องกับการจราจรฝึกสอน 50 รอบ

“สัญญาณไฟแดงแจ้งเตือนให้ผู้ขับขี่หยุด”

รูปที่ 4-58 ตัวอย่างรูปภาพที่ห้าสิบแปดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“ผู้หญิงคนหนึ่งเดินบนถนน”

รูปที่ 4-59 ตัวอย่างรูปภาพที่ห้าสิบเก้ารวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“ผู้หญิงคนหนึ่งเดินไปตามถนนในเมือง”

รูปที่ 4-60 ตัวอย่างรูปภาพที่หกสิบรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“การจราจรที่แสนวุ่นวายในเมืองแห่งหนึ่งที่มีรถยนต์จำนวนมากบนท้องถนน”

รูปที่ 4-61 ตัวอย่างรูปภาพที่หกสิบเอ็ดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“ผู้คนจำนวนมากกำลังเดินข้ามถนนในเมืองใหญ่”

รูปที่ 4-62 ตัวอย่างรูปภาพที่หกสิบสองรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“การจราจรที่แสนวุ่นวายในเมืองแห่งหนึ่งที่เคลื่อนตัวได้อย่างช้าบนถนน”

รูปที่ 4-63 ตัวอย่างรูปภาพที่หกสิบสามรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“รถยนต์หลายคันเคลื่อนตัวได้อย่างช้าบนถนน”

รูปที่ 4-64 ตัวอย่างรูปภาพที่หกสิบสี่รวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“สัญญาณไฟแดงเตือนให้รถยนต์และรถจักรยานยนต์รถยนต์และรถจักรยานยนต์....”

รูปที่ 4-65 ตัวอย่างรูปภาพที่หกสิบห้ารวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

จากรูปที่ 4-65 ที่โมเดล VGG16 + Bidirectional LSTM ที่ทำการฝึกสอนทั้งหมด 50 รอบ ชุดข้อมูลรวม Flickr8k + จราจรที่จัดทำขึ้นเอง โดยในรูปที่ 4-65 โมเดลได้บรรยายออกมาว่า “สัญญาณไฟ

แดงเตือนให้รถยนต์และรถจักรยานยนต์รถยนต์และรถจักรยานยนต์....” สังเกตุเห็นได้ว่ามีคำซ้ำเกิดขึ้นโดยสาเหตุจากกรณีนี้อาจเกิดจาก จำนวนรอบการฝึกสอนที่น้อยเกินไปสำหรับการฝึกสอน VGG16 + Bidirectional LSTM โดยเฉพาะในกรณีของ Bidirectional LSTM ที่เนื่องจากจะต้องมีการประมวลผลเป็นลำดับทั้งไปและกลับแต่การฝึกสอนที่จำนวนรอบน้อยอาจทำให้ Bidirectional LSTM เรียนรู้ได้ไม่มากพอ

4.8.2 ชุดข้อมูลทดสอบ Flickr8k ฝึกสอน 50 รอบ

“นักสโนว์บอร์ดกำลังเล่นสโนว์บอร์ดลงจากเนินเขาที่เต็มไปด้วยหิมะ”

รูปที่ 4-66 ตัวอย่างรูปภาพที่หกสิบหกรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“นักโต้คลื่นกำลังโต้คลื่น”

รูปที่ 4-67 ตัวอย่างรูปภาพที่หกสิบเจ็ดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“ชายหนุ่มกำลังเล่นสเก็ตบอร์ดบนบันได”

รูปที่ 4-68 ตัวอย่างรูปภาพที่หกสิบแปดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“สุนัขสีน้ำตาลกำลังว่ายน้ำในน้ำ”

รูปที่ 4-69 ตัวอย่างรูปภาพที่หกสิบเก้ารวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“นักเล่นสเก็ตบอร์ดกำลังกระโดดขึ้นไปในอากาศ”

รูปที่ 4-70 ตัวอย่างรูปภาพที่เจ็ดสิบรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“นักโต้คลื่นกำลังเล่นเซิร์ฟ”

รูปที่ 4-71 ตัวอย่างรูปภาพที่เจ็ดสิบเอ็ดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“ชายคนหนึ่งปีนหน้าผา”

รูปที่ 4-72 ตัวอย่างรูปภาพที่เจ็ดสิบสองรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“นักปั่นจักรยานกำลังกระโดดข้ามสิ่งกีดขว้าง”

รูปที่ 4-73 ตัวอย่างรูปภาพที่เจ็ดสิบสามรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

จากรูปที่ 4-73 เองก็เช่นเดียวกันสาเหตุอาจเกิดจากจำนวนรอบการฝึกสอนที่น้อยทำให้โมเดล VGG16 + Bidirectional LSTM เรียนรู้ได้ไม่ดีพอ

“นกสีขาวบินโฉบเหนือน้ำ”

รูปที่ 4-74 ตัวอย่างรูปภาพที่เจ็ดสิบสี่รวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

ตารางที่ 4-8 ค่าผลคะแนน BLEU ของชุดข้อมูลรวม Flickr8k และ ชุดข้อมูลการจราจรโมเดล VGG16 + Bidirectional LSTM ที่ถูกฝึกสอน 50 รอบ

BLEU	Test
BLEU-1	0.619851
BLEU-2	0.433974
BLEU-3	0.320100
BLEU-4	0.232916

4.8.3 ชุดข้อมูลทดสอบที่เกี่ยวข้องกับการจราจรฝึกสอน 100 รอบ

“การจราจรที่ติดขัดบนท้องถนนแห่งหนึ่ง”

รูปที่ 4-75 ตัวอย่างรูปภาพที่เจ็ดสิบห้ารวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“ผู้คนจำนวนมากกำลังข้ามทางม้าลายในเมืองใหญ่แห่งหนึ่ง”

รูปที่ 4-76 ตัวอย่างรูปภาพเจ็ดสิบหกรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“การจราจรที่วุ่นวายบนท้องถนน”

รูปที่ 4-77 ตัวอย่างรูปภาพที่เจ็ดสิบเจ็ดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“สัญญาณไฟจราจรสีแดงเตือนให้รถยนต์และรถจักรยานยนต์ทุกคันต้องหยุด”

รูปที่ 4-78 ตัวอย่างรูปภาพที่เจ็ดสิบแปดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“ผู้หญิงคนหนึ่งกำลังเดินข้ามทางม้าลาย”

รูปที่ 4-79 ตัวอย่างรูปภาพที่เจ็ดสิบเก้ารวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“ผู้หญิงคนหนึ่งเดินไปตามถนน”

รูปที่ 4-80 ตัวอย่างรูปภาพที่แปดสิบรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“การจราจรที่แสนวุ่นวายบนท้องถนนแห่งหนึ่ง”

รูปที่ 4-81 ตัวอย่างรูปภาพที่แปดสิบเอ็ดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“รถยนต์หลายคันบนท้องถนนที่มีรถจอดติดอยู่”

รูปที่ 4-82 ตัวอย่างรูปภาพที่แปดสิบสองรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“รถยนต์หลายคันกำลังเคลื่อนตัวเข้าไปในความมืด”

รูปที่ 4-83 ตัวอย่างรูปภาพที่แปดสิบสามรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“ไฟแดงเตือนให้รถจักรยานยนต์และรถยนต์ทุกคันต้องหยุด”

รูปที่ 4-84 ตัวอย่างรูปภาพที่แปดสิบสี่ถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

4.8.4 ชุดข้อมูลทดสอบ Flickr8k ฝึกสอน 100 รอบ

“ชายคนหนึ่งปีนหน้าผา”

รูปที่ 4-85 ตัวอย่างรูปภาพที่แปดสิบห้ารวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“ชายคนหนึ่งกำลังโต้คลื่นในมหาสมุทร”

รูปที่ 4-86 ตัวอย่างรูปภาพที่แปดสิบหกรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“นักปั่นจักรยานกำลังขี่ไปตามถนน”

รูปที่ 4-87 ตัวอย่างรูปภาพที่แปดสิบเจ็ดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“นักสเก็ตบอร์ดกำลังเล่นสเก็ตบอร์ด”

รูปที่ 4-88 ตัวอย่างรูปภาพที่แปดสิบแปดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

สุนัขสีน้ำตาลกระโดดขึ้นไปในอากาศ

รูปที่ 4-89 ตัวอย่างรูปภาพที่แปดสิบเก้ารวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“สุนัขสีน้ำตาลกระโดดขึ้นไปในอากาศเพื่อจับลูกบอล”

รูปที่ 4-90 ตัวอย่างรูปภาพที่เก้าสิบรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“สุนัขสีน้ำตาลและสีขาวกำลังวิ่งอยู่บนพื้นหญ้า”

รูปที่ 4-91 ตัวอย่างรูปภาพที่เก้าสิบเอ็ดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“นักโต้คลื่นกำลังโต้คลื่นในมหาสมุทร”

รูปที่ 4-92 ตัวอย่างรูปภาพที่เก้าสิบสองรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“เด็กผู้หญิงสองคนกำลังเล่นอยู่ในสนามหญ้า”

รูปที่ 4-93 ตัวอย่างรูปภาพที่เก้าสิบสามรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“ผู้ชายคนหนึ่งกำลังกระโดดลงไปในทราย”

รูปที่ 4-94 ตัวอย่างรูปภาพที่ยี่สิบรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

จากรูปที่ 4-75 จนถึง 4-94 ผลลัพธ์การบรรยายภาพของโมเดล VGG16 + Bidirectional LSTM แสดงออกมาได้ค่อนข้างที่ตรงกับความหมายในรูปที่ถึงแม้บางรูปจะมีการขาดรายละเอียดเล็กน้อย อย่างเช่นรูปภาพที่ 4-88 ที่การบรรยายไม่ได้ระบุเพศของผู้เล่น และรูปภาพที่ 4-89 ที่โมเดลไม่ได้บรรยายว่าสุนัขกำลังคาบสิ่งใดอยู่ในปาก รวมถึงรูปภาพที่ 4-92 ที่โมเดลก็ไม่ได้บรรยายถึงเพศของนักโต้คลื่น แต่โมเดลไม่สามารถสรุปผลโดยการดูแค่การบรรยายจากโมเดลว่าเป็นการบรรยายที่ดีหรือไม่ดีได้ดังนั้นจึงนำเมทริกซ์ BLEU เข้ามาวัดประสิทธิภาพในการบรรยายโดยเฉลี่ยของชุดข้อมูลทดสอบของFlickr8k และ ชุดข้อมูลการจราจรเพื่อให้แสดงผลลัพธ์การประเมินที่จับต้องได้

4.8.5 ตัวอย่างชุดข้อมูลฝึกสอนที่เกี่ยวข้องกับการจราจร

ในหัวข้อนี้จะเป็นการแสดงผลลัพธ์ของรูปภาพและการบรรยายจากชุดข้อมูลฝึกสอนเพื่อเป็นการดูความถูกต้องของการฝึกสอนด้วยเพราะจากชุดข้อมูลนี้โมเดลควรแสดงผลลัพธ์ได้ใกล้เคียงกับคำบรรยายอ้างอิงมากที่สุด

“ผู้หญิงคนหนึ่งกำลังเดินข้ามทางม้าลาย”

รูปที่ 4-95 ตัวอย่างรูปภาพที่เก้าสิบห้ารวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“รถคันสีแดงจอดอยู่บนถนน”

รูปที่ 4-96 ตัวอย่างรูปภาพที่เก้าสิบหกรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“การจราจรที่ตัดขัดบนท้องถนนแห่งหนึ่งในกรุงเทพ”

รูปที่ 4-97 ตัวอย่างรูปภาพที่เก้าสิบเจ็ดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

4.8.6 ตัวอย่างชุดข้อมูลฝึกสอนของ Flickr8k

“ชายคนหนึ่งกำลังเล่นสกีลงจากเนินเขาที่เต็มไปด้วยหิมะ”

รูปที่ 4-98 ตัวอย่างรูปภาพที่เก้าสิบแปดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“ผู้หญิงคนหนึ่งกำลังเล่นเทนนิสในสนามเทนนิส”

รูปที่ 4-99 ตัวอย่างรูปภาพที่เก้าสิบเก้ารวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“ผู้หญิงสองคนยิ้มให้กล้อง”

รูปที่ 4-100 ตัวอย่างรูปภาพหนึ่งร้อยหกรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

ตารางที่ 4-9 ค่าผลคะแนน BLEU ของชุดข้อมูลรวม Flickr8k และ ชุดข้อมูลการจราจรโมเดล VGG16 + Bidirectional LSTM ที่ถูกฝึกสอน 100 รอบ

BLEU	Train(6816)	Test(1704)
BLEU-1	0.698177	0.607605
BLEU-2	0.546698	0.420551
BLEU-3	0.440029	0.307364
BLEU-4	0.349506	0.221886

4.9 รูปและคำบรรยายภาพภาษาไทยที่สร้างจากโมเดล ResNet50 + Bidirectional LSTM (ชุดข้อมูลรวม)

```
Model: "model_1"
__________________________________________________________________________________________________
 Layer (type)                Output Shape              Param #   Connected to
==================================================================================================
 input_3 (InputLayer)        [(None, 70)]              0         []

 input_2 (InputLayer)        [(None, 2048)]            0         []

 embedding (Embedding)       (None, 70, 128)           898304    ['input_3[0][0]']

 dropout (Dropout)           (None, 2048)              0         ['input_2[0][0]']

 bidirectional (Bidirection  (None, 256)               263168    ['embedding[0][0]']
 al)

 dense (Dense)               (None, 128)               262272    ['dropout[0][0]']

 dropout_1 (Dropout)         (None, 256)               0         ['bidirectional[0][0]']

 dense_1 (Dense)             (None, 128)               16512     ['dense[0][0]']

 dense_2 (Dense)             (None, 128)               32896     ['dropout_1[0][0]']

 add (Add)                   (None, 128)               0         ['dense_1[0][0]',
                                                                  'dense_2[0][0]']

 dense_3 (Dense)             (None, 128)               16512     ['add[0][0]']

 dense_4 (Dense)             (None, 7018)              905322    ['dense_3[0][0]']

==================================================================================================
Total params: 2394986 (9.14 MB)
Trainable params: 2394986 (9.14 MB)
Non-trainable params: 0 (0.00 Byte)
```

รูปที่ 4-101 สรุปโมเดล ResNet50

4.9.1 ชุดข้อมูลทดสอบที่เกี่ยวข้องกับการจราจรฝึกสอน 50 รอบ

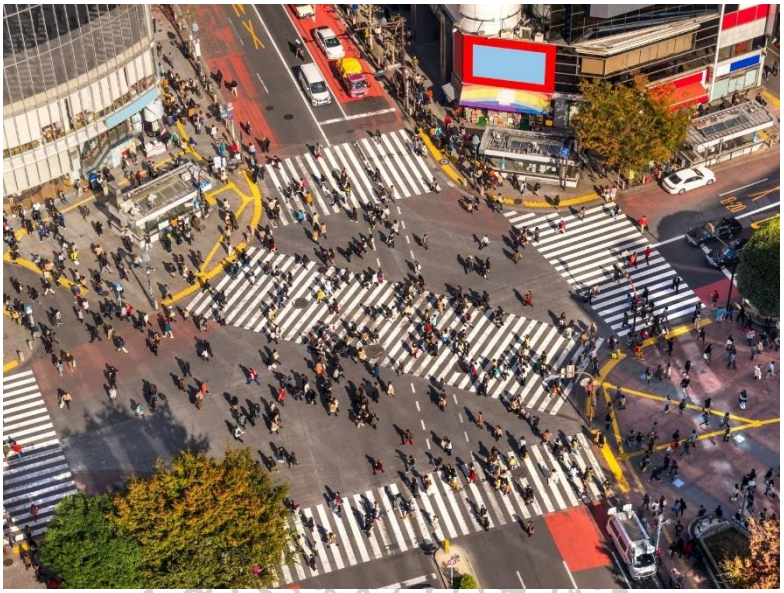
“การจราจรติดขัดบนท้องถนนในกรุงเทพที่เต็มไปด้วยไฟแดงบนท้องถนนในกรุงเทพ”

รูปที่ 4-102 ตัวอย่างรูปภาพที่หนึ่งร้อยสองรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

จะเห็นได้ว่ารูปที่ 4-102 มีการเกิดคำซ้ำซึ่งไม่เป็นผลดีต่อตัวโมเดล

“การจราจรที่แสนวุ่นวายในกรุงเทพ”

รูปที่ 4-103 ตัวอย่างรูปภาพที่หนึ่งร้อยสามรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น



“ผู้คนจำนวนมากกำลังเดินข้ามทางม้าลาย”

รูปที่ 4-104 ตัวอย่างรูปภาพที่หนึ่งร้อยสี่รวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“รถจักรยานยนต์จำนวนมากบนท้องถนนเคลื่อนตัวบนท้องถนนช้า”

รูปที่ 4-105 ตัวอย่างรูปภาพที่หนึ่งร้อยห้ารวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

จะเห็นได้ว่ารูปที่ 4-105 ยังอธิบายได้ไม่ละเอียดพอเพราะในภาพไม่ได้มีแต่รถจักรยานยนต์เท่านั้น

“ไฟแดงเตือนให้ผู้ขับขี่ต้องหยุดไป”

รูปที่ 4-106 ตัวอย่างรูปภาพที่หนึ่งร้อยหกรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

4.9.2 ชุดข้อมูลทดสอบ Flickr8k ฝึกสอน 50 รอบ

“สุนัขสีน้ำตาลกำลังเล่นอยู่ในสนามหญ้า”

รูปที่ 4-107 ตัวอย่างรูปภาพที่หนึ่งร้อยเจ็ดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“สุนัขสีน้ำตาลและสีขาวกำลังวิ่งอยู่บนพื้นหญ้า”

รูปที่ 4-108 ตัวอย่างรูปภาพที่หนึ่งร้อยแปดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“นักโต้คลื่นกำลังโต้คลื่น”

รูปที่ 4-109 ตัวอย่างรูปภาพที่หนึ่งร้อยเก้ารวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“สุนัขสีดำและสีขาวกำลังวิ่งอยู่ในสนามทดสอบความคล่องตัว”

รูปที่ 4-110 ตัวอย่างรูปภาพที่หนึ่งร้อยสิบรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“นกสีขาวกำลังบินอยู่เหนือน้ำ”

รูปที่ 4-111 ตัวอย่างรูปภาพที่หนึ่งร้อยสิบเอ็ดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“สุนัขสีน้ำตาลกำลังว่ายอยู่ในน้ำ”

รูปที่ 4-112 ตัวอย่างรูปภาพที่หนึ่งร้อยสิบสองรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“นักสเก็ตบอร์ดกำลังกระโดดขึ้นไปในอากาศ”

รูปที่ 4-113 ตัวอย่างรูปภาพที่หนึ่งร้อยสิบสามรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“นักปั่นจักรยานกระโดดขึ้นไปในอากาศ”

รูปที่ 4-114 ตัวอย่างรูปภาพที่หนึ่งร้อยสิบสี่รวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“สุนัขสีน้ำตาลวิ่งอยู่ในน้ำ”

รูปที่ 4-115 ตัวอย่างรูปภาพที่หนึ่งร้อยสิบห้ารวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

ตารางที่ 4-10 ค่าผลคะแนน BLEU ของชุดข้อมูลรวม Flickr8k และ ชุดข้อมูลการจราจรโมเดล ResNet50 + Bidirectional LSTM ที่ถูกฝึกสอน 50 รอบ

BLEU	Test
BLEU-1	0.646815
BLEU-2	0.462727
BLEU-3	0.344358
BLEU-4	0.252195

4.9.3 ชุดข้อมูลทดสอบที่เกี่ยวข้องกับการจราจรฝึกสอน 100 รอบ

“ไฟจราจรสีแดงเตือนให้รถจักรยานยนต์และรถยนต์ทุกคันต้องหยุด”

รูปที่ 4-116 ตัวอย่างรูปภาพที่หนึ่งร้อยสิบหกรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“ผู้คนกำลังเดินข้ามทางม้าลาย”

รูปที่ 4-117 ตัวอย่างรูปภาพที่หนึ่งร้อยสิบเจ็ดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“ผู้หญิงคนหนึ่งเดินไปตามถนน”

รูปที่ 4-118 ตัวอย่างรูปภาพที่หนึ่งร้อยสิบแปดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“การจราจรติดขัดเพราะหนาแน่นไปด้วยรถยนต์จำนวนมาก”

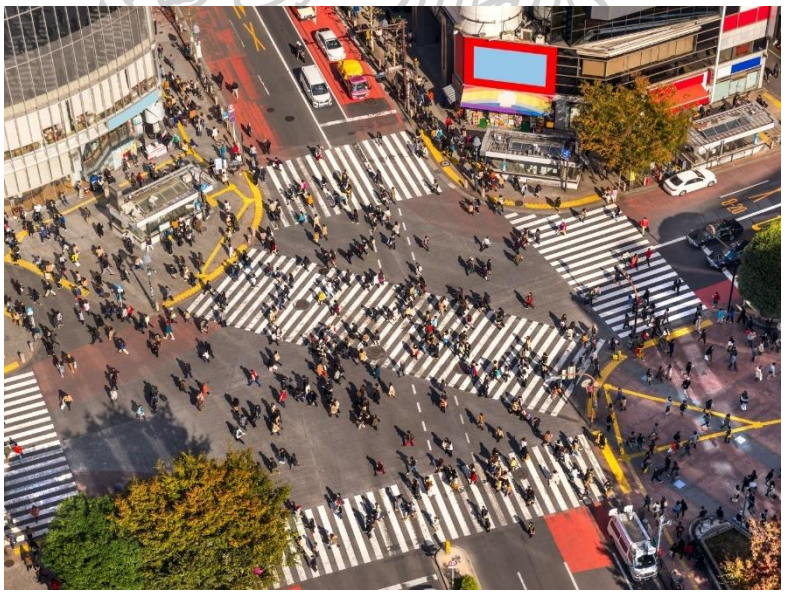
รูปที่ 4-119 ตัวอย่างรูปภาพที่หนึ่งร้อยสิบเก้ารวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“การจราจรติดขัดเพราะหนาแน่นไปด้วยรถยนต์บนท้องถนน”

รูปที่ 4-120 ตัวอย่างรูปภาพที่หนึ่งร้อยยี่สิบรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“การจราจรติดขัดในเมืองแห่งหนึ่งในกรุงเทพ”

รูปที่ 4-121 ตัวอย่างรูปภาพที่หนึ่งร้อยยี่สิบเอ็ดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น



“ผู้คนจำนวนมากกำลังเดินข้ามทางม้าลายในเมืองใหญ่แห่งหนึ่ง”

รูปที่ 4-122 ตัวอย่างรูปภาพที่หนึ่งร้อยยี่สิบสองรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

จากรูปที่ 4-116 จนถึง 4-122 ผลลัพธ์การบรรยายภาพของโมเดล ResNet50 + Bidirectional LSTM แสดงออกมาได้ค่อนข้างที่ตรงกับความหมายในรูปที่ถึงแม้บางรูปจะมีการขาดรายละเอียดเล็กน้อย แต่ในบางรูปก็มีการเพิ่มรายละเอียดเข้าอย่างเช่นรูปที่ 4-121 ที่บรรยายว่าการจราจรติดขัดในเมืองแห่งหนึ่งในกรุงเทพ ที่มีรายละเอียดคำว่า “กรุงเทพ” ที่เป็นคีย์สำคัญเพิ่มเข้ามาจากที่โมเดล VGG16 + Bidirectional LSTM ไม่มี

4.9.4 ชุดข้อมูลทดสอบ Flickr8k ฝึกสอน 100 รอบ

“ชายคนหนึ่งกำลังเล่นสเก็ตบอร์ด”

รูปที่ 4-123 ตัวอย่างรูปภาพที่หนึ่งร้อยยี่สิบสามรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“นักโต้คลื่นกำลังขี่คลื่น”

รูปที่ 4-124 ตัวอย่างรูปภาพที่หนึ่งร้อยยี่สิบสี่รวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“ชายคนหนึ่งกำลังเล่นสเก็ตบอร์ดบนบันได”

รูปที่ 4-125 ตัวอย่างรูปภาพที่หนึ่งร้อยยี่สิบห้ารวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“นักปั่นจักรยานกระโดดขึ้นไปในอากาศโดยมีดวงอาทิตย์เป็นฉากหลัง”

รูปที่ 4-126 ตัวอย่างรูปภาพที่หนึ่งร้อยยี่สิบหกรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“สุนัขสีขาวกำลังวิ่งอยู่ในน้ำ”

รูปที่ 4-127 ตัวอย่างรูปภาพที่หนึ่งร้อยยี่สิบเจ็ดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“สุนัขสีน้ำตาลและสีขาววิ่งอยู่บนชายหาด”

รูปที่ 4-128 ตัวอย่างรูปภาพที่หนึ่งร้อยยี่สิบแปดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

4.9.5 ตัวอย่างชุดข้อมูลฝึกสอนที่เกี่ยวข้องกับการจราจร

“รถยนต์สีแดงคันหนึ่งจอดอยู่บนถนน”

รูปที่ 4-129 ตัวอย่างรูปภาพหนึ่งร้อยยี่สิบเก้ารวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“สัญญาณไฟจราจรสีแดงเตือนให้รถจักรยานยนต์และรถยนต์ทุกคันต้องหยุด”

รูปที่ 4-130 ตัวอย่างรูปภาพที่หนึ่งร้อยสามสิบรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

4.9.6 ตัวอย่างชุดข้อมูลฝึกสอน Flickr8k

“คนเจ็ดคนกำลังเล่นอยู่ในมหาสมุทร”

รูปที่ 4-131 ตัวอย่างรูปภาพที่หนึ่งร้อยสามสิบเอ็ดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“สุนัขสีน้ำตาลเดินหอบอยู่ข้างนอก”

รูปที่ 4-132 ตัวอย่างรูปภาพที่หนึ่งร้อยสามสิบสองรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

ตารางที่ 4-11 ค่าผลคะแนน BLEU ของชุดข้อมูลรวม Flickr8k และ ชุดข้อมูลการจราจรโมเดล ResNet50 + Bidirectional LSTM ที่ถูกฝึกสอน 100 รอบ

BLEU	Train(6816)	Test(1704)
BLEU-1	0.669693	0.611543
BLEU-2	0.511792	0.430566
BLEU-3	0.404764	0.319303
BLEU-4	0.315446	0.233104

4.10 รูปและคำบรรยายภาพภาษาไทยที่สร้างจากโมเดล MobileNetV2 + Bidirectional LSTM (ชุดข้อมูลรวม)

```
Model: "model_1"
__________________________________________________________________________________________________
 Layer (type)                Output Shape              Param #   Connected to
==================================================================================================
 input_3 (InputLayer)        [(None, 70)]              0         []

 input_2 (InputLayer)        [(None, 1280)]            0         []

 embedding (Embedding)       (None, 70, 128)           898304    ['input_3[0][0]']

 dropout (Dropout)           (None, 1280)              0         ['input_2[0][0]']

 bidirectional (Bidirection  (None, 256)               263168    ['embedding[0][0]']
 al)

 dense (Dense)               (None, 128)               163968    ['dropout[0][0]']

 dropout_1 (Dropout)         (None, 256)               0         ['bidirectional[0][0]']

 dense_1 (Dense)             (None, 128)               16512     ['dense[0][0]']

 dense_2 (Dense)             (None, 128)               32896     ['dropout_1[0][0]']

 add (Add)                   (None, 128)               0         ['dense_1[0][0]',
                                                                  'dense_2[0][0]']

 dense_3 (Dense)             (None, 128)               16512     ['add[0][0]']

 dense_4 (Dense)             (None, 7018)              905322    ['dense_3[0][0]']

==================================================================================================
Total params: 2296682 (8.76 MB)
Trainable params: 2296682 (8.76 MB)
Non-trainable params: 0 (0.00 Byte)
__________________________________________________________________________________________________
```

รูปที่ 4-133 สรุปโมเดล MobileNetV2

4.10.1 ชุดข้อมูลทดสอบที่เกี่ยวข้องกับการจราจรฝึกสอน 50 รอบ

“สัญญาณไฟจราจรสีแดงบอกให้รถเตรียมหยุด”

รูปที่ 4-134 ตัวอย่างรูปภาพที่หนึ่งร้อยสามสิบสี่รวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“การจราจรที่หนาแน่นไปด้วยรถยนต์”

รูปที่ 4-135 ตัวอย่างรูปภาพที่หนึ่งร้อยสามสิบห้ารวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“ผู้หญิงในชุดสีแดงกำลังเดินอยู่บนถนน”

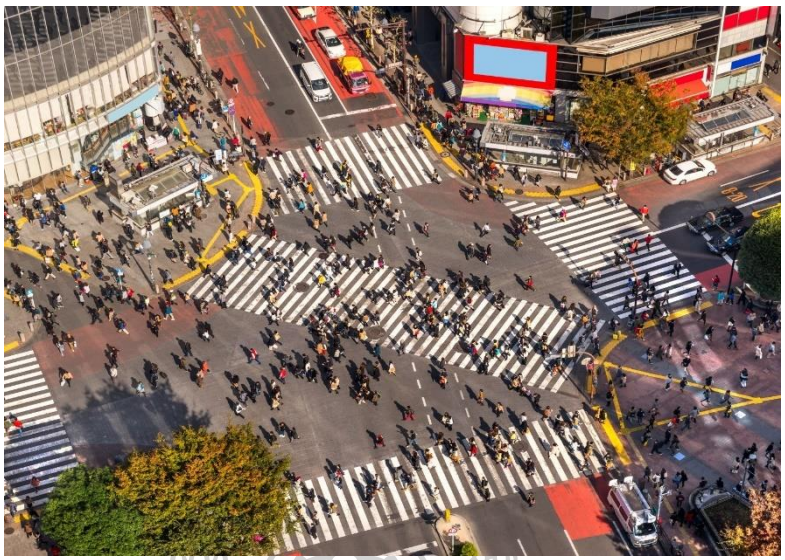
รูปที่ 4-136 ตัวอย่างรูปภาพที่หนึ่งร้อยสามสิบหกรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

จากรูปที่ 4-136 เห็นได้ว่ามีการบรรยายที่ไม่ดีเนื่องจากในรูปผู้หญิงไม่ได้ใส่ชุดสีแดง

“รถยนต์หลายคันบนถนนในเมืองแห่งหนึ่ง”

รูปที่ 4-137 ตัวอย่างรูปภาพที่หนึ่งร้อยสามสิบเจ็ดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

จากรูปที่ 4-137 เห็นได้ว่าโมเดลยังไม่สามารถบรรยายองค์ประกอบแบบเข้าใจโดยรวมได้ทั้งหมด



“ผู้คนจำนวนมากกำลังเดินข้ามทางม้าลายณเมืองแห่งหนึ่ง”

รูปที่ 4-138 ตัวอย่างรูปภาพที่หนึ่งร้อยสามสิบแปดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

4.10.2 ชุดข้อมูลทดสอบ Flickr8k ฝึกสอน 50 รอบ

“สุนัขสีน้ำตาลและสีขาวกำลังวิ่งอยู่บนพื้นหญ้า”

รูปที่ 4-139 ตัวอย่างรูปภาพที่หนึ่งร้อยสามสิบเก้ารวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“นักสโนว์บอร์ดกำลังกระโดดข้ามทางลาด”

รูปที่ 4-140 ตัวอย่างรูปภาพที่หนึ่งร้อยสี่สิบรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“นักโต้คลื่นกำลังโต้คลื่น”

รูปที่ 4-141 ตัวอย่างรูปภาพที่หนึ่งร้อยสี่สิบเอ็ดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“เด็กผู้หญิงสองคนกำลังเดินอยู่บนพื้นหญ้า”

รูปที่ 4-142 ตัวอย่างรูปภาพที่หนึ่งร้อยสี่สิบสองรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“นักโต้คลื่นในชุดดำน้ำสีดำกำลังโต้คลื่นใต้คลื่น”

รูปที่ 4-143 ตัวอย่างรูปภาพหนึ่งร้อยสี่สิบสามรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

จากรูป ที่ 4-143 เห็นได้ว่าโมเดลยังบรรยายออกมาได้ไม่ดีเนื่องจากมีรายละเอียดที่ผิดอยู่

“เด็กผู้ชายคนหนึ่งกำลังเล่นสเก็ตบอร์ดบนทางลาด”

รูปที่ 4-144 ตัวอย่างรูปภาพที่หนึ่งร้อยสี่สิบสี่รวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“นกสีขาวตัวใหญ่บินอยู่เหนือน้ำ”

รูปที่ 4-145 ตัวอย่างรูปภาพที่หนึ่งร้อยสี่สิบห้ารวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“คนกลุ่มหนึ่งนั่งอยู่บนโขดหิน”

รูปที่ 4-146 ตัวอย่างรูปภาพที่หนึ่งร้อยสี่สิบหกรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

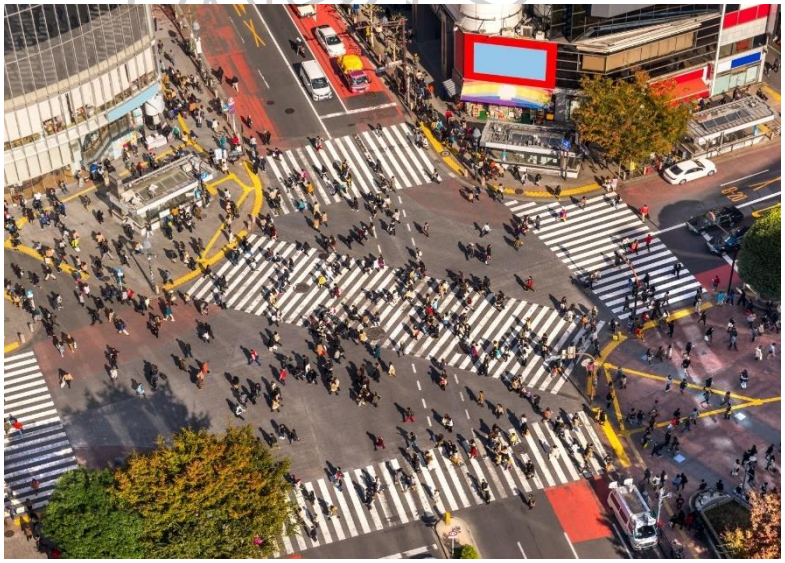
ตารางที่ 4-12 ค่าผลคะแนน BLEU ของชุดข้อมูลรวม Flickr8k และ ชุดข้อมูลการจราจรโมเดล MobileNetV2 + Bidirectional LSTM ที่ถูกฝึกสอน 50 รอบ

BLEU	Test
BLEU-1	0.614416
BLEU-2	0.438282
BLEU-3	0.326279
BLEU-4	0.238734

4.10.3 ชุดข้อมูลทดสอบที่เกี่ยวข้องกับการจราจรฝึกสอน 100รอบ

“รถยนต์หลายคันจอดติดไฟแดงอยู่บนถนน”

รูปที่ 4-147 ตัวอย่างรูปภาพที่หนึ่งร้อยสี่สิบเจ็ดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น



“ผู้คนจำนวนมากกำลังเดินข้ามถนนในเมือง”

รูปที่ 4-148 ตัวอย่างรูปภาพที่หนึ่งร้อยสี่สิบแปดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

4.10.4 ชุดข้อมูลทดสอบ Flickr8k ฝึกสอน 100 รอบ

“นักสโนว์บอร์ดกำลังกระโดดขึ้นไปในอากาศ”

รูปที่ 4-149 ตัวอย่างรูปภาพที่หนึ่งร้อยสี่สิบเก้ารวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“นักปั่นจักรยานกำลังกระโดดขึ้นไปในอากาศบนทางลาด”

รูปที่ 4-150 ตัวอย่างรูปภาพที่หนึ่งร้อยห้าสิบรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

4.10.5 ตัวอย่างชุดข้อมูลฝึกสอน Flickr8k และ ชุดข้อมูลการจราจร

“ผู้คนจำนวนมากกำลังเดินข้ามทางม้าลาย”

รูปที่ 4-151 ตัวอย่างรูปภาพที่หนึ่งร้อยห้าสิบเอ็ดรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

“เด็กชายคนหนึ่งแกว่งชิงช้า”

รูปที่ 4-152 ตัวอย่างรูปภาพที่หนึ่งร้อยห้าสิบสองรวมถึงคำบรรยายภาษาไทยที่โมเดลสร้างขึ้น

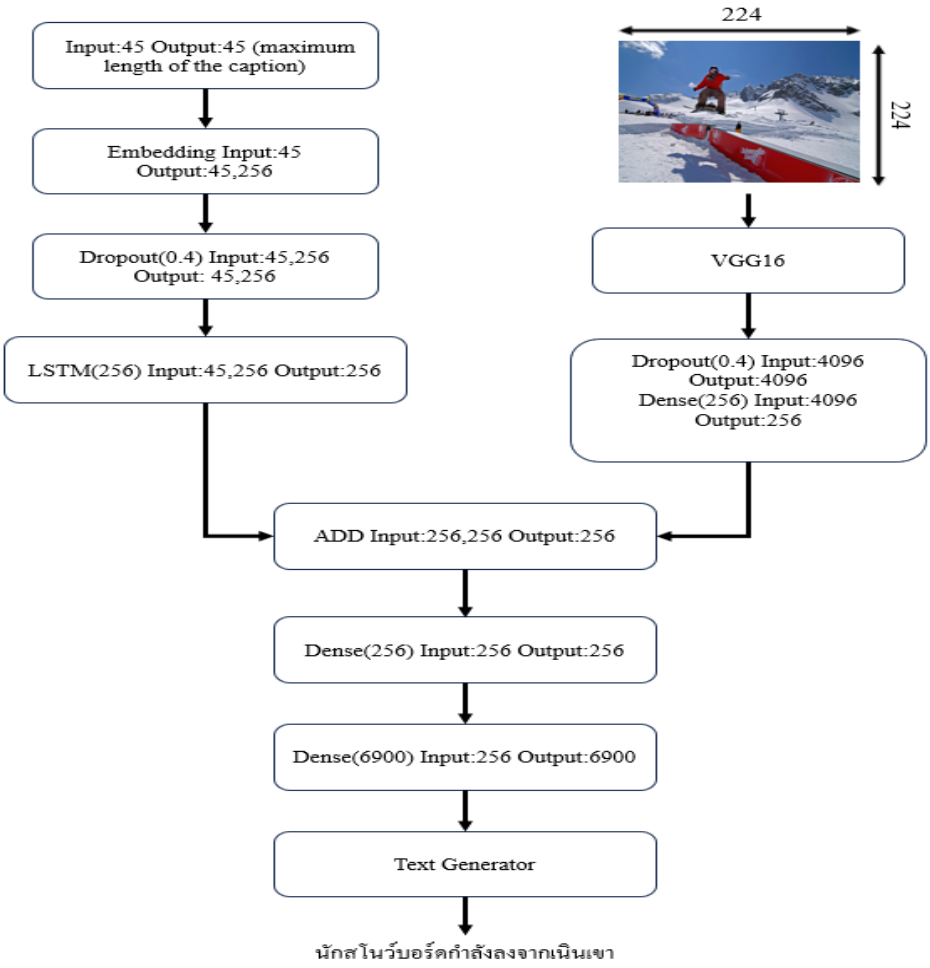
ตารางที่ 4-13 ค่าผลคะแนน BLEU ของชุดข้อมูลรวม Flickr8k และ ชุดข้อมูลการจราจรโมเดล MobileNetV2 + Bidirectional LSTM ที่ถูกฝึกสอน 100 รอบ

BLEU	Train(6816)	Test(1704)
BLEU-1	0.681907	0.622566
BLEU-2	0.527521	0.441552
BLEU-3	0.421062	0.326594
BLEU-4	0.332014	0.237952

4.11 โมเดลที่นำมาเปรียบเทียบ

```
Model: "model_1"
__________________________________________________________________________________________________
 Layer (type)                Output Shape              Param #   Connected to
==================================================================================================
 input_3 (InputLayer)        [(None, 70)]              0         []

 input_2 (InputLayer)        [(None, 4096)]            0         []

 embedding (Embedding)       (None, 70, 256)           1796608   ['input_3[0][0]']

 dropout (Dropout)           (None, 4096)              0         ['input_2[0][0]']

 dropout_1 (Dropout)         (None, 70, 256)           0         ['embedding[0][0]']

 dense (Dense)               (None, 256)               1048832   ['dropout[0][0]']

 lstm (LSTM)                 (None, 256)               525312    ['dropout_1[0][0]']

 add (Add)                   (None, 256)               0         ['dense[0][0]',
                                                                  'lstm[0][0]']

 dense_1 (Dense)             (None, 256)               65792     ['add[0][0]']

 dense_2 (Dense)             (None, 7018)              1803626   ['dense_1[0][0]']

==================================================================================================
Total params: 5240170 (19.99 MB)
Trainable params: 5240170 (19.99 MB)
Non-trainable params: 0 (0.00 Byte)
__________________________________________________________________________________________________
```

รูปที่ 4-153 โมเดลสรุปโมเดลที่นำมาเปรียบเทียบ



รูปที่ 4-154 โครงสร้างโมเดลที่นำมาเปรียบเทียบ

ตารางที่ 4-14 ค่าผลคะแนน BLEU ของชุดข้อมูลรวม Flickr8k และ ชุดข้อมูลการจราจรโมเดลที่นำมาเปรียบเทียบที่ถูกฝึกสอน 100 รอบ

BLEU	Train(6816)	Test(1704)
BLEU-1	0.665669	0.564001
BLEU-2	0.513675	0.379860
BLEU-3	0.414095	0.271847
BLEU-4	0.332521	0.192607

4.12 เปรียบเทียบวิธีการของงานวิจัยกับโมเดลที่นำมาเปรียบเทียบ

ตารางที่ 4-15 เปรียบเทียบโมเดลของงานวิจัยกับวิธีการอื่นในชุดทดสอบ

BLEU	BLEU-1	BLEU-2	BLEU-3	BLEU-4
VGG16 + Bidirectional LSTM 100 รอบ	**0.607605**	**0.420551**	**0.307364**	**0.221886**
ResNet50 + Bidirectional LSTM 100 รอบ	**0.611543**	**0.430566**	**0.319303**	**0.233104**
MobileNetV2 + Bidirectional LSTM 100 รอบ	**0.622566**	**0.441552**	**0.326594**	**0.237952**
Method From [24] 100รอบ	0.564001	0.379860	0.271847	0.192607

โมเดลที่เราออกแบบทั้งสามโมเดลได้ค่าคะแนนที่สูงกว่าโมเดลที่นำมาเปรียบเทียบแสดงให้เห็นถึงประสิทธิภาพในการพัฒนาโมเดล

ตารางที่ 4-16 เปรียบเทียบโมเดลของงานวิจัยกับโมเดลที่นำมาเปรียบเทียบของชุดฝึกสอน

BLEU	BLEU-1	BLEU-2	BLEU-3	BLEU-4
VGG16 + Bidirectional LSTM 100 รอบ	**0.698177**	**0.546698**	**0.440029**	**0.349506**
ResNet50 + Bidirectional LSTM 100รอบ	**0.669693**	**0.511792**	**0.404764**	**0.315446**
MobileNetV2 + Bidirectional LSTM 100 รอบ	**0.681907**	**0.527521**	**0.421062**	**0.332014**
Method From [24] 100 รอบ	0.665669	0.513675	0.414095	0.332521

ตารางที่ 4-17 เปรียบเทียบโมเดล CNN ที่ออกแบบเองกับโมเดลที่นำมาเปรียบเทียบของชุดทดสอบ

CNN ที่วางโครงสร้างเอง + Bidirectional LSTM 50รอบ	VGG16(ไม่มี imagenet) + LSTM 50รอบ	Resnet50(ไม่มี imagenet) + LSTM 50รอบ	MobilenetV2(ไม่มี imagenet) + LSTM 50รอบ
BLEU1 = 0.513043	BLEU1 = 0.482294	BLEU1 = 0.465117	BLEU1 = 0.433215
BLEU2 = 0.283934	BLEU2 = 0.271646	BLEU2 = 0.250106	BLEU2 = 0.229439
BLEU3 = 0.177309	BLEU3 = 0.175115	BLEU3 = 0.158585	BLEU3 = 0.142761
BLEU4 = 0.112540	BLEU4 = 0.113238	BLEU4 = 0.103232	BLEU4 = 0.089530

บทที่ 5
สรุปและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

วิทยานิพนธ์นี้ได้นำเสนอวิธีการพัฒนาการบรรยายภาพภาษาไทยที่ใช้ Convolutional Neural Network (CNN) สำหรับการแยกคุณลักษณะภาพอินพุต และ Bidirectional LSTM สำหรับสร้างคำบรรยายภาพตามรูปภาพและข้อความก่อนหน้า โดย CNN ที่ได้เลือกใช้นั้นมีทั้งหมด 3 ชนิด คือได้แก่ VGG16, ResNet50 และ MobileNetV2 ซึ่ง CNN ทั้ง 3 ชนิดนี้จะทำงานรวมกันกับ Bidirectional LSTM รวมถึงยังมีการวางโครงสร้างของ CNN เองแล้วนำไปทำการทดลองกับ Bidirectional LSTM เช่นกัน แล้วสุดท้ายจะประเมินผลลัพธ์การบรรยายด้วย BLEU โดยซึ่งค่า BLEU จะแสดงค่าคะแนนจาก 0-1 โดย 0 หมายถึงผลลัพธ์ในการบรรยายจากโมเดลไม่ตรงกับคำบรรยายอ้างอิงเลย และ 1 หมายถึงผลลัพธ์จากการบรรยายของโมเดลตรงกับคำบรรยายอ้างอิง

วิธีการทดลองโดยตอนแรกได้ทำการจัดเตรียมชุดข้อมูลสองFlickr8k อย่างเดียวก่อนโดยทำการทดลองกับ CNN ที่วางโครงสร้างเองกับBidirectional LSTM จากนั้นมีการรวมข้อมูลกันคือ Flickr8k และชุดข้อมูลการจราจรที่ได้จัดทำขึ้นเองรวมถึงมีการนำค่าน้ำหนักจาก ImageNet ที่เป็นฐานข้อมูลรูปภาพมากกว่า 14 ล้านรูปเข้ามาช่วยเสริมคุณลักษณะให้แก่ CNN ทั้ง 3 จากนั้นได้ทำการออกแบบโมเดลใหม่จากโมเดลที่นำมาต่อยอดโดยได้เพิ่มชั้น Dense หรือการทำ Fully Connected อีก 1 ชั้น ให้กับฝั่งของการเข้ารหัสรูปภาพรวมถึงได้กำหนด Kernel ใหม่ให้กับทุกชั้น และเปลี่ยนจากการใช้ LSTM แบบปกติไปใช้ Bidirectional LSTM ที่สามารถเรียนรู้หรือประมวลผลได้แบบสองทิศทางคือทิศทางไปข้างหน้าและย้อนกลับทำให้สามารถเรียนรู้คำที่มีความหมายใกล้เคียงกันและคำที่สะกดคล้ายกันแต่ความหมายไม่เหมือนกันได้

ในส่วนของชุดข้อมูลการทดลองสำหรับ Flickr8k เนื่องจากในฐานข้อมูลชุดนี้คำบรรยายเป็นภาษาอังกฤษทั้งหมดดังนั้นจึงต้องทำการแปลเป็นภาษาไทยโดยใช้ Google Translate ก่อน และยังต้องตัดอักขระพิเศษบางตัวออกด้วยเนื่องจากในชุดข้อมูลคำบรรยายเป็นภาษาอังกฤษดังนั้นคำบรรยายบางประโยคหรือบางคำอาจจะมี จุด Full Stop, Question Marks และอักขระพิเศษอื่นติดมาดังนั้นก่อนหรือหลังทำการแปลเป็นภาษาไทยจะต้องลบอักขระพิเศษเหล่านี้ทิ้งไปด้วยเนื่องจากในภาษาไทยไม่มีอักขระพวกนี้อยู่รวมถึงหากไม่ลบออกแล้วนำเข้าไปในโมเดลการฝึกสอนอาจทำให้โมเดลเรียนรู้สิ่งที่ผิดและการบรรยายผลลัพธ์จะออกมาผิดตามไปด้วย จากนั้นนำไปรวมกับชุดข้อมูลการจราจร และสิ่งสำคัญอีกอย่างสำหรับโมเดลคือก่อนที่จะนำข้อความภาษาไทยเข้าไปฝึกสอนจะไม่

สามารถนำข้อความที่เป็นภาษาไทยแบบเรียงชิดติดกันเข้าไปได้เลยพร้อมกันทีเดียวเนื่องจากโมเดลจะต้องมีการเข้ารหัสและถอดรหัสข้อความดังนั้นจึงต้องทำการแยกหรือตัดคำภาษาไทยก่อนโดยได้ใช้ไลบรารี่การตัดคำไทยที่มีชื่อว่า Deepcut ที่สามารถตัดคำไทยแยกออกเป็นคำเดี่ยวแต่ยังสามารถที่จะคงความหมายเดิมได้อยู่

โดยจากการทดลองโมเดล CNN + Bidirectional LSTM ทั้งสามได้แก่ VGG16 + Bidirectional LSTM, ResNet50 + Bidirectional LSTM และ MobileNetV2 + Bidirectional LSTM ในกรณีที่รัน 50และ100รอบสามารถบรรยายข้อความภาษาไทยได้ค่อนข้างตรงกับความหมายภายในรูปโดยได้ค่าคะแนน BLEU ที่สูงกว่าโมเดลที่นำมาเปรียบเทียบและพัฒนาในทั้งกรณีของการฝึกสอนและทดสอบ แต่หากสังเกตการบรรยายข้อความแล้วโมเดลที่ให้ผลลัพธ์การบรรยายได้ดีทั้งชุดข้อมูล Flickr8k และ ชุดข้อมูลการจราจรนั้นได้แก่ VGG16 + Bidirectional LSTM ส่วนโมเดลที่ให้ค่าเฉลี่ย BLEU เยอะที่สุดของทั้ง BLEU-1, BLEU-2, BLEU-3, BLEU-4 ได้แก่ MobileNetV2 + Bidirectional LSTM (ในกรณีที่ฝึกสอน 100 รอบ) และยังพบว่าในบางกรณีการบรรยายในแต่ละโมเดลจะให้ผลลัพธ์การบรรยายที่แตกต่างกันที่บางโมเดลสามารถอธิบายได้ค่อนข้างละเอียดกว่าอีกโมเดลในบางรูปสลับกัน และโมเดล CNN ที่วางโครงสร้างเองกับ Bidirectional LSTM ก็ให้ผล BLEU ที่สูงกว่าวิธีการที่นำมาเปรียบเทียบ

VGG16 + Bidirectional LSTM ที่ 100 รอบ = รถยนต์หลายคันกำลังเคลื่อนตัวเข้าไปในความมืด

ResNet50 + Bidirectional LSTM ที่ 100รอบ = การจราจรติดขัดเพราะหนาแน่นไปด้วยรถยนต์จำนวนมาก

รูปที่ 5-1 แสดงผลลัพธ์เปรียบเทียบการบรรยายที่หนึ่ง

VGG16 + Bidirectional LSTM ที่ 100 รอบ = นักโต้คลื่นกำลังโต้คลื่นในมหาสมุทร

ResNet50 + Bidirectional LSTM ที่ 100 รอบ = นักโต้คลื่นกำลังขี่คลื่น

รูปที่ 5-2 แสดงผลลัพธ์เปรียบเทียบการบรรยายที่สอง

5.2 ปัญหาและข้อเสนอแนะ

1. ในผลลัพธ์การบรรยายบางภาพโมเดลยังมีความเข้าใจผิดเกี่ยวกับองค์ประกอบในรูปภาพ

2. มีข้อจำกัดในเรื่องขององค์ประกอบที่เยอะเกินไปในแต่ละภาพเนื่องจากการบรรยายรูปภาพต้องอาศัยคุณลักษณะจำนวนมาก

5.3 แนวทางการพัฒนาต่อยอด

1. พัฒนาต่อยอดให้โมเดลสามารถบรรยายออกมาในรูปแบบเสียงได้

2. พัฒนาให้สามารถนำไปใช้ได้จริงกับการบรรยายให้กับผู้คนที่เห็นภาพทั่วไปในชีวิตประจำวัน

3. พัฒนาให้สามารถนำไปใช้ได้จริงกับการบรรยายให้กับผู้คนที่สัญจรบนท้องถนน

รายการอ้างอิง

1. สระอุบล, ผ.ด.ก., เรียนรู้ *AI: Deep Learning* ด้วย *Python*. 2565.
2. G, R. *Everything you need to know about VGG16*. 2021; Available from: https://medium.com/@mygreatlearning/everything-you-need-to-know-about-vgg16-7315defb5918.
3. datagen. *ResNet-50: The Basics and a Quick Tutorial*. Available from: https://datagen.tech/guides/computer-vision/resnet-50/.
4. Sharma, N. *What is MobileNetV2? Features, Architecture, Application and More*. 2023; Available from: https://www.analyticsvidhya.com/blog/2023/12/what-is-mobilenetv2/.
5. Anishnama. *Understanding Bidirectional LSTM for Sequential Data Processing*. 2023; Available from: https://medium.com/@anishnama20/understanding-bidirectional-lstm-for-sequential-data-processing-b83d6283befc.
6. GeeksforGeeks. *Bidirectional RNNs in NLP*. 2023; Available from: https://www.geeksforgeeks.org/bidirectional-rnns-in-nlp/?ref=ml_lbp.
7. Santhosh, S. *Understanding BLEU and ROUGE score for NLP evaluation*. 2023; Available from: https://medium.com/@sthanikamsanthosh1994/understanding-bleu-and-rouge-score-for-nlp-evaluation-1ab334ecadcb.
8. Kishore Papineni, S.R., Todd Ward, and Wei-Jing Zhu, *BLEU: a Method for Automatic Evaluation of Machine Translation*. 2002. p. 311-318.
9. Zhou Lei, C.Z., Shengbo Chen, Yiyong Huang and Xianrui Liu, *A Sparse Transformer-Based Approach for Image Captioning*. 2020, IEEE. p. 213437 - 213446.
10. Mookdarsanit, P.M.a.L., *Thai-IC: Thai Image Captioning based on CNN-RNN Architecture*. 2020. p. 40-45.
11. Md. Zakir Hossain, F.S., Mohd Fairuz Shiratuddin, Hamid Laga, *A Comprehensive Survey of Deep Learning for Image*

Captioning. 2018.

12. Bryan A. Plummer, L.W., Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier,

Svetlana Lazebnik, *Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models.* 2016.

13. Min Yang, W.Z., Wei Xu, Yabing Feng, Zhou Zhao, Xiaojun Chen, Kai Lei, *Multitask Learning for Cross-Domain Image Captioning*. IEEE. p. 1047 - 1061.
14. Sarin Watcharabutsarakham, S.M., Kantip Kiratiratanapruk, Pitchayagan Temniranrat, *Image Captioning for Thai Cultures*. 2022, IEEE.
15. Miao Xin, H.Z., Ding Yuan, Mingui Sun, *Learning discriminative action and context representations for action recognition in still images*. 2017, IEEE.
16. Junnan Li, D.L., Caiming Xiong, Steven Hoi, *BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation*. 2022.
17. Ron Mokady, A.H., Amit H. Bermano, *ClipCap: CLIP Prefix for Image Captioning.* 2021.
18. Hanan Nasser Alkalouti, M.A.A.M., *Encoder-Decoder Model for Automatic Video Captioning Using Yolo Algorithm*. 2021, IEEE.
19. Ling Cheng, W.W., Feida Zhu, Yong Liu, Chunyan Miao, *Geometry-Entangled Visual Semantic Transformer for Image Captioning*. 2021.
20. Dzmitry Bahdanau, K.C., Yoshua Bengio, *Neural Machine Translation by Jointly Learning to Align and Translate*. 2015, ICLR.
21. Allen Nie, R.C.-G., Christopher Potts, *Pragmatic Issue-Sensitive Image Captioning*. 2020.
22. Pengchuan Zhang, X.L., Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, Jianfeng Gao, *VinVL: Revisiting Visual Representations in Vision-Language Models*. 2021.
23. Karan Desai, J.J., *VirTex: Learning Visual Representations from Textual Annotations*.
24. aswintechguy. *Deep-Learning-Projects/Image Caption Generator - Flickr Dataset/*. Available from: https://github.com/aswintechguy/Deep-Learning-Projects/tree/main/Image%20Caption%20Generator%20-%20Flickr%20Dataset.

ภาคผนวก

ภาคผนวก ก

ผลงานที่ได้รับการตีพิมพ์

CERTIFICATE OF PRESENTATION

This is to certify that

Witchaphon Tieancho and Sopon Phumeechanya

has successfully presented a paper titled

Enhancing THAI Image Captioning Performance Using CNN and Bidirectional LSTM

at the ECTI-CON 2024
The 21st International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology
May 27-30 2024
KKU Science Park, Khon Kaen University,
Khon Kaen, THAILAND

Lunchakorn Wuttisittikulkij
Chulalongkorn University
TPC Chair

KHON KAEN CITY

ECTI-CON 2024
Khonkaen, Thailand

ECTI Association

IEEE THAILAND SECTION

ECTI-CON 2024
Khonkaen, Thailand

THE 21ST ECTI-CON 2024

MAY 27 - 30, 2024

KKU Science Park, Khon Kaen University, Khon Kaen, Thailand

WELCOME TO ECTI-CON 2024

The 21st International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology or ECTI-CON 2024 is the twenty-first annual international conference organized by Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI) Association, Thailand. The conference aims to provide an international platform to present technological advances, launch new ideas and showcase research work in the field of electrical engineering, electronics, computer, telecommunications and information technology.

TOPICS

- Circuits & Systems
- Communications
- Control Systems
- Medicine & Biology
- Industry Applications
- Robotics and Automation
- Antennas, Electromagnetics, RF/Microwave Components and Circuits
- Computer & Computational Intelligence
- Education
- Geoscience & Remote Sensing
- Energy & Power Electronics
- Signal Processing

PROCEEDINGS

The accepted papers will be submitted for inclusion in IEEE Xplore subject to meeting IEEE Xplore's socpe and quality requriements.

IMPORTANT DATES

Special session proposal	**OCTOBER 31, 2023**
Paper submission deadline	**~~DECEMBER 30, 2023~~ JAN 31, 2024**
Notification of acceptance	**APRIL 12, 2024**
Camera-ready paper submission	**APRIL 19, 2024**

CONTACT US

Email: ecti-con2024@kku.ac.th
Website: https://ecti-con2024.kku.ac.th/
Conference Record: #60892

Enhancing THAI Image Captioning Performance using CNN and Bidirectional LSTM

Witchaphon Tieancho
Department of Electrical Engineering
Faculty of Engineering and Industrial Technology
Silpakorn University
Nakhon Pathom, Thailand
tieancho_w@su.ac.th

Sopon Phumeechanya
Department of Electrical Engineering
Faculty of Engineering and Industrial Technology
Silpakorn University
Nakhon Pathom, Thailand
phumeechanya_s@su.ac.th*
*Corresponding author

***Abstract*— This research has designed a deep learning model to create Thai captions using a convolutional neural network (CNN) in VGG16 format to extract image features, and it is used to procreate captions using bidirectional LSTM. The data warehouse used for training and testing is Flickr8k, which combines customized traffic-related image and caption information. For the first set of data, that is Flickr8k, all subtitles had to be translated from English to Thai using Google Translate, and ways to deal with the data before training were to remove special characters to prevent the Thai language description from being distorted. Then, to evaluate the result of the captions the model produced compared to default captions, the BLEU metric was used to measure the score. The resulting average score was effective because it was higher than the compared models. The score values were paralleled up to 4 grams.**

Keywords—Thai captions, convolutional neural network, bidirectional LSTM, BLEU

I. INTRODUCTION

Image captioning is work related to the description of images. It has linked vision and language to create captions. The main principle of it is to disseminate the elements of the image as much as possible. This is a field of study that is becoming progressively more prevalent right now. Most of the work is in the field of English narrative; it has very little in other languages. We perceived the importance of this; therefore, we came up with the idea of creating Thai image captioning. We recognized the study which we did could be further developed in the future, such as making it into a notification system, so we included a traffic data set because we thought that just presenting the current data set might not be sufficient.

We conduct our lectures with the Flickr8k [1] and custom-collected traffic datasets, respectively. First, visual information and lecture information in English must be translated into Thai using Google Translate [2]. Second, the canvas observation and recital datum will be manually written. By combining the two datasets together, we train them using a generated image captioning model and display the results. However, traffic datasets were generated because we recognized that research in this area could be applied in a variety of ways, such as when the model receives images of people walking across the road, an audible warning can be issued to drivers to be cautious, but this research does not extend beyond creating alerts.

In this paper, we designed our own model using a convolutional neural network (CNN) to extract image features and a bidirectional LSTM to generate subtitles. This is an exceptionally significant process. Finally, the BLEU [3] score is to be used to measure the accuracy of the output compared to the original dissertation. It evaluates all the commentary that is part of the test and illustrates the average consequence. The values that came out had a higher assign score than the primitive model.

II. RELATED WORK

Before designing the Thai language caption model, we analyzed style prototypes from several research papers on English description. Desai and Johnson [4] presented Virtex with a pre-training tutorial using multiple captions to learn how to reveal images and train a convolutional network from the beginning, and then transferred it to downstream recognition. Later, Bahdanau et al. [5] proposed that neural machine translation is a new process for machine translation in which they proposed a strategy for expanding the bottleneck by having the model automatically search for sections of the source sentence, as encoding the source sentence into a fixed-length vector has been determined to be the bottleneck. Also, the dataset used in training is flickr8k [1], and we used imagenet [6] for the pre-training CNN model. After that, Li et al. [7] exhibited BLIP, a new VLP outline that pre-trains multiple encoders and decoders using bootstrapped datasets. Cheng et al. [8] presented a maiden design named Geometry-Entangled Visual Semantic Transformer (GEVST). It consists of four parallel encoders: VV (pure visual), VS (semantic fused to visual), SV (visual fused to semantic), SS (pure semantic) for creating the final subtitles where geometric properties are used visually and semantically in the fusion module and self-attention module.

Currently, there are multiple research studies regarding the creation of captions for images that can be studied, as well as different methods. Like the work of Mokady et al. [9], they used CLIP encryption as the subtitle prefix and a mapping system, then adjusted the language model to create a description. The main idea of this research is to combine the GPT2 pre-trained language model, where the results of the method can effectively generate rich and significant captions. Later, Zhang et al. [10] showed the details of a method to improve visualization for visual language tasks and to develop an object detection model that provides object-centered visualization. They intend to propose that visual features are highly significant in VL models. Nie et al. [11] proffered receptive issues related to image captioning, such as the minor problems that arise in the images.

In addition to the previous studies, there is additional study that is mostly used to establish image captioning models. O'Shea and Nash [12] presented an introduction to

979-8-3503-8155-9/24/$31.00 ©2024 IEEE

convolutional neural network (CNN). They concluded that CNN differs from neural networks in that it focuses on the entire problem domain and knowledge about specific inputs is retained. Staudemeyer and Rothstein Morris [13] stated Long Short-Term Memory is a very powerful ambulatory classifier, including how LSTM-RNN is improved and LSTM self-resetting has been introduced that can increase the available space memory of extraneous information. Mahadevaswamy and Swathi [14] summarized the technical information on sentiment analysis using bidirectional LSTM. Onward, Schuster and Paliwal [15] proposed an amplified RNN for a bidirectional recurrent neural network that could be instructed without limitations on the input data to predefined future frames by training simultaneously in both the forward and backward directions. The bidirectional structure can be easily adjusted to allow a good estimation of conditional exploration feasibility.

Mookdarsanit et al. [16] used a deep learning model to create Thai captions using CNN encoding in VGG16 to extract objects in the image as well as using RNN that have a long short-term memory (LSTM) inside to compose Thai sentences to match the meaning of the pictures, and then the image description was evaluated by the BLEU score. Watcharabutsarakham et al. [17] created an interesting model training structure using Inceptionv3 and an LSTM that generates captions through opportunistic search. Papineni et al. [18] proposed an automatic machine translation estimation method called BLEU. Another significant point is to acknowledge the work of [19], who designed a captivating captioning framework using CNN and LSTM in the method. It demonstrated how to design model structures and make comparisons.

III. The Proposed Method

This section describes the content used in designing the structure of the research and various methods for preparing and manipulating data.

A. Datasets

We have taken the dataset from flickr8k [1]; it consists of a pair of pictures and English subtitles. There are 8091 photos in total, and each photo has five different captions. Combine them with our custom-made traffic image-related data set, which contains 429 images and 5 different Thai subtitles for each image as well. So, there were 8520 images in total and 42600 captions.

B. Data Preparation

Before training the data, we observe the dataset in Flickr8k [1] that mostly consists of images of dogs running on grass, people skiing down mountains, racing cars on the race track, girls playing on the grass, etc. As a result, when adding traffic information, we must provide not only traffic-related elements, but also other relevant elements in detail, such as girls, boys, old people, colors, and so on. Because not only does it add features, but it also reinforces repeated learning. In addition, the captions that accompany Flickr8k images are in English. We have translated all captions into Thai using Google Translate [2]. As for the traffic data set with Thai subtitles, we wrote it ourselves. However, there are several features to consider while translating English captions into Thai for the Flickr8k dataset, such as the usage of Full Stops, Question Marks, Exclamation, Commas, Colons, Apostrophes, Semicolons, Hyphens, and Quotation Marks. Even after translating them into Thai, they still do not disappear. Therefore, before or after translating, we have deleted those things because they affect training and make Thai sentences different. An example of the English subtitles translated into Thai from Fig. 1 is shown in Fig. 3.

1.A lady wearing a helmet holding a bike.
2.A woman in a blue vest and a sky blue helmet stands with her bicycle in traffic.
3.A woman in a helmet rides her bike behind a car.
4.A woman with a helmet and a backpack walks next to her bike.
5.Women with bike and a helmet wait for traffic.

Fig. 1. Example images and captions from the Flickr8k dataset.

1.รถจักรยานยนต์จำนวนมากจอดติดไฟแดงบนถนน
2.รถจักรยานยนต์หลายคันและรถยนต์จำนวนหนึ่งกำลังจอดรอสัญญาณไฟบนถนน
3.รถจำนวนมากจอดรอสัญญาณไฟบนท้องถนน
4.รถจักรยานยนต์และรถยนต์จำนวนมากไม่สามารถเคลื่อนตัวได้เพราะติดไฟแดง
5.รถจักรยานยนต์และรถยนต์จำนวนมากบนท้องถนน

Fig. 2. Example images and captions of custom traffic datasets

1.A lady wearing a helmet holding a bike.
2.A woman in a blue vest and a sky blue helmet stands with her bicycle in traffic.
3.A woman in a helmet rides her bike behind a car.
4.A woman with a helmet and a backpack walks next to her bike.
5.Women with bike and a helmet wait for traffic.

↓

1.ผู้หญิงสวมหมวกกันน็อคจูงจักรยาน
2.ผู้หญิงในเสื้อกั๊กสีน้ำเงินและหมวกกันน็อคสีฟ้ายืนอยู่กับจักรยานของเธอท่ามกลางการจราจร
3.ผู้หญิงสวมหมวกกันน็อคขี่จักรยานอยู่หลังรถยนต์
4.ผู้หญิงสวมหมวกกันน็อคและเป้สะพายหลังเดินข้างจักรยาน
5.ผู้หญิงที่มีจักรยานและหมวกกันน็อครอการจราจร

Fig. 3. Example caption is translated by Google Translate.

C. Model Development Design

The architecture of the caption extraction model in this research uses VGG16 to encode images to extract a total of 4096 features and it also has a total of 16 main layers, it accepts RGB color image of size 224×224 as input data. All the extracted features were combined with training Bidirectional LSTM later. Before combining them, we designed them to have a layer of one dropout layer and two dense layers on the image encoding side. On the other side, bidirectional LSTM also receives values from embedding, where it internally receives a maximum caption length of 70 and vocabulary size of 7018. Then the input of images and text was combined to decode, and the last two density layers were defined before importing the text generator function to create captions. In which the activation function of Dense is used on all layers as Relu with the exception of the last layer, which is used as Softmax. As for compiling the model, set the loss function to categorical_crossentropy and the optimizer to adam.

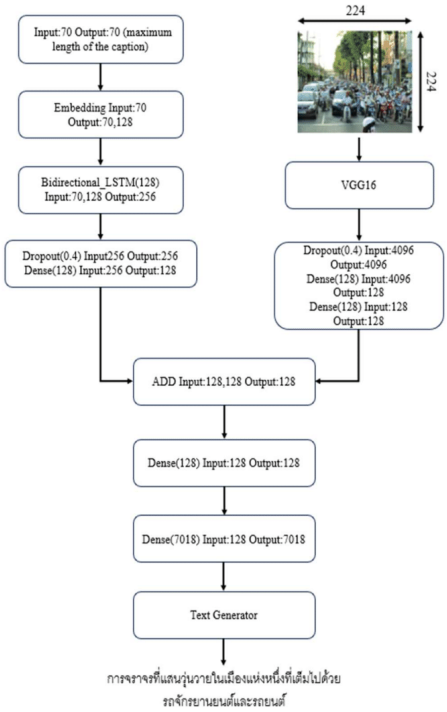


Fig. 4. The model architecture designed and developed.

IV. Experiments

Before doing the image and text training, we divided the data into train 80% and test 20%, respectively. The first consists of non-random Flickr8k images combined with a selected traffic dataset to be used for training. The training is carried out on an A100 GPU. This is because our model has a large number of images and text, with each caption being quite detailed and long. As a result, appropriate resources that give good processing speed must be chosen, and the training model was set at 100 epochs, using the same number of epochs as the model being compared with [19]. Then, once the model completes training, we evaluate the caption which it produces against the original five captions using the BLEU metric. The results are shown in Table 1.

TABLE I. The results of mean BLEU score values of training and testing sets.

	Our method		Method from [19]	
	Train	Test	Train	Test
BLEU-1	**0.698177**	**0.607605**	0.665669	0.564001
BLEU-2	**0.546698**	**0.420551**	0.513675	0.379860
BLEU-3	**0.440029**	**0.307364**	0.414095	0.271847
BLEU-4	**0.349506**	**0.221886**	0.332521	0.192607

From TABLE I, it can be seen that our BLEU average score is higher than the compared models. This shows that the model we have designed performs well when the training parameters are set as described and our highlight is using bidirectional LSTM instead of the normal LSTM by [19]. Where the range of its score will be in the range of 0 to 1. Where 0 means no score at all or the predicted sentence does not match the real sentence, a value of 1 means that the anticipated sentence matches every veracious clause. The n-gram of BLEU is measured at a maximum of 4 grams and has a set weight value in each gram to be (1,0,0,0), (0.5,0.5,0,0), (0.33,0.33,0.33,0) and (0.25.0.25,0.25,0.25) respectively. There are sample images and captions from the test set, according to Fig.5 to Fig.14. Including showing the equation for calculating the BLEU score by (1) from [18],

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \qquad (1)$$

where BP is brevity penalty, w_n is the weight film based on N, N is n-gram sequence and p_n is a n-gram precision.

“สัญญาณไฟจราจรสีแดงเตือนให้รถยนต์และรถจักรยานยนต์ทุกคันต้องหยุด”

Fig. 5. Example 1: Test image and caption related to traffic.

“ผู้หญิงคนหนึ่งกำลังเดินข้ามทางม้าลาย”

Fig. 6. Example 2: Test image and caption related to traffic.

“การจราจรที่ติดขัดบนท้องถนนแห่งหนึ่ง”

Fig. 7. Example 3: Test image and caption related to traffic.

“ผู้คนจำนวนมากกำลังข้ามทางม้าลายในเมืองใหญ่แห่งหนึ่ง”

Fig. 8. Example 4: Test image and caption related to traffic.

“การจราจรที่วุ่นวายบนท้องถนน”

Fig. 9. Example 5: Test image and caption related to traffic.

“ชายคนหนึ่งปีนหน้าผา”

Fig. 10. Example 1: Test image and caption from Flickr8k.

“ชายคนหนึ่งกำลังโต้คลื่นในมหาสมุทร”

Fig. 11. Example 2: Test image and caption from Flickr8k.

“นักปั่นจักรยานกำลังขี่ไปตามถนน”

Fig. 12. Example 3: Test image and caption from Flickr8k.

“นักสเก็ตบอร์ดกำลังเล่นสเก็ตบอร์ด”

Fig. 13. Example 4: Test image and caption from Flickr8k.

Fig. 14. Example 5: Test image and caption from Flickr8k.

The result shows success in the visualization description because it can be communicated to create understanding, but there are some consequences that do not represent enough details in the explanation from the images, such as Fig. 12. and Fig. 13., which we hope will have particulars related to sex, is also described, but none are available.

V. CONCLUSION

The operation of creating Thai captions in this research focuses on the design and experimentation of deep learning-based architectures that include CNN and bidirectional LSTM. By experimenting with Flickr8k combined with a custom traffic dataset, the BLEU metric is used to measure the score of the model compared to the original captions, both for testing and training. The results from the experiment are satisfactory because the average BLEU scores of the model test set in this research are 0.607605, 0.420551, 0.307364, and 0.221886, respectively, which are higher than the average scores of the model that we compared, but there is still a part that needs to be careful. When translating English captions from the Flickr8k dataset into Thai using Google Translate, certain characters must be removed before or after translation, such as Full Stop, Question Marks, Exclamation, Commas, Colons, Apostrophes, Semicolons, Hyphens, and Quotation Marks because the Thai description was mistranslated. Finally, for this research, we appreciate [19] for designing the image caption creation model that was used as a prototype for this research to develop and expand upon.

ACKNOWLEDGMENT

The authors would like to acknowledge the contribution of Department of Electrical Engineering, Faculty of Engineering and Industrial Technology, Silpakorn University Scholarship for funding this research.

REFERENCES

[1] “Flickr 8k Dataset”, kaggle[Online]. https://www.kaggle.com/datasets /adityajn105/flickr8k. (Accessed: Jan. 26, 2024).

[2] “translate 3.1.0”, PyPI [Online]. https://pypi.org/project/translate/3.1. 0/. (Accessed: Jan. 26, 2024).

[3] “bleu 0.3”, PyPI [Online]. https://pypi.org/project/bleu/. (Accessed: Jan. 26, 2024).

[4] K. Desai and J. Johnson, “VirTex: Learning Visual Representations from Textual Annotations,” 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, pp. 11157-11168.

[5] D. Bahdanau, K. Cho and Y. Bengio, “NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE.” arXiv:1409.0473, 2016, pp. 1-15.

[6] “Imagenet”, Image Net [Online]. https://www.image-net.org/. (Accessed: Jan. 26, 2024).

[7] J. Li, D. Li, C. Xiong and S. Hoi, “BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation.” arXiv:2201.12086, 2022.

[8] L. Cheng, W. Wei, F. Zhu, Y. Liu and C. Miao, “Geometry-Entangled Visual Semantic Transformer for Image Captioning.” arXiv:2109.14137, 2021.

[9] R. Mokady, A. Hertz, and A. H. Bermano, “ClipCap: CLIP Prefix for Image Captioning.” arXiv:2111.09734, Nov 2021.

[10] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi and J. Gao, “ VinVL: Revisiting Visual Representations in Vision-Language Models,” IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 5575-5584.

[11] A. Nie, R. Cohn-Gordon and C. Potts, “Pragmatic Issue-Sensitive Image Captioning.” arXiv:2004.14451, 2020.

[12] K. O’Shea, R. Nash, “An Introduction to Convolutional Neural Networks.” arXiv:1511.08458, 2015.

[13] R. C. Staudemeyer, E. Rothstein Morris, “Understanding LSTM – a tutorial into Long Short-Term Memory Recurrent Neural Networks.” arXiv:1909.09586, 2019.

[14] U. B. Mahadevaswamy, P. Swathi, “Sentiment Analysis using Bidirectional LSTM Network.” Procedia Computer Science, Volume 218, 2023, pp. 45-56.

[15] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” IEEE Transactions on Signal Processing, vol. 45, no.11, 1997, pp. 2673-2681.

[16] P. Mookdarsanit, L. Mookdarsanit, “Thai-IC: Thai Image Captioning based on CNN-RNN Architecture.” International Journal of Applied Computer Technology and Information Systems, vol. 10, No. 1, 2020.

[17] S. Watcharabutsarakham, S. Marukatat, K. Kiratiratanapruk and P. Temniranrat, “Image Captioning for Thai Cultures,” 2022 17th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), Chiang Mai, Thailand, 2022, pp. 1-5.

[18] K. Papineni, S. Roukos, T. Ward and W. Jing Zhu, “BLEU: a Method for Automatic Evaluation of Machine Translation.” Proceedings of the 40th Annual Meeting of the Association for Computational (ACL), Philadelphia, July 2002, pp. 311-318.

[19] “Deep-Learning-Projects/Image Caption Generator – Flickr Dataset”, github [Online]. https://github.com/aswintechguy/Deep-Learning-Projects/tree/main/Image%20Caption%20Generator%20-%20Flickr%20Dataset. (Accessed: Jan26, 2024).

ประวัติผู้เขียน

ชื่อ-สกุล	วิชญ์พล เทียนชอ
วุฒิการศึกษา	วศ.บ.วิศวกรรมอิเล็กทรอนิกส์และระบบคอมพิวเตอร์ ภาควิชาวิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์และเทคโนโลยีอุตสาหกรรม มหาวิทยาลัยศิลปากร (2562)
ผลงานตีพิมพ์	Witchaphon Tieancho and Sopon Phumeechanya, "Enhancing THAI Image Captioning Performance Using CNN and Bidirectional LSTM" THE 21st INTERNATIONAL CONFERENCE ON ELECTRICAL ENGINEERING/ELECTRONICS, COMPUTER, TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY (ECTI-CON), KHONKAEN, THAILAND, 2024