การเรียนรู้เชิงลึกสำหรับการตรวจจับและรู้จำคำบรรยายในวีดิทัศน์

วิทยานิพนธ์

ของ

ธนดล สิงขรอาสน์

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

พฤศจิกายน 2564

การเรียนรู้เชิงลึกสำหรับการตรวจจับและรู้จำคำบรรยายในวีดิทัศน์

วิทยานิพนธ์

ของ

ธนดล สิงขรอาสน์

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

พฤศจิกายน 2564

DEEP LEARNING FOR VIDEO SUBTITLE DETECTION AND RECOGNITION

Thanadol Singkhornart

A Thesis Submitted in Partial Fulfillment of Requirements

for Master of Science (Information Technology)

November 2021

คณะกรรมการสอบวิทยานิพนธ์ ได้พิจารณาวิทยานิพนธ์ของนายธนดล สิงขรอาสน์ แล้วเห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ของมหาวิทยาลัยมหาสารคาม

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(รศ. ดร. ธัชพงศ์ กตัญญูกุล)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผศ. ดร. โอฬาริก สุรินต๊ะ)

.....กรรมการ
(ผศ. ดร. สาธิต แสงประดิษฐ์)

.....กรรมการ
(ผศ. ดร. แกมกาญจน์ สมประเสริฐศรี)

มหาวิทยาลัยอนุมัติให้รับวิทยานิพนธ์ฉบับนี้ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ของมหาวิทยาลัยมหาสารคาม

.....	.....
(ผศ. ศศิธร แก้วมั่น)	(รศ. ดร. กริสน์ ชัยมูล)
คณบดีคณะวิทยาการสารสนเทศ	คณบดีบัณฑิตวิทยาลัย

ชื่อเรื่อง การเรียนรู้เชิงลึกสำหรับการตรวจจับและรู้จำคำบรรยายในวีดิทัศน์
ผู้วิจัย ธนดล สิงขรอาสน์
อาจารย์ที่ปรึกษา ผู้ช่วยศาสตราจารย์ ดร. โอฬาริก สุรินต๊ะ
ปริญญา วิทยาศาสตรมหาบัณฑิต **สาขาวิชา** เทคโนโลยีสารสนเทศ
มหาวิทยาลัย มหาวิทยาลัยมหาสารคาม **ปีที่พิมพ์** 2564

บทคัดย่อ

ในปัจจุบันมีวีดิทัศน์จำนวนมากที่ถูกเผยแพร่ผ่านอินเทอร์เน็ตในช่องทางต่าง ๆ เช่น Youtube และ Facebook มีผู้ชมบางส่วนที่มีปัญหาในการรับรู้ข้อมูลจากวีดิทัศน์เนื่องจากปัญหาทางด้านภาษาหรือมีปัญหาด้านการฟัง ดังนั้นคำบรรยายจึงถูกเพิ่มเข้ามาในวีดิทัศน์ ในวิทยานิพนธ์นี้ได้นำเสนอถึงการนำวิธีการเรียนรู้เชิงลึกมาใช้โดยใช้วิธีโครงข่ายประสาทแบบคอนโวลูชัน (CNN) ร่วมกับ วิธีหน่วยความจำระยะสั้นระยะยาว (LSTM) ซึ่งเรียกว่า CNN-LSTM เพื่อที่จะนำมารู้จำคำบรรยายจากวีดิทัศน์ เราได้สร้างตัวอย่างต้นแบบ CNN ที่มีจำนวน 16 ชั้น โดยชั้นสุดท้ายเป็น การย่อขนาดโดยใช้ค่าสูงสุด (Max-pooling) ก่อนที่จะส่งเข้า LSTM โดยในการเรียนรู้นั้นเราได้ใช้รูปภาพคำบรรยายที่มีรูปแบบ ขนาด และพื้นหลังที่หลากหลาย แล้วใช้ การจำแนกการเชื่อมต่อชั่วคราว (CTC loss) ในการคำนวนหาค่า loss และถอดรหัสเป็นผลลัพธ์ สำหรับข้อมูลที่นำมาใช้ในการเรียนรู้นั้นได้มาจากการรวมรวม 24 วีดิทัศน์จาก Youtube และ Facebook ที่มีคำบรรยายภาษาไทย อังกฤษ ตัวเลขไทยและตัวเลขอารบิก ซึ่งมีทั้งหมด 157 ตัวเพื่อนำมาถอดรหัสข้อมูลในชุดรูปภาพนั้นมีทั้งหมด 4,224 รูป ซึ่งได้ค่าเฉลี่ยความผิดพลาดที่น้อยที่สุดคือ 11.06%

คำสำคัญ : การรู้จำคำบรรรยายวีดิทัศน์, โครงข่ายประสาทคอนโวลูชัน, หน่วยความจำระยะสั้นระยะยาว, การจำแนกการเชื่อมต่อชั่วคราว

TITLE	DEEP LEARNING FOR VIDEO SUBTITLE DETECTION AND RECOGNITION		
AUTHOR	Thanadol Singkhornart		
ADVISORS	Assistant Professor Olarik Surinta , Ph.D.		
DEGREE	Master of Science	**MAJOR**	Information Technology
UNIVERSITY	Mahasarakham University	**YEAR**	2021

ABSTRACT

Nowadays, many videos have been published on Internet channels such as Youtube and Facebook. Many audiences, however, cannot understand the contents of the video, maybe due to the different languages and even hearing impairment. As a result, subtitles have been added to videos. In this paper, we proposed deep learning techniques, which were the combination between convolutional neural networks (CNN) and long short-term memory (LSTM) networks, called CNN-LSTM, to recognize video subtitles. We created the simplified CNN architecture with 16 weight layers. The last layer of the CNN was downsampling using max-pooling before sending it to the LSTM network. We first trained our CNN-LSTM architecture on printed text data which contained various font styles, diverse font sizes, and complicated backgrounds. The connectionist temporal classification was then used as a loss function to calculate the loss value and decode the output of the network. For the video subtitle dataset, we collected 24 videos from Youtube and Facebook, consisting of Thai, English, Arabic, and Thai numbers. The dataset also contained 157 characters. In this dataset, we extracted 4,224 subtitle images from the videos. The proposed CNN-LSTM architecture achieved an average character error rate of 11.06%.

Keyword : Video subtitle text recognition, Convolutional neural networks, long short-term memory network, Connectionist temporal classification

กิตติกรรมประกาศ

วิทยานิพนธ์นี้สำเร็จได้ด้วยความกรุณาของอาจารย์โอฬาริก สุรินต๊ะ อาจารย์ที่ปรึกษาวิทยานิพนธ์ที่ได้ให้คำเสนอแนะ แนวคิดตลอดจนแก้ไขข้อบกพร่องต่าง ๆ มาโดยตลอด จนวิทยานิพนธ์เล่มนี้เสร็จสมบูรณ์ ขอกราบขอบพระคุณ ณ ที่นี้

ธนดล สิงขรอาสน์

สารบัญ

หน้า

สารบัญภาพ

สารบัญตาราง

บทที่ 1
บทนำ

1.1 ความเป็นมาของการวิจัย

สื่อประเภทวีดิทัศน์ได้ถูกเผยแพร่ในหลายช่องทางเช่น ยูทูป (Youtube) เฟซบุ๊ก (Facebook) และอินสตาแกรม (Instagram) เป็นต้น ทำให้ผู้ชมสามารถเลือกชมได้อย่างอิสระ ทั้งนี้เพื่อให้เข้าถึงผู้ชมได้หลากหลาย เช่น ชาวต่างชาติ และผู้พิการทางการได้ยิน ทำให้ผู้ผลิตวีดิทัศน์เพิ่มคำบรรยาย (Subtitle) ลงไปในวีดิทัศน์ ทำให้เกิดประโยชน์ต่อผู้ชมที่จะได้รับรู้เนื้อหาที่ถูกต้อง และยังเป็นการเพิ่มจำนวนคนเข้าชมวีดิทัศน์ ทั้งนี้ยังส่งผลให้ผู้ผลิตได้รับรายได้จากการเข้าชมวีดิทัศน์ ดังนั้นการเพิ่มคำบรรยายที่อยู่ในวีดิทัศน์สามารถเป็นได้ทั้งคำบรรยายที่ที่แสดงออกมาตามคำพูดของผู้พูดในวีดิทัศน์และ เป็นคำบรรยายที่ปรากฏขึ้นในจุดที่น่าสนใจ

ทั้งนี้ โดยส่วนมากคำบรรยายที่ปรากฏบริเวณด้านล่างของวีดิทัศน์จะเป็นแบบที่แสดงออกมาตามเสียงของผู้พูด หรือผู้บรรยายในวีดิทัศน์ ทำให้สามารถใช้วิธีการทางด้านการตรวจจับตัวอักษร (Text Detection) เพื่อตรวจจับ บริเวณที่เป็นคำบรรยายที่ปรากฏในวีดิทัศน์ และใช้วิธีการรู้จำเสียง (Speech Recognition) [1] ซึ่งเป็นงานวิจัยทาง ด้านปัญญาประดิษฐ์ (Artificial Intelligence) ดังนั้น ในวิทยานิพนธ์ฉบับนี้ ได้มุ่งเน้นในส่วนของการตรวจจับคำบรรยาย (Subtitle Detection) [2] เพื่อตรวจจับคำบรรยายทั้งแบบที่ปรากฏบริเวณด้านล่างของวีดิทัศน์และปรากฏขึ้นในบริเวณที่น่าสนใจ (ดังภาพประกอบที่ 1) ซึ่งเป็นงานวิจัยที่เกี่ยวข้องกับรูปภาพ (Image) ผลลัพธ์ที่ได้จากการตรวจจับตัวอักษร นอกจากการตรวจจับคำบรรยายที่ปรากฏในวีดิทัศน์ วิทยานิพนธ์ฉบับนี้ ได้ทำวิจัยในส่วนของการรู้จำคำบรรยาย (Subtitle Recognition) ซึ่งเป็นกระบวนการแปลงรูปภาพตัวอักษรให้เป็นข้อความที่อยู่ในรูปแบบของตัวอักษร (Text)

จากที่ได้กล่าวมาข้างต้น ในวิทยานิพนธ์นี้มุ่งเน้นในการนำการเรียนรู้เชิงลึก (Deep Learning) มาเพื่อช่วยในการตรวจจับและรู้จำคำบรรยายวีดิทัศน์ เนื่องจากเป็นวิธีการที่มีความถูกต้องและแม่นยำ ดังนั้น ในกระบวนการของการตรวจจับคำบรรยายวีดิทัศน์ ได้ทดสอบกับ 3 วิธี ประกอบด้วย You only look once version 3 (YoloV3) [3] , Tiny-YOLOv3 [3] และ RetinaNet [4] สำหรับการรู้จำคำบรรยายวีดิทัศน์ ได้เลือกใช้วิธี Convolutional Neural Networks (CNN) ร่วมกับ Long short-Term Memory (LSTM) [5] และ Gated Recurrent Unit (GRU) [6] โดยข้อมูลวีดิทัศน์ที่ใช้ในวิทยานิพนธ์นี้ประกอบไปด้วยวีดิทัศน์ที่มีคำบรรยายเป็นภาษาไทย ภาษาอังกฤษ ตัวเลขไทย และตัวเลขอารบิก ซึ่งได้เก็บรวบรวมจากวีดิทัศน์จำนวนทั้งสิ้น 24 วีดิทัศน์ โดยการวัดประสิทธิภาพในการตรวจจับคำบรรยายวีดิทัศน์ใช้ค่าอัตราความแม่นยำเฉลี่ย (Mean Average

Precision: mAP) [7] และการรู้จำคำบรรยายวีดิทัศน์ ใช้ค่าความผิดพลาดระดับตัวอักษร (Character Error Rate: CER) [8]

1.2 ความมุ่งหมายของการวิจัย

พัฒนากระบวนการในการตรวจจับและรู้จำคำบรรยายภาพ และเปรียบเทียบความถูกต้องการของการรู้จำคำบรรยายแต่ละสถาปัตยกรรม

1.3 ขอบเขตของการวิจัย

1.3.1 ข้อมูลที่ใช้ในการวิจัย

1.3.1.1 *รวบรวมวีดิทัศน์จำนวน 24 วีดิทัศน์*

ที่มีคำบรรยายภาษาไทย ภาษาอังกฤษ ตัวเลขไทย และตัวเลขอารบิก วีดิทัศน์ที่ใช้ได้มาจาก Facebook Thairath - ไทยรัฐออนไลน์(3นาทีคดีดัง), TEP - Thailand Education Partnership ภาคีเพื่อการศึกษาไทย Youtube Bearhug, KLUAYTHAI, KRIT Eighth Floor, Genierock, Taj Tracks, Cakes & Eclairs, 7clouds, DopeLyrics, Equilanora, JEMIN Apollo, San Ko, Tangerine JJY, SNH48 Lyrics, Lemoring, Zaty Farhani

1.3.1.2 *เตรียมข้อมูลเพื่อนำไปใช้ในกระบวนการตรวจจับคำบรรยายวีดิทัศน์*

โดยแปลงคลิปวีดิทัศน์ทั้งหมดเป็นรูปภาพ โดยรูปภาพที่ได้จากการแปลงมีขนาด 1280x720 พิกเซล (Pixel) เพื่อนำไปสร้าง Ground Truth ด้วยโปรแกรม labelImg ซึ่ง Ground Truth คือการกำหนดบริเวณที่เป็นคำบรรยาย (Subtitle) ที่ปรากฏในรูปภาพ โดยรูปภาพที่ใช้ในการทดลองประกอบด้วย 2,700 รูปภาพ

1.3.1.3 *เตรียมข้อมูลเพื่อนำไปใช้ในกระบวนการรู้จำคำบรรยายวีดิทัศน์*

โดยการสร้างคำตอบ (Label) ให้กับรูปภาพของคำบรรยาย ซึ่งคำตอบคือข้อความของคำบรรยายที่ปรากฏในวีดิทัศน์ โดยรูปภาพและคำบรรยายที่ใช้ในการทดลองประกอบด้วย 4,224 รูปภาพ

1.3.2 ตัวอักษรที่ใช้ในการรู้จำคำบรรยายวีดิทัศน์

ตัวอักษรที่ใช้ในการรู้จำคำบรรยายวีดิทัศน์ มีจำนวนทั้งสิ้น 157 ตัวอักษร ตามตารางที่ 1.1

ตารางที่ 1.1 ตัวอักษรที่ใช้ในการรู้จำคำบรรยายวีดิทัศน์

ชนิดของตัวอักษร	ตัวอักษร
พยัญชนะ	ก ข ฃ ค ฅ ฆ ง จ ฉ ช ซ ฌ ญ ฎ ฏ ฐ ฑ ฒ ณ ด ต ถ ท ธ น บ ป ผ ฝ พ ฟ ภ ม ย ร ล ว ศ ษ ส ห ฬ อ ฮ
สระ	ะ ั า ๅ ิ ี ึ ื ุ ู ำ ็ ่ ้ ๊ ๋ เ แ ๎ ๏ ๚ ๛
วรรณยุกต์	่ ้ ๊ ๋ ์
เครื่องหมายวรรคตอน	ๆ ฯ ๎
ตัวเลขไทย	๑ ๒ ๓ ๔ ๕ ๖ ๗ ๘ ๙ ๐
ตัวอักษรภาษาอังกฤษ	A B C D E F G H I J K L M N O P Q R S T U V W X Y Z a b c d e f g h i j k l m n o p q r s t u v w x y z
ตัวเลขอารบิก	1 2 3 4 5 6 7 8 9 0
อักขระพิเศษ	. , ! ( ) - $ ฿ & : ? * “ ‘ (space)

1.3.3 แบ่งข้อมูลที่ใช้ในการทดลองและวิธีการวัดประสิทธิภาพ

1.3.3.1 *ข้อมูลที่ใช้ในการตรวจจับคำบรรยายวีดิทัศน์*

รูปภาพที่มีคำบรรยายวิดีทัศน์จำนวน 2,700 รูปภาพจะถูกแบ่งด้วยวิธี Cross-Validation โดยแบ่งออกเป็น 5-fold เพื่อใช้สำหรับตรวจสอบประสิทธิภาพในการตรวจจับคำบรรยายวีดิทัศน์ ด้วยค่าความแม่นยำเฉลี่ย (mean average precision: mAP)

1.3.3.2 *ข้อมูลที่ใช้ในการรู้จำคำบรรยายวีดิทัศน์*

รูปภาพคำบรรยายวิดีทัศน์จำนวน 4,224 รูปภาพจะถูกแบ่งด้วยวิธี Cross-Validation และแบ่งเป็น 5-fold เพื่อใช้สำหรับตรวจสอบประสิทธิภาพในการรู้จำคำบรรยายวีดิทัศน์ ด้วยค่าความผิดพลาดระดับตัวอักษร (Character Error Rate: CER)

1.3.4 กระบวนการเรียนรู้เชิงลึกสำหรับการตรวจจับและรู้จำคำบรรยายวีดิทัศน์

1.3.4.1 *ตรวจจับคำบรรยายวีดิทัศน์ ด้วยวิธีการเรียนรู้เชิงลึก*

ได้แก่ YoloV3, Tiny-YOLOv3 และ RetinaNet

1.3.4.2 *รู้จำคำบรรยายวีดิทัศน์ ด้วยวิธีการเรียนรู้เชิงลึก*

ได้แก่ CNN, LSTM และ GRU

1.4 ผลที่คาดว่าจะได้รับจากงานวิจัยครั้งนี้

กระบวนการเรียนรู้เชิงลึกสำหรับการตรวจจับและรู้จำคำบรรยายวีดิทัศน์ที่มีความถูกต้องสูงที่สามารถตรวจจับและรู้จำคำบรรยายที่มีขนาดและตำแหน่งแตกต่างกัน รวมไปถึงภาษาที่แตกต่างกัน

1.5 นิยามศัพท์เฉพาะ

การตรวจจับคำบรรยายวีดิทัศน์ (Video Subtitle Detection) คือ การใช้โครงข่ายประสาทคอนโวลูชันในการตรวจจับตำแหน่งของคำบรรยาย โดยใช้วิธีต่าง ๆ เช่น YOLOv3, Tiny-YOLOv3 และ RetinaNet

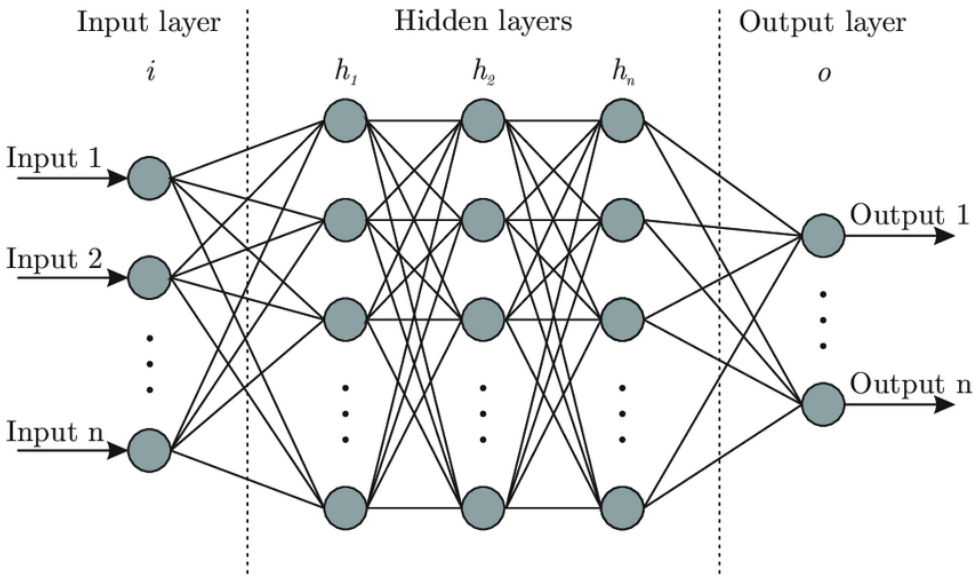
รู้จำคำบรรยายวีดิทัศน์ (Video Subtitle Recognition) คือ การใช้โครงข่ายประสาทแบบคอนโวลูชันในการพยากรณ์ข้อความของคำบรรยายโดยใช้วิธี เช่น VGG16 และ VGG19

การเรียนรู้เชิงลึก (Deep Learning) คือ การเรียนรู้ของคอมพิวเตอร์ที่มีหลักการมาจากการทำงานของสมองมนุษย์ โดยจะเป็นการรับข้อมูลเข้ามา แล้วประมวลผลซ้ำ ๆ ในรูปแบบต่าง ๆ จำนวนมาก เพื่อทำการหาจุดที่จะบ่งบอกคุณลักษณะของข้อมูลที่นำเข้าได้ และแสดงผลลัพธ์ออกมาเป็นกลุ่มต่าง ๆ ว่าข้อมูลที่ใส่เข้าไปนั้นเป็นข้อมูลที่อยู่ในกลุ่มไหน เช่น การใส่รูป สุนัข เข้าไปเมื่อทำการประมวลผล แล้วผลลัพธ์จะออกมาเป็น สุนัข หรือ สัตว์ เป็นต้น

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 การเรียนรู้เชิงลึก (Deep Learning)

การเรียนรู้เชิงลึกเป็นส่วนหนึ่งของการเรียนรู้ของเครื่องจักร (Machine Learning) [9] มีพื้นฐานมาจากโครงข่ายประสาทเทียม (Artificial Neural Network : ANN) เป็นวิธีที่สร้างขึ้นเพื่อทำให้เครื่องจักรสามารถเรียนรู้ได้โดยใช้ต้นแบบมาจากระบบประสาทของมนุษย์ โดยต้องใส่ข้อมูลเข้าไปในชั้นรับข้อมูล (Input Layer) จากนั้นเครื่องจักรจะนำข้อมูลไปประมวลผลในชั้นซ่อน (Hidden Layer) แล้วจะนำเสนอข้อมูลผลลัพธ์ในชั้นแสดงผล (Output Layer) ตามภาพประกอบที่ 2.1 โดยโครงข่ายประสาทเทียมได้มีการพัฒนามาอย่างต่อเนื่องซึ่งวิธีที่นิยมใช้ในปัจจุบันคือ โครงข่ายประสาทคอนโวลูชัน (Convolution Neural Network: CNN) Long Short-Term Memory (LSTM) และ Gated Recurrent Unit (GRU)

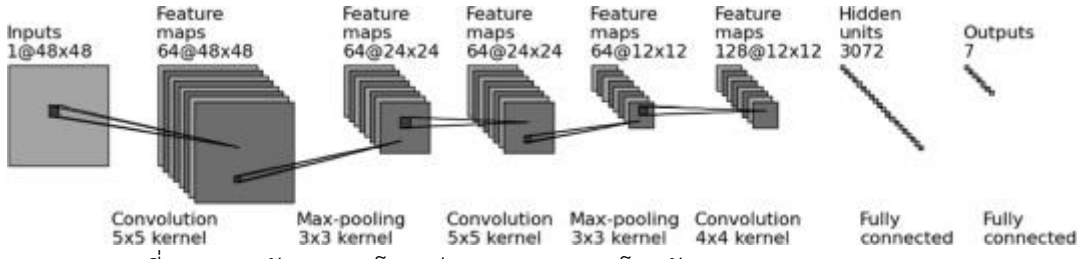


ภาพประกอบที่ 2.1 การเรียนรู้เชิงลึก (Deep Learning) [10]

2.2 โครงข่ายประสาทคอนโวลูชัน (Convolution Neural Network: CNN)

โครงข่ายประสาทคอนโวลูชัน (CNN) เป็นที่รู้จักในเรื่องของการเรียนรู้เชิงลึก (Deep Learning) ซึ่งมีการใช้ในงานวิจัยหลาย ๆ ด้าน เช่น การจัดกลุ่มข้อมูล การแบ่งรูปภาพ และการรู้จำคำพูด ในหลาย ๆ ปีมานี้มีสถาปัตยกรรม CNN จำนวนมากได้ถูกเสนอขึ้น เช่น EfficientNet [24], InceptionResNet [25] และ NASNet [26] และต่อมาในปี 2015 ได้มีการพัฒนา Visual Geometry

Group (VGG) โดย Simonyan และ Zisserman [27] จากมหาวิทยาลัยออกซฟอร์ด จุดประสงค์เพื่อตั้งชั้นของ CNN โดยเรียกว่า VGGNet ซึ่งได้ใช้ชั้นคอนโวลูชัน 16-19 ชั้น มาประมวลผลด้วยตัวกรองขนาดเล็กของชั้นคอนโวลูชัน โดยที่ขนาดของข้อมูลนำเข้า (input) มีขนาด 224x224 พิกเซล โดยที่แต่ละชั้นขนาดของข้อมูลจะถูกลดขนาดลงโดยใช้กระบวนการ max pooling โดยที่ขนาดของชั้นที่เล็กที่สุดมีขนาด 7x7 พิกเซล หลังจากนั้นจะตามมาด้วยชั้นเชื่อมโยงสมบูรณ์ (Fully Connected: FC) ซึ่งมีขนาด 4,096 4,096 1,000 ตามลำดับ และสิ้นสุดที่ softmax เป็นชั้นสุดท้ายตามภาพประกอบที่ 2.2



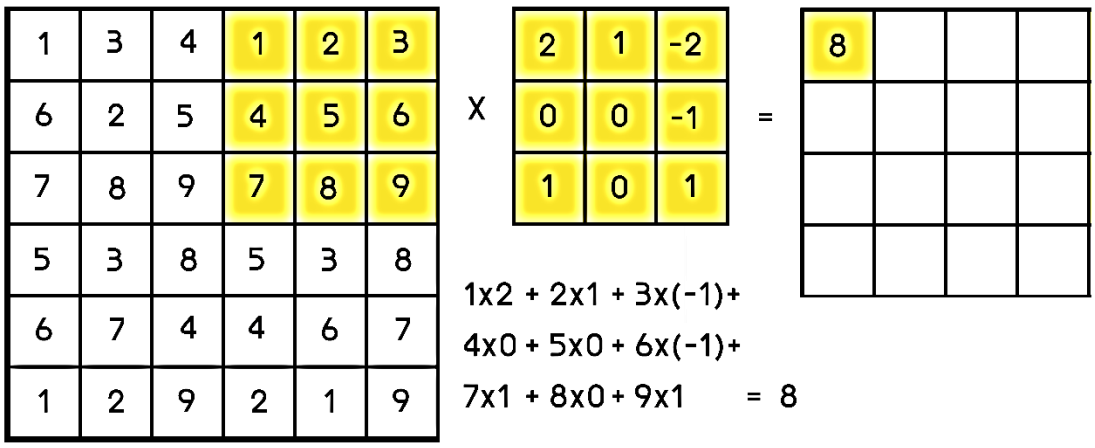
ภาพประกอบที่ 2.2 สถาปัตยกรรมโครงข่ายประสาทคอนโวลูชัน (Convolution Neural Network: CNN) [11]

2.2.1 ชั้นรับข้อมูล (Input Layer)

ชั้นรับข้อมูลเป็นชั้นแรกของระบบการทำงานโดยจะทำหน้าที่ในการรับข้อมูลรูปภาพรับขนาดความสูง (Height) ความกว้าง (Width) และความลึก (Depth) ของรูปภาพ โดยความสูงและความกว้างจะมีหน่วยเป็นพิกเซล และความลึกจะเป็นเลขตามสี เช่น ภาพสีเทา (Gray Scale) จะมีค่าเท่ากับ 1 และภาพสี (BGR) จะมีค่าเท่ากับ 3

2.2.2 ชั้นคอนโวลูชัน (Convolution Layer)

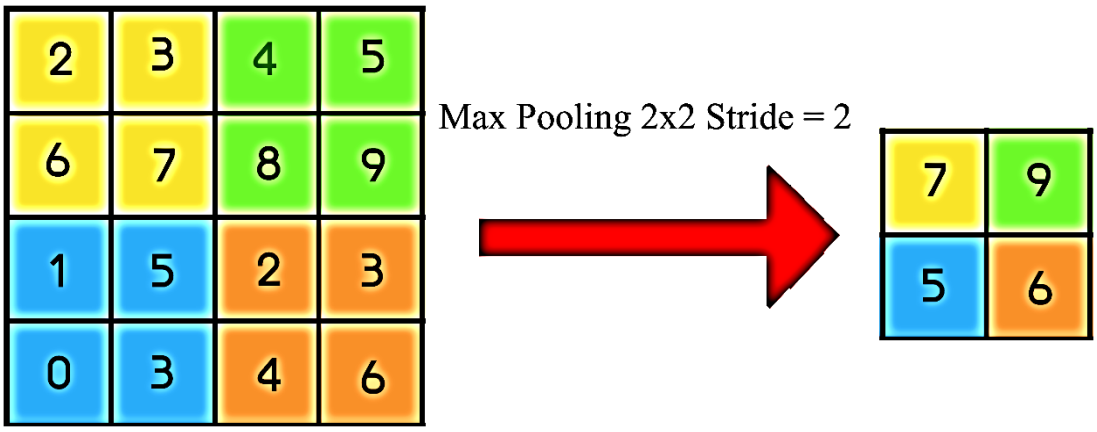
ชั้นคอนโวลูชันเป็นชั้นแรกที่ต่อจากชั้นรับข้อมูล โดยจะทำหน้าที่ในการแยกคุณสมบัติ (Feature Extract) เช่น สี ขอบ รูปทรง จากข้อมูลที่รับมาจากชั้นรับข้อมูล โดยจะทำการเปรียบเทียบรูปภาพที่รับจากชั้นรับข้อมูลกับตัวกรอง (Filter) โดยที่ขนาดของตัวกรองจะสามารถปรับได้ตามความเหมาะสมแต่จะนิยมใช้ที่ขนาด 3 x 3 สำหรับภาพที่มีขนาดไม่ใหญ่มาก 5 x 5 สำหรับภาพที่มีขนาดใหญ่ขึ้นมาโดยจะทำงานโดยการนำตัวเลขในขนาดของ filter มาคูณกับตัวเลขในขนาดของรูปภาพที่ตำแหน่งตรงกับ filter และทำการเลื่อนจากซ้ายไปขวา บนลงล่าง ตามภาพประกอบที่ 2.3



ภาพประกอบที่ 2.3 ตัวอย่างการทำงานของชั้นคอนโวลูชัน ขนาด 3x3 stride=1 [11]

2.2.3 ชั้นพลูลิ่ง (Pooling)

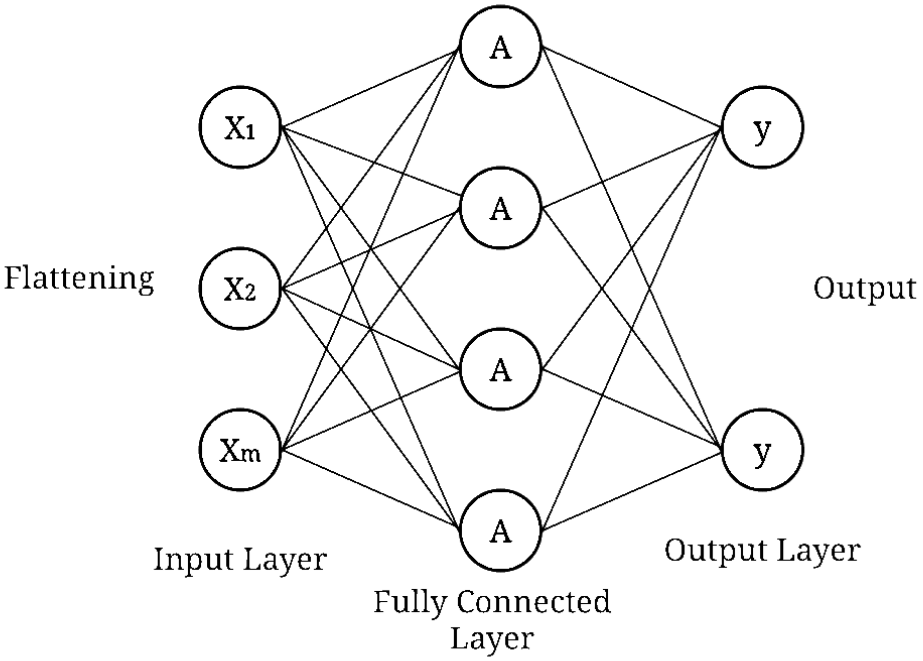
ชั้นพลูลิ่งเป็นชั้นที่จะต่อกับชั้นคอนโวลูชัน [11] ซึ่งจะทำหน้าที่ดึงค่าที่ต้องการคือค่าสูงสุด (Max Pooling) หรือ ค่าเฉลี่ย (Average Pooling) จากชั้นคอนโวลูชันเพื่อลดขนาดของข้อมูลลงตามขนาดดึงค่า (Pool Size) แต่ยังคงไว้ซึ่งลักษณะเด่นของข้อมูล



ภาพประกอบที่ 2.4 ตัวอย่างการทำงานของชั้น Max Pooling [11]

2.2.4 ชั้นเชื่อมโยงสมบูรณ์ (Fully Connected Layer)

ชั้นเชื่อมโยงสมบูรณ์เป็นชั้นที่เชื่อมต่อโดยสมบูรณ์กับชั้นต่าง ๆ ตามภาพประกอบที่ 2.5 ซึ่งจะมีลักษณะเป็น 1 มิติ แล้วนำไปสู่ชั้นผลลัพธ์ดังนั้นจึงเป็นชั้นสุดท้ายเพราะไม่สามารถนำไปเข้าชั้นคอนโวลูชั้น หรือ ชั้นพลูลิ่งได้อีก



ภาพประกอบที่ 2.5 ตัวอย่างการเชื่อมต่อระหว่างชั้นรับข้อมูล ชั้นเชื่อมโยงสมบูรณ์และชั้นผลลัพธ์ [12]

2.2.5 ชั้นผลลัพธ์ (output)

ชั้นผลลัพธ์เป็นชั้นที่ประมวลผลลัพธ์ที่ได้จากชั้นซ่อนออกมาแสดงผลเป็นตัวเลขที่ระบุตามกลุ่มของคำตอบ

2.3 การตรวจจับคำบรรยาย (Subtitle Detection)

การตรวจจับคำบรรยายคือการเรียนรู้ (Train) คำบรรยายจากตัวอย่างรูปภาพ และ นำมาผ่านขั้นตอนต่างๆเพื่อที่จะทำให้สามารถได้รับตำแหน่งคำบรรยายนั้นมา [13] โดยการตรวจจับในอดีตจะมุ่งเน้นไปที่การตรวจจับจากลักษณะของตัวอักษรซึ่งการตรวจจับได้มีการศึกษาและพัฒนามาอย่างต่อเนื่องเพื่อที่จะตอบสนองต่อการตรวจจับสิ่งที่ต้องการจากรูปภาพนั้น ๆ ซึ่งเนื่องจากการตรวจจับคำบรรยายนั้นมีความซับซ้อนและแตกต่างกันมากไม่ว่าจะเป็น มุม ขนาด องศาและพื้นหลังที่ซับซ้อน [2] จึงได้มีการใช้โครงข่ายประสาทแบบคอนโวลูชัน (Convolutional Neural Network: CNN) เข้ามาช่วยในการตรวจจับลักษณะต่าง ๆ โดยใช้วิธีการ (Algorithm) เช่น You only look once Version 3 (YOLOv3), Tiny-YOLOv3 และ RetinaNet ซึ่งทั้ง 3 วิธีนี้จะถูกใช้ในส่วนของการเรียนรู้ตรวจจับตำแหน่งคำบรรยายภายในงานวิจัยนี้

YOLO เป็นสิ่งที่สร้างขึ้นมาเพื่อทำงานเกี่ยวกับการตรวจจับวัตถุรวมไปถึงการจัดกลุ่มวัตถุ โดยใช้วิธีการสร้างกรอบเพื่อกำหนดขอบเขตและจัดกลุ่มให้สิ่งที่อยู่ในขอบเขตนั้น โดย YOLO มีลักษณะคล้ายกับ Fully Convolutional neural network (FCNN) โดยจะทำการส่งรูปภาพผ่านไปสู่ FCNN และได้รับผลลัพธ์ออกมาเป็นพิกัดที่ตรวจจับได้ โดย

2.3.1 YOLOv3

YOLOv3 เป็นวิธีที่พัฒนาต่อมาจาก YOLO โดยเป็นการเพิ่มความเร็วขึ้น 3 เท่า ซึ่ง YOLOv3 จะมีขั้นตอนการทำงานอยู่ 4 ขั้นตอน ขั้นแรกคือการคาดเดากรอบโดยจะคาดเดาออกมาเป็นตำแหน่งของ bounding box โดยมีค่าตำแหน่ง x, y ที่จะทำให้สามารถได้รับความกว้างและความสูงของตำแหน่งนั้น ตำแหน่งซึ่งจะสามารถวาดออกมาเป็นสี่เหลี่ยมเพื่อบ่งบอกตำแหน่ง ขั้นที่ 2 จะเป็นการคาดเดากลุ่มโดยใช้ binary cross-entropy แทนการใช้ softmax เนื่องจาก softmax มักจะเป็นการจำแนกประเภทได้เพียง 1 ประเภทต่อกรอบเท่านั้น แต่วิธีนี้จะสามารถจำแนกประเภทได้หลายประเภทเช่น ภาพที่มีผู้หญิง softmax มักจะเลือกว่าผลลัพธ์จะออกมาเป็นผู้หญิง หรือ มนุษย์ (Using a softmax imposes the assumption that each box has exactly one class which is often not the case) [3] เพียงอย่างเดียว แต่ binary cross-entropy จะจัดให้เป็นทั้งผู้หญิง และมนุษย์ ขั้นที่ 3 เป็นการคาดเดาโดยใช้ขนาดที่แตกต่างกัน 3 ขนาด ขั้นที่ 4 เป็นการใช้ CNN 53 หรือเรียกอีกอย่างว่า Darknet-53 ซึ่งเป็นการใช้ CNN 53 ชั้น ตามภาพประกอบที่ 2.6 ชั้น ในการดึงคุณลักษณะเฉพาะ (Feature Extractor)

	Type	Filters	Size	Output
	Convolutional	32	3 x 3	256 x 256
	Convolutional	64	3 x 3 / 2	128 x 128
1 x	Convolutional	32	1 x 1	
	Convolutional	64	3 x 3	
	Residual			128 x 128
	Convolutional	128	3 x 3 / 2	64 x 64
2 x	Convolutional	64	1 x 1	
	Convolutional	128	3 x 3	
	Residual			64 x 64
	Convolutional	256	3 x 3 / 2	32 x 32
8 x	Convolutional	128	1 x 1	
	Convolutional	256	3 x 3	
	Residual			32 x 32
	Convolutional	512	3 x 3 / 2	16 x 16
8 x	Convolutional	256	1 x 1	
	Convolutional	512	3 x 3	
	Residual			16 x 16
	Convolutional	1024	3 x 3 / 2	8 x 8
4 x	Convolutional	512	1 x 1	
	Convolutional	1024	3 x 3	
	Residual			8 x 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

ภาพประกอบที่ 2.6 สถาปัตยกรรม Darknet-53 [3]

2.3.2 Tiny-YOLOv3

Tiny-YOLOv3 เป็น YOLOv3 ที่ถูกตัดชั้นเชื่อมโยงสมบูรณ์ (Fully connected layer) และชั้นคอนโวลูชันออกไปบางส่วนโดยจะใช้เป็น Darknet19 ตามภาพประกอบที่ 2.7 หรือตัดออกเหลือเพียง 19 ชั้นทำให้ Tiny-YOLOv3 มีความเร็วที่มากกว่า YOLOv3 และมีความต้องการอุปกรณ์น้อยกว่า YOLOv3 ทำให้นิยมนำมาใช้ในระบบที่มีการตรวจจับตำแหน่งสิ่งต่าง ๆ แบบตามเวลาจริง (Real Time) แต่เนื่องจากการตัด ชั้นคอนโวลูชันออกไปบางส่วนทำให้มีประสิทธิภาพในการตรวจจับน้อยลงบ้าง

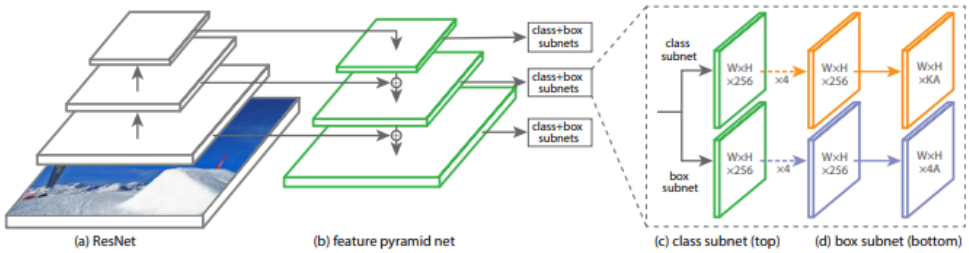
Layer	Type	Filters	Size/Stride	Input	Output
0	Convolutional	16	3 x 3/1	416 x 416 x 3	416 x 416 x 16
1	Maxpool		2 x 2/2	416 x 416 x 16	416 x 416 x 16
2	Convolutional	32	3 x 3/1	208 x 208 x 16	208 x 208 x 32
3	Maxpool		2 x 2/2	208 x 208 x 32	208 x 208 x 32
4	Convolutional	64	3 x 3/1	104 x 104 x 32	104 x 104 x 64
5	Maxpool		2 x 2/2	104 x 104 x 64	104 x 104 x 64
6	Convolutional	128	3 x 3/1	52 x 52 x 64	52 x 52 x 128
7	Maxpool		2 x 2/2	52 x 52 x 128	52 x 52 x 128
8	Convolutional	256	3 x 3/1	26 x 26 x 128	26 x 26 x 256
9	Maxpool		2 x 2/2	26 x 26 x 256	26 x 26 x 256
10	Convolutional	512	3 x 3/1	13 x 13 x 256	13 x 13 x 512
11	Maxpool		2 x 2/1	13 x 13 x 512	13 x 13 x 512
12	Convolutional	1024	3 x 3/1	13 x 13 x 512	13 x 13 x 1024
13	Convolutional	256	1 x 1/1	13 x 13 x 1024	13 x 13 x 256
14	Convolutional	512	3 x 3/1	13 x 13 x 256	13 x 13 x 512
15	Convolutional	255	1 x 1/1	13 x 13 x 512	13 x 13 x 255
16	YOLO				
17	**Route 13**				
18	Convolutional	128	1 x 1/1	13 x 13 x 256	13 x 13 x 128
19	Up-sampling			13 x 13 x 128	13 x 13 x 128
20	**Route 19 8**				
21	Convolutional	256	3 x 3/1	13 x 13 x 384	13 x 13 x 256
22	Convolutional	255	1 x 1/1	13 x 13 x 512	13 x 13 x 256
23	YOLO				

ภาพประกอบที่ 2.7 สถาปัตยกรรม Tiny-YOLOv3 [14]

2.3.3 RetinaNet

RetinaNet เป็นการใช้ Feature Pyramid Network (FPN) [15] ซึ่งเป็นการตรวจจับข้อมูลที่จะปรับขนาดของภาพหลายอัตราส่วนเพื่อที่จะเพิ่มความแม่นยำในการตรวจจับแต่การทำเช่นนั้นทำให้อาจจะใช้เวลาที่นานและต้องใช้หน่วยความจำ (memory) มากกว่าวิธีอื่น ตามภาพประกอบที่ 2.8 เป็นหลักในสถาปัตยกรรม เพื่อสร้างรูปภาพที่มีขนาดแตกต่างกัน โดย RetinaNet จะทำหน้าที่ 2 อย่างคือ การจัดกลุ่มให้กับตำแหน่งที่ล้อมกรอบว่าสิ่งต่าง ๆ ในกรอบคืออะไร และ

วิเคราะห์กรอบที่ได้กับ ground truth ว่าตำแหน่งจุดพิกัดทั้ง 4 ตำแหน่งมีความสัมพันธ์กับ ground truth หรือไม่



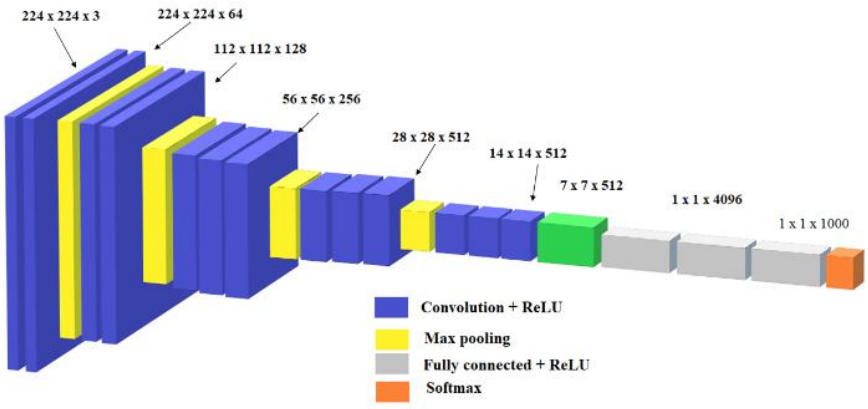
ภาพประกอบที่ 2.8 สถาปัตยกรรม RetinaNet [5]

2.4 การรู้จำคำบรรยาย (Subtitle Recognition)

การรู้จำคำบรรยายคือการเรียนรู้คำบรรยายจากตัวอย่างรูปภาพ และ นำมาผ่านขั้นตอนต่าง ๆ เพื่อที่จะทำให้สามารถได้รับคำบรรยายนั้นมา [13] ซึ่งคำบรรยายที่นำมาเพื่อเรียนรู้ในงานวิจัยนี้มีลักษณะที่เป็นประโยคของข้อความซึ่งจะมีความต่อเนื่องกันของคำบรรยายภายในประโยคดังนั้นจึงได้นำวิธี [16] Recurrent Neural Network (RNN) มาใช้ และนำมาถอดรหัส (Decode) โดยใช้ Connectionist Temporal Classification Loss (CTC Loss) [17]

2.4.1 Visual Geometry Group (VGG)

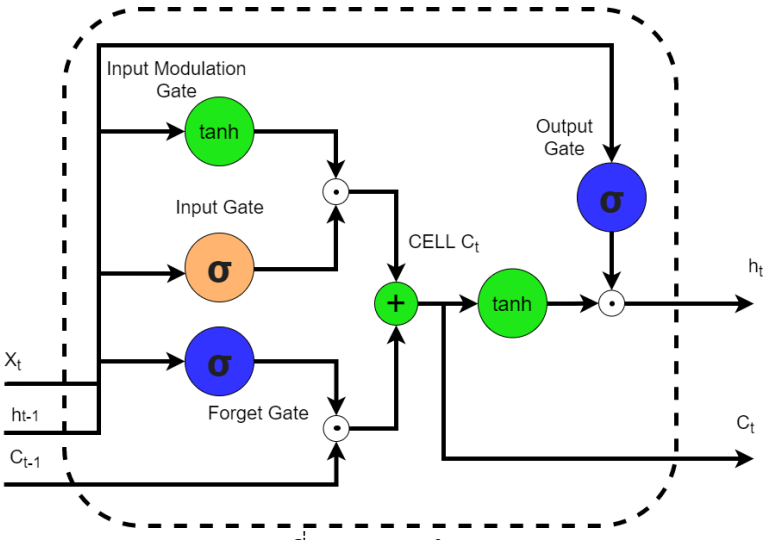
VGG [18] เป็นสถาปัตยกรรมที่นำ CNN ขนาด 3x3 stride=1 max pooling ขนาด 2x2 stride=2 fully connected layer 2 ชั้น และปิดด้วย softmax มาใช้ในการเรียนรู้เชิงลึกโดยที่จะมีชั้นความลึก16-19 ชั้น โดยส่วนมากจะนิยมใช้ที่ 16 ชั้นและ 19 ชั้น หรือเรียกว่า VGG16 และ VGG19 ตามลำดับ โดยที่สถาปัตยกรรม VGG16 สามารถดูได้ที่ ภาพประกอบที่ 2.9



ภาพประกอบที่ 2.9 สถาปัตยกรรม VGG16 [18]

2.4.2 Long Short-Term Memory (LSTM)

Hochreiter และ Schmidhuber [28] ได้นำเสนอถึง Recurrent Neural Network (RNN) ชนิดใหม่ซึ่งมีชื่อว่า Long Short-Term Memory (LSTM) ซึ่งได้นำมาช่วยในการแก้ไขปัญหา Vanishing Gradient ที่ถูกพบเมื่อใช้วิธี RNN กับข้อมูลที่มีความยาว เช่น เสียงพูดหรือ วีดิทัศน์ โดย LSTM ประกอบด้วย input gate, output gate และ forget gate ซึ่งจะเป็นสิ่งที่ควบคุมการไหลของข้อมูล โดยที่เมื่อ LSTM ได้รับข้อมูลมาจากชั้น input เป็นครั้งแรก LSTM จะเข้าสู่ input gate และเข้าสู่ output gate เพื่อตัดสินใจว่าจะเก็บค่าที่ได้ไว้แล้วจะวนซ้ำใน LSTM หรือแสดงผลข้อมูล หากเลือกที่จะวนซ้ำจะนำข้อมูลกลับมาเพื่อเข้าสู่ forget gate ซึ่งจะตัดสินใจว่าจะลบค่าที่เก็บไว้ทิ้งหรือยังคงเก็บค่าไว้ หากเก็บไว้จะไปรอการอัพเดทจาก input gate ซึ่งจะตัดสินใจว่าจะอัพเดทค่านั้นหรือไม่ และจะอัพเดทด้วยค่าอะไร แล้วส่งค่านั้นไป output gate เพื่อตัดสินใจว่าจะนำข้อมูลนั้นออกไปแสดงหรือว่านำข้อมูลกลับไปวนซ้ำอีกรอบ ดังนั้น LSTM จึงสามารถเรียนรู้จากข้อมูลที่เป็นลำดับและเก็บหรือลบข้อมูลทิ้งถ้าข้อมูลนั้นไม่จำเป็นตามภาพประกอบที่ 2.10

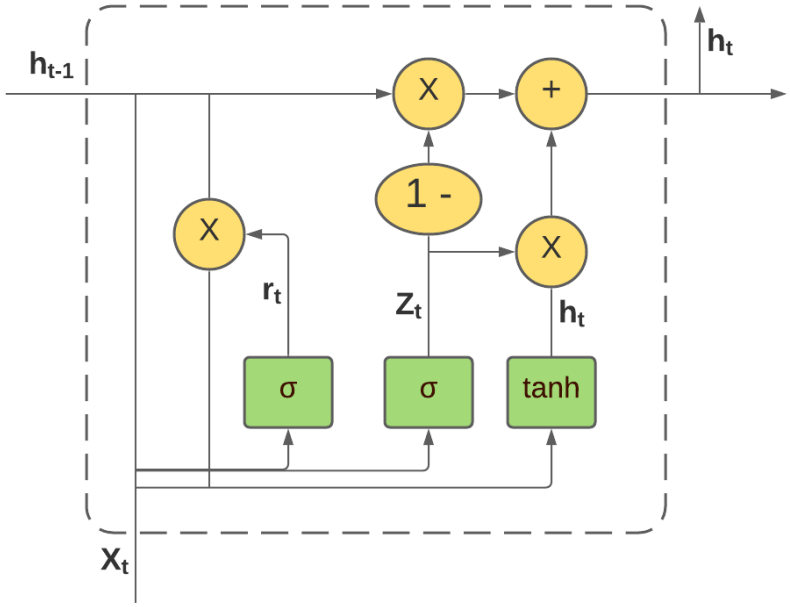


ภาพประกอบที่ 2.10 การทำงานของ LSTM

2.4.3 Gated Recurrent Unit (GRU)

GRU จะมีหลักการทำงานคล้ายกับ LSTM แต่มีความเร็วที่มากกว่าเพราะมีการตัด input และ output gate ออกในการวนซ้ำ เปลี่ยนมาใช้ reset gate และ update gate โดยที่ข้อมูล input เพื่อเข้ามาครั้งแรกจะทำการเก็บค่านั้นไว้แล้วจะติดสินใจว่าจะนำข้อมูลไปวนซ้ำใน GRU หรือแสดงผล หากนำไปวนซ้ำจะเข้าสู่ reset gate เพื่อที่จะตัดสินใจว่าต้องลบข้อมูลไหน และเก็บข้อมูล

ไหนไว้ หลังจากนั้นจะเข้าสู่ update gate เพื่อที่จะตัดสินใจว่าจะอัพเดทข้อมูลไหน แล้วนำไปตัดสินใจว่าจะแสดงผลหรือนำไปวนซ้ำ ตามภาพประกอบที่ 2.11



ภาพประกอบที่ 2.11 การทำงานของ GRU

2.4.4 Connectionist Temporal Classification Loss (CTC Loss)

CTC Loss [17] เป็น loss ฟังก์ชันที่ถูกใช้ในการเรียนรู้ของ LSTM โดยเป็นการถอดรหัสเพื่อที่จะแก้ปัญหาของข้อมูลที่เป็นลำดับ ซึ่งจะสามารถคาดเดาข้อมูลที่ต่อเนื่องกันได้ ในช่วงนี้ได้ถูกใช้ในการจำแนกกลุ่มของข้อมูลที่เป็นลำดับ รวมไปถึงการจดจำลายมือและการจดจำเสียงพูด CTC จะค้นหาว่าข้อมูลนั้นเป็น blank หรือ ไม่มีตัวอักษร ดังนั้นข้อมูลที่เป็น blank หรือ ไม่มีตัวอักษรจะไม่ถูกเปลี่ยนไปเป็นตัวอักษรอื่น

2.5 การตรวจสอบประสิทธิภาพ

วิธีในการตรวจสอบความถูกต้องแม่นยำในการตรวจจับและรู้จำคำบรรยาย

2.5.1 Intersection Over Union (IoU)

$$IOU = \frac{Area\ of\ Overlap}{Area\ of\ Union}$$

IoU เป็นวิธีที่ใช้ในการหาอัตราซ้อนทับกันของตำแหน่งที่ได้รับจากการตรวจจับตำแหน่งกับตำแหน่งที่กำหนดจากการทำ ground truth โดยการนำพื้นที่ที่ทับกัน (Area of

Overlap) ตามภาพประกอบที่ 2.12 มาหารด้วยพื้นที่ทั้งหมด (Area of Union) ตามภาพประกอบที่ 2.13 โดยนิยมใช้ค่า IoU [19] ที่ 0.5 หรือข้อความที่ตรวจจับตรงกับตำแหน่งที่กำหนดใน ground truth มากกว่า 50% ตามภาพประกอบที่ 2.14

ภาพประกอบที่ 2.12 Area of Overlap โดยสีฟ้าคือพื้นที่ Ground Truth สีแดงคือผลลัพธ์ที่ได้การการตรวจจับตำแหน่งคำบรรยาย สีเหลืองคือพื้นที่ซ้อนทับ (Area of Overlap)

ภาพประกอบที่ 2.13 Area of Union ซึ่งเป็นการรวมของพื้นที่ Ground Truth และผลลัพธ์การตรวจจับตำแหน่งคำบรรยาย

ภาพประกอบที่ 2.14 ค่า mAP 62.912 ซึ่งมากกว่ามีค่า IoU มากกว่าร้อยละ 50

2.5.2 อัตราความแม่นยำเฉลี่ย (mean Average Precision: mAP)

เป็นการหาค่าเฉลี่ยความถูกต้อง [7] ซึ่งได้จากการนำค่า Precision และ Recall มาทำเป็นกราฟและพื้นที่ใต้กราฟจะเป็นค่า Average Precision จากนั้นนำไปหาค่าเฉลี่ยจึงได้เป็นค่า mAP โดยที่ค่า Precision และ Recall หาได้ดังนี้

2.5.2.1 *Precision*

$$Precision = \frac{tp}{tp+fp}$$

เป็นค่าการคาดเดาที่ถูกต้องต่อการคาดเดาทั้งหมดโดยที่ tp คือค่าการคาดเดาที่ถูกต้องตาม ground truth fp คือการคาดเดาว่าถูกแต่ผลลัพธ์ผิดจาก ground truth

2.5.2.2 *Recall*

$$Recall = \frac{tp}{tp+fn}$$

เป็นค่าการคาดเดาที่ถูกต้องต่อการคาดเดาทั้งหมดโดยที่ tp คือค่าการคาดเดาที่ถูกต้องตาม ground truth fp คือการคาดเดาว่าที่ผิดและผิดจริงตาม ground truth

2.5.3 ความผิดพลาดระดับตัวอักษร (Character Error Rate: CER)

$$CER = \frac{Insertions+Substitutions+Deletions}{Length(Ground\ truth\ label)}$$

จากสูตรการหาค่า CER [8] ด้านบนโดยที่ค่า Insertions คือจำนวนตัวอักษรที่นำเข้า Substitutions คือจำนวนตัวอักษรที่เปลี่ยนไป Deletions คือจำนวนตัวอักษรที่หายไป และ Length คือจำนวนตัวอักษรทั้งหมดที่ได้มาจากเฉลยของ dataset

2.6 งานวิจัยที่เกี่ยวข้อง

2.6.1 งานวิจัยที่เกี่ยวข้องกับการตรวจจับคำบรรยาย (Text Detection)

Ma et al. [2] ได้สร้างวิธีตรวจจับข้อความแบบใหม่ชื่อว่า ReLaText ซึ่งเป็นเทคนิคใหม่ในการตรวจจับตำแหน่งคำบรรยายโดยใช้การเชื่อมความสัมพันธ์ (link relationship) ในการจัดกลุ่มข้อความ คอนโวลูชั่นแบบกราฟ (Graph Convolutional Netword: GCN) ในการแยกความสัมพันธ์ที่ไม่ควรจะติดกันออก ซึ่ง ReLaText จะมีความแม่นยำในการตรวจจับข้อความที่มีช่องว่างหรือระยะห่างที่ใหญ่ (large inter-character) และแม่นยำในการตรวจจับข้อความที่มีระยะห่างระหว่างบรรทัดน้อย โดยใช RCTW-17, MSRA-TD500, Total-Text,CTW1500 และ DAST1500 ในการทดลอง ผลการทดลองด้วย RCTW-17 ด้วยรูปภาพขนาด 1500 พิกเซล พบว่าค่า Recall ได้ 61.7% F-Score ได้ 68.1% แต่มีค่า Precision 75.9% ซึ่งน้อยกว่าวิธี TextMountain อยู่ 4.9% เนื่องจากงานวิจัยนี้ได้อธิบายถึงการตรวจจับคำข้อความจึงได้นำมาคำอธิบายนั้นมาใช้เพื่อเป็นส่วนหนึ่งในการอธิบายถึงการตรวจจับคำบรรยาย

He et al. [20] ได้คิดวิธีที่จะลดเวลาในการตรวจจับข้อความที่มีหลายขนาด (muti-scale) โดยจะแบ่งออกเป็น 2 ขั้นตอน ซึ่งขั้นตอนแรกจะใช้ Scale-based Region Proposal Network (SRPN) ซึ่งมีประสิทธิภาพในการตรวจจับข้อความแบบกว้าง แล้วตัดพื้นที่ซึ่งไม่ใช่ข้อความออกรวมไปถึงปรับขนาดของข้อความ และเข้าสู่ขั้นตอนที่ 2 เป็นการใช้ เครือข่ายเชื่อมโยงสมบูรณ์ (Fully Convolutional Network) ซึ่งเป็นวิธีที่ใช้ ชั้นเพียงชั้นคอนโวลูชัน และ พลูลิ่ง เท่านั้น โดยจะตัดชั้นเชื่องโยงสมบูรณ์ออก (Fully connected layer) และใช้คอนโวลูชั่น 1x1 แทนในการแปล

ข้อความที่ตรวจจับมาจากขั้นตอนที่ 1 ซึ่งจะมีความแม่นยำสูงกว่าแต่ว่ามีความแคบในการทำงานน้อยกว่าขั้นตอนแรก เนื่องจากพื้นที่ซึ่งไม่ใช่ข้อความถูกนำออก และการปรับขนาดของข้อความทำให้วิธีนี้มีประสิทธิภาพและความเร็วที่สูง ซึ่งมีค่าวัดประสิทธิภาพ (F-measure) 85.40% ที่ 16.5 เฟรมต่อวินาที (frame per second: fps) และประสิทธิภาพในการแข่งขัน (competitive performance) 79.66% ที่ 35.1 fps เนื่องจากมีการอธิบายถึงความจำเป็นของการตรวจจับข้อความที่ต้องสามารถตรวจจับข้อความเดิมแบบที่มีความละเอียดสูงและแบบความละเอียดต่ำ ได้จึงได้หาวิธีที่สามารถตรวจจับแบบหลายขนาดได้

Wang et al. [21] ได้คิดวิธีในการเพิ่มประสิทธิภาพการตรวจจับตำแหน่งข้อความที่อยู่ห่างกันมากหรือที่มีอยู่ชิดกันมาก โดยการจัดกลุ่มของตัวอักษรเพื่อเพิ่มประสิทธิภาพการตรวจจับข้อความที่มีลักษณะเป็นเส้นโค้งด้วยวิธี CNN โดยใช้สถาปัตยกรรม VGG16 เป็นหลักในการตรวจจับ ซึ่งรูปภาพที่นำมาใช้ในการเรียนรู้จะถูกทำ ground truth หลายรอบ โดยการหมุนข้อความตามเข็มและทวนเข็มนาฬิการเพิ่มให้ข้อความแต่ละส่วนตั้งตรงจากการทดลองด้วย dataset DAST1500 ซึ่งเป็น dataset ที่เก็บรูปภาพที่มีลักษณะกระจัดกระจาย โค้ง หรือ แออัด และ dataset อื่น ๆ คือ ICDAR15, MTWL, Totaltext, SCUT-CTW1500 พบว่า 4 ใน 5 dataset วิธีการใหม่นี้มีประสิทธิภาพดีกว่าวิธีการเดิม แต่ยังมีข้อผิดพลาดในการตรวจจับข้อความที่ยากในการระบุ และข้อความที่มีขนาดเล็กเกินไป เนื่องจากที่มีการอธิบายถึงการทำ ground truth เพื่อกำหนดขนาดและตำแหน่งของข้อความตัวอย่างจึงได้มีการนำวิธี ground truth มาใช้

Zhu and Du [22] ได้สร้างวิธีใหม่ชื่อ TextMountain โดยจะเป็นการใช้ข้อมูลตำแหน่งของกรอบและจุดศูนย์กลางโดยการหาความน่าจะเป็นของจุดศูนกลางและกรอบโดยที่จุดศูนย์กลางจะเป็นเหมือยอดภูเขาและส่วนบนและล่างของตัวอักษรจะเป็นตีนเขา และหาทิศทางที่ข้อความจะไปเพื่อเพิ่มประสิทธิภาพในการตรวจจับ โดยได้มีการทดลองใช้ dataset ดังนี้ MLT, ICDAR2015, RCTW-17 และ SCUT-CTW1500 ซึ่งได้ใช้ ResNet-50 เป็นหลักในการทดลองในทุก dataset แต่จะพิเศษที่ ICDAR2015 ที่ได้มีการทดลองใช้ทั้ง ResNet-50 และ VGG-16 เนื่องจากได้มีการกล่าวถึง precision recall และ f-measure จึงได้ไปศึกษาและพบวิธีการแสดงผลการทดลองด้วยค่า mAP

Huang et al. [23] เป็นการเสนอวิธี Multiscale Connectionist Text Proposal Network (MS-CTPN) ที่จะสามารถตรวจจับช้อความแบบหลายขนาดและการเชื่อมโยงของข้อความโดยใช้ Res-VGG16 เป็นพื้นฐาน ทดลองโดยใช้ข้อมูลจาก SynthText ซึ่งมีรูปภาพตัวอย่างทั้งหมด 12,000 รูป โดย 10,000 ใช้เรียนรู้ อีก 2,000 ใช้ทดสอบ ซึ่งผลการทดลองออกมาว่าการใช้ VGG16 + RPN จะใช้เวลาน้อยที่สุดคือ 0.2159 รูปต่อวินาที ค่า IoU 0.5624 หรือตำแหน่งที่ได้รับตรงกับตำแหน่ง ground truth 56.24% และ Res-VGG16 + Bi-RNN + MS-RPN ใช้เวลานานที่สุดแต่ค่า

IoU 0.7635 หรือผลที่ได้ตรงกับ ground truth 76.35% ซึ่งเยอะที่สุด เนื่องจากได้อธิบายเกี่ยวกับ intersection over union (IoU) จึงได้นำไปศึกษาเพิ่มเติมและได้หาวิธีการกำหนด IoU เพื่อที่จะใช้ในการทดสอบ

2.6.2 งานวิจัยที่เกี่ยวข้องกับการรู้จำคำบรรยาย (Text Recognition)

Yan and Xu [24] ได้นำเสนอถึงสถาปัตยกรรม residual network โดยเป็นการใช้ Bi-GRU และ Conncetionist temporal classification (CTC) ในการรู้จำคำบรรยายวีดิทัศน์ทั้งภาษาอังกฤษและภาษาจีน โดยสถาปัตยกรรมที่นำเสนอนั้นได้ทดลองกับ 2 dataset คือ ICDAR2003 และ ICDAR2013 ซึ่งมีค่าความแม่นยำ (Accuracy) 92.3% และ 89.2% ตามลำดับ

Jemni et al. [25] เป็นการตรวจจับและแก้ไขลายมือเขียนอารบิกที่ผิดหรือไม่มีในพจนานุกรม โดยการใช้ MDLSTM และ CNN ในการทดลองกับ dataset KHATT โดยได้แบ่งเป็น 2 ขั้นตอน ในขั้นแรก จะเป็นการตรวจจับตำแหน่งของข้อความ และในขั้นตอนที่สองจะเป็นการแก้ไขข้อความให้ถูกต้องโดยการรู้จำข้อความและเลือกคำจากพจนานุกรม โดยใช้ต้นแบบ WSL, PSL, MSL เนื่องจากมีการทำทั้งตรวจจับข้อความและรู้จำข้อความจึงได้แบ่งการทดลองออกเป็นส่วนของ การตรวจจับคำบรรยายและการรู้จำคำบรรยาย

Gan et al. [16] เสนอถึง 1D-CNN และ Temporal Convolutional Recurrent Network (TCRN) โดยได้ตั้งชื่อว่า 1D-TCRN เพื่อที่จะนำมาใช้ในการรู้จำลายมือตัวอักษรจีนตาม IAHCT dataset โดยนำมาเขียนในอากาศเพื่อที่จะรู้จำลำดับขีด และท่าทางการเขียน ในแบบจำลองนี้นำ 1D residual convolution block มาใช้และนำมาเชื่อมกับ Sequence layer หรือก็คือเป็นการเชื่อมต่อของ CNN, LSTM และ CTC เพื่อนำมารู้จำลายมือตัวอักษรภาษาจีน 2,565 ตัวอักษร

Zhang et al. [26] ได้เสนอเทคนิคใหม่คือ Scale-aware hierarchical attenion network for scene text recognition (SaHAN) ซึ่งได้รับแรงบรรดาลใจจากเครือข่ายพีระมิด (pyramid network) โดยเป็นวิธีที่ใช้ทั้ง โครงข่ายประสาทคอนโวลูชั่น (Convolution Neural Network: CNN) และ โครงข่ายประสาทแบบเกิดซ้ำ (Recurrent Neural Network : RNN) ซึ่งแบ่งเป็น 2 ส่วน ส่วนแรกคือการเข้ารหัสด้วย CNN และ RNN โดยจะใช้ ResNet-45 เป็นหลัก แล้วทำการลดขนาดลง (convolution layer) และอัดให้เหลือ 1 มิติ (fully connected layer) แล้วนำเข้าสู่ bidiractional long short-term memory (BiLSTM) และนำไปถอดรหัสด้วยการใช้ Gated Recurrent Unit (GRU) โดยวิธีนี้ได้ทดลองด้วย dataset 7 อย่าง คือ IIIT5K-Words (IIIT5K), Street View Text (SVT), ICDAR 2003 (IC03), ICDAR 2013 (IC13), ICDAR 2015 (IC15), SVT-Perspective (SVT-P), CUTE80 ได้ผลลัพธ์ดังนี้ IIIT5K ที่ใช้พจนานุกรม 50 คำ มีความแม่นยำ 99.2% ที่ใช้พจนานุกรรม 1000 คำ 97.9% ที่ไม่ใช้พจนานุกรรม 91.2% SVT ที่ใช้พจนานุกรรม 50 คำ 98.5% ที่ไม่ใช้พจนานุกรรม 90.4% IC03 ที่ใช้พจนานุกรรม 50 คำ 99.3% ทั้งหมด 98.4 ที่ไม่ใช้

พจนานุกรรม 95.5 IC13 ที่ไม่ใช้พจนานุกรรม 93.0% IC15 ที่ไม่ใช้พจนานุกรรม 75.0% IC15 แบบ 1811 รูป ที่ไม่ใช้พจนานุกรรม 80.7% SVT-P ที่ใช้พจนานุกรรม 50 คำ 95.2 ทั้งหมด 91.0 ที่ไม่ใช้พจนานุกรรม 82.8 CUTE80 ที่ไม่ใช้พจนานุกรรม 84.4% เนื่องจากมีการอธิบายถึง CNN, RNN, LSTM, BiLSTM และ GRU จึงช่วยลดระยะเวลาในการศึกษาข้อมูลเบื้องต้นทำให้เข้าใจข้อมูลเบื้องต้นเกี่ยวกับ CNN, RNN, LSTM, BiLSTM และ GRU

Chamchong et al. [27] เป็นการรู้จำลายมือเขียนไทยโบราณ โดยใช้ CNN, RNN และ CTC Loss มาสร้างสถาปัตยกรรม6แบบเพื่อเปรียบเทียบกัน โดยนำข้อมูลลายมือเขียนไทยโบราณมาจากห้องสมุดแห่งชาติไทย ซึ่งค่า CER ที่ดีที่สุดคือ 11.9% โดยจะใช้ตัวอักษร 44 ตัว สระ 18 ตัว วรรณยุกต์ 4 ตัว สัญลักษณ์อื่น ๆ 5 ตัว เลขไทย 10 ตัว รวมทั้งสิ้น 81 ตัว ในการทดสอบจะใช้ชุดข้อมูลที่เก็บไว้ก่อน ค.ศ. 1902 โดยมีทั้งหมด 140 รูป จากการทดสอบพบว่าการนำข้อมูลเข้าทดสอบพร้อมกัน 32 (batch 32) จะมีความผิดพลาดน้อยกว่าเข้าทีละ 64 (batch 64) LSTM จะมีความผิดพลาดมากกว่า GRU การมีชั้น CNN มากกว่าทำให้มีความผิดพลาดน้อยลง การใช้ CNN จะให้ผลลัพธ์ที่ดีกว่าการไม่ใช้ CNN เนื่องจากมีการทดลองในการรู้จำลายมือทั้งแบบใช้แต่ RNN รวมถึงใช้ทั้ง CNN และ RNN และได้ข้อสรุปว่า การใช้สถาปัตยกรรมที่มี CNN ให้ผลลัพธ์ที่ดีกว่าจึงได้สร้างสถาปัตยกรรมที่ใช้ CNN ร่วมกับ RNN

เนื่องจากวิทยานิพนธ์นี้ได้ใช้สถาปัตยกรรมที่มีรูปแบบเดียวกันกับของ Chamchog et al. ดังนั้นจึงได้ลองนำสถาปัตยกรรมของ Chamchong et al. มาใช้ในการสร้างต้นแบบและเปรียบเทียบกับต้นแบบที่ใช้ในวิทยานิพนธ์นี้

บทที่ 3
วิธีดำเนินการวิจัย

บทนี้ได้กล่าวถึง ขั้นตอนการดำเนินการวิจัยของการเรียนรู้เชิงลึกสำหรับการตรวจจับและรู้จำคำบรรยายวีดิทัศน์ โดยกระบวนการทำวิจัยได้แบ่งออกเป็น 2 ขั้นตอนดังต่อไปนี้ การตรวจจับคำบรรยายวีดิทัศน์ และการรู้จำคำบรรยายวีดิทัศน์ ซึ่งได้อธิบายถึงรายละเอียดวิธีดำเนินการวิจัย ในหัวข้อดังต่อไปนี้

3.1 ชุดข้อมูลที่ใช้ในการวิจัย

3.2 การตรวจจับคำบรรยายวีดิทัศน์และทดสอบประสิทธิภาพ

3.3 การรู้จำคำบรรยายวีดิทัศน์และทดสอบประสิทธิภาพ

3.1 ชุดข้อมูลที่ใช้ในงานวิจัย

ข้อมูลวีดิทัศน์ที่ใช้ในวิทยานิพนธ์คือวีดิทัศน์จำนวนทั้งสิ้น 24 วีดิทัศน์ ที่เก็บรวบรวมจาก Youtube และ Facebook โดยคำบรรยายที่ปรากฏในวีดิทัศน์ประกอบไปด้วย ภาษาไทย ภาษาอังกฤษ ตัวเลขไทย และตัวเลข อารบิก ซึ่งมีทั้งคำบรรยายที่อยู่ในจัดที่น่าสนใจหรืออยู่ที่พื้นที่ด้านล่างตาม**ผิดพลาด! ไม่พบแหล่งการอ้างอิง**

ก) ข)

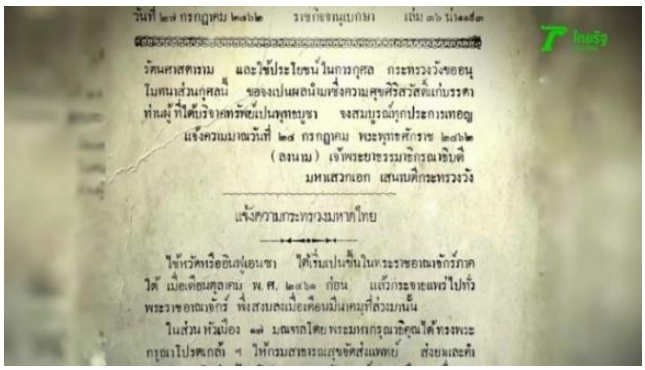
ภาพประกอบที่ 3.1 ตัวอย่างของคำบรรยายที่ปรากฏในวีดิทัศน์ ก) คำบรรยายที่ปรากฏบริเวณด้านล่างของวีดิทัศน์ และ ข) คำบรรยายที่ปรากฏขึ้นในบริเวณที่น่าสนใจ
ที่มา: วีดิทัศน์ประกอบจาก Youtube ช่อง echo และ Mushroom TV

นำวีดิทัศน์ทั้งหมดไปแยกเป็นเฟรม (Frame) เพื่อนำไปทำ Ground Truth สำหรับใช้ในการทดสอบและวัดประสิทธิภาพ โดยเรียกชุดข้อมูลนี้ว่า ชุดข้อมูลรูปภาพคำบรรยาย (Video Subtitle Image Dataset) โดยชุดข้อมูลแบ่งออกเป็น 2 ส่วน ดังนี้

3.1.1 รูปภาพที่ใช้สำหรับตรวจจับคำบรรยายวีดิทัศน์

รูปภาพคำบรรยาย คือรูปภาพที่มีคำบรรยายปรากฏอยู่ 2 รูปแบบคือ คำบรรยายที่ปรากฏบริเวณด้านล่างของวีดิทัศน์ และคำบรรยายที่ปรากฏขึ้นในบริเวณที่น่าสนใจ ซึ่งคำบรรยายนั้นมีทั้งภาษาไทย ภาษาอังกฤษ ตัวเลขไทย และตัวเลขอารบิก รวมทั้งสิ้น 2,700 รูปภาพ ตัวอย่างของรูปภาพคำบรรยายแสดงดังภาพประกอบที่ 3.2

ก)



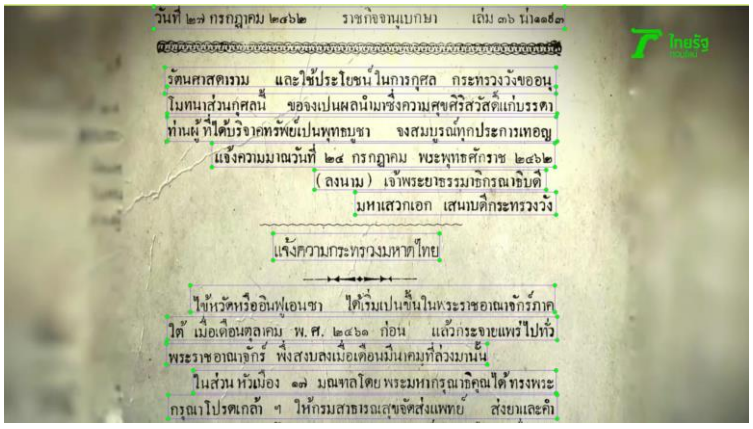
ข)

ค)

ภาพประกอบที่ 3.2 ตัวอย่างรูปภาพที่ใช้สำหรับตรวจจับคำบรรยายวิดีทัศน์ ก) รูปภาพที่ประกอบด้วยภาษาอังกฤษและเลขอารบิก ข) รูปภาพที่ประกอบด้วยภาษาไทยและเลขไทย ค) รูปภาพที่ประกอบด้วยภาษาไทยและเลขอารบิก

ทั้งนี้ รูปภาพทั้งสิ้น 2,700 รูปภาพจะถูกนำไปสร้าง Ground Truth และเก็บ Ground Truth ไว้ในไฟล์รูปแบบเอ็กซ์เอมแอล (.xml) เพื่อใช้สำหรับเปรียบเทียบตำแหน่งที่ค้นพบ และตำแหน่งจริงของคำบรรยาย รูปภาพ Ground Truth ที่แสดงตำแหน่งจริงของคำบรรยายแสดงดังภาพประกอบที่ 3.3

ก)



ข)

ค)

ภาพประกอบที่ 3.3 ตัวอย่าง Ground Truth ที่แสดงตำแหน่งจริงของคำบรรยาย

3.1.2 รูปภาพคำบรรยายวีดิทัศน์ที่ใช้สำหรับการรู้จำ

รูปภาพคำบรรยายวีดิทัศน์ที่ใช้สำหรับการรู้จำ เป็นรูปภาพคำบรรยาย (Subtitle Image) ที่นำมาจาก Ground Truth โดยนำมาเพื่อใช้สำหรับการรู้จำคำบรรยายวีดิทัศน์ (Video Subtitle Recognition) โดยเป็นรูปภาพคำบรรยายที่เป็น ภาษาไทย ภาษาอังกฤษ ตัวเลขไทย และตัวเลขอารบิก จำนวนทั้งสิ้น 4,224 รูปภาพ ตัวอย่างของรูปภาพคำบรรยายวีดิทัศน์แสดงดังภาพประกอบที่ 3.4ก โดยรูปภาพคำบรรยายทุกรูปจะถูกนำไปสร้างคำตอบ (Label) เพื่อใช้สำหรับตรวจสอบความถูกต้องในการรู้จำ และวัดประสิทธิภาพของอัลกอริทึม ตัวอย่างการสร้างคำตอบให้กับรูปภาพคำบรรยาย วีดิทัศน์แสดงดังภาพประกอบที่ 3.4ข

1147_ส่วนคำกล่าวขานที่ว่า

1148_เมื่อได้ลิ้มลองเนื้อมนุษย์แล้ว จะหากินอยู่เรื่อยๆ

ก)

1147_ส่วนคำกล่าวขานที่ว่า	1148_เมื่อได้ลิ้มลองเนื้อมนุษย์แล้ว จะหากินอยู่เรื่อยๆ
1147 *คือหมายเลขลำดับรูปภาพ*	1148 *คือหมายเลขลำดับรูปภาพ*
“ส่วนคำกล่าวขานที่ว่า” *คือคำตอบ (Label)*	“เมื่อได้ลิ้มลองเนื้อมนุษย์แล้ว จะหากินอยู่เรื่อยๆ” *คือคำตอบ (Label)*

ข)

ภาพประกอบที่ 3.4 ตัวอย่างรูปภาพคำบรรยายวีดิทัศน์ ก) รูปภาพคำบรรยายวีดิทัศน์ และ ข) คำตอบ (Label)

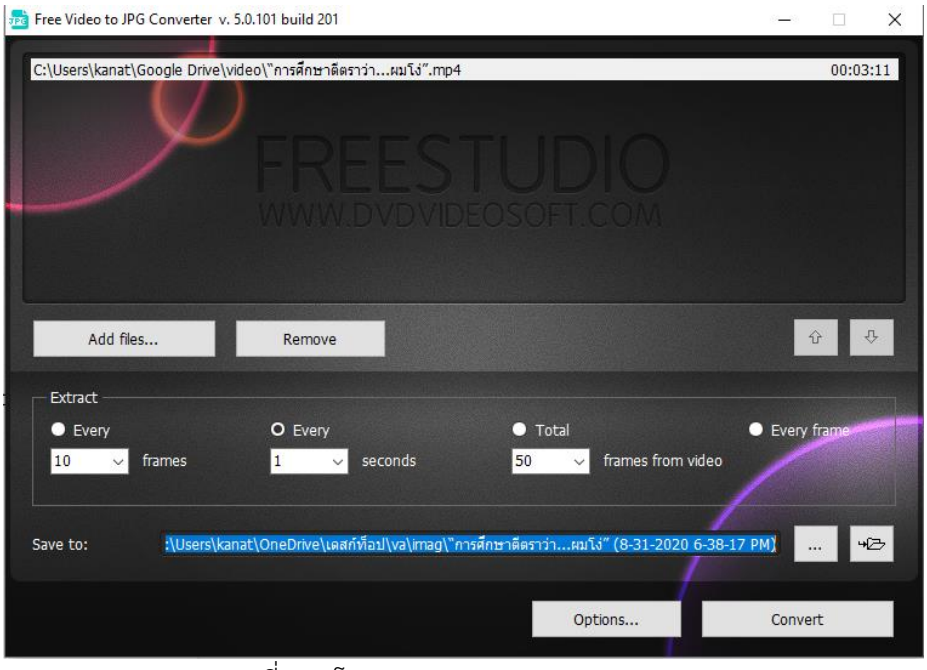
3.2 การเตรียมข้อมูล

การเตรียมข้อมูลแบ่งออกเป็น

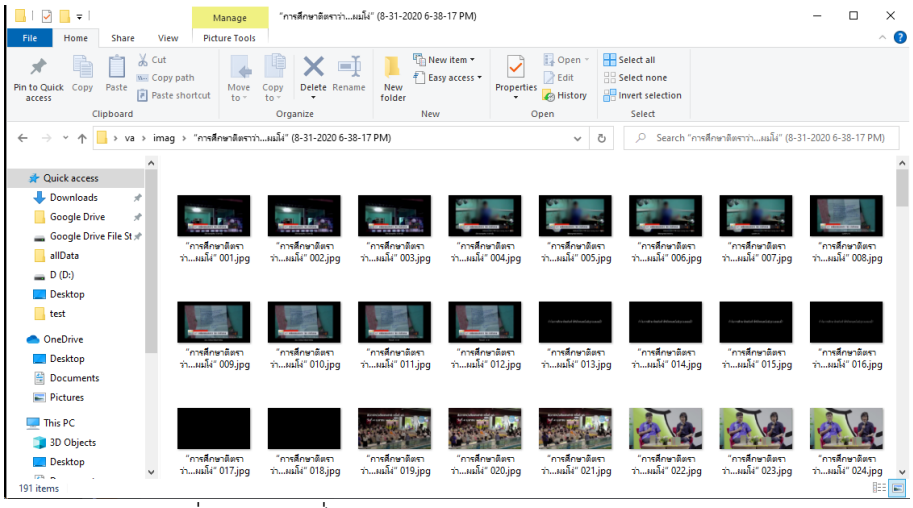
3.2.1 หาตัวอย่างที่มีการฝังคำบรรยายภาษาไทยและอังกฤษ

3.2.2 นำคลิปวีดิทัศน์ไปแปลงเป็นรูปภาพ

ด้วยโปรแกรม Free Video to JPG Converter ภาพประกอบที่ 3.5 ซึ่งเป็นโปรแกรมแปลงวีดิทัศน์เป็นไฟล์รูปภาพ และจะได้ผลลัพธ์ออกมาเป็นรูปภาพ ดังภาพประกอบที่ 3.6



ภาพประกอบที่ 3.5 โปรแกรม Free Video to JPG Converter

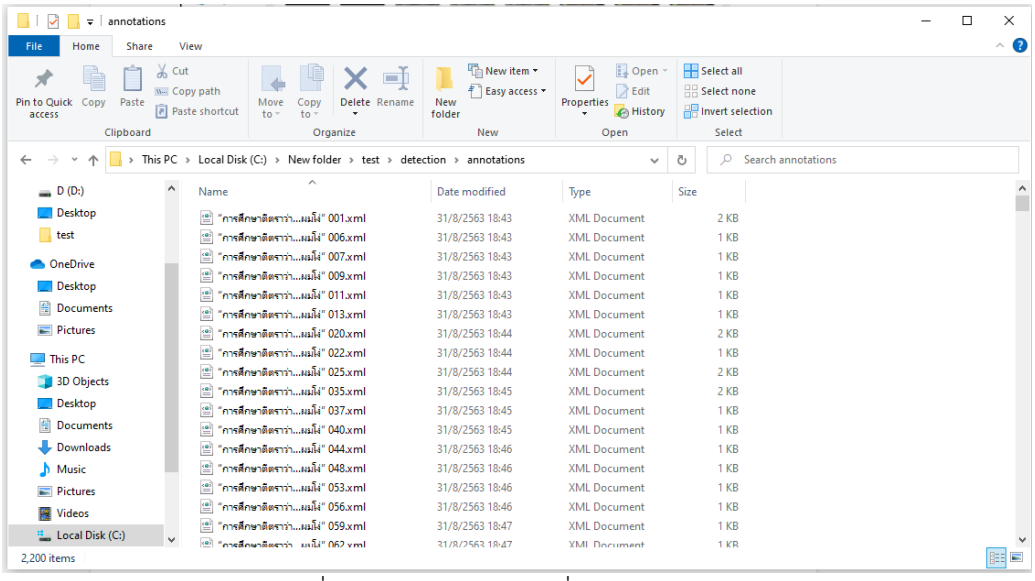


ภาพประกอบที่ 3.6 ภาพที่ได้จากโปรแกรม Free Video to JPG Converter

3.2.3 นำรูปภาพไปสร้าง ground truth

ด้วยโปรแกรม LabelImg ภาพประกอบที่ 3.7 และได้ผลลัพธ์ออกมาเป็นไฟล์ .xml ตามภาพประกอบที่ 3.8

ภาพประกอบที่ 3.7 ภาพโปรแกรม LabelImg



ภาพประกอบที่ 3.8 ภาพไฟล์ .xml ที่ได้จากโปรแกรม LabelImg

3.2.4 นำรูปภาพคำบรรยาย (Subtitle image) มาสร้างไฟล์ข้อความตัวอักษรประกอบ (Text transcription)

โดยการตัดมาเฉพาะพื้นที่ที่ถูกทำ ground truth และตั้งชื่อไฟล์ตามคำบรรยายในรูปภาพ ดังภาพประกอบที่ 3.9

ภาพประกอบที่ 3.9 ตัวอย่างไฟล์ภาพข้อความคำบรรยายภาษาอังกฤษ

3.3 การตรวจจับคำบรรยายวีดิทัศน์

วิทยานิพนธ์นี้ได้ทำการทดลองการตรวจจับและการรู้จำคำบรรยายวีดิทัศน์บน Google Colab ซึ่งประมวลผลด้วยหน่วยประมวลผลภาพ (Graphic Processing Unit: GPU) โดยกระบวนการตรวจจับคำบรรยายวีดิทัศน์ แบ่งเป็นขั้นตอนดังนี้

3.3.1 ข้อมูลที่ใช้ในการทดสอบกระบวนการตรวจจับคำบรรยายวีดิทัศน์

ข้อมูลที่ใช้ในการทดสอบกระบวนการตรวจจับคำบรรยายวีดิทัศน์คือ รูปภาพที่ใช้สำหรับตรวจจับคำบรรยายวีดิทัศน์ จำนวนทั้งสิ้น 2,700 รูปภาพ โดยแบ่งข้อมูลที่ใช้ในการทดสอบด้วยวิธี k fold Cross-Validation โดยกำหนดให้ k=5

3.3.2 ทำการสร้างโมเดลในการตรวจจับคำบรรยายวีดิทัศน์

ทำการสร้างโมเดลในการตรวจจับคำบรรยายวีดิทัศน์ด้วยวิธีการเรียนรู้เชิงลึกด้วยวิธี YoloV3, Tiny-YOLOv3 และ RetinaNet

3.3.3 ทดสอบประสิทธิภาพของโมเดลในการตรวจจับคำบรรยายวีดิทัศน์

ทดสอบประสิทธิภาพของโมเดลในการตรวจจับคำบรรยายวีดิทัศน์ด้วยค่าความแม่นยำเฉลี่ย (Mean Average Precision: mAP) โดยกำหนดให้มีค่า IoU ที่ 0.5 ตามภาพประกอบที่ 3.10 และ ภาพประกอบที่ 3.11

ภาพประกอบที่ 3.10 ตัวอย่างผลลัพธ์ค่า mAP 62.912%

ภาพประกอบที่ 3.11 ตัวอย่างผลลัพธ์ค่า mAP 55.563%

3.4 การรู้จำคำบรรยายวีดิทัศน์

การรู้จำคำบรรยายวีดิทัศน์ แบ่งเป็นขั้นตอนดังนี้

3.4.1 ข้อมูลที่ใช้ในการทดสอบในการรู้จำคำบรรยายวีดิทัศน์

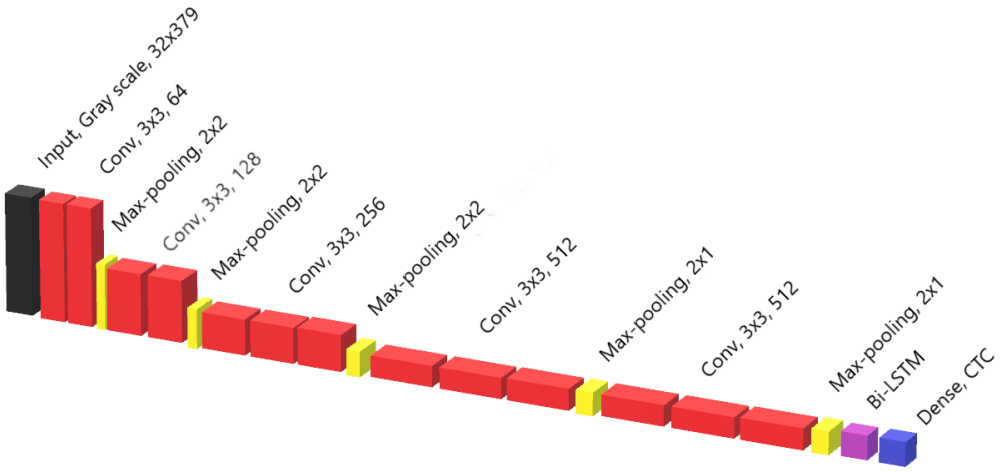
ข้อมูลที่ใช้ในการทดสอบในการรู้จำคำบรรยายวีดิทัศน์คือ รูปภาพคำบรรยายวีดิทัศน์ที่ใช้สำหรับการรู้จำ จำนวนทั้งสิ้น 4,224 รูปภาพ โดยแบ่งข้อมูลที่ใช้ในการทดสอบด้วยวิธี Cross-Validation โดยกำหนดให้ k=5

3.4.2 ทำการสร้างโมเดลในการรู้จำคำบรรยายวีดิทัศน์

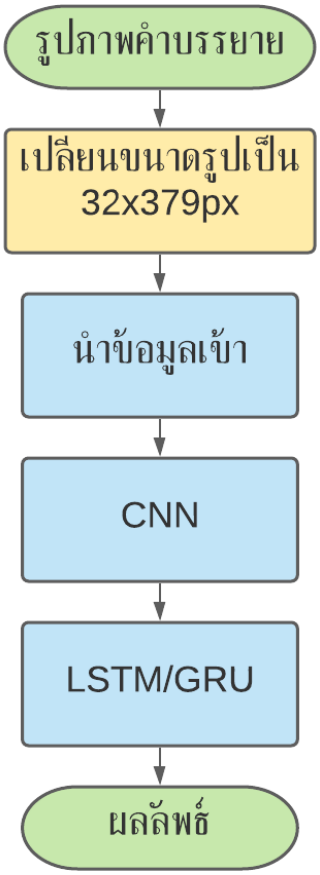
ด้วยวิธีการเรียนรู้เชิงลึกด้วยวิธี Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) และ Gated Recurrent Unit (GRU) ตามตารางที่ 3.1 ภาพประกอบที่ 3.12 และภาพประกอบที่ 3.13 ซึ่งเป็นภาพที่แสดงถึงขั้นตอนการทำงานของ CNN-LSTM ตารางที่ 3.2 ซึ่งเป็นตารางแสดงสถาปัตยกรรมที่ใช้ในวิทยานิพนธ์นี้ที่นำ VGG16 และ VGG19 มาใช้ร่วมกับ LSTM และ GRU ตารางที่ 3.3 ซึ่งเป็นสถาปัตยกรรมของ Chamchong et al. ซึ่งแตกต่างกับสถาปัตยกรรมที่ใช้ในวิทยานิพนธ์ที่จำนวนชั้นที่น้อยกว่าและขนาดของ max pooling ที่ชั้นสุดท้ายของจะใช้เป็นขนาด 5x1 เพื่อที่จะให้ขนาดออกมาตรงกัน

ตารางที่ 3.1 ตัวอย่างสถาปัตยกรรม CNN-LSTM

Stage	Operatiors	Resolution	Channels	Layers
1	Conv 3x3	32x379	64	2
2	Max-pooling 2x2			
3	Conv 3x3	16x189	128	2
4	Max-pooling 2x2			
5	Conv 3x3	8x94	256	3
6	Max-pooling 2x2			
7	Conv 3x3	4x94	512	3
8	Max-pooling 2x1			
9	Conv 3x3	2x94	512	3
10	Max-pooling 2x1			
11	Bi-LSTM	94	256	2
12	Dense & Softmax Function	157		1
13	CTC Loss Function			



ภาพประกอบที่ 3.12 สถาปัตยกรรม CNN-LSTM



ภาพประกอบที่ 3.13 กระบวนการทำงานในการรู้ตำคำบรรยาย

ตารางที่ 3.2 สถาปัตยกรรมที่ใช้ในการสร้างแบบจำลองการรู้จำคำบรรยายวีดิทัศน์

Model A	Model B	Model C	Model D
16 weight layers	16 weight layers	19 weight layers	19 weight layers
Input (gray image), 32x379			
3x3 conv,64	3x3 conv,64	3x3 conv,64	3x3 conv,64
3x3 conv,64	3x3 conv,64	3x3 conv,64	3x3 conv,64
Max-pooling 2x2, stride 2			
3x3 conv,128	3x3 conv,128	3x3 conv,128	3x3 conv,128
3x3 conv,128	3x3 conv,128	3x3 conv,128	3x3 conv,128
Max-pooling 2x2, stride 2			
3x3 conv,256	3x3 conv,256	3x3 conv,256	3x3 conv,256

ตารางที่ 3.2 สถาปัตยกรรมที่ใช้ในการสร้างแบบจำลองการรู้จำคำบรรยายวีดิทัศน์ (ต่อ)

Model A	Model B	Model C	Model D
3x3 conv,256	3x3 conv,256	3x3 conv,256	3x3 conv,256
3x3 conv,256	3x3 conv,256	3x3 conv,256	3x3 conv,256
		3x3 conv,256	3x3 conv,256
Max-pooling 2x1, stride 1			
3x3 conv,512	3x3 conv,512	3x3 conv,512	3x3 conv,512
3x3 conv,512	3x3 conv,512	3x3 conv,512	3x3 conv,512
3x3 conv,512	3x3 conv,512	3x3 conv,512	3x3 conv,512
		3x3 conv,512	3x3 conv,512
Max-pooling 2x1, stride 1			
3x3 conv,512	3x3 conv,512	3x3 conv,512	3x3 conv,512
3x3 conv,512	3x3 conv,512	3x3 conv,512	3x3 conv,512
3x3 conv,512	3x3 conv,512	3x3 conv,512	3x3 conv,512
		3x3 conv,512	3x3 conv,512
BiLSTM	BiGRU	BiLSTM	BiGRU
BiLSTM	BiGRU	BiLSTM	BiGRU
Dense & Softmax Function,157			
CTC Loss Function			

ตารางที่ 3.3 สถาปัตยกรรมตามบทความของ Chamchong et al.

Model 1	Model 2
6 weight layers	6 weight layers
Input (gray image), 32x379	
Conv 3x3, 16	Conv 3x3, 16
Max-pooling 2x2	
Conv 3x3, 32	Conv 3x3, 32
Max-pooling 2x2	
Conv 3x3, 32	Conv 3x3, 32
Max-pooling 5x1	

ตารางที่ 3.3 สถาปัตยกรรมตามบทความของ Chamchong et al. (ต่อ)

Model 1	Model 2
Bi-GRU	Bi-LSTM
Bi-GRU	Bi-LSTM
Dense & Softmax Function, 157	
CTC Loss Function	

ตัวอย่างการทำงานของ Model A โดยเริ่มจากการนำข้อมูลขนาดความสูง 32 พิกเซล และความกว้าง 379 พิกเซล ซึ่งเป็นขนาดที่ผ่านการเปลี่ยนขนาด (Resize) ด้วย cv2 resize ซึ่งจะทำให้สามารถรักษาอัตราส่วนเดิมได้ เข้าไปประมวลผลด้วยชั้น คอนโวลูชันชั้นแรกซึ่งมีขนาด 3x3 และมีค่า filter 64 และนำไปลดขนาดครึ่งหนึ่งโดยการดึงค่าสูงสุดที่ได้จากชั้นคอนโวลูชัน แล้วจึงนำไปเข้าสู่ชั้นคอนโวลูชันและชั้นพลูลิงไปจนได้ขนาด none,93,512 โดยที่เลข 93 หมายถึงความยาวของลำดับ (sequence length) และ 512 คือ filter แล้วนำไปเข้าสู่ชั้น LSTM หรือ GRU ที่มี 2 ชั้น โดย LSTM หรือ GRU จะทำงานเป็นคู่ซึ่งแต่ละคู่จะเรียกว่า BiLSTM หรือ BiGRU ซึ่ง LSTM ชั้นแรกจะทำงานจากซ้ายไปขวา และ LSTM ชั้นที่ 2 จะทำงานจากขวาไปซ้ายแล้วจึงนำผลลัพธ์ที่ได้ไปรวมกันที่ชั้นถัดไปซึ่งก็คือชั้น Dense โดยในชั้นนี้ข้อมูลจะมีลักษณะดังนี้ None,93,157 ซึ่งเลข 93 คือ เลขของความยาวลำดับ (sequence length) และ 157 คือเลขของข้อมูลที่เป็นไปได้ทั้งหมดบวกหนึ่งโดยจะเป็นตัวอักษรภาษาอังกฤษตัวเล็ก ตัวอักษรภาษาอังกฤษตัวใหญ่ พยัญชนะไทย สระ วรรณยุกต์ และ เลขไทย เพื่อใช้ในการหาคำตอบและนำไปถอดรหัสโดยใช้ CTC Loss ซึ่งเป็นหนึ่งในชั้น output โดยจะแปลงตัวเลขที่ได้จากชั้น Dense ให้เป็นตัวอักษรตามที่ได้กำหนดไว้

3.4.3 ทดสอบประสิทธิภาพของโมเดลในการรู้จำคำบรรยายวีดิทัศน์

ด้วยค่าความผิดพลาดระดับตัวอักษร (Character Error Rate: CER) โดยประโยคที่มีตัวอักษรเยอะที่สุดมี 90 ตัวอักษร ตัวอย่างผลลัพธ์ที่ได้จากการทดลอง ตามภาพประกอบที่ 3.14, ภาพประกอบที่ 3.15 และ ภาพประกอบที่ 3.16

Original text = ส่วนคำกล่าวขานที่ว่า

Predicted text = ส่วนคำกล่าวขานที่ว่า

ภาพประกอบที่ 3.14 ตัวอย่างผลลัพธ์มีผลลัพธ์ตรงกับข้อมูลนำเข้าทั้งหมด

ดังนั้นจากสูตรการหาค่า $\mathrm{CER} = \frac{\mathrm{I+S+D}}{\mathrm{N}}$ ดังนั้น $\mathrm{CER} = \frac{0+0+0}{20} = 0.0$ หรือ ผิดพลาด 0%

Original text = หลุมหลบภัยทางอากาศของทหารญี่ปุ่นเท่านั้น

Predicted text = หลุมหลบภัยทางอากาศของทหารญี่ปุ่นแท่านั้น

ภาพประกอบที่ 3.15 ตัวอย่างผลลัพธ์โดยมีการเปลี่ยนจาก “เ” เป็น “แ” 1 ตัวอักษร จากทั้งหมด 40 ตัวอักษร

ดังนั้นจากสูตรการหาค่า $\mathrm{CER} = \frac{\mathrm{I+S+D}}{\mathrm{N}}$ ดังนั้น $\mathrm{CER} = \frac{0+1+0}{40} = 0.025$

หรือผิดพลาด 2.5%

Original text = นำข้อมูลอันเป็นเท็จเข้าสู่ระบบคอมพิวเตอร์

Predicted text = นำข้อมูลอันเป็นเกจเข้าสู้ระบบคอมพิวเตอร์

ภาพประกอบที่ 3.16 ตัวอย่างผลลัพธ์โดยมีการเปลี่ยนจาก “ท” เป็น “ก” ที่ตำแหน่ง 16 “ ็ ” หายไปตรงตำแหน่งที่ 17 ทำให้หลังตำแหน่งที่ 17 จะเปลี่ยนไปทั้งหมดรวมถึง “ ่ ” เป็น “ ้ ”

ดังนั้นจากสูตรการหาค่า $\mathrm{CER} = \frac{\mathrm{I+S+D}}{\mathrm{N}}$ ดังนั้น $\mathrm{CER} = \frac{1+23+0}{41} = 0.59$

หรือผิดพลาด 59%

บทที่ 4
ผลการวิจัย

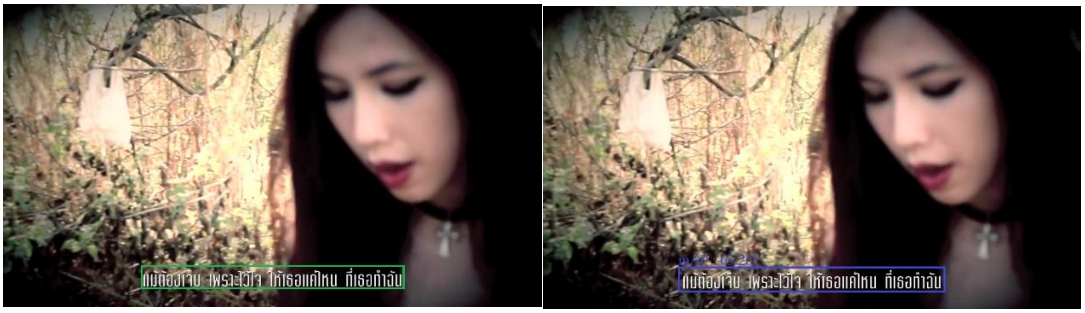
ในบทนี้แสดงถึงผลลัพธ์ที่ได้จากการทดลองโดยการใช้ TensorFlow framework ใน Google Colab และใช้ฟังก์ชันของ Google Colab ในการดูว่า GPU ที่ใช้เป็น GPU อะไร โดยเลือกใช้ GPU Tesla T4 ในการทดลองตรวจจับคำบรรยาย และ GPU Tesla P100 ในการรู้จำคำบรรยาย โดยใช้ dataset ของคำบรรยายวีดิทัศน์ที่ได้มาจากการใช้วิธี 5-fold cross-validation โดยที่มีค่า mean Average Precision (mAP) ที่ Intersect over Union (IoU) = 0.5 ในการวัดประสิทธิภาพการตรวจจับคำบรรยายวีดิทัศน์ ซึ่งหากค่า mAP มีค่ามากแสดงว่าการทดสอบนั้นมีผลลัพธ์ที่ดีกว่าแบบที่มีผลลัพธ์ที่มีค่ารน้อยกว่า และค่า Character Error Rate (CER) เป็นค่าสำหรับวัดประสิทธิภาพของการรู้จำคำบรรยายวีดิทัศน์ ถ้าหากค่า CER มีค่าน้อยแสดงว่าผลลัพธ์นั้นดีกว่า

ตารางที่ 4.1 ผลลัพธ์ค่า mAP โดยกำหนดให้ค่า IoU = 0.5

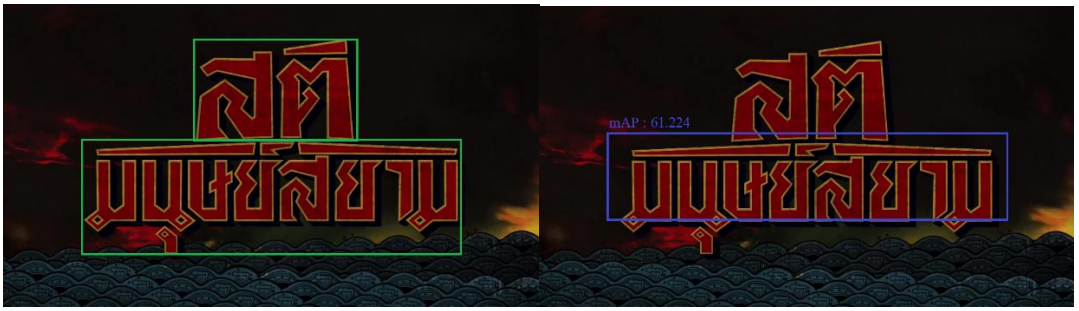
ชุดข้อมูล	**วิธีการที่ใช้ตรวจจับคำบรรยาย**		
	YOLO	**RetinaNet**	**tiny-YOLO**
set 1	90.73	90.39	87.00
set 2	85.67	89.56	86.85
set 3	88.84	91.20	88.33
set 4	91.78	92.67	93.33
set 5	87.25	93.32	91.39
mean	**88.85**	**91.43**	**89.38**
SD	**2.49**	**1.56**	**2.86**

ผลลัพธ์ซึ่งได้จากการทดลองแต่ละวิธีกับชุดของข้อมูลที่แตกต่างกันซึ่งจากการทดลองแสดงให้เห็นว่าผลลัพธ์ที่ดีที่สุด คือ Tiny-YOLO ซึ่งมีค่า mAP 93.33% แต่หากวัดจากค่าเฉลี่ยแล้ววิธีที่ดีที่สุดคือ RetinaNet ซึ่งมีค่า mAP 91.43% สามารถดูตัวอย่างรูปภาพ Ground truth ที่ภาพประกอบที่ 4.1 (ภาพฝั่งซ้าย) และตัวอย่างผลลัพธ์ในการตรวจจับที่ภาพประกอบที่ 4.1 (ภาพฝั่งขวา)

ก)



ข)



ค)

ภาพประกอบที่ 4.1 ภาพผลลัพธ์การตรวจจับตำแหน่งคำบรรยาย รวมไปถึงร้อยละที่ถูกต้องและตำแหน่งของกรอบที่ถูกวาดตามภาพ ก) ภาพผลลัพธ์ที่มีกรอบมากกว่า ground truth, ข) ภาพผลลัพธ์ที่มีกรอบเท่ากับ ground truth และ ค) ภาพผลลัพธ์ที่มีกรอบน้อยกว่า ground truth

ผลลัพธ์ค่า CER ซึ่งได้จากการทดลองแต่ละต้นแบบกับชุดข้อมูลที่แตกต่างกัน จากผลการทดลองพบว่า Model C ที่ทดลองโดยใช้ Batch size 128 มีค่า CER ที่น้อยที่สุดคือ 9.36% และ Model B ที่ Batch size 128 ที่มีค่า CER 10.11% ซึ่งน้อยเป็นอันดับ 2 ตามตารางที่ 4.2 และผลลัพธ์ที่ได้จากการนำสถาปัตยกรรมของ Chamchong et al. มาใช้พบว่า Model 1 ที่ทดลองโดยใช้ Batch size 32 มีค่าน้อยที่สุดคือ 17.35% ตามตารางที่ 4.3 ดังนั้นจึงได้นำข้อมูลค่า CER เวลาที่ใช้ในการเรียนรู้ (Train) และเวลาในการแสดงผลต่อ 1 รูป ของ Model C, Model B และ Model 1 มาเปรียบเทียบกันตามตารางที่ 4.4 ซึ่งพบว่า Model C มีผลลัพธ์ที่ดีกว่าโดยมีค่า CER 9.36%, Model B มีค่า 11.19 และ Model 1 มีค่า CER 19.13% แต่ใช้เวลาเพียง 27.09 นาที ซึ่งน้อยกว่าเวลาที่ใช้ในการเรียนรู้ของ Model A 5 เท่า โดยสามารถดูรูปภาพผลลัพธ์แบบสุ่มโดยเป็นภาพภาษาไทย ภาษาอังกฤษ และผสมอย่างละ 3 รูปที่ได้จากการรู้จำคำบรรยายด้วย Model C, Model B และ Model 1 ได้ที่ตารางที่ 4.5 นอกจากนี้ยังได้มีการทดสอบหาค่า CER และตัวอย่างผลลัพธ์ที่ได้จากการทดสอบกับรูปภาพที่สั้นโดยการสุ่ม 50 รูปที่มีน้อยกว่า 10 ตัวอักษรสามารถดูผลลัพธ์ได้ที่ตารางที่ 4.6 ตารางที่ 4.7 ภาพประกอบที่ 4.2 ภาพประกอบที่ 4.3 ภาพประกอบที่ 4.4 และภาพประกอบที่ 4.5

ตารางที่ 4.2 ค่า CER ที่ได้จากสถาปัตยกรรมที่ใช้วิธี CNN และ LSTM

ขนาดของ Batch Size	**ค่าความผิดพลาด (CER Value (%)) ที่ได้จากการรู้จำของแต่ละโมเดล**			
	Model A	**Model B**	**Model C**	**Model D**
32	22.61	18.57	17.84	13.50
64	12.37	15.34	19.39	17.12
128	18.38	**10.11**	**9.36**	16.53

ตารางที่ 4.3 ค่า CER ที่ได้จากการใช้สถาปัตยกรรมตามบทความของ Chamchong et al.

ขนาดของ Batch Size	**ค่าความผิดพลาด (CER Value (%)) ที่ได้จากการรู้จำของแต่ละโมเดล**	
	Model 1	**Model 2**
32	**35.17**	14.24
64	68.20	87.22
128	39.22	65.26

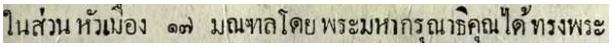
ตารางที่ 4.4 การเปรียบเทียบค่า CER และ เวลาที่ใช้ในการเรียนรู้ (Train)

Model	ขนาดของ Batch Size	CER Value	เวลาที่ใช้ในการเรียนรู้ (Training Time: Min.)	เวลาที่ใช้ในการรู้จำต่อหนึ่งรูป (Sec.)
C	128	**9.36±0.91**	115.43	0.002018
B	128	11.19±1.75	90	0.002027
1	32	19.13**±1.37**	23.12	0.001983

ตารางที่ 4.5 ตัวอย่างผลลัพธ์การทดลองการรู้จำคำบรรยาย

รูปภาพและ Ground truth	ผลลัพธ์การรู้จำ
 **Ground Truth text:** ให้วันเวลาสร้างเราขึ้นใหม่	**Model C:** ให้**ว อา**สร้างเราขึ้นใหม่ **Model B:** ให้วันเวลาสร้างเราขึ้นใหม่ๆ **Model 1:** ใ**ฬ่วรร้าน**เรา**นี่**นใหม่
 **Ground Truth text:** ฆ่าคนไทย 8 หมื่น ทั่วโลก 50 ล้าน!	**Model C:** ฆ่าคนไทย 8 หมื่น ทั่วโลก 50 ล้าน!) **Model B:** ฆ่าคนไทย 8 หมื่น ทั่วโลก 50 ล้าน! **Model 1:** **ข่**าคนไทย **2**ห**มี่**น **ก้**วโลก 50 ล้าน**ป้**

ตารางที่ 4.5 ตัวอย่างผลลัพธ์การทดลองการรู้จำคำบรรยาย (ต่อ)

รูปภาพและ Ground truth	ผลลัพธ์การรู้จำ
 **Ground Truth text:** ในส่วนหัวเมือง ๑๗ มณฑลโดยพระมหากรุณาธิคุณได้ทรงพระ	**Model C:** ในส่วนหัวเมือง ๑๗ มณฑลโดยพระมหากรุณาธิคุณได้ทรงพระ **Model B:** ในส่วนหัวเมือง ๑๗ มณฑลโดยพระมหากรุณาธิคุณได้ทรงพระๆ **Model 1:** ใน**จ้น้**นือง **๑๔ พรเศารษระาทรณ กงยฏ้ทรลนะะ**
 **Ground Truth text:** ทฤษฎีสมคบคิด MH370	**Model C:** ทฤษฎีสมคบคิด MH370 **Model B:** ทฤษฎีสมคบคิด MH370- **Model 1:** ทฤษฎีสมคบคิด MH370
 **Ground Truth text:** WITHIN YOUR GAZE CONTAINS	**Model C:** WITHIN YOUR GAZE CONTAINS **Model B:** WITHIN YOUR GAZE CONTAINS, **Model 1:** WITHIN YOUR GAZE CONTAINS
 **Ground Truth text:** Snow flake melting in my hand	**Model C:** Snow flake melting in my hand Model B: Snow flake melting in my hand, **Model 1:** Snow flake melting in my hand

ตารางที่ 4.5 ตัวอย่างผลลัพธ์การทดลองการรู้จำคำบรรยาย (ต่อ)

รูปภาพและ Ground truth	ผลลัพธ์การรู้จำ
 **Ground Truth text:** Just as thick blood warms a cold blade	**Model C:** Just as thick blood warms a cold blade **Model B:** Just as thick blood warms a cold blades **Model 1:** Just as thick blood warms a cold blade
 **Ground Truth text:** 11 ธ.ค.53	**Model C:** 11 ธ.ค.53 **Model B:** 11 ธ.ค.53. **Model 1:** 11ธ.ค.53
 **Ground Truth text:** ไวรัสปิศาจ	**Model C:** ไวรัสปิศาจ **Model B:** ไวรัสปิศาจๆ **Model 1:** ไวรัสปิ ศาจ

ตารางที่ 4.6 ผลลัพธ์ค่า CER จากการทดลองกับรูปที่มีความยาวน้อย

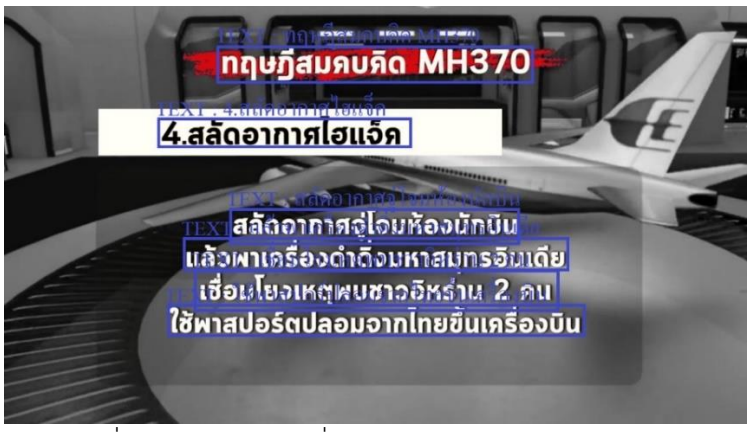
Model	CER
C	**24.88**
B	36.59
1	33.69

ตารางที่ 4.7 ตัวอย่างผลลัพธ์จากการทดลองกับรูปที่มีความยาวน้อย

รูปภาพและ Ground truth	ผลลัพธ์การรู้จำ
 **Ground Truth text:** เออ	Model C: โอ6 Model B: เวๆ Model 1: คอ

ตารางที่ 4.7 ตัวอย่างผลลัพธ์จากการทดลองกับรูปที่มีความยาวน้อย (ต่อ)

 Ground Truth text: MENU	Mode C: MENUN Mode B: MENA, Mode 1: GA
 Ground Truth text: 2562	Model C: 2563 Model B: 266, Mode 1: 2563
	Model C: ครับ Model B: ดรัๆ Model 1: ดรับ
	Model C: นี่ Model B: นี่ๆ Model 1: นี่
	Model C: 9 Model B: 9, Model 1: 9



ภาพประกอบที่ 4.3 ภาพผลลัพธ์ที่ได้จากการตรวจจับและรู้จำคำบรรยาย

Text: ทฤษฎีสมคบคิด MH370

Text: 4.สลัดอากาศไฮแจ็ค

Text: สลัดอากาศจู่โจมห้องนักบิน

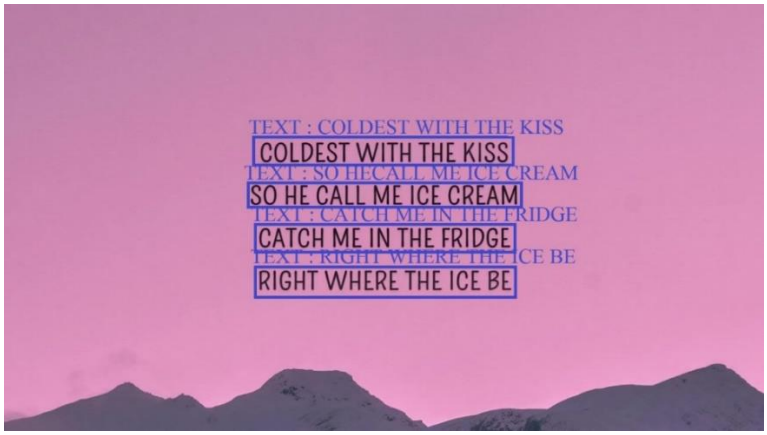
Text: แล้วพาเครื่องดำดิ่งมหาสมุทรอินเดีย

Text: เชื่อมโยงเหตุพบชาวอิหร่าน 2 คน

Text: ใช้พาสปอร์ตปลอมจากไทยขึ้นเครื่องบิน

ภาพประกอบที่ 4.3 ภาพผลลัพธ์ที่ได้จากการตรวจจับและรู้จำคำบรรยาย

Text: ได้ไหม



ภาพประกอบที่ 4.4 ภาพผลลัพธ์ที่ได้จากการตรวจจับและรู้จำคำบรรยาย ภาษาอังกฤษ

Text: COLDEST WITH THE KISS

Text: SO HE CALL ME ICE CREAM

Text: CATCH ME IN THE FRIDGE

Text: RIGHT WHERE THE ICE BE

ภาพประกอบที่ 4.5 ภาพผลลัพธ์ที่ได้จากการตรวจจับและรู้จำคำบรรยาย ภาษาอังกฤษ

Text: YOU CALL ME ON THE TELEPHONE

Text: YOU FEEL SO FAR AWAY

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

งานวิจัยนี้ศึกษาเกี่ยวกับการตรวจจับคำบรรยายโดยใช้ YOLOv3, Tiny-YOLOv3, RetinaNet ซึ่งผลลัพธ์ที่ได้นั้น การเรียนรู้เชิงลึกโดยใช้สถาปัตยกรรมที่เรียกว่า CNN-LSTM ในการรู้จำคำบรรยายจากวีดิทัศน์โดยได้สร้างสถาปัตยกรรม CNN 4 แบบโดยอ้างอิงมาจาก VGG16 และ VGG19 แล้วนำไปเชื่อมกับ Bi-LSTM หรือ Bi-GRU และ ฝึกโดยการใช้ CTC Loss

ผลการทดสอบการตรวจจับคำบรรยาย ผลลัพธ์ที่ดีที่สุดโดยวัดจากค่า mAP ที่ใช้ IoU=0.5 พบว่าวิธี YOLOv3 มีค่า mAP 88.85%, Tiny-YOLOv3 มีค่า 89.38% และ RetinaNet มีค่า 91.43% ดังนั้นจึงวิธี RetinaNet จึงเป็นวิธีที่สุดในการตรวจจับคำบรรยายของ dataset นี้

ผลการทดสอบการรู้จำคำบรรยาย ผลลัพธ์ที่ดีที่สุดโดยวัดจากค่า CER พบว่า model A มีค่า CER 12.37%, Model B 10.11%, Model C 9.36%, Model D 13.50%, Model 1 17.35%, Model 2 22.87% ซึ่งหากเรียงตามลำดับผลลัพธ์จากดีสุดไปแย่สุดได้ดังนี้ Model C, Model B, Model A, Model D, Model 1, และ Model 2 ดังนั้น Model C จึงเป็น Model ที่ดีที่สุดในการรู้จำคำบรรยายของ dataset นี้

ผลลัพธ์การเปรียบเทียบเวลาในการเรียนรู้และเวลาในการแสดงผลการรู้จำต่อ 1 ภาพของ Model C,Model B และ Model 1 ที่ โดยมีจำนวนชั้น CNN 19, 16 และ 4 ชั้นตามลำดับพบว่า Model C ใช้เวลาในการเรียนรู้ 115.43 นาที Model B ใช้เวลาในการเรียนรู้ 90 นาทีและ Model 1 ใช้เวลาในการเรียนรู้ 23.12 นาที ซึ่งน้อยกว่า Model C 92.31 นาที และน้อยกว่า Model B 66.48 นาที และ Model B น้อยกว่า Model C 25.43 นาที โดยเวลาในการแสดงผลการรู้จำต่อ 1 ภาพ สำหรับ Model C คือ 0.002018, Model B คือ 0.002027 และ Model 1 คือ 0.001983

ในส่วนของการรู้จำคำบรรยายที่สั้นนั้น Model ที่มีค่า CER น้อยที่สุดคือ Model C 24.88%, Model B 36.59% และ Model 1 33.69%

5.2 การอภิปรายผล

จากการทดลองการตรวจจับคำบรรยายวิดิทัศน์พบว่าค่า mAP ที่ดีที่สุดที่ใช้ค่า IoU ที่ 0.5 นั้นคือ 91.43% โดยผลลัพธ์ที่ได้นั้นส่วนมากจะมีลักษณะเป็นกรอบที่อยู่ตรงกับตำแหน่งของ ground truth มากกว่า 50% แต่ว่ามีบางภาพโดยเฉพาะรูปภาพที่มีคำบรรยายมากกว่า 1 คำบรรยายซึ่งจะทำให้มี ground truth หลายตำแหน่งทำให้ผลลัพธ์ออกมามีจำนวนกรอบมากกว่าคำบรรยายโดยที่กรอบนั้นจะเป็นกรอบใหญ่ซึ่งจะครอบพื้นที่คำบรรยายทั้งหมดเช่น รูปภาพที่มี 2 คำบรรยาย ทั้ง 2 คำ

บรรยายนั้นจะมีกรอบที่ครอบพอดีหรือมากกว่า น้อยกว่า เล็กน้อยจากตำแหน่งที่กำหนดใน ground และจะมีกรอบขนาดใหญ่มาครอบ 2 คำบรรยายนั้นเข้าด้วยกันเนื่องจากข้อมูลที่นำเข้านั้นอาจจะมีพื้นที่ของ ground truth ที่ซ้อนทับกันดังนั้นจึงอาจจะคิดว่าทั้ง 2 คำบรรยายนั้นเป็นคำบรรยายเดียวกัน ทำให้ผลลัพธ์ที่ออกมามีจำนวนกรอบที่มากกว่าจำนวน ground truth

ในส่วนของกการทดลองการรู้จำคำบรรยายวิดิทัศน์พบว่าค่าที่ดีที่สุดมีค่า CER 9.36% ซึ่งผลลัพธ์ที่ได้นั้นมีหลากหลาย เช่น มีภาษาอังกฤษแทรกในผลลัพธ์ที่มาจากรูปภาพนำเข้าที่มีแต่ภาษาไทย มีผลลัพธ์ภาษาไทยแทรกในผลลัพธ์ที่มาจากรูปนำเข้าที่มีแต่ภาษาอังกฤษ มีตัวเลขไทยแทรกในผลลัพธ์ในภาพนำเข้าที่มีแต่ภาษาอังกฤษ และอื่นๆ ซึ่งสิ่งที่ทำให้ได้ผลลัพธ์เช่นนี้อาจจะมาจาก มีรูปภาพ dataset ที่น้อยเกินไป มีรูปภาพที่มีหลายภาษา คือ ไทย อังกฤษ ทั้งยังมีเลขไทย และเลขอารบิก ฟอนต์ของตัวอักษรที่ไม่เหมือนกัน ขนาดของรูปภาพ และ การปรับขนาดของรูปภาพ เป็นต้น

5.3 ข้อเสนอแนะและงานวิจัยในอนาคต

งานในอนาคต เพื่อที่จะพัฒนาการรู้จำคำบรรยายวีดิทัศน์ ได้มีการวางแผนที่จะทดลองด้วยสถาปัตยกรรม CNN แบบอื่น เช่น Fusion Convolutional Neural Network (Fusion CNN) [28] ที่เป็นการรวม CNN ตั้งแต่ 2 input เข้าด้วยกัน รวมไปถึงวิธีในการเตรียมข้อมูล เช่น การปรับขนาด การลบสิ่งที่ไม่ใช่ข้อความออกจากรูปภาพ การเปลี่ยนสีรูปภาพ [20] การลดรายละเอียดรูปภาพ และการใช้พจนานุกรมในการหาผลลัพธ์ [26] ซึ่งอาจจะสามารถช่วยทำให้การทดสอบมีประสิทธิภาพดีขึ้น

ภาคผนวก

ภาคผนวก ก

ตัวอย่างรูปภาพที่ตัดจากวีดิทัศน์และคำตอบ (Ground truth) แสดงบริเวณที่เป็นคำบรรยาย

ก.1 ตัวอย่างภาพที่มีคำบรรยายภาษาอังกฤษ

สามารถดาวน์โหลดได้จากลิงก์ https://drive.google.com/drive/folders/1w212u-exPzbS-IZTM5s8VDXMx69GFluj?usp=sharing

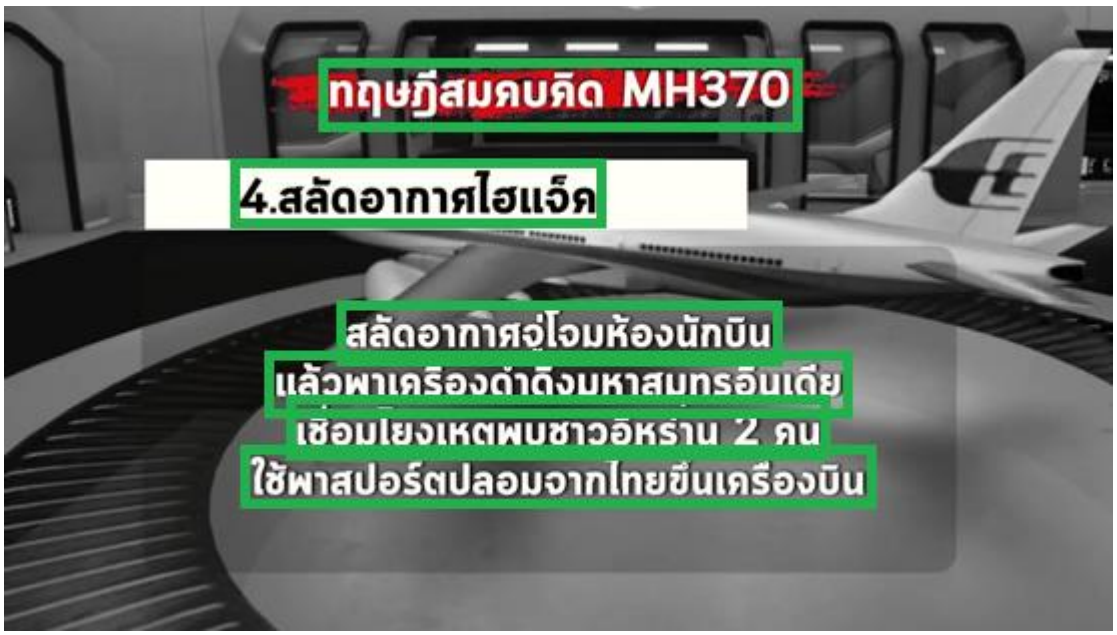
네 곁에서 멀어지지 않아

ne gyeoteseo meoreojiji anha

I'm not getting far away by your side

ก.2 ตัวอย่างภาพที่มีคำบรรยายภาษาไทย

ก.3 ตัวอย่างภาพที่มีคำบรรยายผสมระหว่างภาษาไทย ภาษาอังกฤษ และตัวเลข



ภาคผนวก ข

ตัวอย่างรูปภาพคำบรรยายและป้ายกำกับ (Label)

ข.1 ตัวอย่างภาพคำบรรยายภาษาไทย

สามารถดาวน์โหลดได้จากลิงก์: https://drive.google.com/drive/folders/1fYH01mactGss7r8d9zjvioqkv5l-E_hz

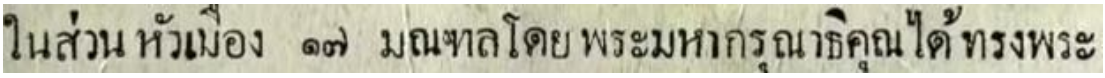
Label: ผู้รักษากฎหมาย

Label: ต้องลงมือเด็ดขาด

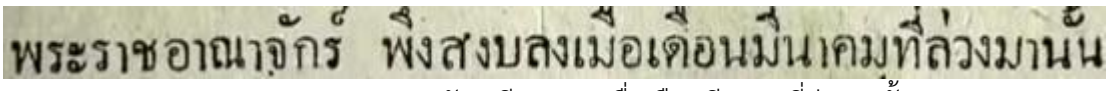
Label: ทอนซิลอักเสบ

Label: ไข้สูง ปวดศีรษะ

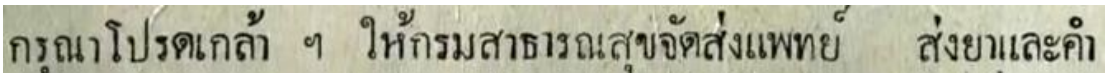
Label: จ้ำเลือดตามผิวหนัง



Label: ในส่วน หัวเมือง ๑๗ มณฑลโดยพระมหากรุณาธิคุณได้ทรงพระ



Label: พระราชอาณาจักร พึงสงบลงเมื่อเดือนมีนาคมที่ล่วงมานั้น



Label: กรุณาโปรดเกล้า ฯ ให้กรมสาธารณสุขจัดส่งแพทย์ ส่งยาและคำ

Label: ถ้าเป็นภาษาหนัง “ผมโดนเป็นสิบๆ เทคเลย”

Label: ถ้าอย่างใดอย่างหนึ่งผิดก็ต้องเริ่มใหม่

Label: ตอนนั้นผมเรียนเพาะช่างไปเจอสาวคนหฎึงบนรถเมล์

Label: ต่อมา ศาลชั้นต้นพิพากษาประหารชีวิต

Label: อาจตกอยู่ในสภาพโคม่าไม่ช้าไม่นาน

ข.2 ตัวอย่างภาพคำบรรยายภาษาอังกฤษ

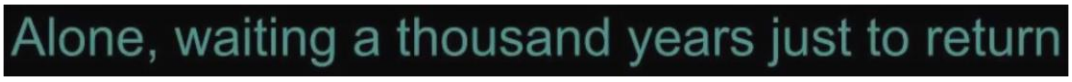
สามารถดาวน์โหลดได้จากลิงก์: https://drive.google.com/drive/folders/1JU_2Ff6Bx6xrhCZEIkZ0yZWSD11yMtkI?usp=sharing

Label: COMMANDER

Label: I REALLY LOVE

Label: CAUSE ONLY THOSE

Label: WILL EVER REALLY KNOW ME



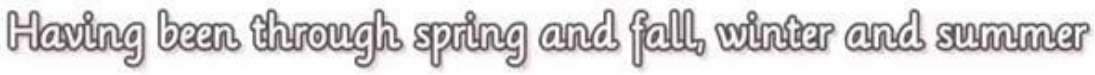
Label: Alone, waiting a thousand years just to return

Label: Dreaming of sentiments, feelings spreading a long distance

Label: Deep sadness, tears flow because of this love story

Label: Ju Jing Yi

Label: Ancient Painting



Label: Having been through spring and fall, winter and summer

Label: HOW COULD I JUST GLEEFULLY COMPLY PLACIDLY?

Label: YOU CAN ONLY LOSE WHAT YOU CLING TO

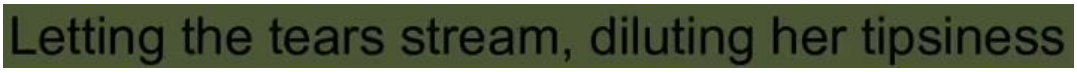
Label: INSTANTLY WHEN I THINK OF YOU – I JUST

Label: GET IT FREE LIKE WILLY

Label: SNOW CONE CHILLY

Label: IN THE JEANS LIKE BILLIE

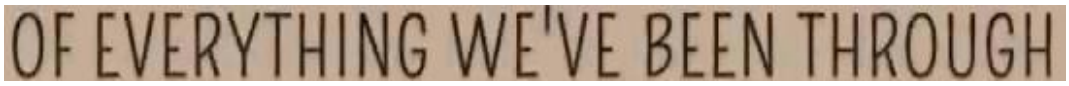
Label: Throughout the entire vacation, reading One Hundred Years of Solitude



Label: Letting the tears stream, diluting her tipsiness

Label: Playing the songs she loved as a girl is enough to make her dance

Label: TOAST TO THE ONES HERE TODAY



Label: OF EVERYTHING WE’VE BEEN THROUGH

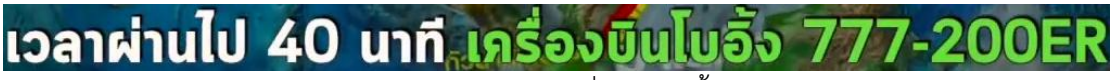
Label: TOAST TO THE ONES THAT WE LOST ON THE WAY

Label: I DON’Y GIVE A FUCK ABOUT YOU ANYWAYS

ข.3 ตัวอย่างภาพคำบรรยายที่ผสมระหว่างภาษาไทย ภาษาอังกฤษอังกฤษ และ ตัวเลข

สามารถดาวน์โหลดได้จากลิงก์ https://drive.google.com/drive/folders/1R1q_-i-U-WNXNLFIWXkJq9YaDg_lG4W9?usp=sharing

Label: 31 ธ.ค. 2019 จีนรายงาน WHO



Label: เวลาผ่าไป 40 นาที เครื่องบินโบอิ้ง 777-200ER

Label: ปี 2016 โลกได้พบชิ้นส่วนปีกของเครื่อง MH370

Label: MH370 ไม่มีคนคุมระหว่างตก เพียงแต่บินไปเรื่อยๆ

บรรณานุกรม

[1] Z. Zhu, W. Dai, Y. Hu, and J. Li, "Speech emotion recognition model based on Bi-GRU and focal loss," *Pattern Recogn Lett,* vol. 140, pp. 358-365, Dec 2020, doi: 10.1016/j.patrec.2020.11.009.

[2] C. Ma, L. Sun, Z. Zhong, and Q. Huo, "ReLaText: Exploiting visual relationships for arbitrary-shaped scene text detection with graph convolutional networks," *Pattern Recogn,* vol. 111, pp. 1-15, Mar 2021, doi: 10.1016/j.patcog.2020.107684.

[3] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *Computing Research Repository,* pp. 1-6, Apr 2018, doi: arxiv:1804.02767.

[4] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans Pattern Anal Mach Intell,* vol. 42, no. 2, pp. 318-327, Feb 2020, doi: 10.1109/TPAMI.2018.2858826.

[5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation,* vol. 9, no. 8, pp. 1735-1780, Nov 1997, doi: 10.1162/neco.1997.9.8.1735.

[6] K. Cho, B. Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *Computing Research Repository,* pp. 1-15, Sep 2014, doi: 10.3115/v1/D14-1179.

[7] S. M. Beitzel, E. C. Jensen, and O. Frieder, "MAP," in *Encyclopedia of Database Systems*, L. Liu and M. T. ÖZsu Eds., 1st ed. Boston, MA: Springer US, 2009, pp. 1691-1692.

[8] P. Wang, R. Sun, H. Zhao, and K. Yu, "A new word language model evaluation metric for character based languages," in *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, Berlin, Heidelberg, M. Sun, M. Zhang, D. Lin, and H. Wang, Eds., Jan 2013: Springer Berlin Heidelberg, pp. 315-324, doi: 10.1007/978-3-642-41491-6_29.

[9] B. P. Woolf, "Chapter 1 - Introduction," in *Building Intelligent Interactive Tutors*, B. P. Woolf Ed. San Francisco: Morgan Kaufmann, 2009, pp. 1-20.

[10] F. Bre, J. M. Gimenez, and V. D. Fachinotti, "Prediction of wind pressure coefficients on building surfaces using artificial neural networks," *Energy and Buildings,* vol. 158, pp. 1429-1441, Jan 2018, doi: 10.1016/j.enbuild.2017.11.045.

[11] A. D. Torres, H. Yan, A. H. Aboutalebi, A. Das, L. Duan, and P. Rad, "Patient facial emotion recognition and sentiment analysis using secure cloud with hardware acceleration," *Intelligent Data-Centric Systems,* pp. 61-89, Aug 2018, doi: 10.1016/B978-0-12-813314-9.00003-7.

[12] A. Ram and C. Reyes-Aldasoro, "The relationship between fully connected layers and number of classes for the analysis of retinal images," *Computing Research Repository,* pp. 1-10, Apr 2020, doi: https://arxiv.org/abs/2004.03624.

[13] Y. Xu *et al.*, "End-to-end subtitle detection and recognition for videos in East Asian languages via CNN ensemble," *Signal Process-Image,* vol. 60, pp. 131-143, Feb 2018, doi: 10.1016/j.image.2017.09.013.

[14] He, Huang, Wei, Li, and Guo, "TF-YOLO: An improved incremental network for real-time object detection," *Appl Sci-Basel,* vol. 9, no. 16, pp. 1-16, Aug 2019, doi: 10.3390/app9163225.

[15] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017, pp. 936-944, doi: 10.1109/CVPR.2017.106.

[16] J. Gan, W. Wang, and K. Lu, "In-air handwritten Chinese text recognition with temporal convolutional recurrent network," *Pattern Recogn,* vol. 97, pp. 1-15, Jan 2020, doi: 10.1016/j.patcog.2019.107025.

[17] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, *Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural 'networks*. Association for Computing Machinery, 2006, pp. 369-376.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, Apr 2015, pp. 1-14.

[19] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in

Conference on Computer Vision and Pattern Recognition (CVPR), Jun 2019, pp. 658-666, doi: 10.1109/CVPR.2019.00075.

[20] W. He, X.-Y. Zhang, F. Yin, Z. Luo, J.-M. Ogier, and C.-L. Liu, "Realtime multi-scale scene text detection with scale-based region proposal network," *Pattern Recogn,* vol. 98, pp. 1-14, Feb 2020, doi: 10.1016/j.patcog.2019.107026.

[21] Y. Wang, L. Peng, and S. Wang, "A multi-stage method for Chinese text detection in news videos," *Procedia Computer Science,* vol. 96, pp. 1409-1417, Sep 2016, doi: 10.1016/j.procs.2016.08.186.

[22] Y. Zhu and J. Du, "TextMountain: Accurate scene text detection via instance segmentation," *Pattern Recogn,* vol. 110, pp. 1-12, Feb 2021, doi: 10.1016/j.patcog.2020.107336.

[23] M. Huang, C. H. Lan, W. Huang, and Y. Tao, "Natural scene text detection based on multiscale connectionist text proposal network," *J Eng-Joe,* vol. 2020, no. 13, pp. 326-329, Jul 2020, doi: 10.1049/joe.2019.1154.

[24] H. Y. Yan and X. Xu, "End-to-end video subtitle recognition via a deep residual neural network," *Pattern Recogn Lett,* vol. 131, pp. 368-375, Mar 2020, doi: 10.1016/j.patrec.2020.01.019.

[25] S. K. Jemni, Y. Kessentini, and S. Kanoun, "Out of vocabulary word detection and recovery in Arabic handwritten text recognition," *Pattern Recogn,* vol. 93, pp. 507-520, Sep 2019, doi: 10.1016/j.patcog.2019.05.003.

[26] J. X. Zhang, C. J. Luo, L. W. Jin, T. W. Wang, Z. Y. Li, and W. Y. Zhou, "SaHAN: Scale-aware hierarchical attention network for scene text recognition," *Pattern Recogn Lett,* vol. 136, pp. 205-211, Aug 2020, doi: 10.1016/j.patrec.2020.06.009.

[27] R. Chamchong, W. Gao, and M. D. McDonnell, "Thai handwritten recognition on text block-based from Thai archive manuscripts," in *International Conference on Document Analysis and Recognition*, Sep 2019, pp. 1346-1351, doi: 10.1109/ICDAR.2019.00217.

[28] Y. Lavinia, H. H. Vo, and A. Verma, "Fusion based deep CNN for improved large-scale image action recognition," in *IEEE International Symposium on Multimedia*, Dec 2016, pp. 609-614, doi: 10.1109/ISM.2016.0131.

ประวัติผู้เขียน

ชื่อ	ธนดล สิงขรอาสน์
วันเกิด	13 เมษายน 2541
สถานที่เกิด	จังหวัดมหาสารคาม
สถานที่อยู่ปัจจุบัน	5 ถนนริมคลองสมถวิล ตำบลตลาด อำเภอเมือง จังหวัดมหาสารคาม 44000
ตำแหน่งหน้าที่การงาน	ลูกจ้างชั่วคราว
สถานที่ทำงานปัจจุบัน	กองแผนงาน มหาวิทยาลัยมหาสารคาม
ประวัติการศึกษา	2549-2552 โรงเรียนสาธิตมหาวิทยาลัยมหาสารคาม (ฝ่ายประถม) 2553-2558 โรงเรียนสาธิตมหาวิทยาลัยมหาสารคาม (ฝ่ายมัธยม) 2559-2562 เทคโนโลยีสารสนเทศ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม