ระบบแปลคำศัพท์ภาษามือไทยโดยการเรียนรู้เชิงลึกบนข้อมูลบางส่วน

THAI SIGN LANGUAGE TRANSLATION SYSTEM THROUGH FEW SHOT LEARNING

ณัฏยา เปลี่ยนวงษ์[1]

ดร.ธนภัทร ฆังคะจิตร[2]

บทคัดย่อ

ผู้พิการทางการได้ยินมีมากเป็นอันดับสองจากจำนวนผู้พิการทั้งหมดในประเทศไทย อุปสรรคหลักในการสื่อสารผ่านทางภาษามือกับผู้พิการทางการได้ยินเกิดจากคำศัพท์บางคำมีท่าทางเฉพาะที่ยากต่อการคาดเดา อย่างไรก็ตามมีงานวิจัยที่นำเสนอการรู้จำคำศัพท์ภาษามือไทยจากวิดีโอตัวอย่างโดยใช้เทคนิคการเรียนรู้เชิงลึก LSTM ซึ่งให้ความแม่นยำสูง แต่ใช้กับคำศัพท์ภาษามือจำนวนน้อยเพียง 5 คำและใช้วิดีโอจำนวนมากถึง 100 วิดีโอในการสอนแบบจำลอง ทำให้เป็นการยากในการสร้างแบบจำลองเพื่อเรียนรู้คำศัพท์ที่มีจำนวนมากขึ้นเนื่องจากต้องใช้ผู้เชี่ยวชาญในการบันทึกวิดีโอตัวอย่าง ดังนั้นงานนี้จึงนำเสนอแนวทางในการสร้างระบบแปลคำศัพท์ภาษามือไทยในชีวิตประจำวัน 47 คำโดยใช้เทคนิคการเรียนรู้เชิงลึกจากตัวอย่างจำนวนน้อย (Few Shot Learning) สำหรับวิดีโอสอนระบบมาจากการสร้างขึ้นของผู้วิจัยโดยอ้างอิงจากวิดีโอของผู้เชี่ยวชาญ แล้วทำการสกัดคุณลักษณะในแต่ละเฟรมของวิดีโอสอนระบบจากท่าทางจากตำแหน่งของมือและใบหน้า รวมถึงการทำมุมของแขนท่อนบนกับท่อนปลาย เพื่อนำมาให้แบบจำลองเรียนรู้รูปแบบ จากนั้นจึงนำแบบจำลองมาทดสอบด้วยวิดีโอของผู้เชี่ยวชาญซึ่งพบว่าแบบจำลองให้ความแม่นยำสูงถึง 74% โดยมีคำศัพท์ที่ทำนายผิดเกิดจากองค์ประกอบของท่าทางที่คล้ายกันหรือมีท่าทางซ้ำกันในตำแหน่งเดิมหลายครั้ง ดังนั้นในการเพิ่มประสิทธิภาพของแบบจำลองในอนาคตสามารถทำโดยใช้วิดีโอจำนวนน้อยจากผู้เชี่ยวชาญมาเพื่อสอนระบบแล้วจึงนำแบบจำลองไปใช้ในระบบจริง

คำสำคัญ: การรู้จำคำศัพท์ภาษามือไทย การเรียนรู้เชิงลึก การเรียนรู้เชิงลึกจากตัวอย่างจำนวนน้อย ตำแหน่งของมือและใบหน้า

---

[1] นักศึกษาหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมข้อมูลขนาดใหญ่

[2] ที่ปรึกษาสารนิพนธ์หลัก

ABSTRACT

In Thailand, the amount of deaf is the second-largest in total numbers of disabled people. The main obstacle to communicate with the deaf is that some words in sign language have specific gestures that are difficult to predict. The former research in Thai sign language recognition provided high accuracy on a few sign-language words using the deep learning technique. Contrastingly, a hundred sample videos per word were input to the LSTM classifier. An extensive effort from experts to produce training videos is required to build a model to learn a larger number of words. Therefore, this work presents a system to translate everyday-life Thai sign words using the few-shot learning technique. Firstly, the few training videos per word are constructed based on the experts' videos. Then, input features on each frame of the videos are extracted from hand and face positions and the angle of the upper and lower arms. Once the model is constructed, experts' videos are used as testing datasets. The experimental results show that our model produces high accuracy (74%) on experts' videos. The misclassified words seem to have similar postures or repeated gestures in the same position. Furthermore, improving the model's performance requires a small number of accurate training videos from experts.

Keywords: Thai Sign Language, Translation System, Few Shot Learning

บทนำ

สถานการณ์ด้านคนพิการในประเทศไทย ปี 2564 จำนวน 2,096,931 คน (ร้อยละ 3.17 ของประชากรทั้งประเทศ) พบว่าเป็นความพิการประเภท ทางการได้ยินหรือสื่อความหมาย 393,826 (ร้อยละ 18.78 ของจำนวนผู้พิการทั้งประเทศ) จัด เป็นอันดับ 2 ของประเภทความพิการทั้งหมด รองจากความพิการทางด้านการเคลื่อนไหว จะเห็นได้ว่ามีผู้พิการจำนวนมากที่ไม่สามารถสื่อสารกับบุคคลทั่วไปได้อย่างปกติ หากบุคคลเหล่านั้นไม่มีความรู้ทางด้านภาษามือ ซึ่งในบางครั้งคำศัพท์ภาษามือสามารถคาดเดาหรือสื่อความหมายได้ แต่ก็มีคำศัพท์จำนวนมากที่ค่อนข้างซับซ้อน จึงต้องอาศัยล่ามภาษามือในการสื่อสารระหว่างคนทั่วไปกับผู้พิการ แต่เนื่องจากทรัพยากรของล่ามภาษามือมีจำนวนไม่เพียงพอและยากต่อการรับบริการ อีกทั้งยังมีปัญหาอุปสรรคในการใช้ล่ามภาษามือ เช่น ประเด็นความน่าเชื่อถือ และไว้วางใจในล่ามภาษามือ ในบางครั้งผู้พิการอยากสื่อสารในเรื่องที่

ค่อนข้างละเอียดอ่อน หรือเป็นเรื่องสำคัญ ทำให้ไม่ไว้วางใจหากไม่ใช่ ญาติ หรือคนสนิท ดังนั้นหากมีการพัฒนาอุปกรณ์ หรือ ช่องทางในการสื่อสารระหว่างผู้พิการและบุคคลทั่วไปได้ จะช่วยลดช่องว่างของการสื่อสาร และแก้ไขปัญหาเบื้องต้นนี้ลงได้

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาระบบแปลคำศัพท์ภาษามือไทยในชีวิตประจำวัน โดยวิธีการที่สนใจได้แก่การเรียนรู้เชิงลึก เทคนิค Few-Shot Learning เนื่องจากปัจจุบันฐานข้อมูลภาษามือไทยยังไม่มีข้อมูลให้ใช้ได้อย่างแพร่หลาย ส่วนใหญ่จะเป็นภาพวาด หรือรูปถ่าย ที่ใช้ในการสะกดคำ ซึ่งค่อนข้างยากหากจะนำมาเรียนรู้ หลังจากสืบค้นข้อมูลพบว่า สมาคมคนหูหนวกแห่งประเทศไทยได้จัดทำระบบฐานข้อมูลภาษามือไทย รวบรวมคำศัพท์ในหนังสือภาษามือไทยเล่ม 1-6 ของสมาคมคนหูหนวกแห่งประเทศไทย (ไม่น้อยกว่า 1,000 คำ) ในแต่ละคำจะมีตัวอย่างท่าทางเป็นวิดีโอ 6 ตัวอย่าง

เนื่องจากข้อมูลที่มีอยู่อย่างจำกัด หากใช้วิธีการเรียนรู้แบบอัตโนมัติด้วยการเลียนแบบการทำงานของโครงข่ายประสาทของมนุษย์ (Deep Learning) อาจไม่เหมาะสมเพราะวิธีการดังกล่าวต้องใช้ตัวอย่างข้อมูลเป็นจำนวนมาก นำไปเรียนรู้ซ้ำ ๆ เพื่อหารูปแบบของคำตอบ อีกทั้งยังมีวิธีการซับซ้อน และต้องใช้ทรัพยากรในการประมวลผลมาก จึงไม่เหมาะสมกับงานวิจัยในครั้งนี้ ดังนั้นผู้วิจัยจึงเลือกวิธีการเรียนรู้แบบไม่กี่ตัวอย่าง (Few-Shot Learning) ในการสร้างระบบแปลคำศัพท์ภาษามือไทยในชีวิตประจำวัน ทำการสกัดคุณลักษณะในแต่ละเฟรมของวิดีโอสอนระบบจากท่าทางจากตำแหน่งของมือและใบหน้า รวมถึงการทำมุมของแขนท่อนบนกับท่อนปลาย เพื่อนำมาให้แบบจำลองเรียนรู้รูปแบบ

วัตถุประสงค์

1. เพื่อพัฒนาระบบแปลคำศัพท์ภาษามือไทยในชีวิตประจำวัน โดยที่แปลคำศัพท์ภาษามือไทยในชีวิตประจำวันได้เบื้องต้น
2. เพื่อสร้างโมเดลโดยใช้ตัวอย่างข้อมูลจำนวนน้อย

ขอบเขตของการวิจัย

1. สร้างโมเดลที่เหมาะสมเพื่อแปลคำศัพท์ภาษามือไทยโดยแบ่งออกเป็น 3 หมวด ได้แก่ หมวดความรู้สึกจำนวน 21 คำ, หมวดรสชาติจำนวน 9 คำ และหมวดคำกริยาจำนวน 17 คำ รวมทั้งหมด 47 คำ

2. คำศัพท์จากฐานข้อมูลภาษามือไทย โดยสมาคมคนหูหนวกแห่งประเทศไทย

ประโยชน์ที่คาดว่าจะได้รับ

1. ช่วยลดช่องว่างในการสื่อสารระหว่างผู้พิการ กับบุคคลทั่วไป และคู่สื่อสารไม่จำเป็นต้องมีความรู้ภาษามือ
2. ลดปัญหาความน่าเชื่อถือ และความไว้วางใจในล่าม
3. สามารถเข้าถึงได้ง่าย ลดปัญหาล่ามไม่เพียงพอ

แนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้อง

Careaga, C., Hutchinson, B., Hodas, N. & Phillips, L., (2019). Metric-Based Few-Shot Learning for Video Action Recognition งานวิจัยนี้กล่าวถึงการรู้จำท่าทางต่าง ๆ โดยใช้ชุดข้อมูล Kinetics 600 โดยเริ่มต้นนำวิดีโอท่าทางมาทำ RGB และOptical Flow หลังจากนั้นทำ pre-trained CNN ด้วย RestNet-18 และ Alext ตามลำดับ งานวิจัยนี้ได้ทดลองใช้ Aggregation Techniques 4 วิธี เพื่อนำแต่ละเฟรมมารวมกัน ได้แก่ Pooling, LSTM, ConvLSTM และ 3D Convolutions และในการทำ Few-Shot Learning ได้ใช้ โมเดลที่ได้รับความนิยม 3 โมเดลมาเปรียบเทียบกัน ได้แก่ Matching Networks, Prototypical Networks และ Learned distance metric

จากผลการวิจัยแสดงให้เห็นว่าการใช้โมเดล Prototypical Networks และAggregation Techniques LSTM เพื่อรวมแต่ละเฟรมเข้าด้วยกัน และจัดการความยาวของแต่ละวิดีโอ ให้ผล Accuracy ที่ 84.2 % ในการทดสอบด้วยข้อมูลปกติ และ 59.4 % ในชุดข้อมูลที่มีความท้าทาย ความท้าทายในกรณีนี้หมายถึงข้อมูลที่มีความคล้ายกันมาก ๆ

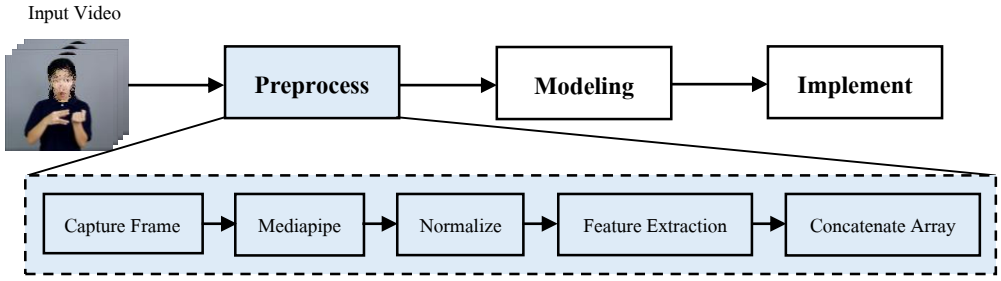
Chaikaew, A., Somkuan, K., Yuyen, T. (2021). Thai Sign Language Recognition: an Application of Deep Neural Network สำหรับงานวิจัยนี้ได้นำเสนอการรู้จำคำศัพท์ภาษามือไทยโดยสร้างชุดข้อมูลด้วยตนเอง จำนวน 5 ท่าทาง ท่าทางละ 100 วิดีโอ รวมทั้งสิ้น 500 วิดีโอ และทำการมาร์คจุดสำคัญบนมือด้วย Mediapipe ได้ผลลัพธ์ออกมาเป็น ตำแหน่งพื้นที่แต่ละจุด ซึ่งในวิจัยนี้มาร์คเฉพาะจุดบนฝ่ามือจำนวนข้างละ 21 จุด หลังจากนั้น แคปเจอร์เฟรมจากวิดีโอ นำออกเป็นไฟล์ CSV แล้วนำไปเทรนโมเดล โดยใช้ LSTM, BLSTM และGRU นำผลลัพธ์มาเปรียบเทียบกัน พบว่า LSTM ให้ประสิทธิภาพดีที่สุด มี Accuracy ที่ 97 % และ Loss ที่ 6 %

Snell, J., Swersky, K. & Zemel, R. (2017). Prototypical Networks for Few-shot Learning สำหรับงานวิจัยนี้ นำเสนอแนวคิดแบบจัดกลุ่ม และถูกนำไปใช้เพื่อทำนายป้ายกำกับของข้อมูลในขณะที่โมเดลการทำคลัสเตอร์นี้จะได้รับการฝึกอบรม (train) สำหรับแต่ละตอน (episode) และคำนวณการสูญเสีย (loss)

ในทุกตอนโมเดลจะถ่ายโอนทั้ง support set และquery set ไปยังเลเยอร์การ embedded เซนทรอยด์หรือจุดศูนย์กลางของกลุ่ม (เช่น c1, c2 และ c3 ในภาพ) คือค่าเฉลี่ยของ label ที่เกี่ยวข้อง ในขณะที่ query set (เช่น x ในภาพ) จะถูกจัดให้อยู่ในประเภทเดียวกันกับคลัสเตอร์ที่ใกล้ที่สุด เช่น c2 ในภาพ แทนที่จะใช้การหาระยะห่างแบบโคไซน์ (Cosine distance) Prototypical Networks ใช้ระยะห่างแบบยุคลิด (Euclidean distance) เพื่อคำนวณหาความแตกต่างของเซนทรอยด์ระหว่าง support set และquery set

ระเบียบวิธีวิจัย

การศึกษาวิจัยครั้งนี้เป็นการนำเสนอระบบแปลคำศัพท์ภาษามือไทยโดยการเรียนรู้แบบเชิงลึกด้วยเทคนิค Few-Shot Learning โดยมีแนวทางการวิจัยดังนี้



ภาพที่ 1 การทำงานของระบบแปลคำศัพท์ภาษามือไทยโดยการเรียนรู้แบบ Few-Shot Learning

1. การเก็บข้อมูล (Input Video)

1.1 ข้อมูลที่ใช้ในการศึกษา

ในการเก็บข้อมูล ทำการบันทึกคลิปวิดีโอโดยใช้ Resolution หรือความละเอียดคมชัดของวิดีโอ 1080p HD ที่ 30 fps (Frame Per Second) โดยตั้งกล้องถ่ายในพื้นหลังแบบหยุดนิ่ง และถ่ายเพียงแค่ท่อนบนของร่างกายตั้งแต่เอวขึ้นไป ซึ่งบันทึกทั้งหมด 47 คำ คำละ 6 วิดีโอ

ภาพที่ 2 ตัวอย่างรูปภาพในแต่ละเฟรม

2. การเตรียมข้อมูลสำหรับสอนระบบ (Preprocess)

2.1 Capture Frame

เมื่อได้รับ Input Video เข้ามาจะใช้ OpenCV ในการ Capture Frame ในแต่ละ Video ให้เป็นเฟรมภาพ และเก็บไว้ในโฟลเดอร์ โดยเฉลี่ยแล้วความยาวต่อคำอยู่ที่ 4 วินาที ดังนั้นจึงกำหนดให้วิดีโอสูงสุดมีความยาว 4 วินาที ใน 1 วินาทีสามารถ Capture Frame ได้ 30 Frame ดังนั้นจะได้ Frame ทั้งหมด 120 Frame แต่จากการทดลองปรับจำนวนการ Capture Frame ลงพบว่าหากลดลง 50 % หรือ 15 Frame ต่อวินาทีทำให้โมเดลมีประสิทธิภาพมากที่สุด

2.2 Mediapipe

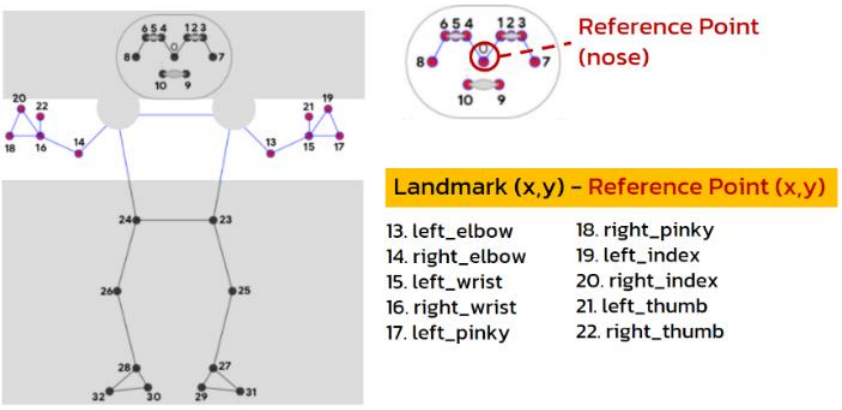
หลังจากดำเนินการขั้นตอนที่ 2.1 เรียบร้อยแล้ว และทำการระบุตำแหน่งของ Landmark หรือจุดสำคัญต่าง ๆ ของมือ และท่าทางโดยใช้ Mediapipe เก็บตำแหน่งของแต่ละจุด ค่าที่ได้เป็นพิกัดจุด 3 มิติ (x, y, z) บอกพิกัดตาม % ของขนาดภาพ สำหรับ x จะเทียบกับความกว้าง และ y จะเทียบกับความสูง เช่นภาพขนาด 1080 x 1080 หากได้ x: 0.50 และ y 0.10 หมายถึงพิกัดที่ x = 0.5 * 1080 และ y = 0.10 * 1080 หรือ (540, 108) ในพิกัด (x, y) ซึ่งจุด 0,0 เริ่มต้นที่มุมซ้ายบน และมีค่า x, y ที่มากที่สุดที่มุมขวาล่าง ซึ่งในงานวิจัยนี้จะพิจารณาตำแหน่งใน 2 มิติเท่านั้น ได้แก่ ค่าในแนวแกน x และค่าในแนวแกน y

เริ่มนับเฟรมแรกของวิดีโอเมื่อมีการรับค่าตำแหน่งของมือซ้ายหรือขวาในเฟรมนั้น ๆ เพื่อนำค่าไปทำ Normalize และFeature Extraction ต่อไป หลังจากเริ่มนับเฟรมแรกไปแล้ว เมื่อทำ Landmark แต่ละเฟรม แล้วไม่พบตำแหน่งที่ต้องการจะแทนค่าด้วยเมตริก 0 ตามขนาดเดิมในตำแหน่งนั้นแทน

2.3 Normalize Landmark

เนื่องจากสรีระร่างกาย หรือส่วนสูงของผู้ใช้งาน ไม่เท่ากันดังนั้นจึงจำเป็นต้องทำการ Normalize จุดต่าง ๆ ที่ได้จาก Landmark โดยผู้วิจัยจะถือให้ Pose [0] หรือตำแหน่งของจมูก เป็น

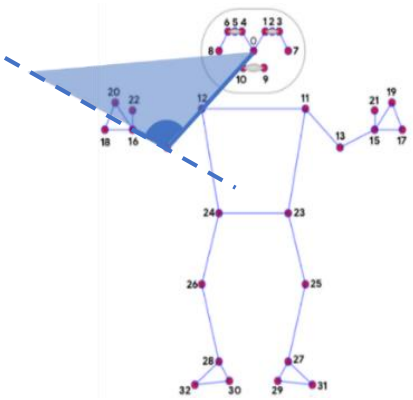
จุดอ้างอิงของแต่ละบุคคล หลังจากนั้นจะนำจุดต่าง ๆ มาลบด้วยตำแหน่งจมูก โดยตำแหน่งของร่างกายที่นำมาใช้ทั้งหมด 10 จุดดังภาพที่ 3



ภาพที่ 3 ตัวอย่างตำแหน่งอ้างอิงของจมูก และตำแหน่ง Pose ของ Mediapipe

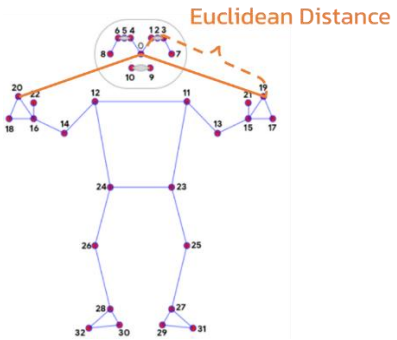
2.4 Feature Extraction

การทำมุมระหว่างแขนท่อนบนกับแขนท่อนปลาย เช่นการทำมุมของแขนด้านขวาใช้สร้างเวกเตอร์ระหว่างตำแหน่งที่ 12 หรือไหล่ ไปจนถึงตำแหน่งที่ 14 หรือศอก และสร้างเวกเตอร์ระหว่างตำแหน่งที่ 16 หรือข้อมือ ไปจนถึงตำแหน่งที่ 14 หรือศอก เป็นแขนของมุม กำหนดให้ตำแหน่งของศอกเป็นจุดยอดมุม หามุมที่กระทำระหว่างแขนด้านซ้ายเช่นเดียวกันกับด้านขวา



ภาพที่ 4 ตัวอย่างการทำมุมระหว่างแขนท่อนบนกับแขนท่อนปลายของแขนขวา

การหาระยะห่างระหว่างนิ้วชี้ของมือทั้งสองข้างกับจมูก โดยใช้การหาระยะทางแบบ Euclidean Distance จากตำแหน่งจมูกไปยังตำแหน่งปลายนิ้วชี้โดยใช้พิกัดจาก Pose



ภาพที่ 5 ตัวอย่างระยะห่างระหว่างนิ้วชี้กับจมูก

2.5 Concatenate Array

เมื่อดำเนินการ Normalize Landmark และ Feature Extraction เรียบร้อยแล้วจะได้ Array ของแต่ละตำแหน่ง กรณีเป็นพิกัดจุดจะแปลงให้เป็น Array มิติเดียวโดยใช้ Flatten หลังจากนั้นนำ Array มา Concatenate กันตามลำดับดังตารางที่ 1 จะได้ Array สำหรับ 1 เฟรมได้แก่ Array 1 มิติ มีสมาชิกจำนวน 39 Array

ตารางที่ 1 ตัวอย่างลำดับการ Concatenate Array ของตำแหน่งด้านขวา

ลำดับ	ตำแหน่ง	พิกัด	Shape (Array)
1	การทำมุมระหว่างแขนท่อนบนกับแขนท่อนปลาย	Pose(12,14) ทำมุมกับ Pose(14,16)	4
2	ระยะห่างระหว่างนิ้วชี้กับจมูก	ระยะห่างระหว่างPose(0) กับPose(20)	2
3	ตำแหน่งของท่าทาง	Pose(14,16,18,20,22)	33
รวม		Pose(มุมแขน) + Pose(ระยะห่างจมูกกับมือ) + Pose (ตำแหน่ง Pose - ตำแหน่งจมูก)	39

เนื่องจากข้อมูลวิดีโอคำศัพท์เป็นข้อมูลประเภท Sequence ดังนั้นการจะนำข้อมูลไปใช้ในโมเดลนั้นต้องนำ Array ของเฟรมทั้งหมดมารวมกันเป็น Input สำหรับ 1 วิดีโอ โดยนำ Array มาสร้างเป็น list แล้วจึงนำเข้าไปสอนโมเดล

3. การสร้างโมเดล (Few-Shot Learning Modeling)

ข้อมูลที่ได้จากข้อ 2.5 จะถูกใช้เป็นข้อมูลเพื่อสอนระบบ (Training Data) ทั้งหมด 282 วิดีโอ แบ่งการชุดการอบรมในการสอนระบบแต่ละงาน(Task / Episode) ดังนี้ N-Way = 47, K-Shot = 6, Query = 6 โดยสอนทั้งหมด 600 Task และทดสอบ 100 Task

หลังจากศึกษางานวิจัยที่เกี่ยวข้องพบว่ามีโมเดลที่นิยมใช้ในการทำ Few-Shot Learning อยู่ 3 โมเดล ได้แก่ Prototypical Networks[10], Relation Networks[11] และ Matching Networks[12]ดังนั้นจึงได้ทดลองใช้โมเดลทั้ง 3 โมเดลในการแปลคำศัพท์ เพื่อเลือกโมเดลที่มีประสิทธิภาพมากที่สุด ไปใช้งานจริง พื้นฐานของโมเดล Few-Shot Learning ทั้ง 3 แบบ คือ neural network ที่เป็น CNN แต่จะแตกต่างกันที่วิธีการจัดกลุ่ม หรือจัดประเภท ส่วนวิธีการสอน หรือหลักการในส่วนอื่น ๆ คล้ายกันเกือบทั้งหมด การสร้างโมเดลในงานวิจัยนี้ประยุกต์มาจาก package ของ sicara[6] ซึ่งมี Library ให้ใช้ตามความต้องการ

สำหรับโมเดลที่เลือกใช้ได้แก่ Prototypical Networks ขั้นแรก เลือก backbone เป็น ResNet50 เพื่อทำ Feature Extraction และเป็นโครงสร้างหลักโดยผ่านการสอนมาจาก ImageNet เนื่องจาก ResNet มีรูปแบบ Input เป็น 3 มิติได้แก่ (height, width, channel) เช่น (n_sample, 224, 224, 3) แต่ Input ที่ได้จากการ Concatenate Array เป็น 2 มิติ ดังนั้นจึงจำเป็นต้องจัดข้อมูลจากเดิมคือ (จำนวนเฟรม, Position ต่าง ๆ) ทำการเพิ่มมิติที่สามเป็นจำนวนวิดีโอ ที่มีอยู่ใน Dataset จะได้เป็น (จำนวนเฟรม, Position ต่าง ๆ, จำนวนวิดีโอ) และตัด layer สุดท้ายของโมเดลออกหรือชั้น output และต่อด้วย Flatten layer จากนั้นนำค่าจาก ResNet50 เข้าสู่โมเดล Prototypical Networks ซึ่งหลักการ Predict class ของ Prototypical Networks มาจากระยะทาง (คำนวณจาก Euclidean distance) ของต้นแบบ(support set) เทียบกับ input (query) ที่ใส่เข้าไป และใช้ Optimizer คือ Adam หลังจากนั้นก็นำมาทดลองปรับเปลี่ยนตัวแปร Backbone และเพิ่ม task ในการสอน และตำแหน่งที่นำเข้าไปเป็น input เพื่อปรับปรุงโมเดลให้ดีขึ้น โดยกำหนดตัวแปรดังตารางที่ 2 ซึ่งสามารถวัดประสิทธิภาพของโมเดล Prototypical Networksได้ Accuracy = 0.781, Loss = 0.27

ตารางที่ 2 การกำหนดตัวแปรของโมเดล

Parameter	
Model	Prototypical Networks
Training Task	600 Task
Backbone	ResNet50 (imageNet)
Input	Pose(มุมแขน) + Pose(ระยะห่างจมูกกับมือ) + Pose (ตำแหน่ง Pose - ตำแหน่งจมูก)
FPS	60 fps
Optimizer	Adam
Distance	Euclidean distance
Loss Function	Categorical Cross Entropy

4. การนำโมเดลไปใช้กับระบบ (Implement)

นำโมเดล Prototypical Networks ซึ่งเป็นโมเดลที่มีประสิทธิภาพมากที่สุด ไปใช้ในการแปลคำศัพท์ภาษามือไทย โดยผ่าน GUI (Graphical User Interface) อย่างง่ายที่พัฒนาขึ้น ประกอบไปด้วยการทำงาน 3 ส่วนดังนี้

ส่วนที่ 1 จอแสดงผลวิดีโอ ซึ่งสามารถเชื่อมต่อกับกล้องของโน้ตบุ๊กแบบ Real ในกรณีแปลคำศัพท์จากกล้อง

ส่วนที่ 2 แสดงผลลัพธ์ของคำศัพท์ที่แปลได้ โดยอ้างอิงจากคลังคำศัพท์ที่สอนระบบ

ส่วนที่ 3 เมนูการใช้งาน ประกอบด้วย เมนูแปลคำศัพท์จากกล้องแบบ Real time ระบบจะแปลคำศัพท์จากการจับภาพผ่านกล้องของเครื่องโน้ตบุ๊ก เมนูแปลศัพท์จากวิดีโอระบบสามารถแปลคำศัพท์จากวิดีโอที่นำเข้าได้ และเมนูสุดท้ายคือเมนูออกจากระบบ เพื่อปิดการทำงานของระบบ

ภาพที่ 6 ตัวอย่างโปรแกรมรูปแบบ GUI ในการนำไปใช้จริง

ผลการวัดประสิทธิภาพความถูกต้องของโมเดล

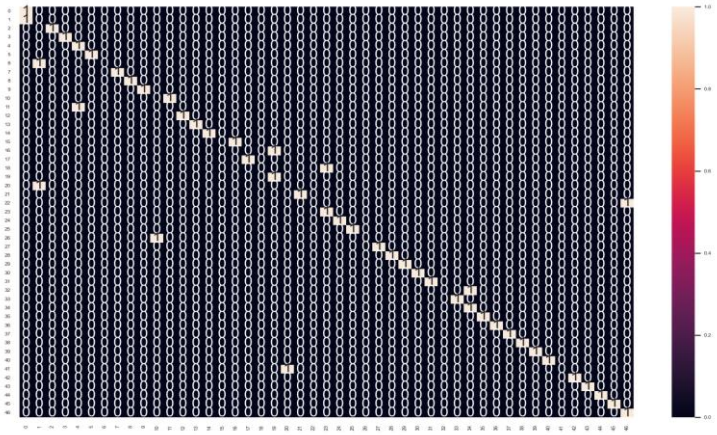
จากการทดสอบพบว่าการกำหนดตัวแปร และฟีทเจอร์ตามตารางที่ 2 ให้ผลการทดสอบให้ค่าความถูกต้องที่ดีที่สุด ดังนั้นจึงทำการสอนโมเดลด้วย ดังนี้ N-Way = 47 ,K-Shot = 6, Query = 6 ได้ค่า Accuracy = 0.781 และLoss = 0.27

ผลจากการนำไปใช้งานกับวิดีโอโดยสมาคมคนหูหนวก

เนื่องจาก Data Set mสร้างขึ้นนั้นเกิดจากการบันทึกวิดีโอโดยอาสาที่ไม่มีความเชี่ยวชาญในภาษามือโดยเฉพาะดังนั้นผู้วิจัยจึงทำการทดสอบระบบแปลคำภาษามือไทยโดยใช้ข้อมูลตัวอย่างวิดีโอของฐานข้อมูลภาษามือไทยที่จัดทำขึ้นโดยสมาคมคนหูหนวกแห่งประเทศไทย

ผลการทดสอบ

ผลการทดสอบ N-Way = 47, K-Shot = 6, Query = 1 ได้ค่า Accuracy = 0.74 จะเห็นได้ว่ามีประสิทธิภาพใกล้เคียงกับการทดสอบด้วยวิดีโออาสาที่สร้างขึ้นโดยผู้วิจัย ซึ่งได้ Confusion Matrix ดังภาพที่ 7



ภาพที่ 7 Confusion Matrix ทดสอบกับวิดีโอโดยสมาคมคนหูหนวก

ผลการทดสอบจากวิดีโอที่สร้างโดยอาสา และวิดีโอที่สร้างโดยสมาคมคนหูหนวกพบว่าคำศัพท์ส่วนใหญ่ที่แปลไม่ถูกต้องนั้น เป็นคำศัพท์ที่มีท่าทางที่คล้ายกันมาก ตำแหน่งของมือ องศาของแขน อยู่ในตำแหน่งที่ใกล้เคียงกัน ดังตารางที่ 3

ตารางที่ 3 เปรียบเทียบการแปลคำที่คล้ายคลึงกันจากวิดีโออาสา และวิดีโอสมาคมคนหูหนวก

วิดีโอที่สร้างโดยอาสา		**วิดีโอที่สร้างโดยสมาคมคนหูหนวก**	
Actual	**Predicted**	**Actual**	**Predicted**
กิน	ขี้เกียจ/ ดื่ม/ พูด/ อิ่ม	กิน	กลัว
ขี้เกียจ	คิด/ ลืม	คิดถึง	ขม
ซื้อ	ดีใจ/ ปรึกษา/ มั่นใจ	ดม	ดีใจ
ยิงปืน	ทะเลาะ/ เกลียด/ โกนหนวด/ ไม่ชอบ	ดีใจ	ขี้เกียจ
สอบ	ดม/ ดื่ม/ เสียใจ	พูด	ฟ้อง
อธิบาย	ทะเลาะ/ วิ่ง	ฟ้อง	รัก
		ยิงปืน	สงสัย
		ร้องไห้	กิน
		วิ่ง	ไม่แน่ใจ
		หิว	ดม
		เกลียด	เบื่อ
		โกรธ	ร้องไห้

สรุปผลการศึกษา

การพัฒนาโมเดลที่มีประสิทธิภาพในการแปลคำศัพท์ภาษามือไทย ประกอบด้วยขั้นตอนต่อไปนี้ ขั้นตอนแรกเตรียมข้อมูลวิดีโอที่เก็บมาโดยการ Capture frame เป็นรูปภาพจำนวน 60 FPS (15 frame per second) หลังจากนั้นใช้ Mediapipe ในการระบุตำแหน่งของร่างกาย และมือขั้นตอนที่สอง นำค่าพิกัดที่ได้มา Normalize ซึ่งถือให้จมูกของแต่ละบุคคลเป็นจุดอ้างอิง นำตำแหน่งพิกัดมาลบด้วยจุดอ้างอิง และนำพิกัดส่วนอื่นมาทำ Feature Extraction เมื่อได้ Feature ที่ต้องการก็นำมา Concatenate เป็น 1 เฟรม และนำแต่ละเฟรม มารวมกันเป็น 1 วิดีโอ จัดเก็บในรูปแบบ Numpy array โดยมีการทำ Feature Extraction หาองศาแขนท่อนปลายทำมุมกับท่อนบน มีศอกเป็นจุดยอดมุม และหาระยะห่างระหว่างนิ้วชี้กับจมูก หลังจากนั้นดำเนินการขั้นตอนสุดท้ายนำข้อมูลที่ได้เข้าสู่โมเดล Prototypical Networks เพื่อแปลคำศัพท์ภาษามือไทย ผลการทดลองให้ความแม่นยำในการแปลคำศัพท์ภาษามือไทยทั้ง 47 คำ วัดค่า Accuracy ได้ถึง74%

ข้อเสนอแนะ

1. ควรเก็บข้อมูลตัวอย่างในแต่ละคำเพิ่ม (K-shot) เพราะจำนวน K-shot ส่งผลต่อประสิทธิภาพโมเดล หาก K-shot น้อยจะทำให้ความแม่นยำน้อยลง แต่หาก K-shot มากจะทำให้โมเดลแม่นยำขึ้น ดังนั้นจึงควรเก็บข้อมูลเพิ่มขึ้น
2. สร้างแอปพลิเคชันเพื่อรองรับการใช้งานบนมือถือ
3. การเรียงคำเป็นประโยค การสื่อสารด้วยภาษามือ มีหลายประเภท หนึ่งในนั้นคือการเลือนแบบประโยคในการพูดโดยใช้คำศัพท์มาเรียงต่อกัน ทำให้สามารถนำคำมาเรียงต่อกันเพื่อเป็นประโยคได้
4. สามารถเพิ่ม text to speech เพิ่มเติมเมื่อทำนายคำศัพท์ได้ เพิ่มความสะดวกต่อผู้ที่สื่อสาร
5. อาจมีการทดลองเพิ่มเติม (Future Work) ในขั้นตอนของการสอนระบบ โดยทดลองปรับการ

    5.1 สร้าง Data set ด้วยบุคคลที่หลากหลายใน 1 class เนื่องจากตำแหน่งมือ สัดส่วนร่างกายหรือท่าทางของแต่ละบุคคลแตกต่างกัน จะทำให้โมเดลยืดหยุ่นขึ้น หรือเก่งขึ้น

    5.2 นำโมเดลที่เก่งด้าน Sequence มาใช้รวม Aggregate เฟรมแต่ละเฟรมรวมทั้งจัดการความยาวของวิดีโอ

    5.3 เพิ่ม Feature Extraction โดยวิธีอื่น ๆ เช่น Angular Features, Geometric Features

บรรณานุกรม

ภาษาไทย

กนิษฐา หงส์พรรคมนุญ,วันทนีย์ จันทร์สมปอง และรัตติยากร ทองจุนเจือ.(2551).ระบบตรวจสอบความบกพร่องของชิ้นงานแบบอัตโนมัติโดยวิธีประมวลผลภาพ.สืบค้น 8 กรกฎาคม 2564,จากhttp://www.lib.buu.ac.th/st/Engineering/Electrical/2551/EE-51-26.pdf

การลงโปรแกรมVisual Studio 2010และopenCV.(2555).สืบค้น 1 มิถุนายน 2564,จาก http://kwangee1245.blogspot.com/2012/04/visual-studio-2010-and-opencv.html

ธัญพร มัทวานุกูล.(2564).สถานการณ์คนพิการ 30 มิถุนายน 2564 (รายไตรมาส).สืบค้น 3 กรกฎาคม 2564,จาก https://dep.go.th/th/law-academic/knowledge-base/disabled-person-situation/สถานการณ์คนพิการ-30-มิถุนายน-2564-รายไตรมาส

MediaPipe Holistic อุปกรณ์ที่สามารถจับการเคลื่อนไหวของใบหน้า มือ และท่าทางได้ในเวลาเดียวกัน.(2021).สืบค้น 7 กรกฎาคม 2564,จาก https://www.sertiscorp.com/2021-01-15

ปกรณ์ กัญชลี.(2562).Confusion Matrix เครื่องมือสำคัญในการประเมินผลลัพธ์ของการทำนาย ใน Machine learningสืบค้น 7 กรกฎาคม 2564,จาก https://medium.com/@pagongatchalee/confusion-matrix-เครื่องมือสำคัญในการประเมินผลลัพธ์ของการทำนาย-ในmachine-learning-fba6e3f9508c

ภาษาต่างประเทศ

Bennequin, E. (2021). *Easy Few-Shot Learning*. Retrieved from https://github.com/sicara/easy-few-shot-learning

Careaga, C., Hutchinson, B., Hodas, N. & Phillips, L., (2019). *Metric-Based Few-Shot Learning for Video Action Recognition*.

Chaikaew, A., Somkuan, K., Yuyen, T. (2021). *Thai Sign Language Recognition: an Application of Deep Neural Network*.

MediaPipe Hand. Retrieved from https://google.github.io/mediapipe/solutions/hands.html

Snell, J., Swersky, K. & Zemel, R. (2017). *Prototypical Networks for Few-shot Learning.*

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. & Hospedales, T.(2018*). Learning to Compare: Relation Network for Few-Shot Learning*.

Vinyals, O., Blundell, C., & Lillicrap, T. (2017). *Matching Networks for One Shot Learning*.