Progress Report #2

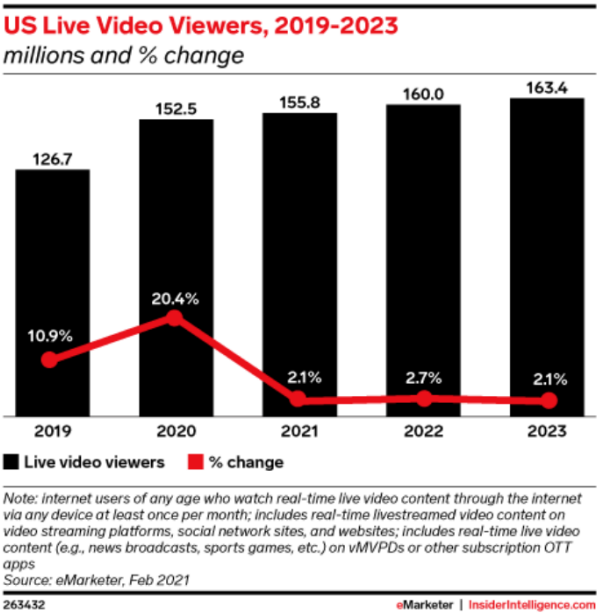
A real-time sign language translator using depth camera for live streaming applications

จิรายุ หาญวงษ์ 6130082121, ธนกฤต นามวงษ์ 6130208021 และ วริศรา กาญจน์วีระโยธิน 6130484821

อ.ดร.ณัฐพล ดำรงค์พาลาสิทธิ์ (อาจารย์ที่ปรึกษา)

Scope of Live Streaming Industry

เนื่องจากเล็งเห็นว่าธุรกิจที่เกี่ยวข้องกับการ Live Streaming นั้นกำลังเป็นที่นิยมจากผู้คนมากขึ้นเรื่อย ๆ ส่วนมากในหมู่วัยรุ่น เนื่องจากสถาณการณ์โควิด 19 ทำให้มีการขยายตัวของจำนวนแพลตฟอร์มการไลฟ์อย่างต่อเนื่อง เริ่มต้นจาก YouTube, Facebook, Twitch, Instagram หรือแพลตฟอร์มน้องใหม่อีกมากมาย อ้างอิงจาก eMarketer [1] เมื่อวันที่ 9 กุมภาพันธ์ 2021 มีการทำการคาดคะเนการเติบโตของผู้เข้าชมการไลฟ์ของประเทศสหรัฐอเมริกา ปี 2019-2023



รูปที่ ก. US Live Video Viewers from 2019-2023

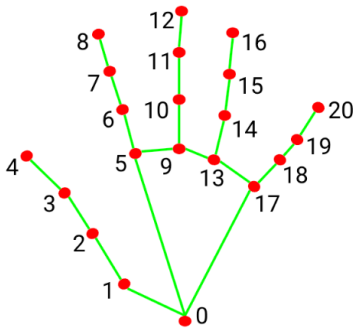
เห็นได้ว่าผู้รับชมการไลฟ์นั้นมากขึ้นอย่างเห็นได้ชัด รายได้ของไลฟ์สตรีมเมอร์หรือบุคคลที่ทำการไลฟ์เป็นอาชีพมีโอกาสถึงเดือนหลายแสนบาท ดังนั้นสโคปของโปรเจคนี้จึงเพ่งเล็งไปที่การสร้างโอกาสทางอาชีพให้กับผู้พิการทางการพูดหรือได้ยิน โดยที่จะใช้กล้องจับความลึกตรวจจับท่าทางของผู้พิการ และนำคำที่ตรวจจับได้มาเรียงเป็นประโยค และใช้ฟังชัน text-to-speech เพื่อให้มีเสียงที่พูดเป็นคำให้ผู้ชมได้ยิน การทำเช่นนี้จะทำให้ผู้ที่พิการสามารถแสดงออกท่าทาง สีหน้า อารมณ์ที่ต้องการจะสื่อได้ และทุกคนสามารถเข้าใจได้ ทั้งนี้ก็เพื่อสร้างโอกาสให้กับผู้ที่มีบุกคลิกที่น่าสนใจและเป็นที่ชื่นชอบให้มีรายได้

Signing Exact English (SEE)

Signing Exact English (SEE) หรือการลงนามภาษาอังกฤษที่แน่นอนนั้น เป็นภาษามือประเภทหนึ่งที่ใช้กันแพร่หลายในหมู่คนที่มีปัญหาด้านการได้ยินที่ใช้ภาษาอังกฤษ นอกจากนี้ยังมีภาษามืออีกประเภทหนึ่งที่มีชื่อว่า American Sign Language (ASL) ซึ่งเป็นภาษามือที่ใช้กันแพร่หลายที่สุด และเป็นต้นแบบของภาษาต่างๆ รวมถึง Signing Exact English อีกด้วย ความแตกต่างระหว่างภาษามือสองประเภทนี้คือหลักไวยากรณ์ในการใช้ ASL จะมีหลักไวยากรณ์เป็นของตัวเอง โดยประโยคจะเริ่มจากเวลา ตามด้วยหัวข้อ และคำอธิบาย กล่าวง่าย ๆคือแทบจะเป็นอีกภาษาหนึ่งโดยสิ้นเชิง ต่างกับ SEE ตรงที่ SEE จะใช้หลักไวยากรณ์ตามภาษาอังกฤษโดยตรง ยกตัวอย่างประโยคเช่น “The class was great this morning” หากเป็น ASL จะเริ่มสื่อสารประมาณว่า “This morning, the class, was great” และ SEE จะเป็น “The class was great this morning” ซึ่งคนที่ใช้ SEE ในชีวิตประจำวัน ส่วนมากจะเป็นผู้ที่ไม่ได้สูญเสียการได้ยินตั้งแต่เกิด หรือไม่ได้เป็นผู้พิการแต่ใช้เพื่อจุดประสงค์อื่น ซึ่งกลุ่มคนเหล่านี้จะเข้าถึงการไลฟ์สตรีมและเข้าใจมันได้ดีกว่า

Data Collecting

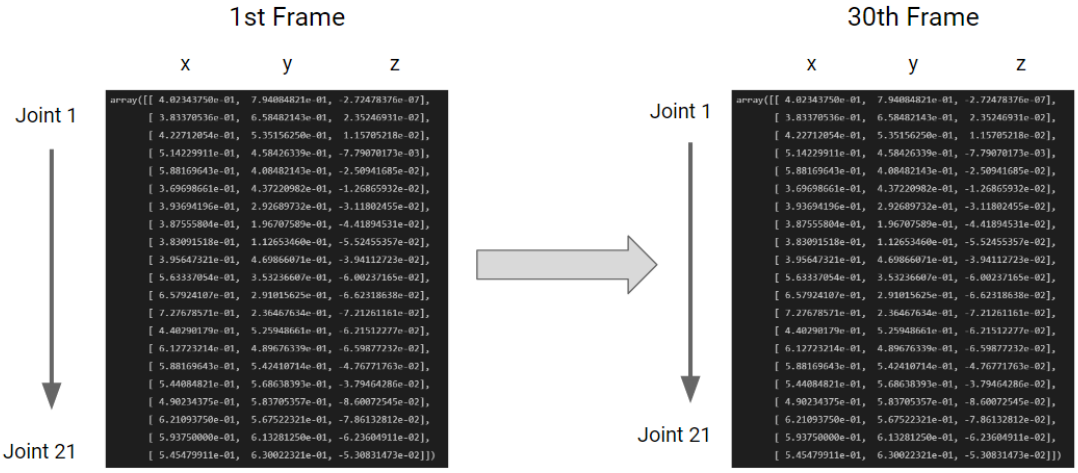
ในขั้นตอนการสร้าง machine learning model เราเริ่มต้นจากการเก็บข้อมูลซึ่งก็คือ ตำแหน่ง x y z ของแต่ละจุดบนมือซึ่งใช้ open-source ของ Mediapipe ในการทำ Skeletal tracking ซึ่งใน 1 คำ เราจะแยกเป็น 30 frames และเก็บ 60 รอบ เพื่อเพิ่มความแม่นยำในโมเดล ซึ่งในตอนนี้เก็บข้อมูลไปแล้ว 20 คำด้วยกัน เช่นคำว่า I, you, am, is, are เป็นต้น



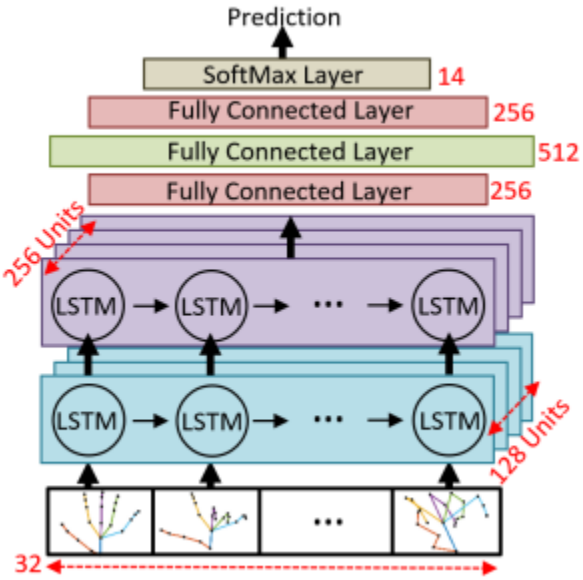
0. WRIST
1. THUMB_CMC
2. THUMB_MCP
3. THUMB_IP
4. THUMB_TIP
5. INDEX_FINGER_MCP
6. INDEX_FINGER_PIP
7. INDEX_FINGER_DIP
8. INDEX_FINGER_TIP
9. MIDDLE_FINGER_MCP
10. MIDDLE_FINGER_PIP
11. MIDDLE_FINGER_DIP
12. MIDDLE_FINGER_TIP
13. RING_FINGER_MCP
14. RING_FINGER_PIP
15. RING_FINGER_DIP
16. RING_FINGER_TIP
17. PINKY_MCP
18. PINKY_PIP
19. PINKY_DIP
20. PINKY_TIP

รูปที่ ข. Label ข้อต่อต่างๆ

รูปที่ ค. Skeletal Tracking



รูปที่ ง. การเก็บข้อมูลของแต่ละมือในรูปแบบ array



รูปที่ จ. เลเยอร์ต่าง ๆ ของโมเดล

(Source : **CNN+RNN Depth and Skeleton based Dynamic Hand Gesture Recognition)**

LSTM Model

โมเดลที่ใช้ในการ classifiy ข้อมูลนั้น อาศัยเป็น LSTM ซึ่งเป็นโมเดลใน library ของ Tensorflow Keras ซึ่งจากการศึกษาบทวิจัยจาก University of Calgary, Alberta, Canada (Yanushkevich) ทำให้เราเลือกใช้ neural network model ซึ่งมี 6 layer ได้แก่ input layer hidden layer 4 layer และ output layer ซึ่ง hidden layer ประกอบไปด้วย lstm 2 layer และ fully connected layer 2 layer

Preliminary Result

ผลที่ได้เบื้องต้นจากโมเดลที่ทำขึ้นมา สามารถจับคำได้ประมาณ 60% ของคำทั้งหมดที่ได้เก็บข้อมูลไปโดยที่มี confidence level มากกว่า 0.9 ปัญหาหลังจากที่ได้ผลลัพธ์ออกมาแล้วมีอยู่ด้วยกันหลัก ๆ 3 ข้อด้วยกันคือ

1. โมเดลทำนายคำที่ไม่ได้ทำ
2. โมเดลไม่สามารถทำนายบางคำได้
3. บางคำมีการทำท่าที่คล้ายกัน และยากที่จะแยกออก

Assumptions and Solutions

ปัญหาแรกเนื่องจากโมเดลได้ทำนายคำที่ไม่ได้ทำ หรือทำแล้วทายออกเป็นคำอื่น สมมุติฐานคาดว่า Interval ของการ evaluate ของ model ไม่ตรงกับท่าทางที่เราทำ ยกตัวอย่างคือ โมเดลอาจจะ evaluate ทุก ๆ 30 frame คล้ายกับตอนที่เก็บข้อมูล แต่เมื่อเราลองทำคำต่าง ๆต่อกัน บางคำไม่ได้เริ่มที่ frame ที่ 0 แต่อาจเริ่มที่ frame ที่ 15 ท่าเริ่มต้นของคำที่เราทำ อาจจะไปคล้ายกับท่าใน frame ที่ 15 ของคำอื่น ดังนั้นโมเดลจึงทำนายผิดพลาดไป วิธีแก้อาจจะเป็นการลองให้โมเดล evaluate 30 frame ย้อนหลังว่า confidence level ของคำไหนมากที่สุด และแสดงผลของคำนั้น แต่อาจจะพบ lag ในระบบ

ปัญหาที่สองคือการที่โมเดลไม่สามารถทำนายบางคำที่เราได้เก็บข้อมูลไปได้ คาดว่าคำที่ทำนายไม่ได้น่าจะไปคล้ายกับคำอื่นพอสมควร โมเดลจึงไม่มีความมั่นใจในการทำนายออกมา ดังนั้นหากเราลด threshold ของความมั่นใจลง เช่น มั่นใจ 50% ขึ้นไปให้แสดงผล โมเดลก็อาจจะทำนายคำออกมาได้ แต่ก็อาจจะมีคำที่เป็น noise รวมอยู่ด้วย

ปัญหาสุดท้ายที่คำคล้ายกันและยากที่จะแยกออก ในส่วนที่คิดว่าหากควบคุม environment และ interval ในการ detect ให้เหมาะสม รวมถึงเก็บ data เพิ่มจะสามารถแยกท่าทางที่คล้ายกันออกจากกันได้

Bibliography

Yanushkevich, K. L. (n.d.). *CNN+RNN Depth and Skeleton based Dynamic.* Retrieved from https://arxiv.org/pdf/2007.11983.pdf