From Gestures to Words: American Sign Language End-to-End Deep Learning Integration with Transformers and Mediapipe

Hiep Nguyen, Department of Computer Science, Texas Christian University

Advisor: Dr. Bingyang Wei

INTRODUCTION

- **The global prevalence of hearing loss is increasing, with an estimated 432 million adults and 34 million children currently affected, projected to rise to over 700 million in 30 years [1].**
- **Despite the rising demand for workplace accessibility using American Sign Language (ASL), current software solutions have yet been able to integrate Deep Learning in transcribing ASL to text in real-time.**
- **Our research aims to address this gap, proposing an innovative solution to bridge the communication divide and foster a more inclusive society.**

OBJECTIVES

- **Develop a real-time prediction Transformer model that utilizes Google ASL Finger Spelling Dataset [2] to accurately interpret gestures into text**
- **Establish a robust server architecture using Flask and Python, encapsulating Transformer model for efficient prediction of finger spelling gestures with high load**
- **Implement a comprehensive web application for video conferencing that integrates real-time communication via AgoraRTC, and Flask server for sign language predictions**

METHODS: ASL FINGER SPELLING RECOGNITION VIA TRANSFORMER ARCHITECTURE AND MEDIAPIPE

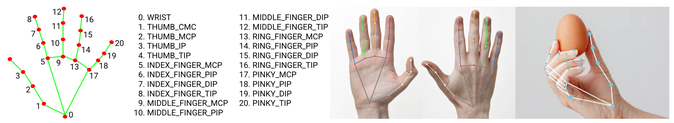

Fig. 1. Google Mediapipe's hand landmark model, detecting 21 different landmark coordinates [2][3].

Key classes of the model:

- **Frames: The initial input layer, holding numerical sequence data from parquet files**
- **Phrase: Secondnput layer, containing label sequences for full phrase representations**
- **Masking: Ignores nulls in the frames layer**
- **Embedding: Normalizes data and creates embedded representations of frames with positional information for pattern recognition**
- **Encoder: Converts sequences into vector representations, using attention and normalization to highlight key input features**
- **Decoder: Uses embeddings from the Encoder to produce the final output sequence, applying attention and normalization to ensure relevance and accuracy**

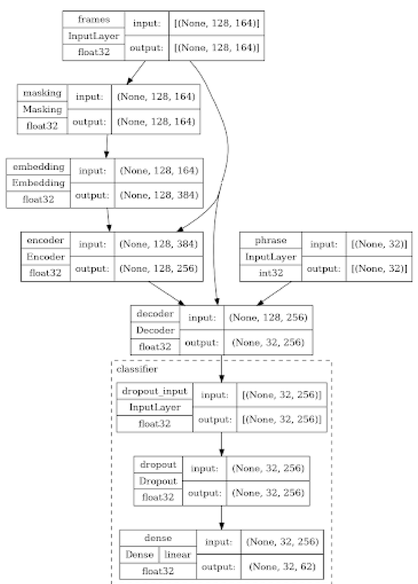

Fig. 2. Transformer model architecture with Embedding, Encoder, and Decoder classes, developed upon Mark Wijkhuizen's research work [4].

METHODS: APPLICATION PROGRAMMING INTERFACE (API) DEVELOPMENT FOR REAL TIME INFERENCE USING FLASK

- **API Implementation and Workflow: Utilizes Flask for an API to interpret ASL in video calls, involving frame capture, base64 encoding, Mediapipe and TFLite processing for prediction, and sending a base64-encoded image with the predicted ASL character back to the client.**
- **Testing and Functionality Assurance: Developed "live_test.py" Python script for local testing to capture and send frames to the Flask server, ensuring functionality and enabling the client to reconstruct the predicted ASL alphabet image from received base64 data.**

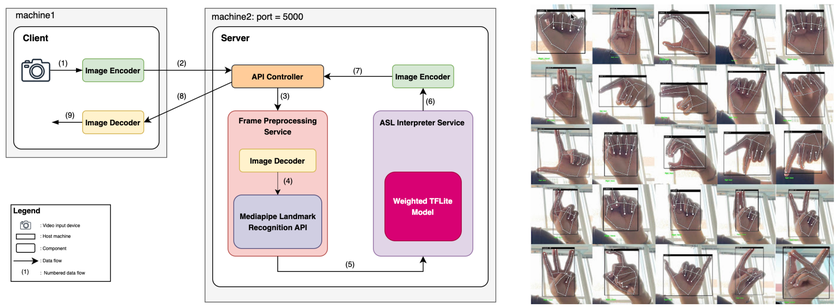

Fig. 3. Back-end Architecture for American Sign Language Real-time Inferencing Flask Server.

METHODS: VIDEO CONFERENCING APPLICATION AND MODEL INTEGRATION USING AGORARTC

- **Developed a video conferencing application with AgoraRTC SDK, enabling real-time video, audio, and chat, chosen for its flexibility over other SDKs like Zoom for better control over video layers**
- **Integrated a Flask server for ASL recognition with the video conferencing app, allowing asynchronous ASL text overlay on video streams without disrupting the main video feed, and facilitating a switch between standard and ASL-assisted modes**

Fig. 4. End-to-end ASL Recognition Model Integration with Video Conferencing Application.

RESULTS

Fig. 5. Model loss visualized graph on training dataset and validation dataset over 150 epochs.

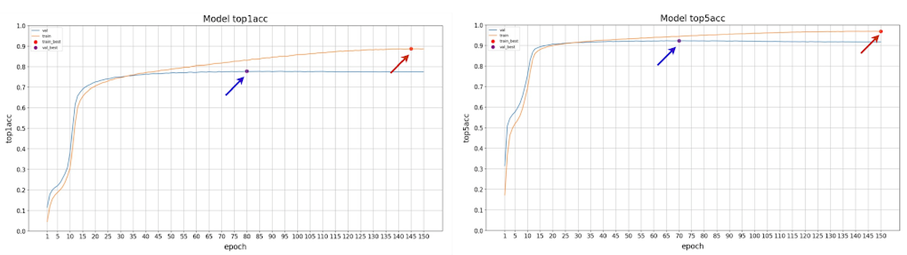

Fig. 6. Model accuracy over 150 epochs. Top one accuracy (left), and top five accuracy (right) are displayed for both training and validation dataset, as well as the peak of both.

RESULTS

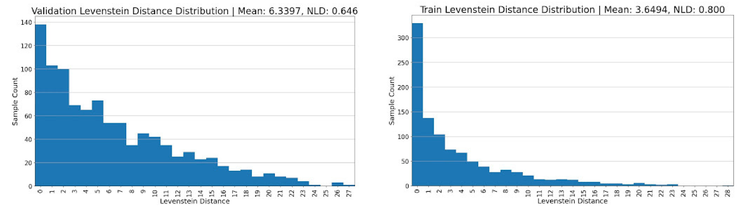

Fig. 7. Levenstein Distance Distribution for predicted phrase on train dataset (left) and validation dataset (right).

- **Model Loss: Decreases consistently across 150 epochs in training, with validation loss plateauing around epoch 65**
- **Top-One Accuracy: Reached a peak of 88% on the training dataset by epoch 145 and 78% on the validation dataset by epoch 80, then stabilized**
- **Top-Five Accuracy: Achieved its highest at 96% on the training dataset by epoch 150, and peaked at 92% on the validation dataset by epoch 70 before stabilizing**
- **Levenstein Distance: On the training dataset, the mean distance is 3.6494 with over 380 phrases correctly predicted. For the validation dataset, the mean distance is 6.3397 with nearly 140 phrases correctly predicted, indicating a promising result for full phrase prediction accuracy**

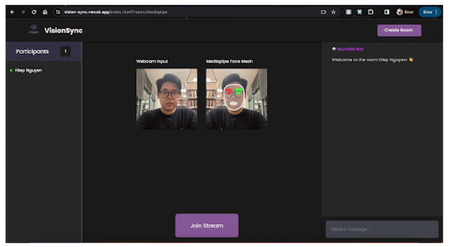

Fig. 8. End-to-end video conferencing web application using AgoraRTC with ASL support.

FUTURE WORK

- **Continue refining the Transformer model with varied lighting, background, and distance settings to develop a more resilient model capable of accurately recognizing complex finger spelling gestures**
- **Enhance the reliability and stability of data transfer from video input to the Flask server by implementing a queue for sequential frame processing**
- **Expand the application's capabilities to support concurrent processing, thereby enabling multiple users to perform inferencing simultaneously, ensuring the system's scalability and efficiency in handling increased load without compromising performance or speed**
- **Integrate advanced analytics to monitor and optimize the model's performance in real-time, and identify bottlenecks or inaccuracies in gesture recognition.**

REFERENCES

- **1. World Health Organization. "Deafness and Hearing Loss." World Health Organization, 28 Feb. 2024, http://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss, last accessed 2024/02/28.**
- **2. MediaPipe Solutions guide, https://developers.google.com/mediapipe/solutions/guide, last accessed 2024/02/21.**
- **3. Hand landmarks detection guide, https://developers.google.com/mediapipe/solutions/vision/hand_landmarker, last accessed 2024/02/21.**
- **4. Wijkhuizen, M. (n.d.). ASLFR Transformer Training Inference. Kaggle. Retrieved November 29, 2023, from https://www.kaggle.com/code/markwijkhuizen/aslfr-transformer-training-inference**