Crafting Sign Language From Speech

By Cutlassfish: Amaris Grondin, Bahar Birsel, Halleluiah Girum, Javier Fernandez Garcia, Nya Haseley-Ayende

Introduction

Myriad models translate sign language to speech (SLT), supporting one-way communication between the deaf and hard of hearing and others. However, to enable two-way communication, we must also translate speech into sign language (SLP). Our project aims to bridge this communication gap by developing a tool that translates sequences of English speech into sequences of American Sign Language.

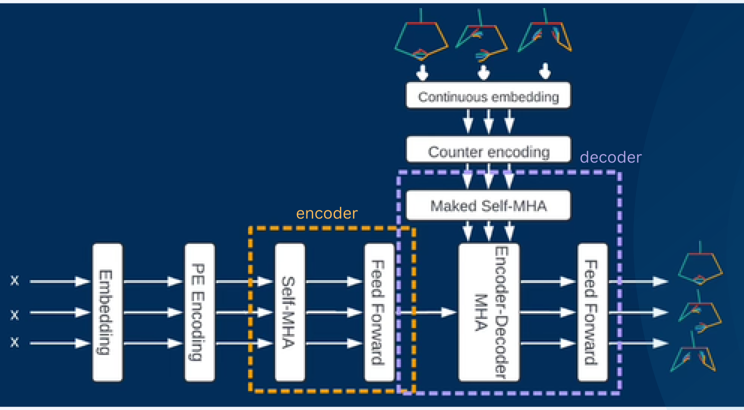
Methodology

Dataset

We use the How2Sign dataset which contains over 80 hours of video translating English sentences to ASL. In particular, we use already extracted body-face-hands keypoints extracted from the videos that describe the posture of the signers at each frame.

Training

The data was preprocessed into a source field of tokenized english sentences and a target field of their corresponding key-point values describing the pose translation. The tokenized sentences were generated using a mapping for every words in our vocabulary, and then returned as 2 tensors: 1 representing the padded sequences with 0s, and another representing the original lengths of the sentences. Once preprocessed, the Progressive Transformer was trained for 500 epochs, using an Adam optimizer and a dynamic learning rate between 0.001 and 0.0002. Training was performed on an Apple M2 Chip CPU, but the code uses a GPU when available.



Model

We use a Progressive Transformer which consists of a Symbolic Encoder and a Progressive Decoder. The Symbolic Encoder uses multi-headed self-attention to transform textual input into a high-dimensional symbolic representation. The Progressive decoder employs a counter decoding technique to generate continuous 3D sign language poses. This counter tracks the progression of sign generation to produce variable-length sequences of American Sign Language.

Discussion

- **Lessons learned:**
  - Data is scarce and difficult to work with, understanding it and preprocessing it appropriately is crucial.
  - Metrics are essential to success. Our model may have found a workaround to minimise loss without improving quality.
- **Lingering problems / limitations:**
  - A more suitable loss function is needed, and the bleu and rouge scores must be made compatible for SLP.
  - The counter decoding technique was also oversimplified.
- **Future work:**
  - Including a face in our poses as ASL also uses the face for communication and generating video of people not just poses.

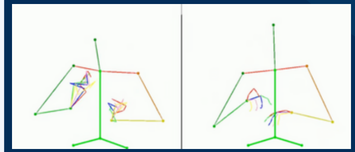
Expected Results

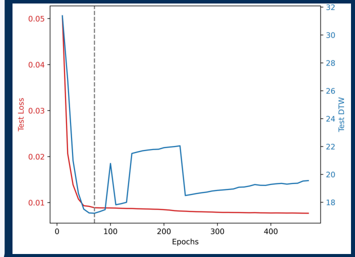
The testing loss captured across 500 epochs of training was ~0.007.This low loss indicates that the model is learning to minimize the MSE loss and fit the data.

Along with the loss, we also report a dynamic time warping (DWT) score which effectively measures the disparity between two sequences - the predicted and ground truth sequences. Across 500 epochs, we see that the best DTW score is ~16. This score seems to increase (worsen) for epochs after 70 while the loss continues to decrease which indicates that the model may be learning to minimise the MSE loss without getting better qualitatively. Thus, this points to MSE loss not being the best suited function for video our video output.

This is also evident in the training and testing videos, which reflect that the model is not accurate at predicting. Improvement to the results could come from using loss functions better adapted to video outputs as well as a more complex counter-decoding technique, and better hyperparameter tuning.

(The results above are based on the Progressive Transformers Paper, we are still finalizing our own results.)




Test Loss
Test DTW
Epochs