American Sign Language Recognition

Gianno Gomez, Mason Paradeza, Marianno Reynoso, Jingye Xu, Yuntong Zhang, Kevin Desai, PhD

The University of Texas at San Antonio, San Antonio, TX 78249

ABSTRACT

According to NIDCD, approximately 15% of Americans live with some form of hearing difficulty. For people categorized with extreme hearing loss, their primary form of communication is through American Sign Language and through lip reading. Given the recent pandemic and the popular use of masks for viral protection, it has become much more difficult for people who rely on reading lips to communicate.

For the above reason, we sought to create an application which translates static images of sign language into their corresponding English letter. For our application we used a novel encoder-decoder network architecture combined with a vision transformer. To Train and test our application we used the MNIST sign language data set.

BACKGROUND

American Sign Language is a difficult language to learn. It relies upon thousands of hand gestures which can communicate single letters or entire words and phrases as well as being dependent on body language and lip reading.

Our original plan was to create a translator which can read entire sentences from video feed. We quickly realized that using video would incur too high of a training cost and so we chose to go with static image classification.

One of the first methodologies we referenced was the VGG16 architecture. This did not produce the results we wanted and we then chose a novel transformer network with which to implement our application.

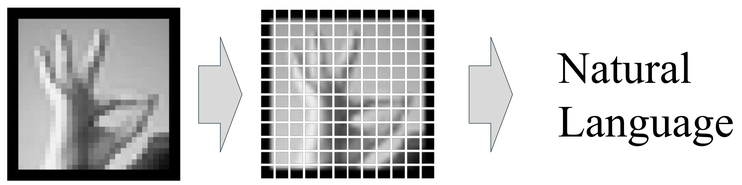
PURPOSE

Given the difficulties with which those with hearing loss must endure, we hoped to create something to improve their communication capabilities.

Our final methodology takes in a grayscale 28*28 image and outputs the corresponding English letter (J and Z cannot be conveyed statically and so they were omitted from this application).

METHODOLOGY

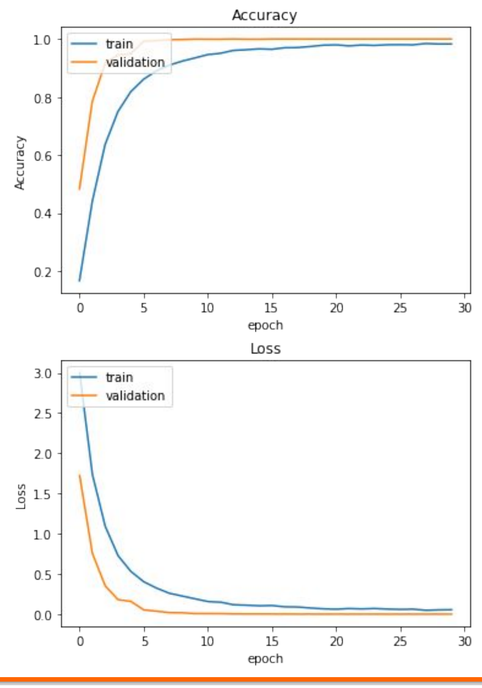
- We were originally going to use the CNN model, but our accuracy stalled at ~60%. The revised methodology is as follows.
- Use a Transformer Network, an encoder-decoder network architecture that uses self-attention and multi-head-attention which is effective for natural language processing.
- Use a Vision Transformer that takes an image input, divides it into “patches” and passes the patches to the Transformer Network as a natural language letter.



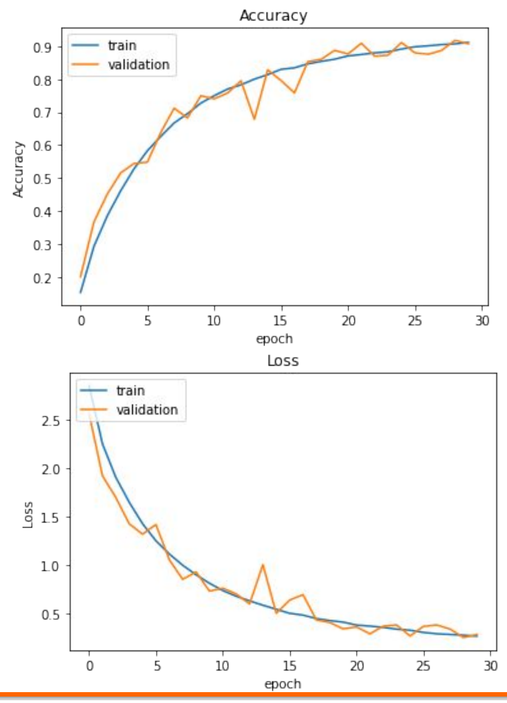
RESULTS

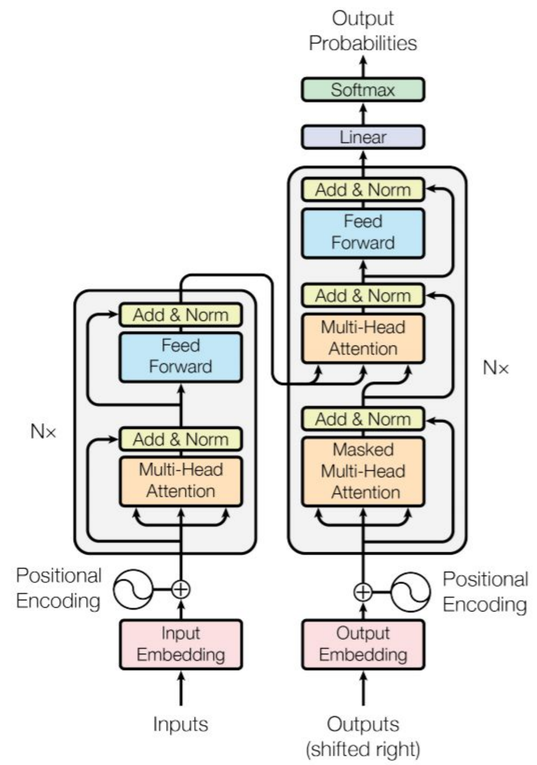
- The Transformer Network and the use of parallel processing gave us a ~+6% in accuracy compared to an existing CNN model.
- Along with an increase in accuracy, our methodology provided a more stable accuracy growth and loss decrease.

Transformer Network Learning Curve



Simple CNN Learning Curve





Transformer Network Architecture

SUMMARY

- Hearing impaired individuals have difficulty using technology to the fullest, our team developed a method to turn images of American Sign Language into readable natural language.
- Using a transformer network combined with a vision transformer, our network takes in an image and outputs its English letter.
- To train and test our network we used the MNIST sign language data set.
- We saw an accuracy increase of +~%6 and a stabilization of accuracy and loss results from using our methodology.

REFERENCES

https://www.kaggle.com/datasets/datamunge/sign-language-mnist
https://keras.io/examples/vision/image_classification_with_vision_transformer/
https://arxiv.org/abs/2010.11929
https://haleyso.github.io/projects/ASL.pdf

UTSA Computer Science