American Sign Language (ASL) is the first language of 500,000 people in the U.S. and Canada [1]. Yet, the automatic translation of ASL into text has never been accomplished. As a step towards achieving this task, we have restricted our attention to translating fingerspelling. Fingerspelling refers to the use of a hand to spell words and acronyms with the ASL alphabet, which has a one-to-one correspondence with the English alphabet.

There are many challenges to accomplishing this task. The first challenge is locating high-quality data. When fingerspelling, there is some variety in the exact placement of fingers. In order to account for this variation, the dataset must be large enough. There is also the issue of the signs for “j” and “z”, which involve movement to express the sign. We source data from GitHub and Kaggle [2][3][4]. The size of our training set is 26500 images.

In this work, we create a convolutional neural network which takes static images of ASL alphabet signs and translates them with a validation accuracy of about 85%. There are two steps to our process. First, the data is preprocessed by using MediaPipe to identify the location of the hands in each image and to produce a “hand graph” containing the landmarks of the hand. The data is then cropped to the hand, and augmentations are applied. Augmentations include reflections, inclusion of noise, changes in brightness, and rotations.

The second step is to classify the preprocessed images. The neural network consists of three sets of 2D convolutional layers with max pooling, using ReLU as the activation function. The final set uses batch normalization between the convolutional layer and the pooling layer to help generalize the model. After the convolutional layers, we flatten the tensors and then use three fully connected linear layers with ReLU activation. 20% dropout is applied to each layer to improve generalization. Lastly, the model includes one fully connected linear layer with SoftMax activation to act as a classifier. The model is allowed to train for 40 epochs with a batch size of 32. Validation accuracy was chosen as an appropriate metric of the model’s performance because the data was evenly distributed among all letters.

We could further improve our 85% accuracy by fine-tuning the model hyperparameters after performing a grid search. Hyperparameters to tune include the learning rate, dropout rate, and the batch normalization hyperparameters. Accuracy might also be improved by building off a deep convolutional neural network backbone, such as ResNet50. This capability is already incorporated into the Keras library, so it would be a readily available next step.

One avenue of future research is to create an architecture which can convert video input of fingerspelling into text. This would require the use of models, such as Recurrent Neural Networks, which take advantage of the temporal information available in videos. Even beyond fingerspelling, there is much work to be done to translate full ASL. This would be a monumental task, requiring researchers to first compile a suitable database of the full language, in addition to finding novel machine learning techniques to address the unique problems posed by the language.

1. https://www.startasl.com/american-sign-language/
2. https://github.com/mon95/Sign-Language-and-Static-gesture-recognition-using-sklearn
3. https://github.com/good-soul/fuzzy-octo-guacamole/tree/main/Datasets/Dataset2
4. https://www.kaggle.com/datasets/ayuraj/asl-dataset