The University of Texas at Austin
Cockrell School of Engineering

Convolutional Neural Networks For ASL Hand Sign Recognition

Connor Fritz

TEXAS

Introduction

American Sign Language (ASL) fingerspelling is a widely used form of communication in the deaf community. In fingerspelling, the letters of English alphabet are each represented by a distinct hand sign. Therefore, fingerspelling can be used to communicate between hearing and non-hearing people. However, most hearing people cannot understand ASL. This communication gap could be overcome with a real-time fingerspelling to English translation system. Previous attempts have been made to use convolutional neural networks (CNNs) to classify hand signs. However, these approaches have suffered from low classification accuracies, and large network sizes, which would make real time implementation more difficult[1].

Objective

The purpose of this project was to create a method for identifying ASL hand signs that is computationally efficient and robust in regards to variations in skin tone, hand placement, and backgrounds. Previous approaches to classifying ASL fingerspelling signs rely on datasets that only include photos taken with uniform skin tones, hand placements, lightings, and backgrounds[2,3]. Consequently, these approaches are not robust to the non-uniform conditions that an implemented translation system would be subjected to. Additionally, other approaches use large image sizes and implement complex architectures and preprocessing techniques[1,2]. These design decisions slow down classification, and would make real-time implementation more difficult. In addition to building a computationally efficient, robust classifier, this project also explored how heterogeneous datasets can improve CNN classification accuracy.

Methods

A deep Convolutional Neural Network (CNN) was built to classify ASL hand sign images. The network was built with as few layers as possible and input images were down sampled to increase the speed and decrease the memory usage of the network. The structure of the CNN is shown in figure 1. A combined dataset, amalgamated from six hand sign datasets was created. Each dataset consisted of around 10000 hand signs from a single user. The combined dataset included photos with a wide variety of skin tones, hand placements, backgrounds, and lightings, and was used to train the classification network. Additionally, six additional CNN's were created, each was trained using five out of the six datasets, with the sixth dataset used to measure the post-training accuracy of each network. The networks were trained using the AdaGrad optimization algorithm, and the categorical cross entropy loss function. Additionally, each convolution layer was followed by an RELU activation function and max-pooling layer, which are not shown in Figure 1.

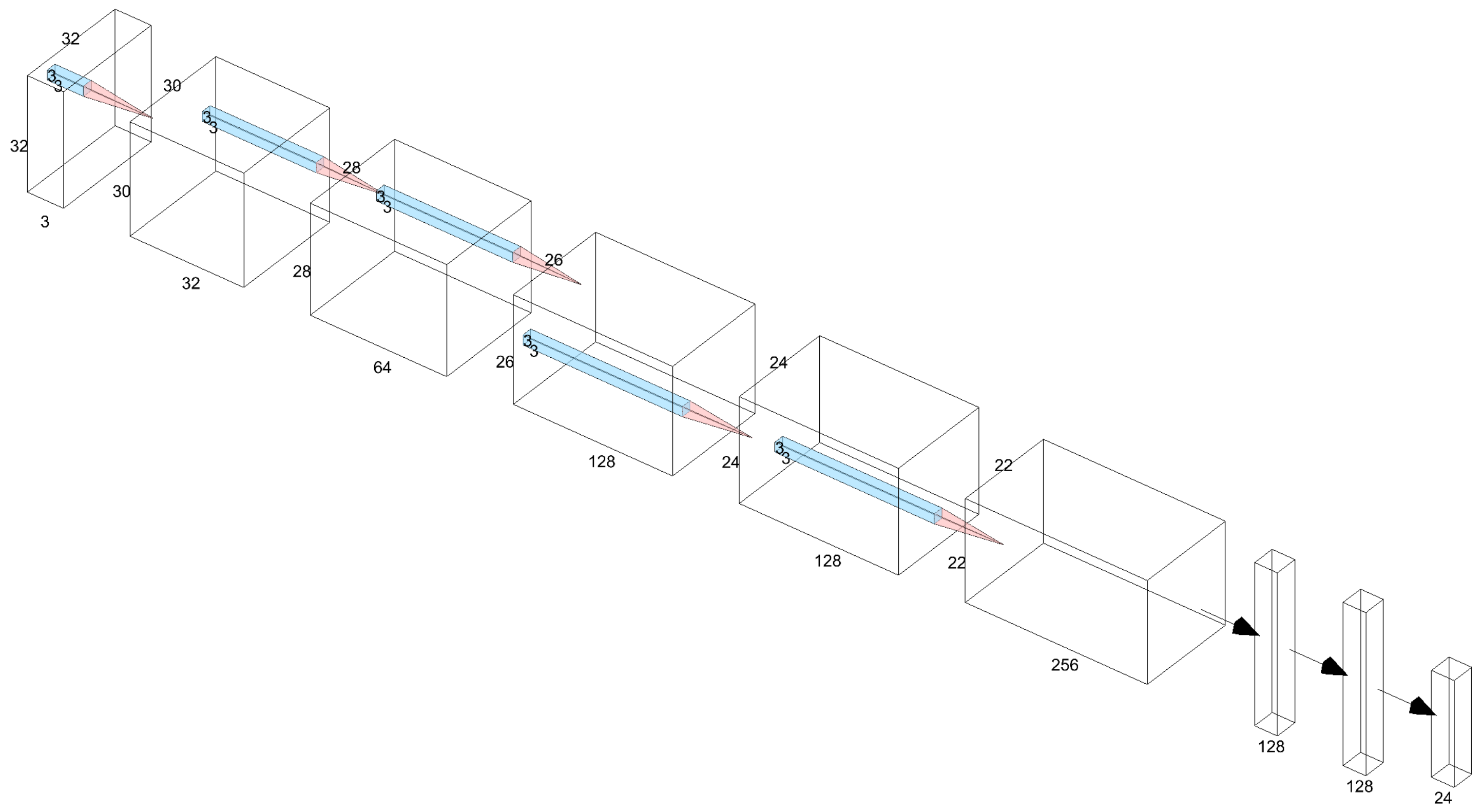

Figure 1 – Structure of the Convolutional Neural Network.
The network is composed of five convolutional layers with 32, 64, 128, 128 and 256 filters respectively. These layers are followed by two dense layers, each with 128 nodes.

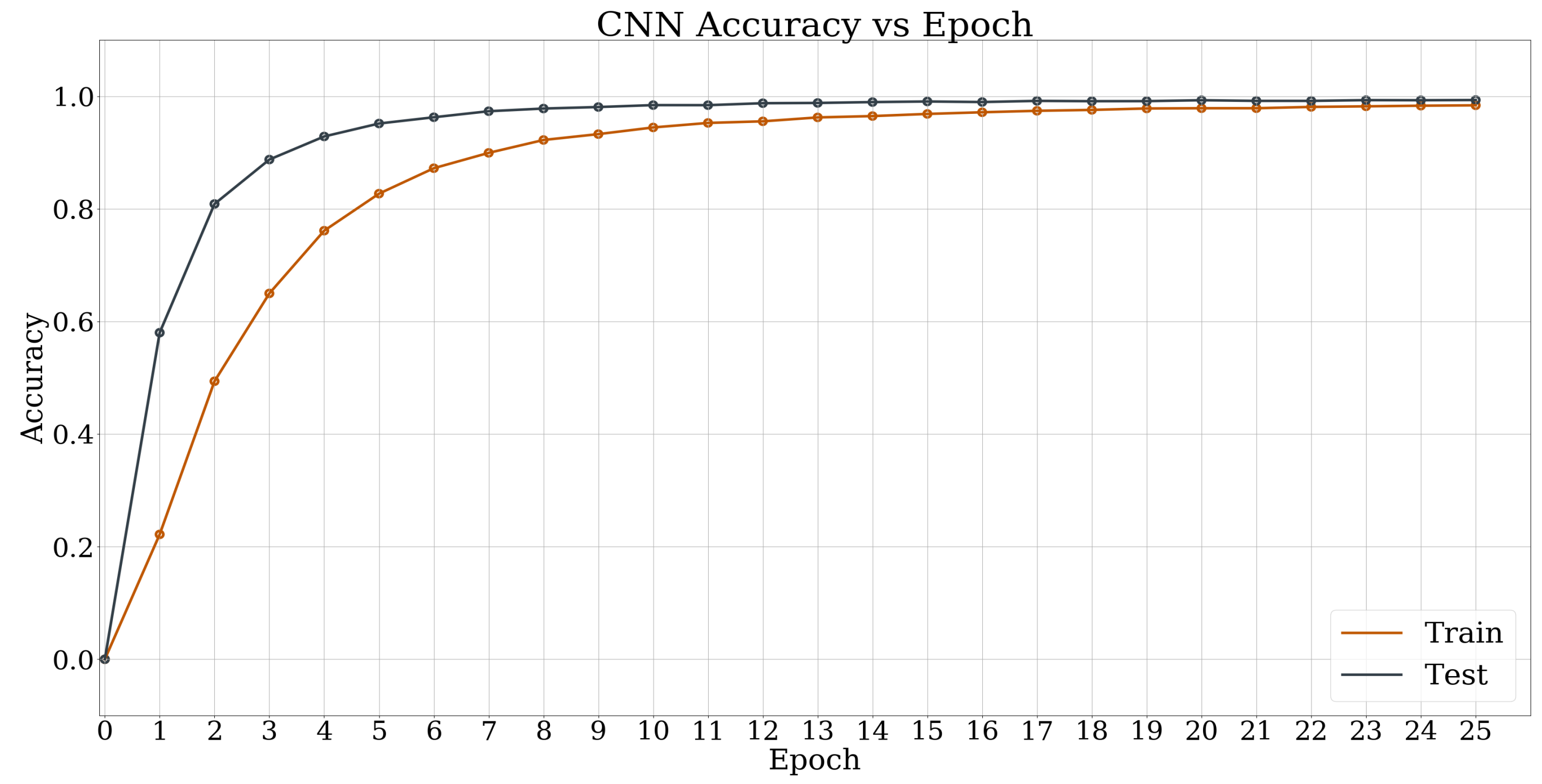

Figure 2 – Convolutional Neural Net Train and validation accuracy vs Epoch.
The network was trained for 25 epochs, after which the training accuracy was 98.4 % and the validation accuracy was 99.3 %

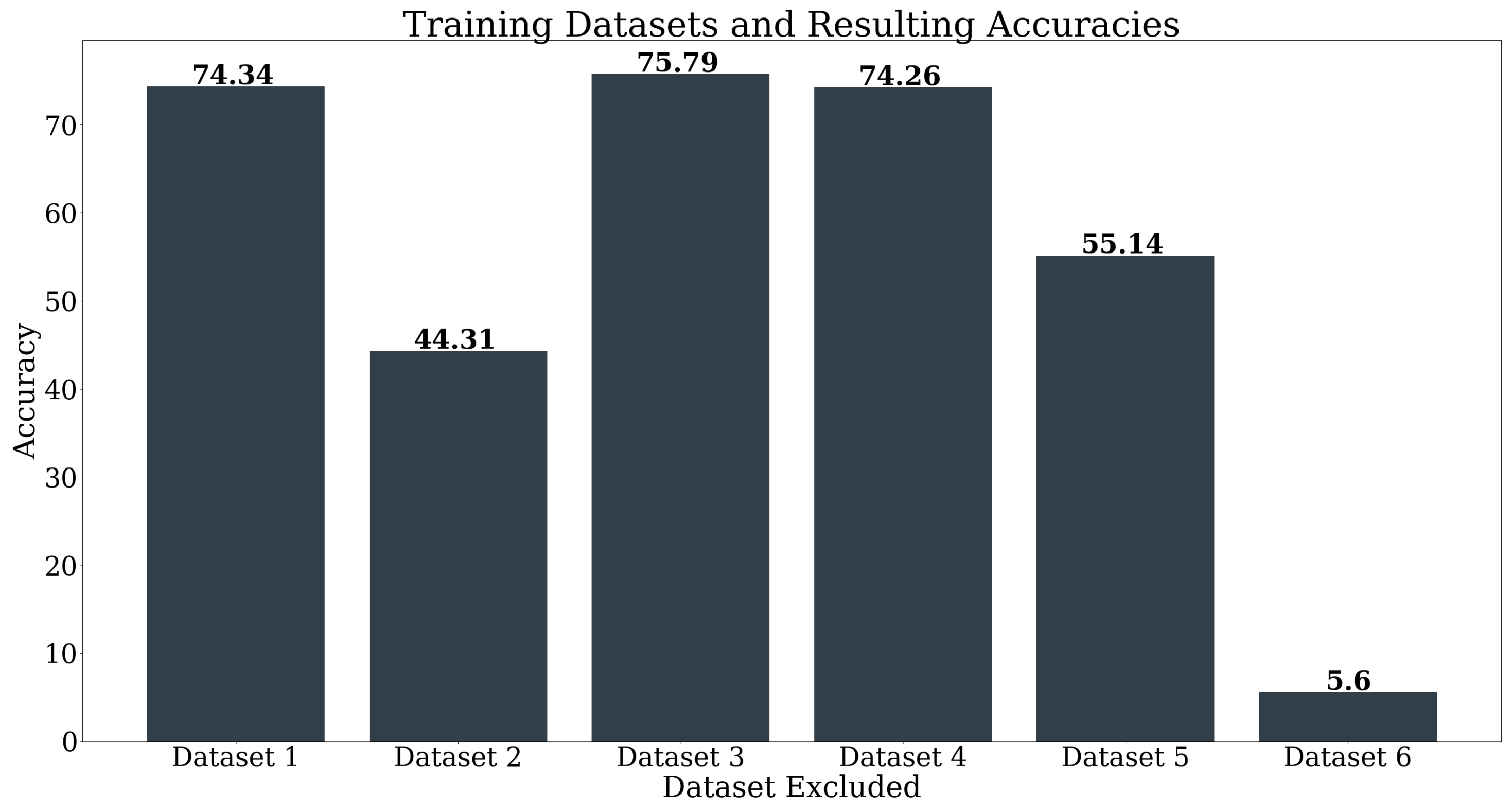

Figure 3 – Test Accuracies with Different Training Datasets
The excluded dataset for each model is included on the x-axis, with the resulting test accuracy on the excluded dataset on the y-axis.

Results

The original CNN was able to achieve 99.3 % accuracy when classifying the withheld test dataset. As shown in Figure 2, Validation accuracy was higher than training accuracy throughout training, which indicates the model underfit the data. The six separate CNNs tested using excluded datasets had accuracies ranging from 5.60% to 74.34%. These results are shown in Figure 3. Validation accuracy was also higher than training accuracy throughout the training of each of these networks, indicating underfitting. The original CNN was able to classify images at an average rate of .0038 seconds per image on a NVIDIA GeForce GTX 960M. Additionally, the CNN only takes up about 4 MB in memory.

Conclusion

The CNN was able to do exceptionally well at classifying ASL hand sign images when trained and tested on all six datasets. Additionally, it was able to classify hand signs quickly, while taking up a relatively small amount of memory. This suggests that uncomplicated architectures are viable for real time applications when their use does not cause any substantial loss in accuracy. When trained and tested different ASL users, the accuracies are less impressive, but still in league with other, more complicated architectures[1,2,3]. However, the CNN which excluded dataset six during training did marginally better than random classification. This is probably because datasets one through five were created from signers with lighter skin tones, while the signer who created dataset six had a darker skin tone. Excluding the dataset six led to a loss in diversity in the combined dataset, and a subsequent loss in generalization. In the future I would like to collect more ASL hand sign data, and build a real time classifier capable of performing accurate classification without user specific calibration.

Acknowledgements

I would like to thank Dr. Christine Julien for reviewing my project design, and Jacob Grohman for reviewing my poster.

Citations

[1] Garcia, B., & Viesca, S. A. (2016). Real-time American Sign Language Recognition with Convolutional Neural Networks. *Convolutional Neural Networks for Visual Recognition*.
[2] Ranga, V., Yadav, N., & Garg, P. (2018). American Sign Language Finger Spelling Using Hybrid Discrete Wavelet Transform-Gabor Filter and Convolutional Neural Network. *Journal of Engineering Science and Technology*, *13*(9)
[3] V. Bheda, D. Radpour, (2017). Using Deep Convolutional Networks for Gesture Recognition in American Sign Language, *CoRR*, vol. abs/1710. 06836