ASL Recognition

Overview

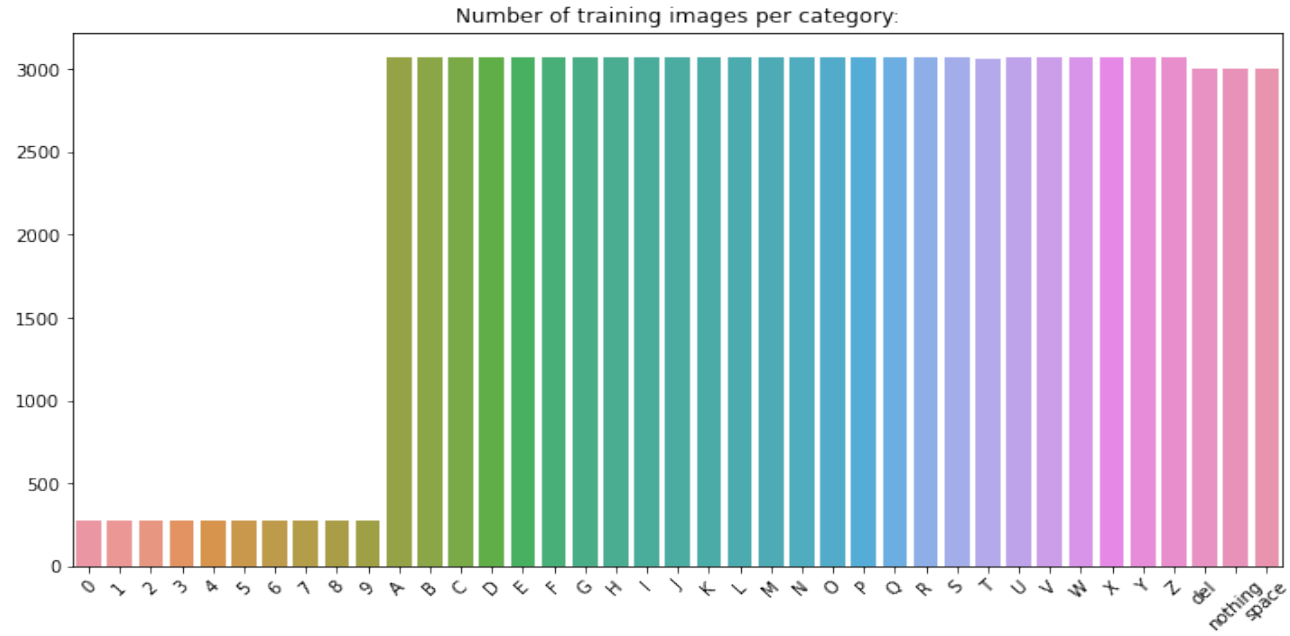
This paper summarizes outlines using Transfer Learning and Data Augmentation to create a deep learning model for an ASL dataset.

Data

The network was trained on a combination of the Kaggle datasets of ASL Alphabet and Numbers and the Massey University Gesture Dataset. The dataset comprises of 91937 images which are 200x200 pixels, divided into 39 classes (26 English letters, numbers 0-9 and 3 additional signs of SPACE, DELETE and NOTHING).

To train the model for more realistic scenarios since this project would involve real-time predictions from a live webcam feed, the data is augmented using brightness, zoom and rotation shifts.

Data Summary



There is a large discrepancy in the availability of images for numbers as opposed to the alphabet.

Transfer Learning

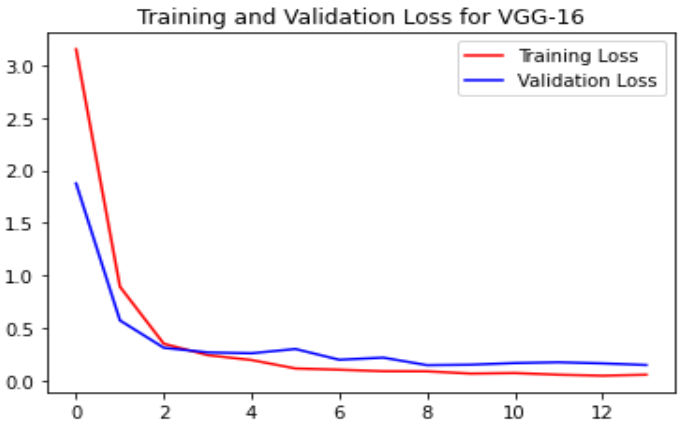
3 base models were used to train the network: Google's Inception v3, VGG-16 and Resnet-50. With each model, the latter blocks of layers were unfrozen to improve upon the training and model performance. A set of Fully Connected layers and a Dropout Layer and added after the inception network so as to conform the neural network for this application (consists of 2 Fully

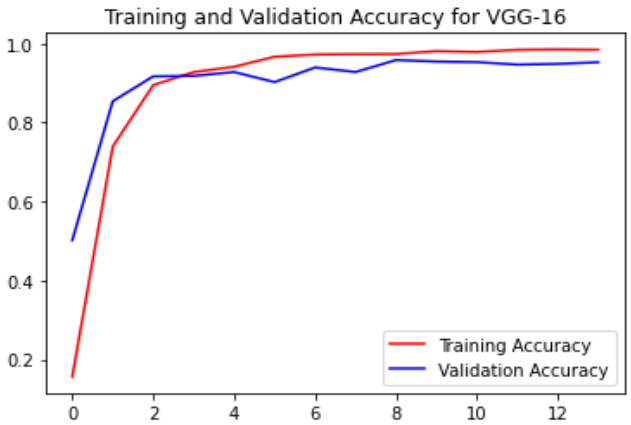
Connected layers, one consisting of 512 ReLu units and the other of 39 Softmax units for the prediction of 39 classes). The model is then trained on the dataset of images.

Results

Model	InceptionV3	VGG16	Resnet50
Training Accuracy	0.9795	0.9842	0.9844
Validation Accuracy	0.8231	0.9525	0.8609
Training Loss	0.0798	0.0549	0.0538
Validation loss	0.7341	0.1471	0.5475
Training Time (h:mm:ss)	1:05:20	27:16	39:13

VGG-16 is the clear choice of network architecture moving forward since it provides the highest validation accuracy and lowest validation loss as well as training time.





Model Application

OpenCV is used to capture frames from a video feed. The application provides an area or green rectangle where the signs are to be detected and recognized. The signs are then captured in frames, which are then processed and fed into the model. Predictions are outputted based on 3 tiers of certainty:

- If the model predicts a sign with a confidence greater than 75%, the prediction is presented as '[Sign] – [Confidence %'].
- Predictions that fall in between 75% and 25% are presented with a 'Maybe … [Sign] - [Confidence %]'
- Predictions below 25% is presented as 'Nothing'.

References:

https://github.com/atul-lanka/asl_recognition – Includes additional plots for training and validation accuracy and loss.