Lip Reading Word Classification Using CNN + LSTMs

Zoe Robert
Abiel Gutiérrez
Spring 2017

Introduction

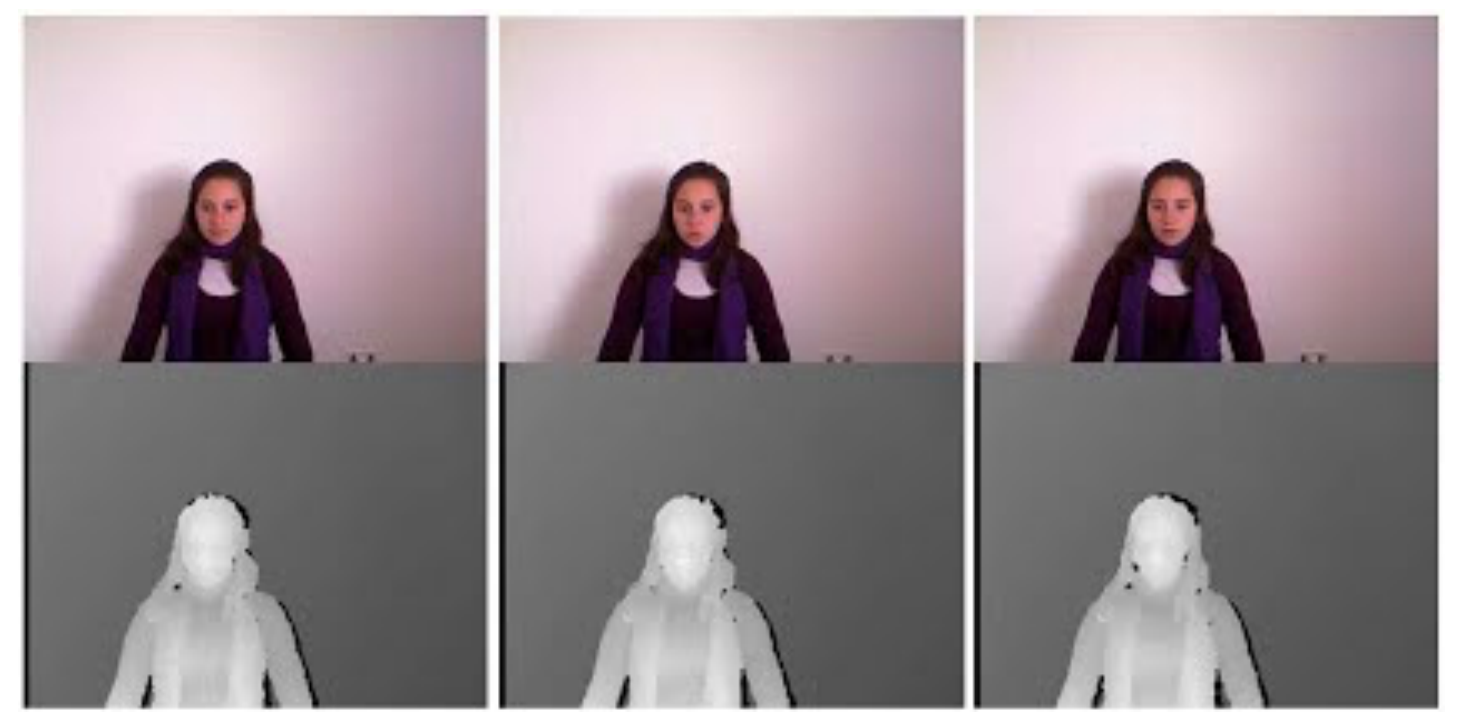
We worked on speech recognition from video without audio. This is interesting given that video traffic is growing at a high rate throughout the web, and this task could help us extract data and process it to gain interesting insights. Applications include profanity detection on social media, news broadcast transcriptions, and on-the-go lip reading mobile apps.

There's been only few word classifier models developed in recent years, and the first phrase classifier was developed by Google's DeepMind just a few months ago. This is thus a field with much to be explored, making it an interesting and exciting topic.

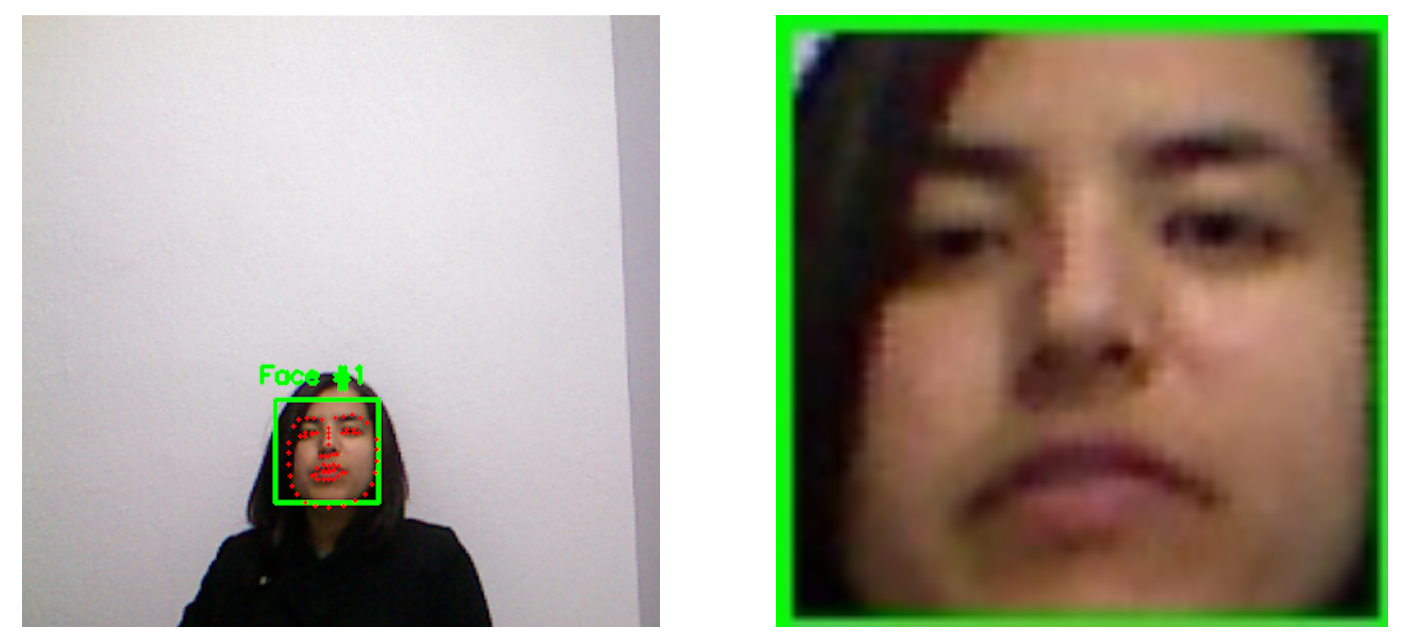
Dataset

MIRACL-V1

Contains frames of 15 people uttering 10 words and 10 phrases 10 times each, for a total of 3000 samples. Each sample is a sequence of color and depth images of size 640 x 480 pixels.



Words Included: *Begin, Choose, Connection, Navigation, Next, Previous, Start, Stop, Hello, Web.*



Preprocessing

Jittering (Augmentation)

We preprocessed the data with OpenCV and dlib's facial landmark library, resulting in 90x90 pixel facial crops. We plan to augment the dataset with horizontally flipped samples and pixel jittering, and will also try transfer learning with pretrained models.

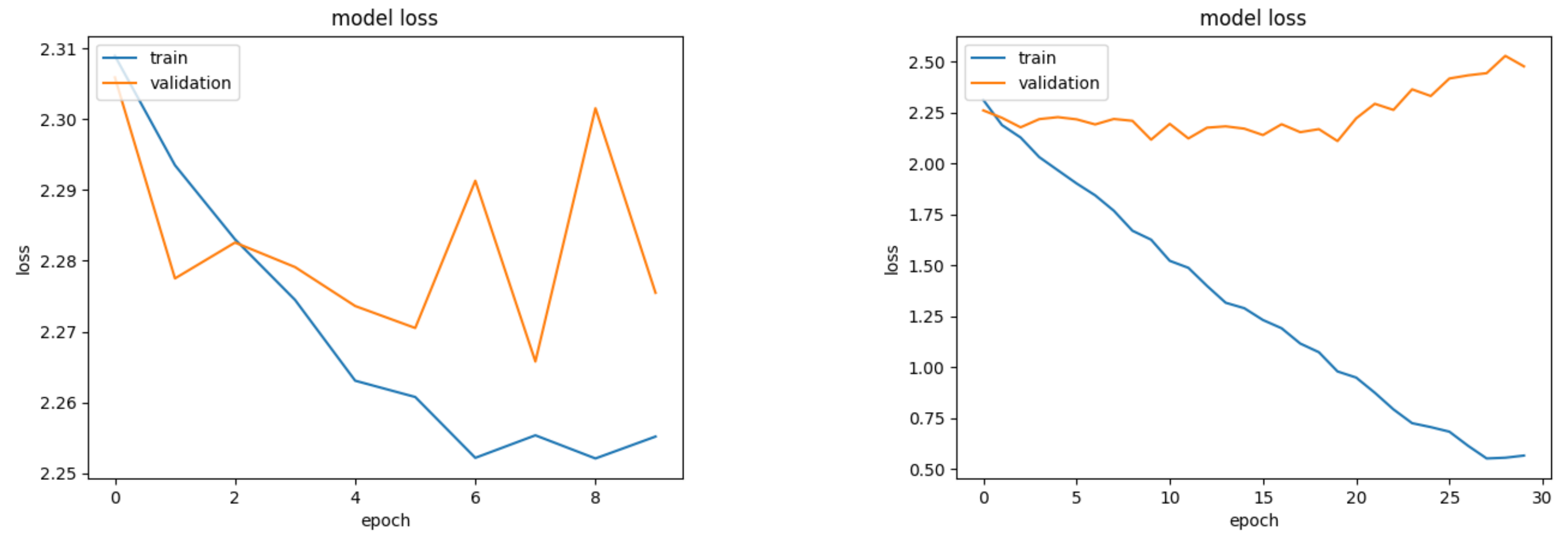
Problem Statement

We are building a model that classifies words uttered by people into one of ten classes. We will use the MIRACL-V1 dataset, and employ a combination of CNNs, RNNs, max pool and dense layers to train our model. We will evaluate by segmenting our data set into training, validation, and test sets, and evaluating our test predictions against ground truths — resulting in an accuracy percentage that will be comparable to other published models.

Evaluation

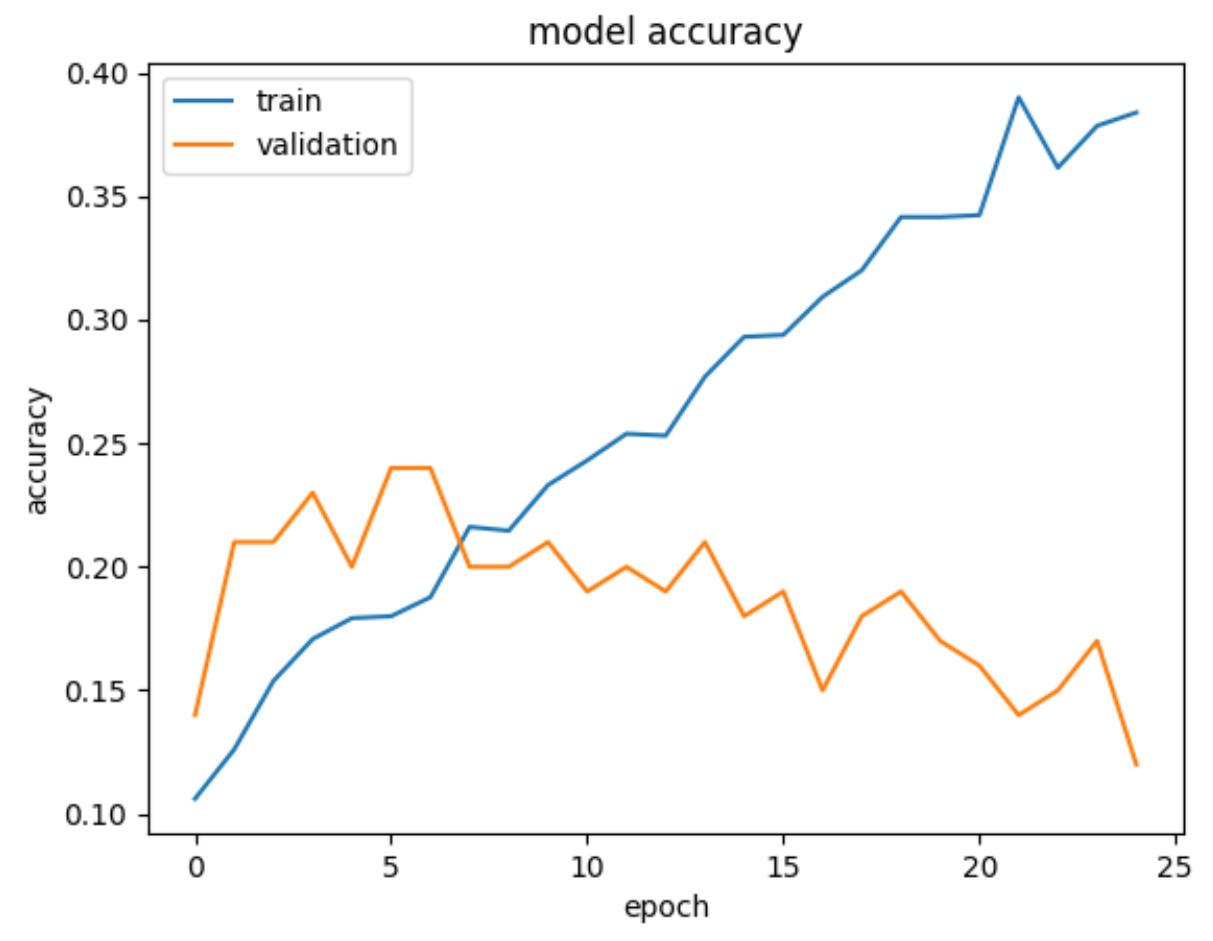
We are analyzing validation results on both seen and unseen subjects during training time, taking the percentage of correctly classified words. We have found that validation accuracy is much higher for seen subjects. We have a lot of work to do in reducing overfitting.

Baseline & Current Model Loss Functions

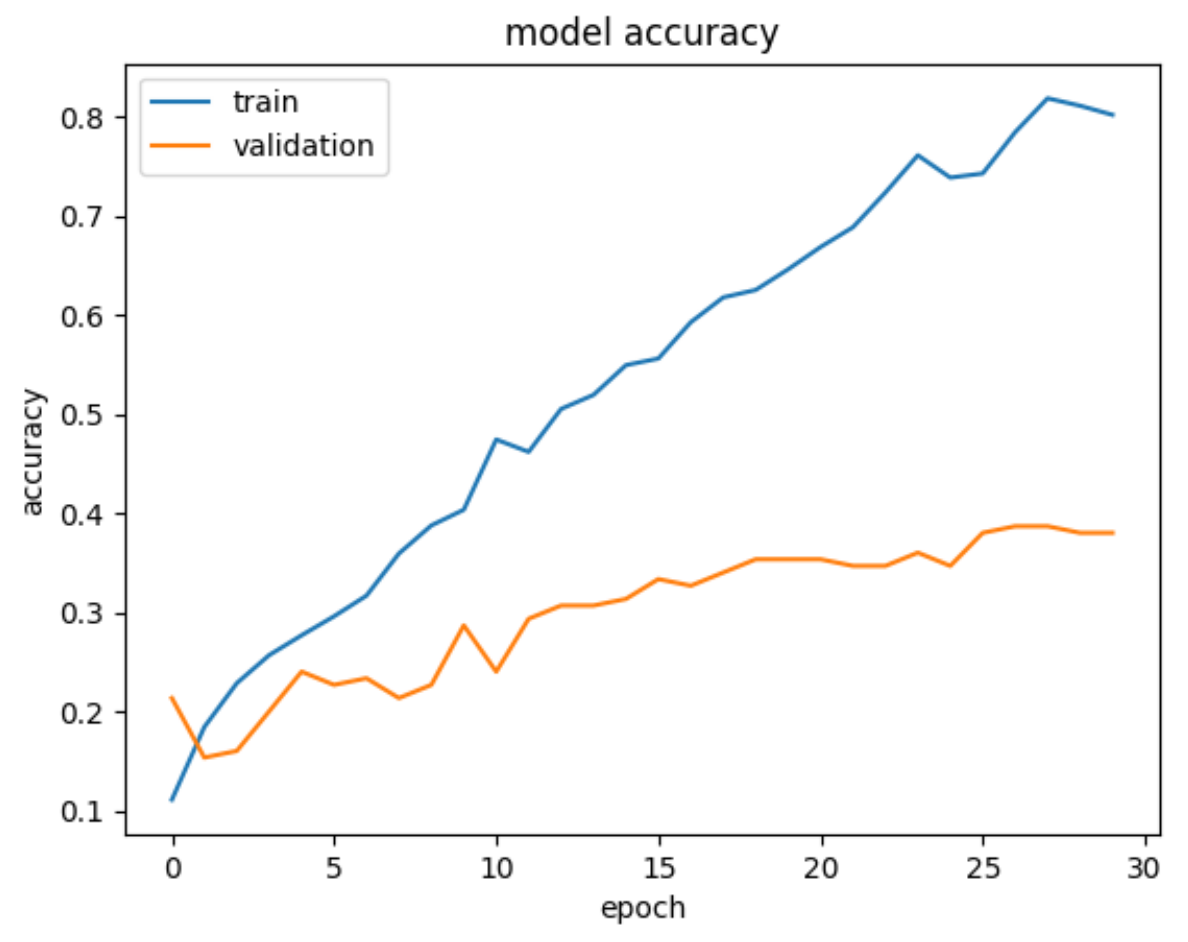


We see a big decrease in loss in the new model, but a big increase in overfitting. We plan to implement dropout in our next trials.

Current Model Accuracies (unseen vs seen subjects)



Training: 38% Validation: 12%
Validation has unseen subject

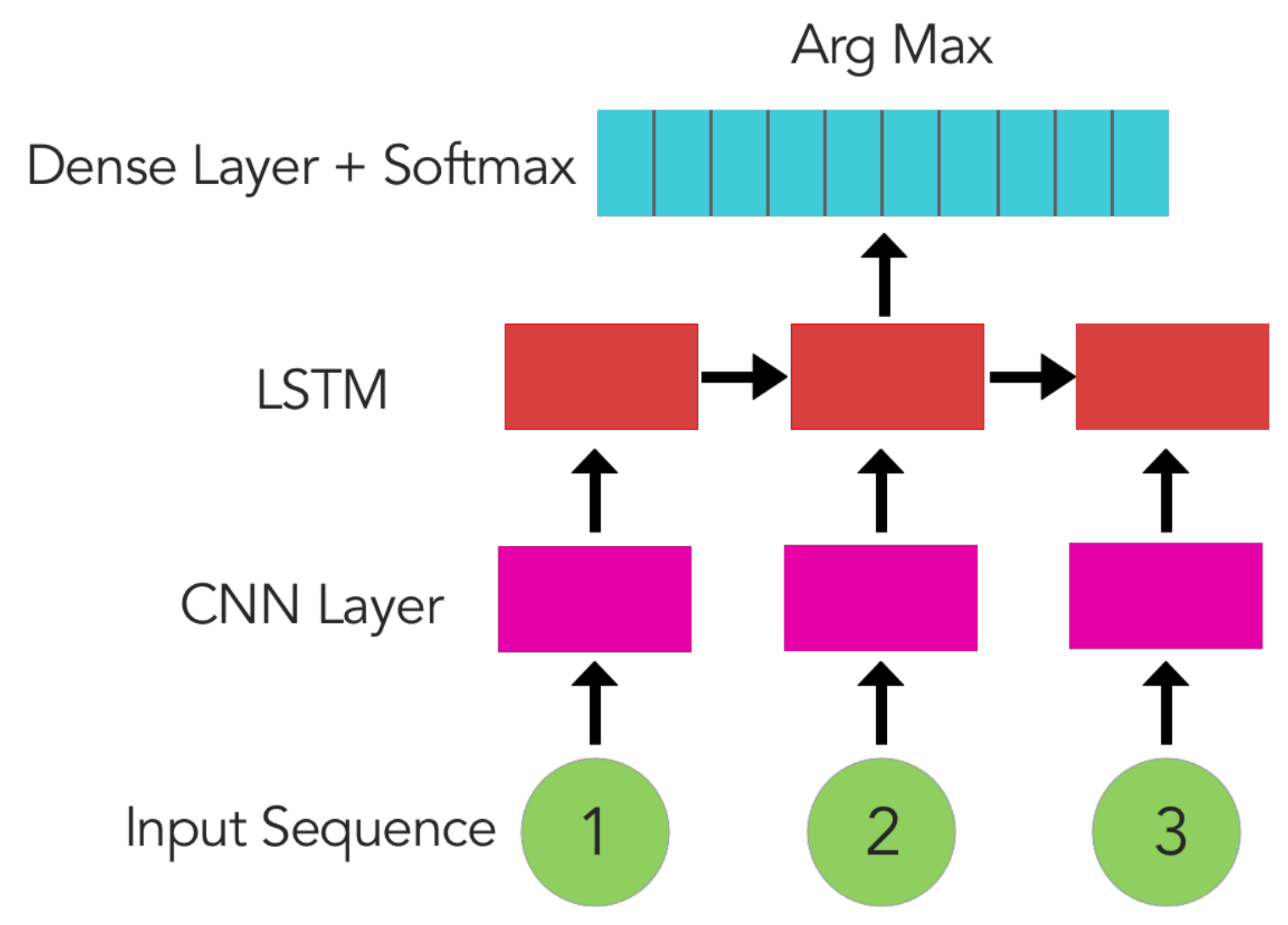


Training: 80% Validation: 38%
Validation has seen subject

Models

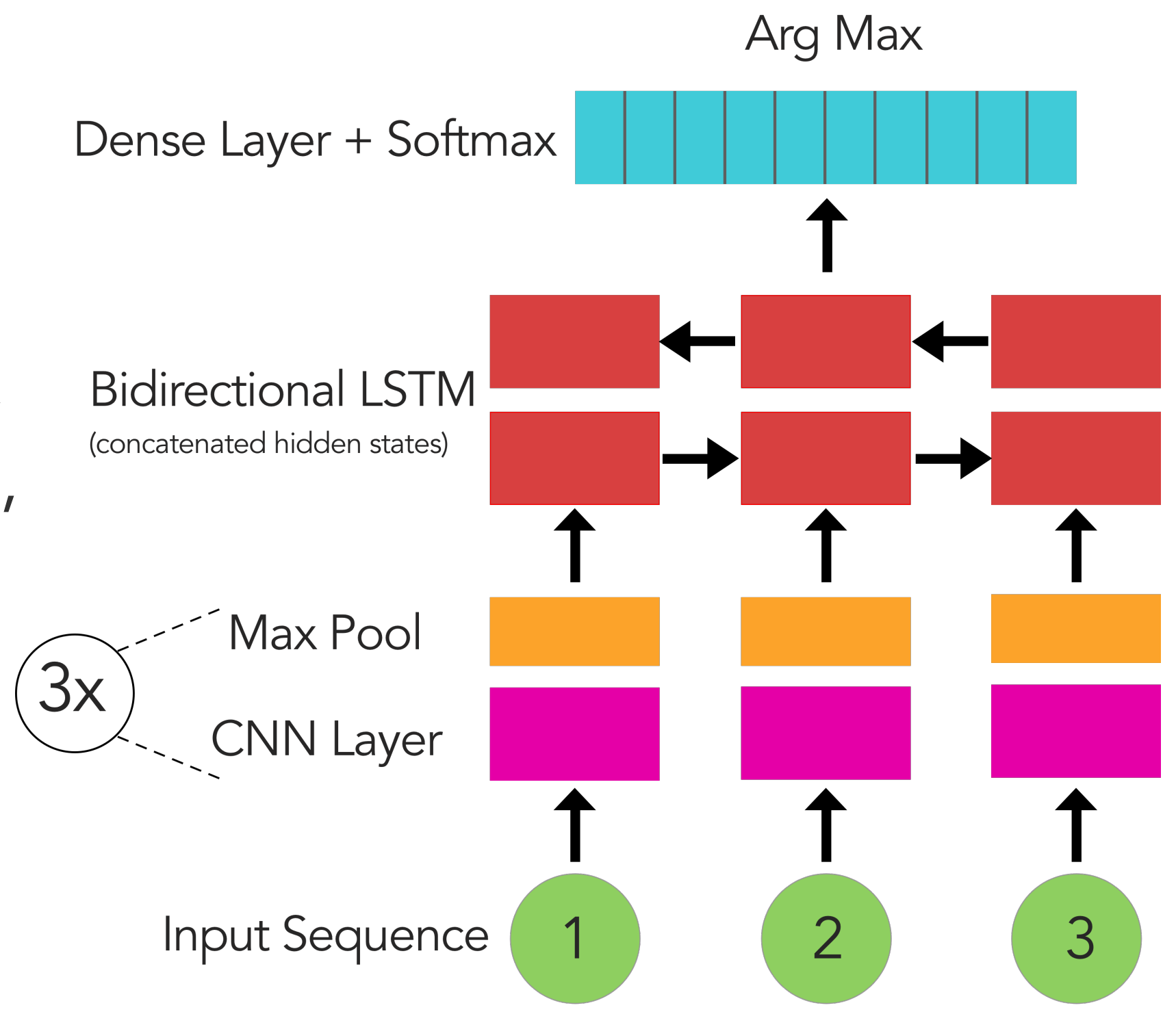
Baseline

Image sequences fed through CNNs and into LSTM, with 1 dense layer and softmax activation. No hyperparam tuning.



Current Model

Image sequences fed through 3 layers of CNNs with max pooling, and through a bidirectional LSTM, with 1 dense layer and softmax activation.



Used Adam Optimizer with cross entropy loss. Performed hyperparamer tuning on LR, batch size, and number of epochs. **Implemented with Tensorflow's Keras high level API.**

Future Directions

We anticipated higher accuracies, but we are also working with a 10-class classifier, so our baseline accuracies start low. We have yet to incorporate any pre-trained facial recognition CNN models as some of our inspiration papers did. Additionally, we will train on a larger data set by using the augmented (jittered, flipped) data.

Generally, we found it very difficult to avoid overfitting with unseen people. This project is easily extendible and raises the question of how to perform visual speech recognition on a much larger corpus (perhaps the entire english dictionary). Is it easier to understand speech from video of a single word being spoken or entire phrases and sentences? How could the addition of audio data improve our ability to interpret the video as text?