Computer Vision Lip Reading

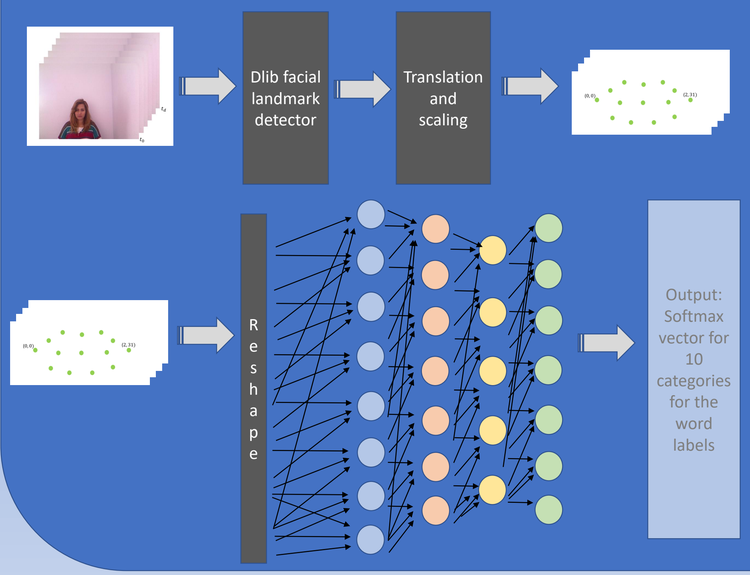
Grace Tilton - gtilton@stanford.edu

Predicting

For this project I used the MIRACL-VC1 dataset to teach a model to read lips. This is a very initial starting point where the problem has been boiled down to a classification problem for the model to predict which word (out of 10 possible words) is being spoken from an input which is a timeseries of snapshots of a person saying that word.

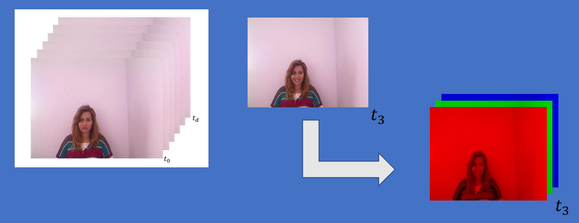
Models

This project actually leverages 2 separate models. First, I used the dlib facial landmark coordinates model (which was trained on the iBUG 300-W dataset) to pre-process the image data into time-series data of the facial landmarks. Then I developed a small neural network and trained on the facial landmark data.



Data & Features

Raw Data:
Each data sample is time series images of a word being spoken. Each set of .jpg files is a series of snapshots of video while one of 15 people speak one of 10 words. There are 10 samples for each word spoken by each person totaling 1500 data samples each classified to one of 10 words.

$x_i =$ 

Note that at each time step in each data point the image is a tuple of RGB values.

Initial Processing:
While the videos were all taken with the same camera and in the same position, there are slight variances in the size/location of a persons mouth in each image. In order to account for this I used the dlib library to identify just the portion of the image that is the mouth as a set of 21 (x,y) coordinates. Then, normalized each of those points by adjusting them proportionally to cover the same area.

This nicely drops the color dimension from the sample data so now each sample is represented by a set of (x,y,t) tuples. Lastly, we had to make sure all data samples were the same length by repeating the last frame for shorter samples.

Results

I broke my data set into 80% Training and 20% Test so 1200 samples were trained on and 300 samples were tested.
With this methodology I was able achieve an accuracy of 76.4% on the training data and 73.8% on the test data. I belive this is because a lot of fidelity is lost in the conversion from image to facial landmark coordinates.

Discussion & Future

What I like about this method: This method strips out of most personal markers of the speaker. One issue I was trying to overcome that the MIRACL team struggled with is that their models had good accuracy when isolated to a specific speaker but significantly reduced results across the speakers.

What I don't like: This solution feels a little piece-meal. I have a feeling that given a different CNN, I could trust the model to convert directly from image series to word classification.
I definitely need some more time to play with the model. I don't necessarily believe it is optimal.

Cool follow-up ideas: Recognizing pauses to delimit words, adding an NLP statistical model to compensate for low uncertainty, a higher fidelity facial landmark identification method

References

Ben-Hamadou, A. (2019). *MIRACL-VC1 - Achraf Ben-Hamadou*. [online] Sites.google.com. Available at: https://sites.google.com/site/achrafbenhamadou/-datasets/miracl-vc1 [Accessed 5 Dec. 2019].

Rosebrock, A. (2019). *Facial landmarks with dlib, OpenCV, and Python - PyImageSearch*. [online] PyImageSearch. Available at: https://www.pyimagesearch.com/2017/04/03/facial-landmarks-dlib-opencv-python/ [Accessed 5 Dec. 2019].Lip

Project Video Link: https://youtu.be/rEo3Arno0Ho