READ MY LIPS

A lip reading deep learning project

Xiaotong Chen, Hao Mao

Stanford University

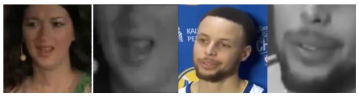
Introduction

When watching NBA games, we see players talking on the field but what are they really talking about? To answer that question, we designed an end-to-end deep neural network that takes grayscale video that contains mouth area as input and outputs the corresponding sentence. The process of recognizing speech based on lips movements is also known as lip reading.

Data

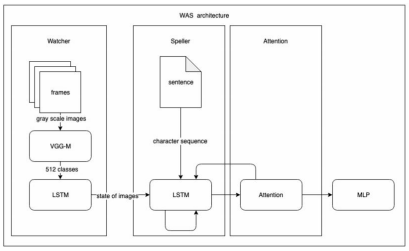
We use the LRS3 dataset and the NBA games we downloaded from the official site.

We also processed the data by cropping the mouth area to exclude noises like background, hair style, eyes and so on. Below are some examples of our dataset.



Architecture

Our model has three components: **Watcher** runs frames through VGG to understand the lip motion and then feed the feature of lip posture to LSTM; **Speller** takes output of Watcher from LSTM along with character sequence and runs them through LSTM; **Attention** aligns the character and lip posture sequences.



We use cross entropy as the loss function since predicting a character is classification.

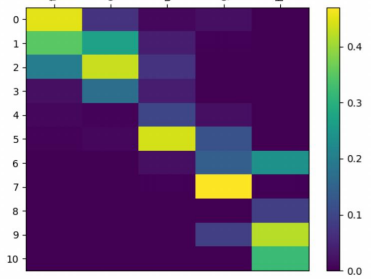
To evaluate the performance, we use the edit distance per character. Because, the prediction usually has different length than the grant truth and directly comparing them doesn't make sense.

Result

With 150+ hours training on a NVIDIA GeForce GTX 2080 GPU, we were able to overfit the training set. EDPC of dev and test set are pretty close to that of training set.

When it comes to the real NBA dataset, the result isn't as good as we would like, though it could get some words or characters right.

Dataset	Size (1000 hours)	Error (EDPC)
Train	96%	0.34
Dev	2%	0.36
Test	2%	0.40
NBA	0.5 hrs	0.68



The word "tough" from the sentence "it was a tough game" by Stephen Curry.

Error Analysis

we realized that our model only works for the frontal view, because there aren't many video from profile view in our dataset.

NBA players don't pronounce as clear as how speakers do in our dataset.

Future Work

It's super slow to train this end-to-end model. We definitely need to train the model longer.

We also need to feed the model with more data from different angles. (not just the frontal view)