REPORT
LIP READING

Utkarsh Sharma 21378 and Satyajeet Shashwat 21243

Abstract

Decoding text from a speaker's mouth movement is known as lipreading. We worked on a model, trained totally end-to-end, that uses spatiotemporal convolutions, a recurrent network, and the connectionist temporal classification loss to map a variable-length series of video frames to text. On the GRID corpus, we can achieve an accuracy of about 90 percent but due to a shortage of resources, we couldn't evaluate the performance to its potential which hopefully can be done when we continue to work on this project further.

Introduction

Lipreading is vital for human communication, but it's challenging due to latent visual cues and context ambiguity. Human lipreading performance is notably poor, motivating the need for automation. Machine lipreaders, with applications in hearing aids, silent dictation, security, and more, face difficulties in extracting spatiotemporal features. Existing deep learning approaches aim for end-to-end feature extraction, but most focus on word classification rather than sentence-level prediction. The goal is to improve automation in lipreading to enhance practical applications.

Related works

(i) J. S. Chung and A. Zisserman uses "Lip Reading in the Wild" dataset and got accuracy of 94.1 percent.

(ii)LIPNET: end-to-end sentence-level lipreading done by Yannis M. Assael, Brendan Shillingford, Shimon Whiteson and Nando de Freitas. They have used GRID Corpus dataset and got the accuraccy of 95.1 percent.

Datasets

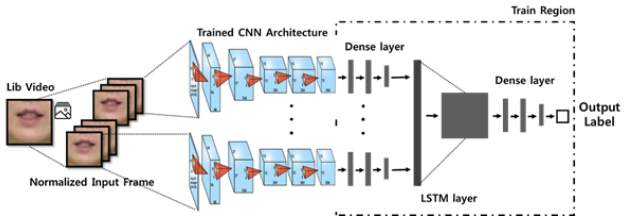
We have used GRID Corpus dataset that is available on University of Sheffield. GRID is a large multitalker audiovisual sentence corpus to support joint computational-behavioral studies in speech perception. In brief, the corpus consists of high-quality audio and video (facial) recordings of 1000 sentences spoken by each of 34 talkers (18 male, 16 female). Sentences are of the form "put red at G9 now". The corpus, together with transcriptions, is freely available for research use.

https://spandh.dcs.shef.ac.uk/gridcorpus/

Architecture

It is a neural network architecture for lipreading that maps variable-length sequences of video frames to text sequences, and is trained end-to-end. First we have extracted the frames from the video (24 frames for 1 second of video) and extracted the lip region. We have mapped the frames with the alignment. After that we have applied the different layers of CNN and then LSTM.

We have applied 3 layers of CONV3D, each followed by a a layer of Activation and spatial max-pooling layer. CONV3D applies a 3 dimentional filter to the dataset and the filter



moves 3-direction (x, y, z) to calcuate the low level feature representations. An activation layer in a CNN is a layer that serves as a non-linear transformation on the output of the convolutional layer. Pooling layer reduces the height and width of the image.

After CNN we applied Bidirectonal LSTM two times and then a dense layer. Bidirectional LSTM or BiLSTM is a term used for a sequence model which contains two LSTM layers, one for processing input in the forward direction and the other for processing in the backward direction. Dense Layer is simple layer of neurons in which each neuron receives input from all the neurons of previous layer.

We have also written a code(function) for CTC Loss.The connectionist temporal classification (CTC) loss eliminates the need for training data that aligns inputs to target outputs. Given a model that outputs a sequence of discrete distributions over the token classes (vocabulary) augmented with a special "blank" token, CTC computes the probability of a sequence by marginalising over all sequences that are defined as equivalent to this sequence. This simultaneously removes the need for alignments and addresses variable-length sequences.

Performance Evaluation

To measure the performance of LipNet and the baselines, we compute the word error rate (WER) and the character error rate (CER), standard metrics for the performance of ASR models. We produce approximate maximum-probability predictions from LipNet by performing CTC beam search. WER (or CER) is defined as the minimum number of word (or character) insertions, substitutions, and deletions required to transform the prediction into the ground truth, divided by the number of words (or characters) in the ground truth. Note that WER is usually equal to classification error when the predicted sentence has the same number of words as the ground truth, particularly in our case since almost all errors are substitution errors.

For both unseen and overlapped speakers we have evaluated the performance. This model exhibits a 2.1× higher performance in the overlapped compared to the unseen speakers split.

Limitation

In lip reading, a fundamental limitation arises from homophemes, where different words sound distinct but involve identical lip movements, making them visually indistinguishable. For instance, phonemes like 'p,' 'b,' and 'm' in English result in homophemes, such as 'mark,' 'park,' and 'bark.' This inherent challenge is compounded by intra-class variations like accents and speaking speed, as well as adversarial imaging conditions such as poor lighting, shadows, motion, resolution, and foreshortening.

Individual Contribution

We both first read some Lipreading research papers, **LIPNET: END-TO-END SENTENCE-LEVEL LIPREADING**, **Lip Reading in the Wild**. Then for extracting lip region we worked on two different ways, Utkarsh extracted the lip region by using the index of array of image. This is working on our dataset but this may or may not work on different image. Satyajeet extracted the lip region using Dlib Face detector which works for all the images. Satyajeet wrote the fuctions for loading videos, alignments, data, and CTC Loss. Utkarsh wrote the code to save the model. Utkarsh worked on CNN part and Satyajeet Worked on LSTM.

References

1) LIPNET: https://arxiv.org/pdf/1611.01599.pdf

2) Lip Reading in the wild: https://www.robots.ox.ac.uk/ vgg/publications/2016/Chung16/chung16.pdf

3) Convolutional Neural Netwojhkjhrk (CNN): https://www.geeksforgeeks.org/convolutional-neural-network-cnn-in-machine-learning/

4) Long Short Term Memory(LSTM) Network: https://www.geeksforgeeks.org/long-short-term-memory-networks-explanation/

5) TensorFlow : https://www.tensorflow.org/api$_{d}ocs/python/tf/all_{s}ymbols$