Project Report
Advancing Lipreading using Spatiotemporal Convolutions and Recurrent Networks

Vishal Bharti 21310
Computer Vision (DSE-312) Course Project
Email: vishal21@iiserb.ac.in

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 spatiotempo ral convolutions, a recurrent network, and the con nectionist 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 re sources, 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 con text 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 work

Lip Reading in the Wild: Chung et al. (CZ16) developed a lip reading model using the "Lip Reading in the Wild" (LRW) dataset, a large-scale dataset for word-level classification from unconstrained video data. Their model, built using a 3D convolutional network and trained end-to-end, achieved a significant performance boost over prior models, reaching an accuracy of 94.1% on the LRW dataset. This work emphasizes word-level classification, leaving room for improvements at the sentence level.

LIPNET: LIPNET is a pioneering end-to-end sentence-level lipreading model that uses the GRID corpus for training. Assael et al. (ASWdF16) incorporated a spatiotemporal convolutional neural network followed by a recurrent neural network to achieve a 95.1% accuracy on the GRID dataset. This system outperforms earlier models in sentence-level prediction, which is more complex due to varying lip movements and sentence structures.

Visual Speech Recognition: Wand et al. (WKS16) employed Long Short-Term Memory (LSTM) networks to recognize visual speech. The paper introduced a sequence learning approach using LSTMs to classify spoken phrases from video frames. Their work highlights the effectiveness of temporal modeling in lipreading.

Deep Lip Reading with Self-Supervised Learning: Afouras et al. (ACZ20) explored a self-supervised learning approach for lipreading by leveraging large amounts of unlabeled data. Their approach reduces reliance on expensive labeled datasets and achieves competitive results on various benchmarks by training a model to predict speech from visual input using an unsupervised method. This work underscores the potential of leveraging unlabeled data in improving lipreading models.

Cross-Attention for Lipreading: Zhou et al. (ZSJ$^+$19) proposed a cross-attention mechanism to enhance lipreading performance. Their system integrates audio and visual data through a cross-modal attention framework, allowing the model to effectively learn correlations between lip movements and speech audio. While this approach achieved strong results on several datasets, it is designed for multimodal scenarios.

Dataset

We have used the GRID corpus dataset, which consists of over 34 talkers speaking 1000 sentences each. This data is publicly available and includes synchronized audio-visual data. The structure of the dataset is suitable for sentence-level lipreading, where sentences like "put red at G9 now" are used to train and validate the model.

Model 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 ex tracted the lip region. We have mapped the frames with the alignment. After that we have applied the different layers of CNN and then LSTM.

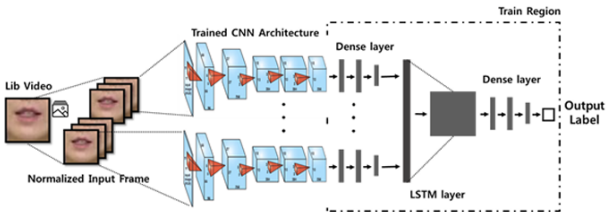

Figure 1: Model architecture

We have applied 3 layers of CONV3D, each followed by a a layer of Activation and spa tial 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.

Results and 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 over lapped compared to the unseen speakers split.

Limitation

Homophemes (e.g., 'bark' and 'mark') pose a significant challenge, as different words often exhibit the same lip movements. Additionally, factors such as accent variations, lighting conditions, and speaking speed can affect the performance of the model.

References

[ACZ20] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Self-supervised learning for audiovisual speech recognition. *arXiv preprint arXiv:2010.09753*, 2020.

[ASWdF16] Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando de Freitas. Lipnet: End-to-end sentence-level lipreading. *arXiv preprint arXiv:1611.01599*, 2016.

[CZ16] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. *arXiv preprint arXiv:1611.01599*, 2016.

[WKS16] Michael Wand, Jan Koutnik, and Jürgen Schmidhuber. Lipreading with long short-term memory. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6115–6119. IEEE, 2016.

[ZSJ+19] Yuchen Zhou, Chuang Shi, Jun Jiao, Jingdong Zhang, and Feng Wu. Cross-attention for audio-visual speech recognition. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2247–2255, 2019.