Automatic Video Speech Recognition System

Ashwin Sateesh Kumar
sateeshkumar.a@northeastern.edu

Basil K Varghese
varghese.b@northeastern.edu

1 Introduction

The advancement of deep learning models and architectures has shown great potential in replicating and learning various aspects of human life, leading to the development of numerous life-changing tools, such as self-driving cars and speech recognition systems. Our goal in this project was to develop an automated video speech recognition system (or a lip-reading model) that can detect and generate words spoken in a short video using only visual cues, with the aim of improving the quality of life for individuals with hearing impairment. A lip-reading model could also have other applications, such as enhancing speech recognition in noisy environments, improving security and surveillance, and enabling people with disabilities to control devices using lip movements. To achieve this, we used pre-trained convolutional models to detect facial features in each image frame, which were then fed into sequence model layers such as LSTM and Transformer to learn how a sequence of image frames produces a certain word. These models have shown great potential in various natural language processing (NLP) tasks, particularly in processing sequential data, which is critical for lip-reading. We used the pre-trained Haar Cascade Classifier to detect the mouth and lips in each image, and cropped and resized the lip section of each image. To extract features from each image, we used VGG16 and Resnet50 convolutional models, which have been trained on large datasets for image classification tasks and have been shown to perform well in various computer vision tasks. Our model was trained on the MIRACL-VC1 dataset, which contains lip-reading images of 15 individuals saying a set of 10 words and 10 phrases consisting of a total of 3000 instances. We used accuracy as the model performance metric since the dataset target label distribution was well-balanced. We also tested our model on pre-recorded video to evaluate its ability to detect words in faces not seen during training. Although our model is a rudimentary prototype due to the limited availability of labeled images for the dataset, our project provides a starting point for future work. Further research could focus on training the model on larger datasets and improving its ability to work on streaming video with continuous sentences.

2 Related Work

Face Detection. There is a large body of work on face detection using deep learning methods. [1] presents an object detection algorithm that works by scanning the image with a sliding window at different scales and positions. Each of the cascade of classifiers is applied sequentially, and features are extracted if it passes the threshold. Another state-of-the-art method for face detection [2] is based on multi-task cascaded convolutional neural networks (MTCNN). This generates candidate windows for faces, a refinement network that applies bounding box regression to refine the locations and sizes of the candidate windows, and a landmark regression network that predicts the facial landmarks. This framework allows joint face detection and alignment in a single network.

Feature Extraction. The process of identifying and extracting relevant information from images is a fundamental step in the computer vision tasks. The ResNet architecture in [3] introduces the concept of residual blocks, which contain skip connections that allow the network to learn residual functions instead of full functions. The authors also introduce a bottleneck design that reduces the number of parameters and computation required to process the features. The Inception

network in [4] contains multiple parallel branches of convolutional layers of different sizes, and a pooling layer which is called the inception module, is another state-of-the-art method that uses 1x1 convolutional layers to reduce the computational cost of the model while maintaining its accuracy.

Lip Reading. [5] uses VGG network to extract the features from the lip after obtaining key frames from the method proposed. Later fused with attention-based LSTM to learn the sequence information. This achieves an accuracy of 88.2% on ten different words. [6] proposes a deep learning model for lip reading in unconstrained environments. The model consists of a 3D convolutional neural network (CNN) for feature extraction and a bidirectional long short-term memory (LSTM) for sequence modeling. The authors propose a new data augmentation method, called frame dropping, to improve the robustness of the model to temporal variations in the video. The proposed model achieves state-of-the-art performance on the LRW dataset. [7] Discusses the traditional lip reading steps: face detection, lip localization followed by feature extraction and recognition. It also speaks about pixel and model-based methods with their performances in extracting the features

Speech Recognition. The neural network presented in [8] learns to transcribe speech utterances to characters using a sequence-to-sequence (Seq2Seq) architecture. The encoder takes the acoustic features extracted from the speech signal and generates a fixed-length vector of hidden state which is then fed to the decoder along with the attention mechanism to generate the sequence of characters representing the transcription of the speech.

3 Method

The MIRACL-VCI dataset used consists of video frames of 10 words and 10 phrases spoken 10 times by 15 individuals giving us a total of around 37000 images. The face from each of these images was detected using pre-trained Haar Cascade Classifier. The lip region was cropped out from these images by changing the parameters of the bounding box and the resultant images were of size 25x58x3.

The lip images were processed using the OpenCV library. Since the images were from videos of people speaking respective words and phrases, each had different frame lengths from 7 to 22 for words and 7 to 27 for phrases. Hence we padded the image arrays with zero vectors to maintain the same dimensions for each instance. Also, the sequence of each of the video frames for each instance of the word and phrase spoken was preserved

The preprocessed lip images (1500 instances for words and 1500 for phrases) were fed into a pre-trained Resnet50 and VGG16 networks for feature extraction. The last softmax layer was removed for these networks to obtain just the spatial feature vectors. The outputs from Resnet50 and VGG16 were 2048 and 512 dimensional vectors.

The feature vectors from each video frame for each of the single instance were sequentially in the order that was initially generated to LSTM, LSTM with Attention and Transformer sequentially, to obtain the probability of a video being classified as a particular word/phrase for each of the respective instances. The predicted class of the word/phrase was mapped to generate the text of word/phrase and added to the video as a caption.

We used two types of positional encoding for the transformer. The traditional sine and cosine based positional encoding and the learned positional encoding that uses 1D convolutional neural networks along with single dense layer at the end to learn the positions. The architecture of the proposed method is shown in Figure 1.

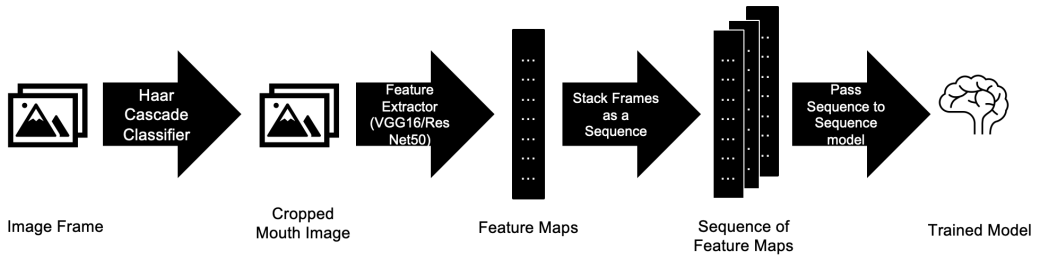

Figure 1: Architecture of the proposed method

4 Experiments

We used multiple methods and ran multiple experiments to train a model that can Identify words from a sequence of image frames. Haar Cascade Classifier was used as to detect and crop the mouth region of each image. VGG16 and ResNet50 pre-trained models were used to extract features from the mouth images by converting each image to 512 and 2048 dimension vectors respectively using Global Average Pooling to convert the last convolutional layer to flattened vectors. We trained LSTM, LSTM with attention, and Transformer models to classify words and phrases separately. Hyperparameter tuning was done for each combination of CNN model and sequence model and the best model was selected based on accuracy. LSTM model with Resnet50 feature maps was used as the baseline model and other models were compared to that Table 2 depicts the results of our experiments.

5 Results

The results of this project show that it is possible for sequence models to learn a sequence of images and accurately classify them into different words and phrases that they represent. This project represents a minor step towards the development of an automated video speech recognition system or a lip-reading model. VGG16 model with Transformer performed the best for detecting words and phrases. This highlights the potential of using pre-trained convolutional models and sequence model layers to process sequential data for lip-reading applications. Expanding the model's capability by training it on a larger vocabulary would be a valuable direction for future work. This could be achieved by scraping videos from the internet and extracting the words and frames using subtitles. Scaling the project would have the potential to replace or enhance many audio-based transcription systems. Overall, this project demonstrates the potential for using deep learning models and architectures to develop automated video speech recognition systems or lip-reading models. With continued research and development, such models could significantly improve the quality of life for individuals with hearing impairments and have a wide range of other applications in fields such as security, surveillance, and device control.

Table 1: Results

Model	Accuracy (Words)	Accuracy (Phrases)
RESNET50 LSTM	71.3	72.3
RESNET50 LSTM Attn	86.1	75.4
RESNET50 Transformer (traditional PE)	82.4	88.3
RESNET50 Transformer (learned PE)	78.7	82.3
VGG16 LSTM	89.1	86.6
VGG16 LSTM Attn	91.3	77.6
VGG16 Transformer (traditional PE)	86.6	87.6
VGG16 Transformer (Learned PE)	82.3	80.6

References

[1] Paul Viola, & Michael Jones (2001) Rapid Object Detection using a Boosted Cascade of Simple Features, *Conference on Computer Vision AND Pattern Recognition*

[2] Kaipeng Zhang, Zhanpeng Zang, & Zhipeng Li (2016) Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks *IEEE*

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun (2015) Deep Residual Learning for Image Recognition, *arXiv*

[4]Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, & Andrew Rabinovich (2014) Going deeper with convolutions, *NeurIPS*

[5] Yuanyao Lu, Hongbo Li (2019) Automatic Lip-Reading System Based on Deep Convolutional Neural Network and Attention-Based Long Short-Term Memory, *International Journal of Applied Science*

[6] Joon Son Chung, Andrew Senior, Oriol Vinyals & Andrew Zisserman (2017) Lip Reading Sentences in Wild, *arXiv*

[7] Soundarya B, Krishnaraj R, & Mythili S (2021) Visual Speech Recognition using Convolutional Neural Network, *IOP Conf. Series: Materials Science and Engineering*

[8] William Chan, Navdeep Jaitly, Quoc V. Le, & Oriol Vinyals (2015) Listen, Attend, and Spell, *NeurIPS*

[9] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei (2016) Video Captioning with Transferred Semantic Attributes, *arXiv*

[10] Sooraj V, Hardhik M, & Nishanth S Murthy (2020) Lip reading technique - A Review, *International Journal of Scientific & Technology*

[11] Oriol Vinyals, Alexander Toshev, Samy Bengio, & Dumitru Erhan (2015) Show and tell: A Neural Image Caption Generator, *IEEE Conference on Computer Vision and Pattern Recognition*

[12] Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Feipeng Cai, Georgios Tzimiropoulos, & Maja Pantic (2018) End-to-End Audio Visual Speech Recognition, *arXiv*