Proceedings of the Fifth International Conference on Communication and Electronics Systems (ICCES 2020)
IEEE Conference Record # 48766; IEEE Xplore ISBN: 978-1-7281-5371-1

Visual Speech Recognition: A Deep Learning Approach

Navin Kumar Mudaliar[1], Kavita Hegde[2], Anand Ramesh[3], Dr. Varsha Patil[4]
Dept. of Computer Engineering
SIES Graduate School of Technology
Navi Mumbai, India
[1]nk.mudaliar2@gmail.com
[2]kavita.hgd98@gmail.com
[3]anandstark0703@gmail.com
[4]varshasp2977@gmail.com

***Abstract*— A machine being able to perform lip-reading would have been deemed impossible a few decades ago. However, the exponential growth of machine learning in the past few years has made it possible for a machine to understand human speech based on visual inputs alone. Numerous research studies infer that a very less percentage of the English language can be comprehended through visual data alone, i.e. lip reading. Visual speech recognition experts can only infer about 3-4% of words spoken through lip-reading after viewing videos (without audio) multiple times. These experts also examine other parameters such as body language, facial cues, habits, and context to some extent. This task is very tedious (or exhausting). The proposed visual speech recognition approach has used the concept of deep learning to perform word-level classification. ResNet architecture is used with 3D convolution layers as the encoder and Gated Recurrent Units (GRU) as the decoder. The whole video sequence was used as an input in this approach. The results of the proposed approach are satisfactory. It achieves 90% accuracy on the BBC data set and 88% on the custom video data set. The proposed approach is limited to word-level only and can easily be extended to short phrases or sentences.**

Keywords*— *Lip reading, visual speech recognition, computer vision, convolutional neural network, deep learning.

I. INTRODUCTION

Lip-reading is a method to understand speech by observing and interpreting the motion of the lips, face, and other social cues. Speech recognition is extremely difficult in noisy environments and visual speech recognition can pave the way in creating assisting technologies for the same.

What makes Lipreading a formidable problem is the plethora of accents, skin color, the pace of uttering words, and facial features especially in the absence of context [1-3]. Transcription of messages in real-time, handling multiple speakers, and improving the performance of current speech recognition systems are all applications of VSR. The two major aspects of Visual Speech Recognition are Natural Language Processing (NLP) and Computer Vision (CV). While advancements in a single field are very impressive, knowledge and advancements from seemingly two distinct sectors can coalesce into a breakthrough of sorts [4]. This paper consists of related works, methodologies used, implementation details, and the results. The proposed approach consists of 3 phases which are preparing the dataset, lip detection, and feature extraction and prediction on unseen visual data. The input is given as a video from where the model extracts the spatial and temporal features for prediction.

It has been determined as it stack more and more layers in a deep learning model, the vanishing gradient problem arises, i.e. backpropagation to the starting layers makes the gradients extremely small. Thus, ResNet architecture came into existence that skipped one or more layers. This was termed as identity shortcut connection [5]. LSTM (Long Short Term Memory) and GRU (Gated Recurrent Units) networks have been designed to solve the problem of losing the input's information when it becomes too long for the model to carry it in further layers [6,19]. It employed encoder-decoder architecture for our approach where the ResNet model was our encoder that extracted the spatial-temporal features from the videos and the GRU part was the decoder that accepted the output of the ResNet and predicted the word.

The reason to choose BBC's Lip Reading in the Wild (LRW) dataset [7] is that it is one of the largest lipreading datasets containing 500 words that is publicly available. The model selects 15 similar words, for example 'Again' and 'Against', 'British' and 'Britain' and words that were completely different, for example, 'About' and 'Spending' where 90% accuracy is obtained on the test dataset using our model. It is also tested out on separate videos of our own where it performed very well. The model can be improved further by making use of the dataset with more variety and instances.

II. RELATED WORKS

Lip Reading is one of the major problems that many researchers are trying to solve which led to the creation of many large datasets like BBC's Lip Reading Sentences (LRS) and the LRW dataset [11, 7]. The inspiration is taken from the audiovisual and visual speech recognition systems [8, 9] where ResNet blocks are used for the audio as well as video aspect of word-level prediction. CTC models coupled with LSTMs is another approach to this problem where character-

level recognition is done by performing predicting labels frame-wise. LipNet [10] is the most popular implementation of this approach.

Another approach that is greatly followed these days is using sequence-to-sequence models which is an encoder-decoder approach using LSTM blocks [11]. This approach's shortcoming is that it fails in longer input sequences. To overcome this, the attention mechanism is implemented, mixing both the audio and video input sequences. Another approach is the use of Convolutional Auto Encoders for feature extraction and coupling it with LSTMs for decoding the hidden vectors as explained in this paper [12]. This is a similar approach to ours where ResNet is used as a feature extractor in place of a simple convolutional neural network.

III. METHODOLOGY

A. Characteristics and Preparation of Dataset

The collection and preparation of the dataset are always one of the most crucial aspects of any machine learning study. However, since the use of Visual Speech Recognition systems has already been determined, there are some substantial and well-defined datasets available at this time. In this study, Lip Reading in the Wild (LRW) [7] dataset is employed. It contains videos of short durations from various BBC shows. It contains over 1000 diverse speakers with many variations in the lighting conditions.

It is a dataset containing 500 words which is greater than any other dataset currently available. However, the video contains a phrase or sentence where the word is uttered generally in the middle of the video. This makes it a challenge to identify the frames containing the word of interest. This is where the metadata comes into play. It contains the video duration and the timestamp of the word in the video. The dataset also has similar words or plural forms of the words but it is avoided using all of them because it would complicate our problem and it is necessary to implement the model on real-time video inputs and for testing purposes. Since the video contains words other than the proposed label, which have to deal with co-articulation of the label with other words.

OpenCV and a machine learning model is used for the preparation of the dataset. Here, the main focus is to extract the Region of Interest (ROI) i.e. the area around the lips. The facial landmark detector implemented inside dlib library [16] produces 68 (x, y)-coordinates that map to specific facial structures. The coordinates were obtained using iBUG 300-W [13,14] dataset while training it on a shape predictor.

All the frames were extracted from the video and each frame was fed to the model for the detection of lips. Then, all the frames of the video containing only lips are added to an array. This array was given as input to the model. This approach is analogous to the image-captioning approach [17] where a set of words is associated with a single image. This approach saves a lot of time in computation as compared to sending batches of overlapping frames to the model as input, one at a time.

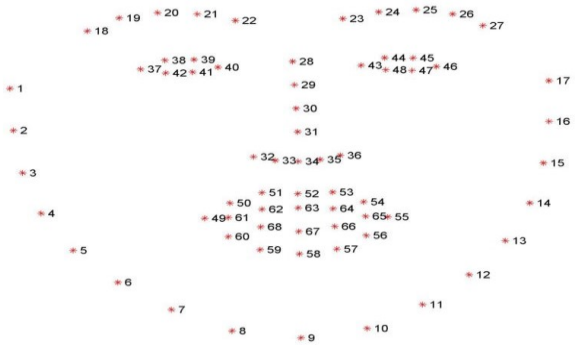

Fig. 1. Facial Landmark

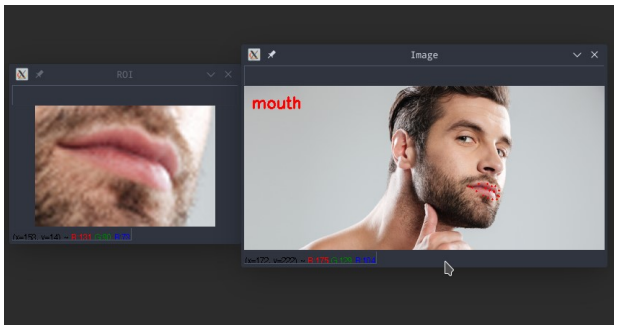

Fig. 2.Mouth localization

B. Model Architecture

One of the bottlenecks of VGG [15] is the degradation of the model's performance as the depth of the model increases because the gradients vanish during backpropagation. ResNet can solve this as the gradients can flow directly from lateral layers to preliminary layers through the skip connections [18]. The ResNet layer is modified by adding 3D convolutional layers to extract the spatial and temporal features of the video inputs. The rest of the model consisted of 2D convolutional layers that were stacked by more layers of similar behavior. This comprised the encoding part of our architecture that extracted and learned all the features which were then given to the decoder i.e. the GRUs for prediction of the labels.

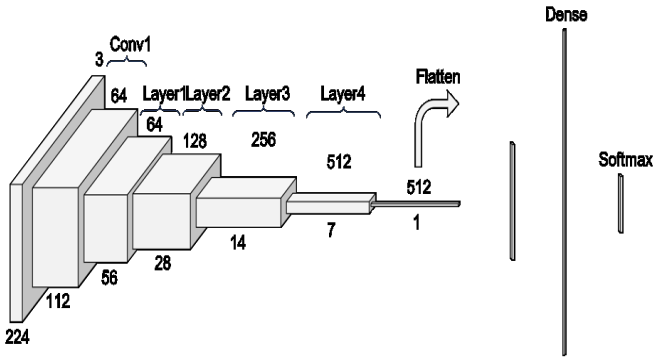

Fig. 3.General ResNet Architecture

Whenever it has been tried to predict something, all the available information are utilized. However, sometimes it needs only recent information to predict something. This is because the context/information needed to comprehend the current sequence is found in the recently seen sequences. Simple recurrent neural networks can handle such situations. However, it fails when the gap between the information is too large. Thus, to retain information for a longer duration, Long Short Term Memory is used as the main characteristics to retain information for a long amount of time.

The Bi-directional Gated Recurrent Units [20,21] (GRU) is a variation of Long Short Term Memory where the information not only flows from front to back but also back to front. It doesn't contain a separate memory unit as it is integrated along with the network. It also has fewer parameters to train compared to Long Short Term Memory networks.

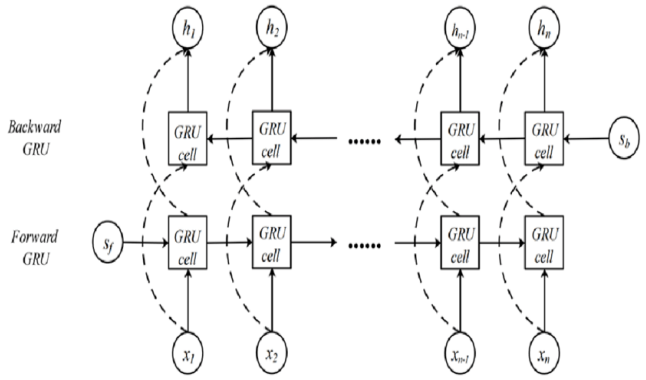

Fig. 4. GRU Network

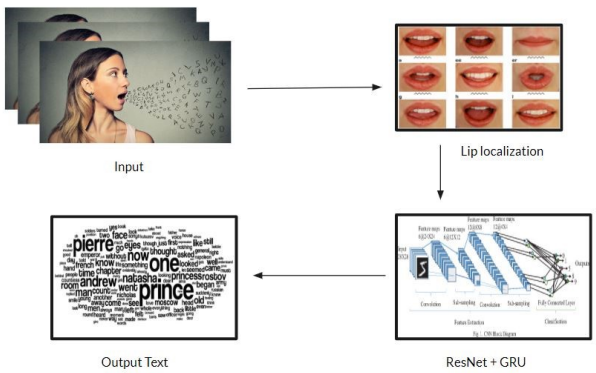

Fig. 5. Proposed System

IV. IMPLEMENTATION

A. Training

The model was trained for over 12 hours on a machine with the GTX 1080 graphics card. The model was fed with the pre-processed data from the training dataset which were 1000 videos per word. First, the encoder was trained on the input, then froze it, followed by the training of the decoder. After that, both the encoder and decoder are trained together. The model progressively was trained for over 30 epochs. It has been on 15 words, few which were very similar, for example 'Again' and 'Against', 'British' and 'Britain' and words that were completely different, for example, 'About' and 'Spending'. The results obtained were great with the training accuracy of over 99%. The validation set consisted of 200 videos out of which our model was able to accurately predict 90%.

B. Testing

The accuracy of the testing dataset was also 90%. These were the video inputs that the model had never seen before. A custom dataset consisting of 25 videos of proposed labels are developed. It was able to identify 22/25 videos perfectly. These videos even consisted of faces with thick facial hair. It has been noticed that the accuracy of the model for each video dropped sometimes even if it identified the label correctly. This was the consequence of videos with facial hair. It is also noticed that if the lips were not localized properly and consistently in each frame, the accuracy of the model dropped. This is a clear indication that there is room for improvement which can be done by adding a greater variety of videos to the dataset library. The comparison with the existing models can be seen in Table 1, where it is clear that the existing models such as LSTM-5, DenseNet with 3D convolution layers and (D3D) and DenseNet with 3D and 2D convolution layers [22,23] weren't able to handle the irregularities of the custom dataset that were mentioned before

TABLE I. RESULTS AND COMPARISON

Models	Dataset	Accuracy (in %)
Existing models		
LSTM-5	LRW	66
LSTM-5	Custom	25.76
D3D	LRW	78
D3D	Custom	34.76
3D + 2D	LRW	83
3D + 2D	Custom	38.19
Our Model		
ResNet with 3D Conv. Layers + GRU	LRW	90.00
	Custom	88.00

V. CONCLUSION

The objective of this work was to give an end-to-end approach to this problem of identifying the words spoken by someone using visual input alone. It has been tried to make a model that can be trained easily and is low on computations,

thus chose the ResNet model along with GRU networks which learned the spatial-temporal features and gave accurate labels. The proposed model has been deployed on a local server to test various video inputs. The model is further trained by extending the vocabulary. This may also be considered as a proof of concept. The accuracy of the model is over 90% on the BBC dataset and 88% on our custom dataset. It doesn't perform well with a video containing subjects with facial hair. Thus, more exploration can be done by adding videos to the dataset containing people with facial hair. The results were spot on for many unseen videos but the positive deviation i.e. by differentiating the similar words, shown was true.

ACKNOWLEDGMENT

We are extremely grateful to Dr. Varsha Patil for her valuable and constructive suggestions during the preliminary stages and her inputs throughout the development of this research study. We are grateful to her and also appreciate the time she spent nurturing and guiding us.

REFERENCES

[1] VSR: https://youtube.com/playlist?list=PLXkuFIFnXUAPIrXKgtIpctv2NuSo7xw3k

[2] G. Fisher. Confusions among visually perceived consonants. Journal of Speech, Language, and Hearing Research, 11(4):796–804, 1968.

[3] https://www.technologyreview.com/2014/09/11/12564/the-challenges-and-threats-of-automated-lip-reading/

[4] Multi-Modal Methods: Visual Speech Recognition (Lip Reading) https://medium.com/mlreview/multi-modal-methods-part-one-49361832bc7e

[5] K. He, X. Zhang, S. Ren, and J. Sun. Identity Mappings in Deep Residual Networks. arXiv preprint arXiv:1603.05027v3,2016.

[6] Illustrated Guide to LSTM's and GRU's: A step by step explanation https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21

[7] J. S. Chung and A. Zisserman, "Lip reading in the wild," in ACCV. Springer, 2016, pp. 87–103.

[8] Petridis, Stavros & Stafylakis, Themos & Ma, Pingchuan & Cai, Feipeng & Tzimiropoulos, Georgios & Pantic, Maja. (2018). End-to-End Audiovisual Speech Recognition. 10.1109/ICASSP.2018.8461326.

[9] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with lstms for lipreading," in Interspeech, 2017.

[10] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "Lipnet: Sentence-level lipreading," arXiv preprint arXiv:1611.01599, 2016.

[11] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in Proc. CVPR, 2017.

[12] Parekh, Dharin & Gupta, Ankitesh & Chhatpar, Shharrnam & Kumar, Anmol & Kulkarni, Manasi. (2018). Lip Reading Using Convolutional Auto Encoders as Feature Extractor.

[13] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, M. Pantic. A semi-automatic methodology for facial landmark annotation. Proceedings of IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR-W), 5th Workshop on Analysis and Modeling of Faces and Gestures (AMFG 2013). Oregon, USA, June 2013.

[14] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, M. Pantic. 300 Faces in the Wild Challenge: The first facial landmark localization Challenge. Proceedings of IEEE Int'l Conf. on Computer Vision (ICCV-W), 300 Faces in-the-Wild Challenge (300-W). Sydney, Australia.

[15] Simonyan, Karen & Zisserman, Andrew. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv 1409.1556.

[16] https://pypi.org/project/dlib/

[17] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, Jiebo Luo; The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4651-4659.

[18] He K., Zhang X., Ren S., Sun J. (2016) Identity Mappings in Deep Residual Networks. In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9908. Springer, Cham.

[19] M. Wand, J. Koutn´ık, and J. Schmidhuber, "Lipreading with long short-term memory," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 6115–6119.

[20] A. Graves, S. Fernandez, and J. Schmidhuber, "Bidirectional ´ LSTM networks for improved phoneme classification and recognition," in International Conference on Artificial Neural Networks. Springer, 2005, pp. 799–804.

[21] Chung, Junyoung et al. "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling." ArXiv abs/1412.3555 (2014): n. pag.

[22] S. Yang et al., "LRW-1000: A Naturally-Distributed Large-Scale Benchmark for Lip Reading in the Wild," 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), Lille, France, 2019, pp. 1-8, doi: 10.1109/FG.2019.8756582.

[23] Joon Son Chung and Andrew Zisserman. Learning to lip read words by watching videos. Computer Vision and Image Understanding, pages 1–10, 2018.