Lip Reading Using Neural Networks and Deep Learning

N.Durga Sri, R.Akhil, S. Venkat Durga Prasad, V.Jayanth

K.Krishna Jyothi (Guide)

Hyderabad Institute of Technology and Management

---***---

Abstract: Lip reading is a method for understanding words or speech without the use of audio through visual interpretation of face, mouth, and lip movement. This activity is challenging since different dictions and speech articulations are used by different people. Using deep learning and other techniques, this research validates the usage of machine learning. using neural networks to create a system for automatically reading lips. Two different CNN architectures were trained on a portion of the dataset. On the basis of their ability to accurately predict words, the trained lip reading models were assessed. A online application for real-time word prediction using the top performing model. Keywords: lip reading, computer vision, deep learning, convolutional neural network, web application, object detection.

1.INTRODUCTION

Even experienced lip readers have found lip reading to be troublesome in recent years. There is potential for lip reading to be solved utilizing several machine learning techniques. An important skill to have is the ability to read lips. The potential for improved speech recognition in noisy or loud circumstances grows as lip reading technology is improved. The advancement of hearing aid technologies for those with hearing problems would be a notable benefit. Similar to this, a lip reading system can be used for speech analysis for security reasons to identify and anticipate information from the speaker when the audio is damaged or not present in the video. Given the diversity of languages used worldwide, the variations in word diction and relative articulation, and phrases.

Developing a computer software that automatically and reliably translates spoken words based only on the speaker's visible lip movement becomes extremely difficult. Even the most skilled lip readers can only guess every other word. Consequently, two architectures were trained and assessed using the capabilities of neural networks and deep learning algorithms. The better-performing model was further tweaked to increase accuracy based on the evaluation. A web application that used the model architecture with overall greater accuracy was created to create a real-time lip-reading system.

2. IMPLEMENTATION

A common multi-layered network contains a class of neural network system known as the convolutional neural network (CNN). The layers are made up of one or more layers connected in a succession of multiple connections. The local connection of high dimensional datasets, such as those made up of photos and videos, can be used by CNN. CNN may be used for speech recognition and computer vision thanks to this characteristic. Convolutional layer, activation function, pooling layer, and fully connected layer are the four key components of a fundamental CNN model. The parameters of the convolutional layer are learned from the input data using a collection of learning filters. The activation function defines the output of one node, which is subsequently the input for the following node in a non-linear transformation. Neuronal layer. The number of parameters and computations in the network are decreased in the pooling layer to control over fitting by reducing the spatial size of the network. Convolutional or pooling layer input volume is used by fully connected layer to change the output of the feature learning section. Using models like Hidden Markov Models and Recurrent Neural Networks (RNN), which are less able to adapt to the motion of the image, this work analyses the use of temporal sequences. The trained models' predicted accuracy suffers as a result. CNN's suitability for use in moving subjects with more precision is up for discussion, though.

2.1 METHODOLOGY

There are various phases and factors to take into account while creating a neural network for lip reading. Here is a potential approach:

1.Data gathering: Compile a sizable database of videos of speakers speaking in a range of settings, including lighting, camera angles, and languages. To take into account changes in dialects, facial characteristics, and lip motions, it's critical to have a diverse dataset.

2.Pre-processing: Crop the areas near the mouth after extracting the frames from the videos and aligning them. Use methods like feature extraction, filtering, and normalizing to reduce noise and improve the pertinent visual cues.

3.Develop an architectural framework for a neural network that can predict spoken words or phonemes from preprocessed photos. The design may incorporate recurrent layers for temporal modeling, convolutional layers for feature extraction, and methods for focusing attention on the lip area.

4.Training: To increase the model's precision and generalizability, train the neural network using the preprocessed dataset and methods such as cross-validation, regularization, and optimization. Employ a loss function that is appropriate for measuring the gap between anticipated and ground-truth labels.

5.Evaluation: Using metrics like accuracy, precision, recall, and F1-score, assess the lip-reading model's performance on a different test set. Analyze the model's advantages and disadvantages by contrasting the findings with those obtained using current cutting-edge techniques. Choose a device or platform that can handle real-time video streams and output anticipated text or speech when deploying the lip-reading model. Optimize the model's size, speed, and energy use using methods like transfer learning, pruning, and quantization. techniques for drawing attention to the lips.

6.Training: Train the neural network using the pre-processed dataset and techniques like cross-validation, regularization, and optimization to improve the model's accuracy and generalizability. Use a loss function that is suitable for gauging the difference between predicted and actual label values.

2.2 EVALUATION

Evaluate the lip-reading model's performance on a different test set using metrics including accuracy, precision, recall, and F1-score. Compare the results with those attained using the most recent cutting-edge methods to evaluate the model's benefits and drawbacks.

When implementing the lip-reading model, pick a device or platform that can process real-time video streams and generate predicted text or speech. Use techniques like transfer learning, pruning, and quantization to optimize the model's size, speed, and energy consumption.

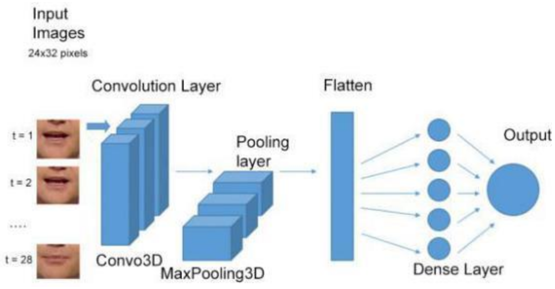
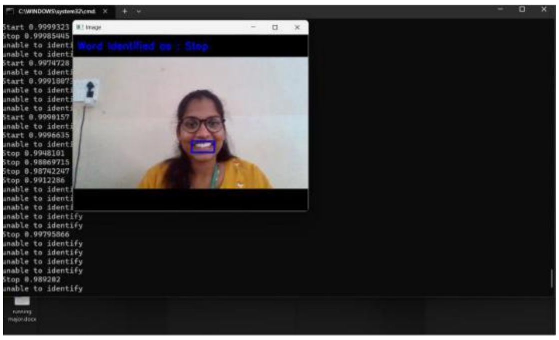


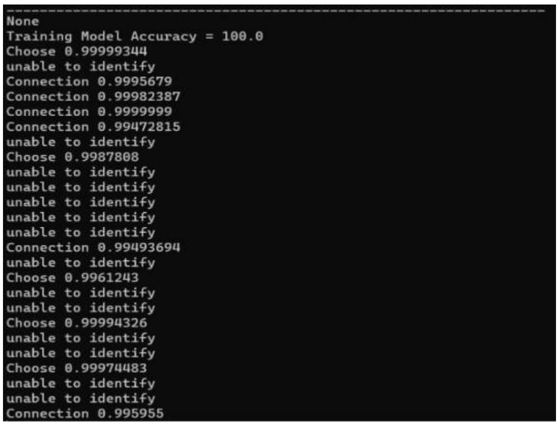
Fig:1 Model architecture



```
Start 0.9999323
Stop 0.99985445
unable to identi
unable to identi
Start 0.9974728
unable to identi
Start 0.9991887
unable to identi
unable to identi
unable to identi
Start 0.9990157
unable to identi
Start 0.9996635
unable to identi
Stop 0.9948101
Stop 0.98869715
Stop 0.98742247
Stop 0.9912286
unable to identi
unable to identi
unable to identi
unable to identify
unable to identify
Stop 0.99795866
unable to identify
unable to identify
unable to identify
unable to identify
Stop 0.989202
unable to identify
```

Word identified as : Stop

Fig:2 Front end web application



```
--------------------------------------------------------------------------------
None
Training Model Accuracy = 100.0
Choose 0.99999344
unable to identify
Connection 0.9995679
Connection 0.99982387
Connection 0.9999999
Connection 0.99472815
unable to identify
Choose 0.9987808
unable to identify
unable to identify
unable to identify
unable to identify
unable to identify
Connection 0.99493694
unable to identify
Choose 0.9961243
unable to identify
unable to identify
Choose 0.99994326
unable to identify
unable to identify
Choose 0.99974483
unable to identify
unable to identify
Connection 0.995955
```

Fig: 3 Frame Generation

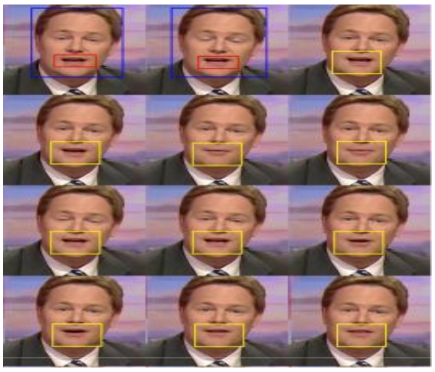


Fig:4 Haar-classification frame

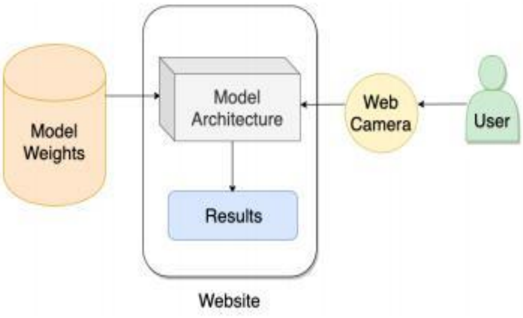


Fig:5 Design of web application

3. CONCLUSIONS

In order to develop a web application for automated real-time lip reading, this project investigated a number of tools and technologies, including object detection using the Haar Feature-Based Cascade classifier, convolutional neural networks, the Keras neural network library, and the Keras JavaScript library. Similar to how the EF-3 design underperformed when trained on the dataset, the lightweight model architecture performed better. Models A1, A2, B, C, and D were constructed on top of the lightweight model architecture after each model had been reviewed, customized, and improved. Model D had the highest accuracy and best performance among the taught architectures. In order to associate the model weights in the web application, this model was built.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to our mentor Mrs. K.Krishna Jyothi for providing us with the fantastic chance to work on this amazing project on this subject. This project also enabled us to do extensive research, which allowed us to learn

a lot of brand-new information. They have our sincere appreciation.

As a follow-up, we also want to express our gratitude to my friends, who greatly aided us in completing this job in the allotted time.

REFERENCES

[1] Martn Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, and others. 2016. A system for extensive machine learning is called Tensorflow. Operating Systems Design and Implementation (OSDI 16), 12th USENIX Conference, pages 265–283.

[2] Adrian Kaehler and Gary Bradski. O'Reilly Media, Inc., "Learning OpenCV: Computer vision with the OpenCV library."

[3] https://github.com/transcranial/keras-js L. Chen, 2016.

2015 Keras documentation by

[4] François Chollet is available at keras.io (2015).

[5] Joon Son Chung and Andrew Zisserman. 2016. Lip reading in the wild. In *Asian*

Conference on Computer Vision. Springer, 87–103.

[6] Dan Hammerstrom. 1993. Neural networks at work. *IEEE spectrum* 30, 6 (1993),

26–32.

[7] Ahmad BA Hassanat. 2011. Visual Speech Recognition, Speech and Language

Technologies, Prof. Ivo Ipsic (Ed.), ISBN: 978-953-307-322-4, InTech.

[8] Simon Haykin. 1994. *Neural networks: a comprehensive foundation*. Prentice Hall

PTR.

[9] Hynek Hermansky. 1990. Perceptual linear predictive (PLP) analysis of speech.

the Journal of the Acoustical Society of America 87, 4 (1990), 1738–1752.

[10] Sanghoon Hong, Byungseok Roh, Kye-Hyeon Kim, Yeongjae Cheon, and Minje

Park. 2016. Pvanet: Lightweight deep neural networks for real-time object

detection. *arXiv preprint arXiv:1611.08588* (2016).