IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⧖ Impact Factor 7.39 ⧖ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.11715

A Machine Learning Approach to Lip Reading Automation System

Ms. Indumathi J[1], Ms. M Mounica[2], Ms. Neha Shivananjaiah[3], Ms. Niha Tarannum A[4], Mr. Rajesh L[5]

[1-4]Information Science and engineering, SJB Institute of Technology, Bengaluru, India

[5]Assistant Professor, Information Science and engineering, SJB Institute of Technology, Bengaluru, India

Abstract: Artificial Intelligence is extensively used to detect the movement of lips. It is observed that there is a high Correlation between the visual motion of mouth and corresponding audio data. This fact has been utilized for lip reading and for improving speech recognition. A Convoluted Neural Network would detect the movement of lips and determine the words spoken. The words that are spoken in the video would be detected by the Trained CNN and displayed in the text format. The CNN relies on information provided by the context, knowledge of the language, and any residual hearing. The aim is to verify whether the use of artificial intelligence methods, namely Deep Neural Network, is a suitable candidate for solving this problem. Practically, the focus is on presenting the results in terms of the accuracy of the trained neural network on test data.

Keywords: Artificial Intelligence, Lip Reading, Deep Neural Network, CNN, Machine learning.

I. INTRODUCTION

Machine Learning is extensively used to detect the movement of lips. It has been observed that the data generated through visual motion of mouth and corresponding audio are highly correlated. This fact has been utilized for lip reading and for improving speech recognition. We propose a system that makes use of CNN (Convolutional Neural Network) which would be trained to detect the movement of lips and predict the words being spoken. This trained CNN will be able to detect the words that are spoken within the video and display it in a text format. We hope to learn whether the utilization of machine learning, more specifically the DNN (Deep Neural Network), could also be an appropriate choice for solving the problem of lip reading. The main aim of our project is to accurately recognize the phrases being spoken through automated lip reading.

Lip reading is also an extremely difficult task because several different words can be spoken with almost indistinguishable lip movements. Therefore, the problem of lip reading provides unique challenges. This has led to numerous advancements in the field of automated speech recognitions systems using machine learning. Several models have been developed to improve hearing aids, for silent dictation in noisy public environments, identification for security purposes etc. However not until the use of Deep Learning did the accuracy of these models increase. Deep

Learning and deep neural networks have revolutionized the quality of automated lip-reading systems due to the large amounts of data sets that can be used.

Audio speech recognition may be difficult in noisy environments. In such cases, recognition through Visual Lip-reading might play an important role. The art of lip reading has various applications, for example it can be used to help people with hearing disabilities, or possibly by security forces in situations where it is necessary to identify a person’s speech when the audio records are not available. However, like speech recognition through audio, visual lip-reading systems also face numerous challenges due to several differences in the inputs. These inputs can include skin color’s, speaking speeds, facial features and different accents. It is almost impossible to manually create a computer algorithm that will be reading completely accurately from the lips. Even human professionals in this field can correctly estimate nearly every other word and can do so only under ideal conditions. Therefore, the complex task of lip reading is suitable candidate for extensive research in the field of deep learning.

II. LITERATURE SURVEY

Michael Wand and et al. [1] contributed a worked on a lipreading system which yields an end-to-end trainable system which consumes an infinitesimal number of frames of un-transcribed target data to recondition the recognition accuracy on the target speaker by making use of domain-adversarial training for speaker independence. It is integrated into the lipreader’s advancement based on a stack of feedforward and LSTM (Long Short-Term Memory) recurrent neural networks. The main goal is to push the network to learn an intermediate data representation which is domain- agnostic i.e., it should be independent whether input data is obtained from target speaker or a source speaker. TensorFlow's Momentum Optimizer is applied using the stochastic gradient descent in order to minimize the multi-class cross- entropy hereby achieving optimization.

This paper [2] describes the study conducted by Sujatha and Krishnan, their paper made use of image- based detection to extract the lip region. The Advantage of this method is that the lip Region of Interest could be extracted without utilizing the geometric properties like corners and edge detection procedures. They managed to achieve high-level accuracy by localization of the lip ROI.

In this paper [3], They have shown through their work, the way to transform a raw video into a word sequence. The primary component of this method may be a processing pipeline used to create the Large-Scale Visual Speech Recognition (LSVSR) dataset utilized in this work, distilled from YouTube videos and comprising of phoneme sequences including video clips of faces speaking. Their approach was first to mix a deep learning-based phoneme recognition model with production-grade word-level decoding techniques. By decoupling phoneme prediction and word decoding as is usually wiped-out speech recognition, hence it's possible to arbitrarily extend the vocabulary without retraining the neural network.

In this paper [4], They have detailed the recent sequence-to-sequence (encoder-decoder with attention) translator architectures that have been developed for speech recognition and machine translation. In this paper the dataset developed is established from thousands of hours of BBC television broadcasts which have speaking faces along with subtitles of what is being said. Their model is devised in such a way that it can operate over dual attention mechanism that can operate over visual input only, audio input only, or both. They have an image encoder, audio encoder and character decoder in place to achieve what is called lipreading. With or without the audio the goal was to recognize the phrases spoken by the talking face.

A. Challenges

Some of the challenges include the following points
1. Usage of geometry-based techniques to determine the landmark points around the lips.
2. Usage of simple machine learning techniques such as K-means clustering and HMM led to poor accuracy.
3. Use of fuzzy clustering method led to improper lip shape cropping.
4. Single feature based Active Shape Models(ASM) may achieve good performance, however it was found to face problems in noisy conditions such as presence of beard, wrinkles, poor textures etc.

B. Proposed System

We propose a system that makes use of CNN (Convolutional Neural Network) which would be trained to detect the movement of lips and predict the words being spoken, a trained model to translate the video sample to a subtitled video, a new, faster and an efficient way for recognition of lip movement appearance and predicting the words or phrases spoken in English language based on the video data fed as an input and This system can be employed in various fields like forensics, film processing, aid to the deaf and dumb, security, etc.

III. METHODOLOGY

Our system contains the following modules.

Module 1 - Pre-Processing
The video is first divided into frames of images and converted into greyscale
- Input: Video in RGB color format
- Output: Set of frames in greyscale

Module 2 - Face detection and cropping
In this module, the system detects the full-frontal view of the subject using face detector and landmarks the face region and crops the ROI.
- Input: Still frames of grayscale images.
- Output: Faces with ROI cropped and saved as a NumPy array.

DOI: 10.17148/IJARCCE.2022.11715

Module 3 -Feature extraction and normalization
In this module, the spatio temporal features are extracted and the frames are normalized to provide uniformity and fed into the CNN

- Input: NumPy array of ROI images
- Output: Normalized frames ready with features extracted

Module 4 - Text classification and decoding
The words spoken in the video data sample is decoded and predicted by the CNN model and is displayed for the user.

- Input: Image data containing normalized frames
- Output: Predicted text played in parallel to the video

IV. System Architecture

Convolution Neural Network (CNN) usually contains three-layer types: Convolution (CONV), Pooling (POOL), and Fully Connected (FC) layers, where CONV and POOL layers are ordinarily repeated several times to from Deep NN and extracted high level feature. A Fully Connected layer is a normal Multi-Layer Perceptron that uses a soft max activation function in the output layer. The architecture of the proposed system is based on working of a CNN. The system makes use of an input layer, three hidden layers and an output layer. The hidden layers consist of 32, 64 and 96 neutrons in subsequent layers respectively. The features extracted from each initial layer is fed into the next layer. This way the number of parameters goes on decreasing as the neural network goes deeper. The whole architecture is then trained end to end by backpropagation where filters/weights are constantly updated and adjusted. The system is tested using both 3 hidden layer architecture yet because the 5 hidden layer architecture, but the 3-layer architecture is given more priority keeping in mind the computation problems for 5-layer architecture The representation of a CNN is shown in the Fig 1.

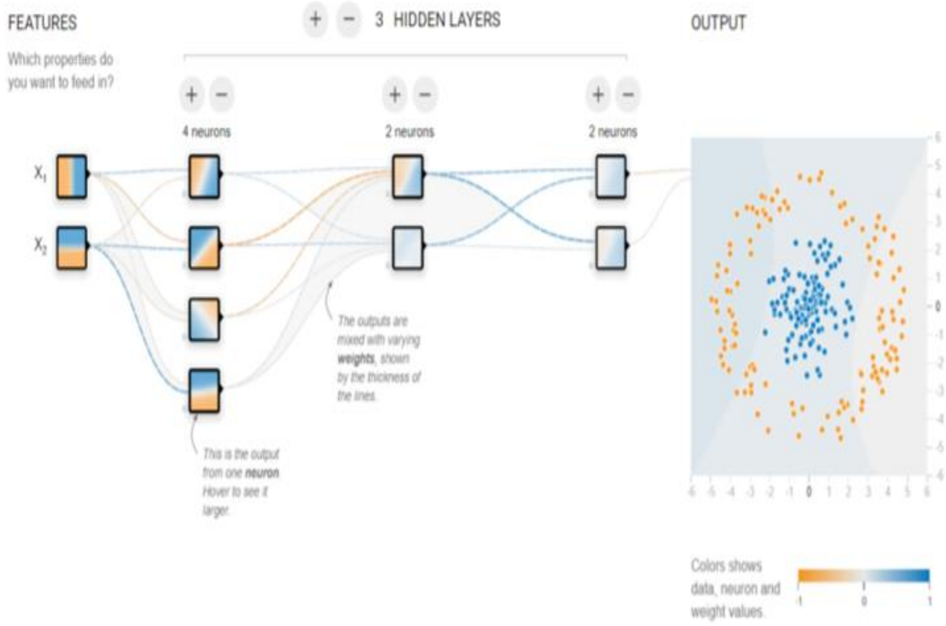

FEATURES
Which properties do you want to feed in?
+ − 3 HIDDEN LAYERS
OUTPUT
+ −
4 neurons
2 neurons
2 neurons
X_1
X_2
The outputs are mixed with varying weights, shown by the thickness of the lines
This is the output from one neuron Hover to see it larger
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6
Colors shows data, neuron and weight values
-1
0
1

Fig 1: Architecture

V. Algorithms Used

A. Viola Jones Algorithm for Face Detection

The Viola-Jones algorithm is a widely used algorithm for detection of objects. The training is usually slow but the detection is fast in this algorithm. The algorithm makes use of the Haar basic feature filters.

DOI: 10.17148/IJARCCE.2022.11715

$$II(y,x) = \sum_{p=0}^{y} \square \sum_{q=0}^{x} Y(p,q)$$

The efficiency of this algorithm can be increased by generation of the integral image. The haar extraction can be done using the integral image integration by adding the four numbers.

The actual detection takes place inside what is called as a detection window. Three parameters are defined, namely: minimum window, maximum window, and a sliding step. Detection happens inside a detection window. The detection window is then moved by the sliding step defined.

- Fix the parameters mentioned
- For the selected window size, the window is moved horizontally and vertically with each step. A collection of face recognition filters is applied on every step. The filter returns a positive value when a face is detected within the current window.
- The procedure is stopped when the window is equal to the maximum window size defined. Else, the window dimensions and step size are increased and the procedure from step 2 is repeated.

Every face recognition filter will have a collection of cascades that is connected to classifiers. Each classifier looks at an oblong subset of the detection window and determines if it's sort of a face. If it does, the following classifier is applied. If all classifiers provide a positive answer, the filter gives a positive answer and also the face is recognized. Otherwise, the filter within the set of N filters is made to run. Every classifier has Haar feature extractors that is weak classifiers. The Weighted sum of 2-D integrals of small rectangular areas attached to each other is the Haar feature. The weights may take values ±1. Fig.2 shows examples of Haar features relative to the enclosing detection window. Gray areas consist of a positive weight and white areas consist of negative weight. Based on the detection size of a window Haar feature extractors are scaled.

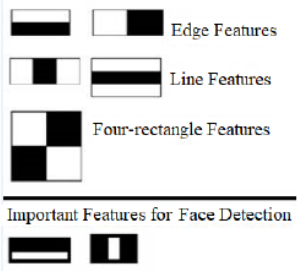


Fig 2: In the example the rectangular feature show the window enclosing detection

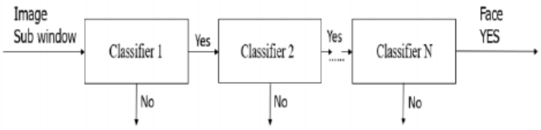


Fig 3: Viola-Jones filter used for Object

The classifiers with the fewest features are placed in the beginning of the cascade which makes the architecture of a cascade efficient. AdaBoost is most used algorithm for training the features. This algorithm has four stages:

1. Haar Feature Selection
2. Creating an Integral Image
3. Adaboost Training
4. Cascading Classifiers

The features that are matched by this algorithm are rectangle features. Rectangle features are the Value is equal to the difference in the sum of pixels in black area and the sum of pixels in white area. 3 types: 2-, 3-, 4-rectangles, two rectangle features was used by Viola & Jones. There is a relation between each feature and a special location in the Sub-window.

B. Convolution Neural Network

A convolutional neural network (CNN, or ConvNet) in deep learning is a class of deep neural networks which is most frequently applied to analyze visual imagery. They are standard versions of multilayer perceptron. A Multilayer perceptron consists of fully connected networks, that is, every neuron in one layer is connected to all the neurons in the next layer. Overfitting of the data might occur due to the "fully-connectedness" nature of these networks General ways of regularization is done by adding magnitude measurement of weights to the loss function.

However, the regularization approach in CNNs is different: they take make use of the hierarchical pattern in data. By using smaller and simpler patterns it assembles even more complex patterns. As a result, the scale of connectedness and complexity of CNNs are on the lower extreme.

Feature Layers: Operations of the three layers are convolution, pooling, or rectified linear unit (ReLU). Convolution: In this layer a set of convolutional filters are put through the input images, which will activate specific features from the images. Pooling layer will perform nonlinear down sampling to simplify the output. This helps the network by reducing the number of parameters it needs to learn about. Rectified linear unit (ReLU) layer will map negative values to zero and maintains positive values which enables faster and more effective training These three operations are repeated over tens and many more layers, in which each layer learns to detect the different features.

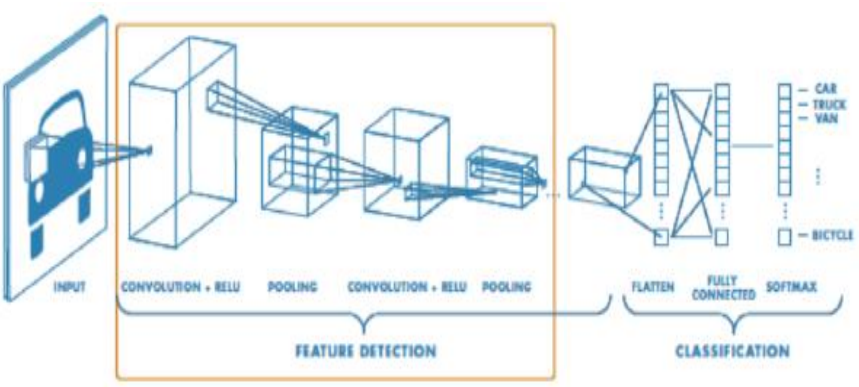


Fig 4: CNN Example

Classification Layers: The architecture of a CNN shifts to classification after feature detection. Fully connected layer (FC) is next to-last layer. FC layer will output a vector of K dimensions, K is the number of classes that the network can predict. The probabilities for each class of any image being classified contains within the vector. SoftMax function is used by the last layer of the CNN architecture to provide the classification output.

VI. Results and Discussions

The system has been trained with the dataset called GRID CORPUS. The system shows variable accuracy between 70-80 % on the test dataset.

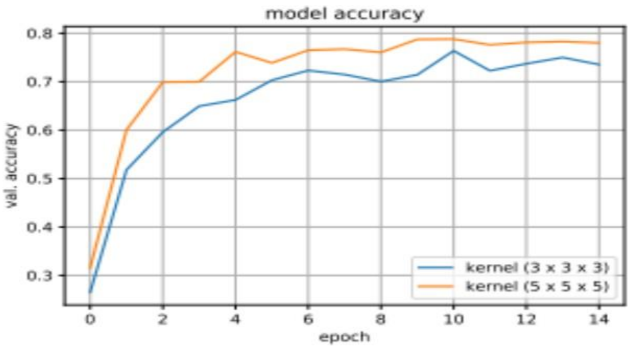
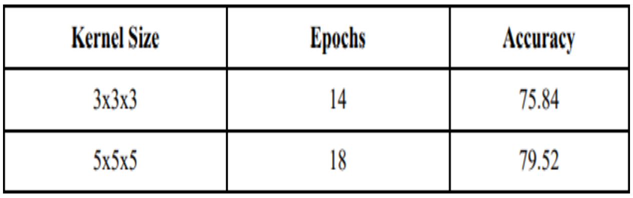


Fig 5: Showing model accuracy

It is evident from the above figure that while increasing the kernel size of CNN from 3X3X3 to 5X5X5 the accuracy increases significantly subject to number of epochs. The Accuracy achieved is depicted in the above figure while comparing the kernel sizes.



Kernel Size	Epochs	Accuracy
3x3x3	14	75.84
5x5x5	18	79.52

Fig 3: Kernel Size and Accuracy

CONCLUSION

Our proposed system is uses artificial intelligence to predict the text from a video sample. It can be confidently said that with minimal requirements the model designed reaches an accuracy oddly close to two-fold higher than that of a human lip reader. Proposed Lip Reading Automation System translate the silent video sample to a subtitled video. Employing a trained CNN, the accuracy fluctuates between 70% to 80% and it supported different video samples. This system can be used in several fields like forensic sciences, photographic processing, as an aid to the deaf and dumb, and in many more applications.

REFERENCES

[1] Michael Wand and Jurgen Schmidhuber. Improving Speaker Independent Lipreading with Domain Adversarial Training, August 2017

[2] P. Sujatha and M. R. Krishnan., Lip feature extraction for visual speech recognition using Hidden Markov Model, April 2012

[3] Brendan Shillingford, Yannis Assael, Matthew W. Hoffman, Thomas Paine, Cían Hughes, Utsav Prabhu, Hank Liao, Hasim Sak, Kanishka Rao, Lorrayne Bennett, Marie Mulville, Ben Coppin, Ben Laurie, Andrew Senior and Nando de Freitas, Large-scale visual speech recognition, 2018

[4] Joon Son Chung, Andrew Senior, Oriol Vinyals and Andrew Zisserman, Lip Reading Sentences in the Wild, 2016

[5] Brendan Shillingford , Yannis Assael, Matthew W. Hoffman, Large-Scale Visual Speech Recognition, 2018.

[6] Gregory J. Wolff, K. Venkatesh Prasad, Lipreading by neural networks: Visual preprocessing, learning and sensory integration, 1993

[7] Yuanyao Lu and Hongbo Li, Automatic Lip-Reading System Based on Deep Convolutional Neural Network and Attention-Based Long Short-Term Memory, 2019.

[8] Lele Chen, Zhiheng Li, Lip Movements Generation at a Glance, 2018

[9] N. Rathee, International Conference on Computing, Communication and Automation (ICCCA), Investigating back propagation neural network for lip reading, 2018.

[10] A. H. Kulkarni and D. Kirange, Artificial Intelligence: A Survey on Lip-Reading Techniques, 10th International Conference on Computing, Communication and Networking Technologies, 2019

[11] T. Saitoh and R. Konishi, Profile Lip Reading for Vowel and Word Recognition, 20th International Conference on Pattern Recognition, 2020, pp. 1356-1359, doi: 10.1109/ICPR.2010.335, 2020.

[12] S. Fenghour, D. Chen, K. Guo and P. Xiao, Lip Reading Sentences Using Deep Learning With Only Visual Cues , in IEEE Access, vol. 8, pp. 215516-215530, 2020, doi: 10.1109/ACCESS.2020.3040906, 2020.

[13] M. H. Rahmani and F. Almasganj, Lip-reading via a DNN-HMM hybrid system using combination of the image-based and model-based features, 3rd International Conference on Pattern Recognition and Image Analysis (IPRIA), 2017, pp. 195-199, doi: 10.1109/PRIA.2017.7983045, 2017.

[14] S. NadeemHashmi, H. Gupta, D. Mittal, K. Kumar, A. Nanda and S. Gupta, A Lip Reading Model Using CNN with Batch Normalization, Eleventh International Conference on Contemporary Computing (IC3), 2018, pp. 1-6, doi: 10.1109/IC3.2018.8530509, 2018.

[15] Lip Reading Sentences Using Deep Learning With Only Visual Cues, Souheil Fenghour, Daqing Chen, Kun Guo, Perry Xiao, 2020.

[16] Convolutional Neural Networks for Predicting Words: A Lip-Reading System, PV Sindhura, S J Preethi, Krupa B Niranjana, 2018.