International Journal of Computer Trends and Technology
ISSN: 2231-2803 / https://doi.org/10.14445/22312803/IJCTT-V67I4P117
Volume 67 Issue 4, 76-78, April 2019

Original Article

An Improvised Word Recognition System using CNN in a Non-Isolated Environment

Neethu Mohan[1], Arul V H[2]

[1,2] *Electronics and Communication Engineering, APJ Abdul Kalam Technological University*
Thejus Engineering College, Kerala, India

Abstract *- This paper mainly focuses on developing a word recognition system using the CNN structure. Several advancements have been made in the Automatic Speech Recognition (ASR) technology that enables the machine to understand the natural language. The main constrain rise is the nature of the input speech signal, which makes it difficult to retain the original information. The noisy speech signal is initially passed through the pre-processing stage and converted to the spectrogram to extract the feature. To extract the features, these spectrograms are fed to CNN's layers to feature extract and then train the model. The vectors are now cross-matched at the testing phase, and the maximum close weighted value from the fully connected layers leads to the output. The system performs with an efficiency of 88.20% in a non-isolated environment.*

Keywords *- ASR,CNN, spectrogram*

I. INTRODUCTION

Studies on speech recognition and its processing have been carried out for more than five decades. The goal of an automatic speech recognition system is the transcription of human speech into spoken words[1].

Commonly used Automatic Speech Recognition(ASR) features still rely on the spectral envelope as prime property for classifying different linguistic characteristics of spoken words. The spectral envelope of a speech signal is very sensitive to distortions such as additive and convolutional noise[2].

DNN- based acoustic models have been shown by many groups to outperform the conventional Gaussian mixture model(GMM) on many ASR tasks. Recently several sites have reported some successful results using a deep convolutional neural network instead of standard fully connected DNN's[3].

DNN improves speech recognition performance over the conventional Gaussian mixture model due to its ability to model complex correlations in speech features. But CNN outperforms this by allowing further error rate reduction. CNN can be regarded as a variant of a standard neural network. Instead of using fully connected hidden layers, the CNN introduces a special network structure that consists of convolution, pooling, and a fully connected layer[4]. Neurons in a fully connected layer have full connections to all activations in the previous layer as seen in regular neural networks and work similarly [9]. Neural networks are adaptive, which means that they modify themselves as they learn from their initial training phase. CNN is a type of artificial neural network used in speech recognition and processing specifically designed to process pixel data[4].

Moreover, supervised learning is fairly common in classification problems because the goal is often to get the computer to learn a classification system that we have created. Compared to a number of classification algorithms like linear classifiers, logistic regression, naive Bayesian networks, multi-layer perceptron, support vector machines, K- means, decision trees, etc., a convolutional neural network is seen to perform better in terms of its classification accuracy, speed of learning, speed of classification, tolerance to missing values, tolerance to irrelevant attributes, tolerance to noise, etc. [5].

Moreover, CNN merges the two stages of feature extraction and classification. The first part of the network uses 2-D convolution to extract features, while the second part classifies these features. Then the whole thing is trained to have good performance when classifying. So the objective of this paper is to introduce a system that does word recognition using CNN in a non-isolated environment[6].

The following sections are arranged: section 1 includes model architecture, section 2 includes results and discussions, and section 3 concludes the paper.

II. MODEL ARCHITECTURE

CNN provides a better method of using the information in the training set to build multiple layers of non-linear feature detectors[7].

A. *Database*

The database (set 1) used in the proposed system consists of sound waves of different pitch and frequency of people in the age group 25 to 40 who speaks the word "stop" and "tree." This database is taken from Tensor Dataflow. Seven hundred samples of word tree and 500 samples of word stop are taken for testing and training. The sound waves are recorded in a non-isolated environment. The database (set 2) consists of the same sound waves ("stop" and "tree") recorded in real-time. Five hundred samples of word tree and 300 samples of word stop are recorded in real-time in a non-isolated environment. Each sound wave is then converted to its corresponding spectrogram image and stored as greyscale images in the database. In the pre-processing stages of word recognition, the noise is removed by framing and windowing, hence smoothening the spectrogram. These spectrograms are then fed into the CNN.

B. *Convolutional Neural Network*

A convolutional neural network can be regarded as a variant of a standard neural network. In using CNN for pattern recognition, the input data need to be organized as a number of feature maps to be fed into CNN. The database consists of audio waves of the word "stop" and "tree" taken from Tensor dataflow. These audio waves are then converted into their corresponding spectrogram, i.e., frequency domain. This is done since CNN can process visual stimuli, i.e., images. Now the spectrogram K is the input to CNN.

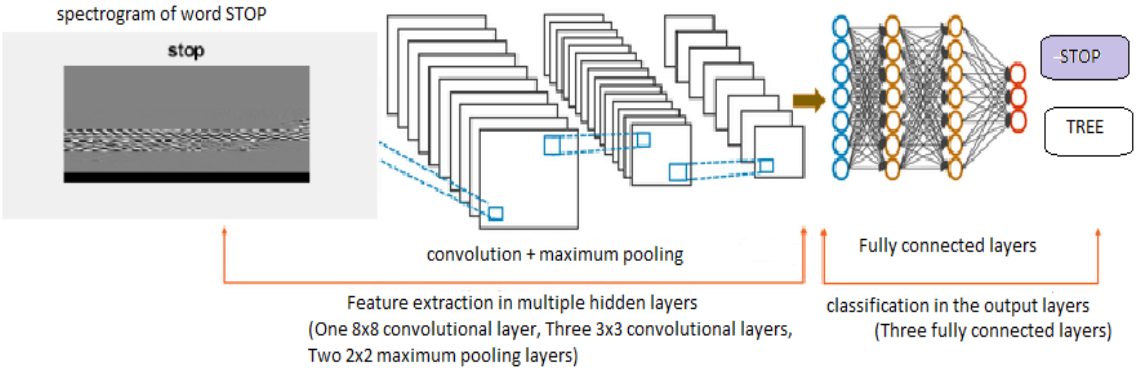

Fig. 1 CNN Architecture

Suppose K is the total number of feature maps. Each hidden layer of CNN has a number of units, each of which takes all outputs of the lower layer as input, multiplies them by a weight vector, sums the result, and passes it through a non-linear activation function such as sigmoid or tanh as given below[4]:

$$O_i^{(l)} = \sigma\left(\sum_j O_j^{(l-1)} w_{j,i}^{(l)} + w_{o,i}^{(l)}\right) \quad (1)$$

$O_i^{(l)}$ denotes the output of the i-th unit in the *l*-th layer. $w_{j,i}^{(l)}$ denotes connecting weights from the j-th unit in layer *l*-1 to the i-th unit in the *l*-th layer. $w_{o,i}^{(l)}$ is a bias added to the i-th unit and σ (x) is a non-linear activation function[1].

O_i for (i = 1,2,3……K), is connected to many feature maps; say J number of feature maps. O_i is also connected to $w_{i,j}$ (i = 1,2,3….K and j = 1,2,3….J). Each unit of one feature map in the convolution layer can be computed as;

$$q_{j,m} = \sigma\left(\sum_{i=1}^{K}\sum_{n=1}^{F} O_{i,n+m-1} w_{i,j,n} + w_{o,j}\right) \quad (2)$$

(j = 1,2,3……J)[1].

The purpose of the pooling layer is to reduce the resolution of feature maps. The pooling function is independently applied to each convolution feature map [1]. In the proposed system, maximum pooling is done. The max-pooling is done as

$$P_{j,m} = max_{n=1}^{G} q_{i,(m-1)*s+n} \quad (3)$$

Here G is the pooling size, and S is the stride size(overlapping adjacent pooling windows)[4].

For classification using CNN, the test and train data are first prepared, and then the design of CNN layers is done. In the approach proposed, four convolutional layers, two pooling layers, and three fully-connected layers are constructed. Figure 1 represents the CNN architecture proposed in this paper.

II. RESULTS AND DISCUSSIONS

The spectrograms of audio waves "stop" and "tree" are greyscaling images with pixel values ranging between 0- 255. These feature maps are again stored as greyscale images of size 200x88. This is first given to the convolutional layer of CNN having a

convolving filter of size 8x8 and 16 layers. This is to capture major changes in images. Further, 3 more convolutional layers are patterned, each of which has a 3x3 filter with 32 layers, 64 layers, and 128 layers, respectively. Between the four convolutional layers, two pooling layers of size 2x2 are patterned for maximum pooling. Three fully connected layers (200,100, and 2, respectively) are designed to provide information about convolutional layers and pooling layers. The third fully connected layer gives the classified output. The accuracy of iteration is obtained as 88.20%.

Table 2 shows the classification accuracy of different recognition modules in speech recognition using CNN. Hence, the result shows that the accuracy is not more than 83.9% for different recognition modules. The system proposed in this paper gives a classification accuracy of 88.20%, which is comparatively better than other experimental results.

Table 1. Iteration results

Validation Accuracy	88.20%
Training finish mode	Manual
Training cycle epoch	42 of 1000
Iterations per epoch	251
Iterations	10480 of 251000
Maximum iterations	251000
Validation frequency	20 iterations
Time elapsed till the present iteration	527 minutes 55 seconds

Table 2. Classification accuracy of different recognition modules[9]

Recognition Module	C180	C360	C720
Classification Accuracy	82.2%	83.4%	83.9%

III. CONCLUSION

Vast advances have been made in ASR technology, and its crucial objective is to make the machine understand natural language. CNN is a deep learning technology that reduces the pre-processing requirements by merging two signal processing stages, i.e., feature extraction and classification. The system proposed in this paper uses these advantages of CNN and iterates with a validation accuracy of 88.20%.

REFERENCES

[1] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu, "Convolutional Neural Networks for Speech Recognition," IEEE/ACM transactions on audio, speech, and language processing, vol. 22, no. 10, October 2014.

[2] Niko Moritz, JörnAnemüller, Birger Kollmeier; "Amplitude Modulation Spectrogram Based Features for Robust Speech Recognition in Noisy and Reverberant Environments," Conference Paper in Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on May 2011.

Jui-Ting Huang, Jinyu Li, and Yifan Gong, "An Analysis of Convolutional Neural Networks for Speech Recognition," Microsoft Corporation, One Microsoft Way, Redmond, WA 98052.

[3] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu, "Convolutional Neural Networks for Speech Recognition," IEEE/ACM transactions on audio, speech, and language processing, vol. 22, no. 10, October 2014.

[4] Osisanwo F.Y, Akinsola J.E.T, Awodele O, Hinmikaiye J.O, Olakanmi O, Akinjobi J, "Supervised Machine Learning Algorithms: Classification and Comparison," International Journal of Computer Trends and Technology (IJCTT) – Volume 48 Number 3 June 2017.

[5] https://www.quora.com/How-is-convolutional-neural-network-algorithm-better-as-compared-to-other-imageclassification-algorithms.

[6] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, NavdeepJaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Brian Kingsbury, Tara Sainath, "Deep Neural Networks for Acoustic Modeling in Speech Recognition**,"** IEEE Signal Processing Magazine | November 2012, Vol 29: pp. 82-97.

[7] Harsh Pokarana, "Explanation of Convolutional Neural Network," IIT Kanpur.

[8] Tao Wang, David J. Wu, Adam Coates Andrew Y. Ng; "End-to-End Text Recognition with Convolutional Neural Networks"; Stanford University, 353 Serra Mall, Stanford, CA 94305