VOL 4 (2020) NO 2

e-ISSN : 2549-9904
ISSN : 2549-9610

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

Speech Command Recognition using Artificial Neural Networks

Sushan Poudel[#], Dr. R Anuradha[#]

[#] Department of Computer Science and Engineering, Sri Ramakrishna Engineering College, NGGO Colony,Vattamalaipalayam,
Coimbatore, India
E-mail: sushanpoudel57@gmail.com, anuradha.r@srec.ac.in

***Abstract*— Speech is one of the most effective way for human and machine to interact. This project aims to build Speech Command Recognition System that is capable of predicting the predefined speech commands. Dataset provided by Google's TensorFlow and AIY teams is used to implement different Neural Network models which include Convolutional Neural Network and Recurrent Neural Network combined with Convolutional Neural Network. The combination of Convolutional and Recurrent Neural Network outperforms Convolutional Neural Network alone by 8% and achieved 96.66% accuracy for 20 labels.**

***Keywords*— Speech Command Recognition, Recurrent Neural Network (RNN), Convolutional Neural Network (CNN).**

I. Introduction

Due to the rapid development of the mobile devices, interacting with the machines using the speech is becoming increasingly popular and effective. Apple's Siri and Google's Google Now are very popular speech interface to the mobile devices. This project aims to build a speech command recognition system that is capable of recognizing predefince speech commands. The required dataset is provided by the Google's TensorFlow and AIY teams, it has 65,000 WAVE speech files of each lasting 1 second. Convolutional Neural Network and Recurrent Neural Network model are found to be very effective for this application.

The Deep Neural Network (DNN) demonstrated great improvement in speech feature extraction and recognition Integration of convolution and serial pooling into a neural network gives escalation to Convolution Neural Networks (CNN). Deep CNN can be formed by stacking up a CNN with a fully connected DNN or to one or more CNN, where it an perforom a robust success for images and speech recognition [1].

Recurrent Neural Networks (RNN) are the neural network type with recurrent connections. Data with varying size have been researched widwly and used for various time series analysis and mostly used for automatic speech recognition [2].

II. literature Review

Several studies have been investigated in the literature to implement speech command recognition system.

In [3], the authors proposed a low-latency speech command recognition system that was efficient to detect predefined keywords. They includes methods of Vanilla Single layer SoftMax where it can't produce the expected results, but can work fast. To prevent overfitting in DNN method, the model showed accurate results at the cost of more memory footprint and higher computational cost. CNN model outperforms the outher two models and achieves relatively very low false positive rate which is the desirable property.

In [4], a recurrent end-to-end neural network classifier for digit recognition that user Long Short-Term Memory (LSTM) to manage length of the speech utterances. The extracted feature are enclosed as a fixed size vector by a RNN and the resultant vector was given to a multilayer preceptor to classify the word spoken. The results show that the feature of MFCC was efficient to characterize a speech signal.

In [5], paper have been made for Bangla speech recognition. The short speech command of Bangla was taken as samples and three different types of CNN was designed to recognize. For dataset, 10 different words of utterances in real life noisy conditions are considered. Experimental results showed that the model using MFCC showed better accuracy but predicting the raw audio was complex. Finally, this system was able to identify single syllable word with excellence in results but found difficulties in recognizing multi-syllable commands.

In [6], the neural network-based speech recognition using the Mel-frequency cestrum co-efficient two speakers the network is trained in such a way that it can identify only one particular person with command and terminate the program of another. This method works significantly when a particular test command is given, it first compares the test data with training data and if it matches, the network decide to recognize the particular speaker.

In [7], the method Very Deep Convolutional Neural Networks (VDCNN) architecture for robust speech recognition has been implemented. This proposed model has the layers of convolutional networks without any fully connected layers. The VDCNN systems have been shown to improve the recognition accuracy compared to CNN. When same has been experimented with MGB-3, the results showed the best results was with VDCNN based method**.**

III. Methodology

A. *Convolution Neural Networks*

Convolutional architecture is implemented with two key layers as: Convolutional layer (for feature extraction), pooling layer(dimensionality reduction of input data for over fitting reduction) and a Fully-connected layer (for final softmax prediction). Input log spectrogram of the speech command is passed through convolution layer for feature extraction and then to pooling layer for dimensionality reduction and finally to the fully-connected softmax for the prediction. Fig. 1. Illustrates the structure of the implemented CNN architecture.

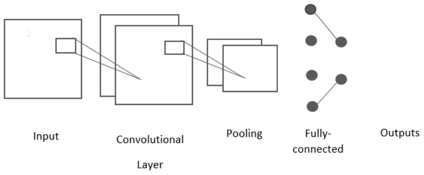

Fig. 1. Structure of CNN Architecture

B. *Convolution and Recurrent Neural Networks combined (CNN-RNN)*

This network is implemented with combination of convolution layer followed by pooling layer and RNN and finally fully-connected layer for softmax prediction. Input audio command is translated into log spectrogram, which is then given as input to the convolutional layer for feature extraction and then to pooling layer for dimensionality reduction. The input data after the pooling operation has small size with important feature intact. The output of pooling layer is then given input to the RNN layer and finally to fully-connected layer for softmax prediction.

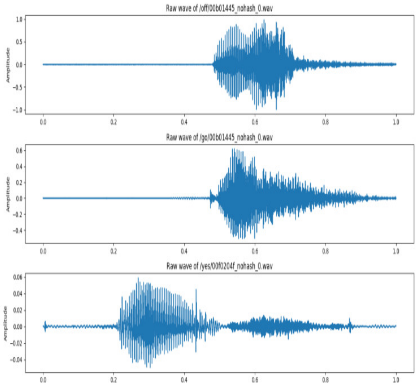

Fig. 2. Amplitude-time representation of input audio signal

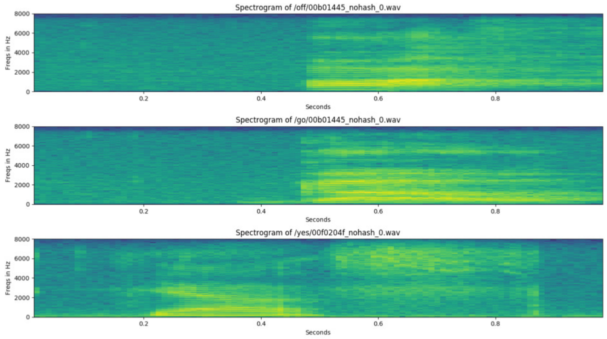

Fig. 3. Log-spectrogram of input audio signal

In RNN, unlike in feed forward neural network, network has the memory and its decision in the present is influenced by what it has learnt in the past. The network is called recurrent because they perform the same task repeatedly for all input sequence. The formula for the current state can be written as $h(t) = f(h(t-1) + x(t))$, where $h(t)$ is the current state, $h(t-1)$ is the previous state and $x(t)$ is current input.

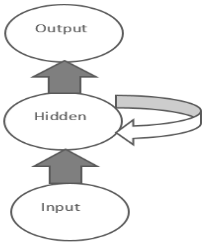

Fig. 4. RNN diagram

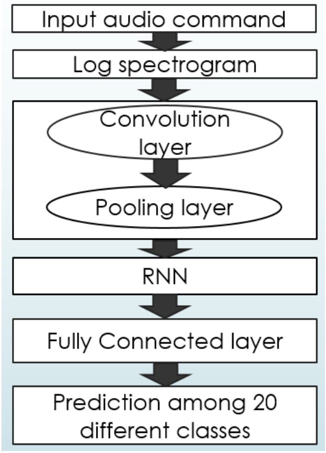

Fig 5. CNN and RNN combined network

IV. Dataset

Speech Commands Dataset is provided by the Google's TensorFlow and AIY teams, which consists of 65,000 WAVE audio files of people's speech of thirty different words of each one second in length. The data set is divided into 80% training set, 10% validation set and 10% test set, and each subset of speech audio is classified as either silence, unknown word, or predefined keywords, which are attached different labels respectively.

V. Experiments And Results

A. *Training Details*

Training Environment

Python 3.7 is used as a programming language and google Tensorflow 2.0.0 and keras as a framework. CPU with clock speed of 3.3 GHZ was used.

Initialization

Weights are initialized to very close to zero but not exactly zero because with proper data normalization it is reasonable to assure that approximately half of the weights will be positive and half negative.

Batch size

Batch size is a trade-off between performance and hardware imitation. If too small, might result in over fitting and if too large, it becomes computationally demanding. So the batch size of 32 is used.

Learning rate

Learning rate of 0.001 if found to be ideal for our model until epoch 11 and after that it is reduced to 0.0001.

Regularization

Dropout method is used as regularization techinique which reduce over fitting in neural networks.

B. *Results and Discussion*

Experimental result showed that among models trained on 20 different speech commands, CNN in combination with the RNN achieved 94.79% validation accuracy, 96.66 test accuracy and 0.117 loss whereas, CNN alone model achieved 85% validation accuracy, 88.44% test accuracy and 0.152 loss. Combined model was able to improve the accuracy of prediction by 8% on test set. Even though training time is slightly increased in the combined model, good improvement over the prediction accuracy compensates for it.

Model	Validation Accuracy	Test Accuracy	Loss
CNN	85.27%	88.44%	0.152
CNN-RNN	94.79%	96.66%	0.117

VI. Conclusion

In this project, we sued the CNN alone and CNN in combination with the RNN as our models. Experiment result shows that, CNN in combination with the RNN outperforms the CNN alone model by achieving improvement of 8% on prediction accuracy. CNN alone model achieves 85% accuracy whereas CNN in combination with the RNN produces 96.66% accuracy. From this experiment we can concolud that CNN and RNN combined model is better implementation for speech command recognition than CNN model alone

References

[1] L.Deng, O. Abdel-Hamid, and D. Yu, “A Deep Convolutional Nerual Network using Hetreogenous Pooling for Trading Acoustic Invaraince with Phonetic Confusion,” in Proc.IEEE Int. Conf. Acous., Speech, Signal Process.(ICASSP), pp. 6669-6673,2013.

[2] Arpita Gupta and Akshay Joshi, “Speech Recognition using Artifical Neural Network”, pp. 0068–0071, April 2018.

[3] Xuejiao Li, Zixuan, “Speech Command Recognition with Convollutional Neural Network”.

[4] Naima Zerari, Samir Abdelhamid, Hassen Bouzgou, Christian Raymond, “Bi-directional Recurrent End-to-End Neural Network Classifier for Spoken Arab Digit,” in IEEE, 2018.

[5] Shakil Ahmed Sumon, Joydip Cowdhury, Sujit Debnath, Nabeel Mohammed, “Bangla Short Speech Commands Recognition Using Convolutional Neural Networks”, in ICBLASP, September 2018.

[6] R. Nicole, “Neural Network based Recognition of Speech using MFCC Features”, IEEE, 2014.

[7] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Simplyfying very Deep Convolutional Nerual Network Architectures for Robust Speech Recogition”, IEEE, pp. 236–243.