Spoken Digit Classification using Deep Learning Algorithms

K. Vaishnavi[1], G. SudhaSadasivam[2]

Department of Computer Science and Engineering, PSG College of Technology, Anna University, Coimbatore, India

E-mail: [1]kvs.cse@psgtech.ac.in, [2]gss.cse@psgtech.ac.in

Abstract

The deep learning technique uses speech recognition in many different applications, including voice assistants, voice authentication, audio transcriptions, etc. Children who are dyslexic, blind persons and those with impairments can all benefit from spoken digit recognition. The goal of this paper is to create spoken digit recognition for the categorization of digits from 0 to 9 utilizing Convolution Neural Networks (CNN) and Long Short -Term Memory neural networks. With the addition of autoencoders, the performance of the CNN model is assessed. Finally, a comparative analysis is performed on the performances of the models based on the performance metrics.

Keywords: Autoencoders, Convolution Neural Network, Deep learning, Long Short -Term Memory Neural Network, Spoken digit recognition.

1. Introduction

The simplest and most effective form of interpersonal communication is speech. Human mind is expressed through speech in all languages. The importance of speech technology research has increased recently for purposes such as accent and dialect recognition, mood and stress detection, and more [5]. In the field of speech recognition, which also encompasses various speech processing technologies, artificial intelligence, big data analysis, and deep learning are frequently used. Simple vocal commands could be used to share information in air traffic control, make phone calls, choose music from players, and control robots. Additionally, they are employed in the medical sector to do front-end speech recognition and to give speech therapy to children who are speech impaired. The speech recognition system is included to help the blind, the disabled, and pupils who have writing

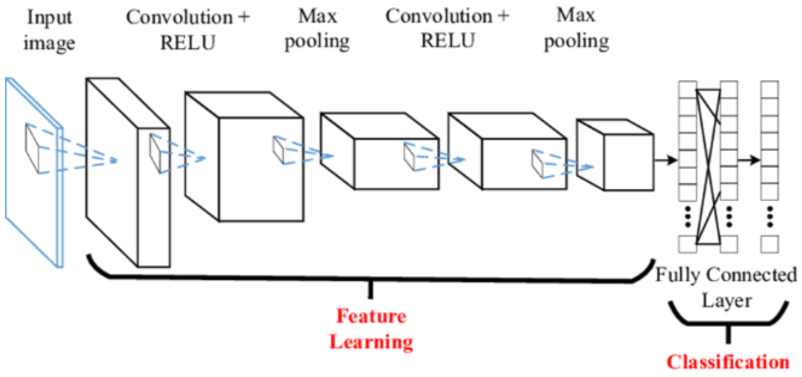
DOI: https://doi.org/10.36548/jucct.2022.4.005

Received: 02.12.2022, received in revised form: 06.01.2023, accepted: 20.01.2023, published: 30.01.2023

difficulties. Several methods for speech recognition utilizing machine learning and deep learning algorithms have been proposed by researchers. In spoken digit recognition, the model is trained on particular spoken data before being put to the test to determine whether it can recognize inputs in real-time.

1.1 Convolution Neural Networks

In deep learning, a Convolution Neural Network is a particular kind of Artificial Neural Network that is frequently employed for image recognition and classification. It works well when processing the image's pixel information. Convolution layer, pooling layer, flatten layer, and dense layer are the four different types of layers used to construct it. The model begins with an input convolution layer that accepts input images of the specified size and has a certain number of neutrons. In CNN, the pooling or flattening layers provide a summary of the characteristics retrieved from the convolution layers before moving on to the following set. Depending on the application, there can be any number of convolution and pooling layers. In CNN, the flatten layer is crucial. The 2D and 3D feature maps are flattened into 1D data by this layer. The completely connected, thick layers are then used to transmit the flattened data. The classification is provided by the thick layers, which map the features to the appropriate target classes. In CNN, the dropout function is employed as a regularization method to stop over fitting of the data. The final thick layer transfers the features to one of the target classes based on the activation function.



.

Figure 1. Basic Architecture of CNN *(Source: https://medium.com/analytics-vidhya/introduction-to-convolutional-neural-network-6942c189a723)*

1.2 Autoencoders

An artificial neural network called an autoencoder is used to extract features from data coding, unsupervised. In order to reduce the dimensionality of a set of data, an autoencoder seeks to learn a representation (encoding). It teaches the network to ignore visual signal noise, and the autoencoder strives to produce a representation (decoding) from the reduced encoding that is as similar as feasible to its original input. Many practical issues can be solved with autoencoders, including picture recognition and learning the semantic meaning of words.

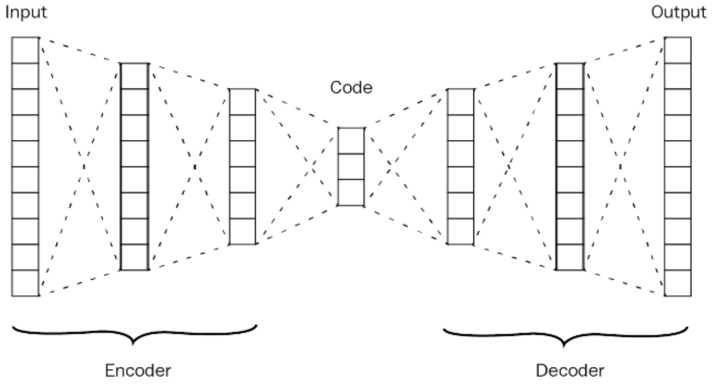


Figure 2. Basic Architecture of Autoencoder *(Source: https://towardsdatascience.com/applied-deep-learning-part-3-autoencoders-1c083af4d798)*

1.3 Long Short -Term Neural Networks

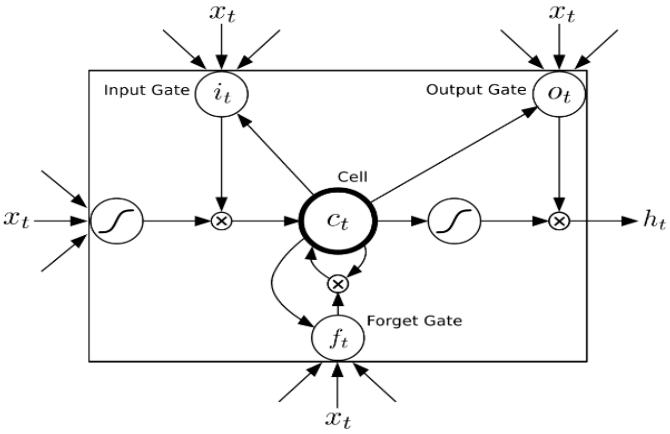


Figure 3. Basic structure of an LSTM Memory cell with forgets, input, and output gates *(Source: https://www.researchgate.net/figure/A-simple-LSTM-gate-with-only-input-output-and-forget-gates_fig7_304066008)*

A unique kind of Recurrent Neural Networks called LSTM network is able to learn long-term dependencies. The long-term reliance issue with conventional RNNs is specifically addressed in the design of LSTMs. The problem of the vanishing/exploding gradient is likewise solved using LSTMs. Although they use a different structure for the repeating modules, LSTMs duplicate the same chain-like topology. The forget gate (sigmoid), input gate (sigmoid), input gate (tanh), and output gate are the four neural network layers that make up an LSTM (sigmoid). After receiving the h (t-1) state's output, the forget gate assists in making decisions regarding what must be eliminated, thus retaining only the information that is important. A sigmoid function surrounds it and changes the input between [0, 1]. The input gate scales the current cell state while adding fresh information from the current input. The tanh layer builds a vector of potential new candidates to add to the current cell state, while the sigmoid layer selects which values should be modified. The output gate's sigmoid function makes decisions based on the state of the current cell.

1.4 Problem Statement

The purpose of the study is to develop LSTM neural networks and CNN for spoken digit recognition for the classification of digits from 0 to 9. The effectiveness of the CNN model is further evaluated with the use of autoencoders. Autoencoders offer a practical method to significantly decrease the noise of input data, increasing the effectiveness of deep learning models.

The format of the paper is as follows: A brief summary of prior research in this field is provided in section 2. The recommended approach is illustrated in section 3. The description of the data gathering and comprehensive implementation information are also elaborated. The conclusion is summarized in section 5 once the findings are tallied in section 4.

2. Related Works

The existing works in speech recognition are presented in this section. In the research on spoken digit recognition by A. Khemani [1], the performance of linear neural networks and convolutional neural networks for recognising the spoken data of digits was examined. While CNN's accuracy was 99%, the Linear NN's accuracy was only 62%. Diverse datasets can also be employed to enhance the model's learning capabilities.

For automatic recognition of spoken digits, P. Sarma et al. [2], suggested a Linear Prediction filter Coefficients using Artificial Neural Network and then used Principal Component Analysis. 82% accuracy was achieved using the approach. A study on Bengali Spoken Digit Classification using Deep Learning Approaches was carried out by Riffat Sharmin et al. [3]. The suggested approach utilised a CNN for speech recognition feature learning and classification. The strategy was successful in achieving 98.37% accuracy.

According to Amirhossein Tavanaei, et al. [4], Support Vector Data Description for Spoken Digit Recognition employed Feature Vector recovered using Mel Frequency Discrete

Wavelet Coefficients (MFDWC) and the Mel Frequency Cepstral Coefficients (MFCC). While MFDWC's accuracy was 92.25 percent, MFCC's accuracy was 92 percent. The MFDWC feature vectors outperformed the MFCC, and the SVDD-based approach with weighted polynomial kernel function method outperformed the other digit identification techniques, according to the results.

3. Proposed Work

For spoken digit recognition, both an LSTM neural network model and a CNN model are used. The performance of the CNN model with the inclusion of autoencoders is examined. The performances of the models are compared in the last stage of the analysis.

3.1 Dataset

The dataset includes 2000 recordings of four English-speaking people speaking 50 of each digit. The audio files are kept as 8 kHz wav files. The silence in the audio files is removed and transformed into spectrogram images, which are audio images made from audio files that are kept in a folder. The transformed audio files are kept in a directory, and to make processing simpler, the photos are further converted to NumPy arrays. In order to speed up image retrieval, the dataset is split into a train set with 1800 recordings and a test set with 200 recordings and are saved in NumPy files. The model is developed by using Google Colab. The model is developed using Keras and various python libraries like matplotlib, numpy etc.

3.2 Convolution Neural Networks

The spoken Digit Classification task is carried out by CNN architecture. The followings are the detailed specifications of the CNN model architecture implemented. Figure 3 provides the workflow of the CNN model.

- Input size :1800 images of (64 x 64 x 3 channel)
- Two Convolution 2D layers followed by Maxpool layer with size (2,2)
- Flatten layer
- Two Dense layers with 25% dropout
- Final classification layer with 10 output classes
- Final layer activation function: softmax

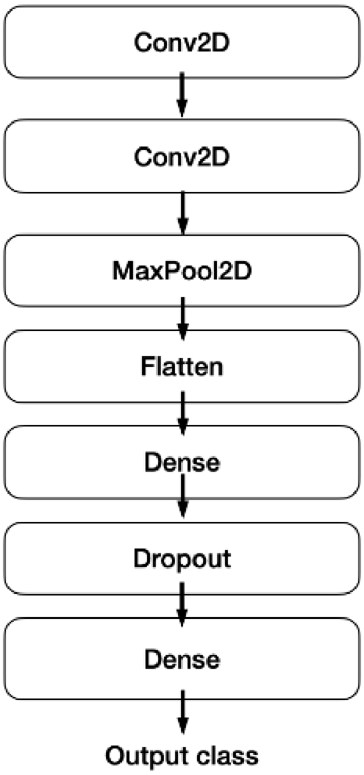


Figure 4. Workflow of CNN

3.3 Autoencoders

Three conv2D convolution layers with ReLU activation function and two max-pooling layers are used to create the encoder for the Autoencoder. Two up-sampling layers, two convolution layers, and a decoding conv2D layer with a sigmoid activation function make up the decoder.

3.4 Long Short -Term Neural Networks

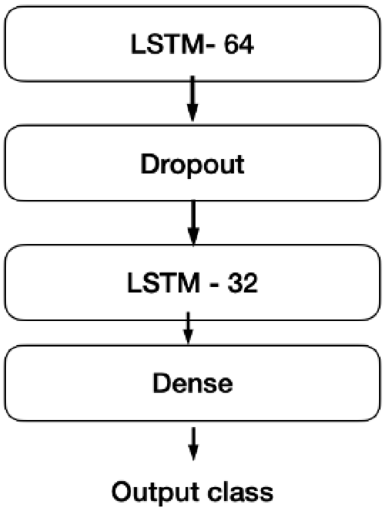


Figure 5. Workflow of LSTM

Recurrent neural networks include neural networks with LSTM. Speech signal classification can use LSTM since time-sequenced data is the main focus of speech recognition. The implemented LSTM model architecture's detailed specifications are listed below. The LSTM model's workflow is shown in Figure 4.

- Input size: 1800 samples of size (64 x 192)
- Input LSTM layers layer with 64 neutrons
- Dropout layer of 25%
- Second LSTM layer with 32 neurons
- Final Dense layer with 10 output classes
- Final Layer activation function: softmax activation

4. Results and Discussion

Accuracy is used as a performance criterion since the data in each output class are balanced. The table below provides the training and testing accuracy gained in various models. The table shows that CNN performs better than the LSTM model in both training and testing accuracy. The CNN model's prediction accuracy has been somewhat enhanced by autoencoders, making it simpler to extract the crucial information from Mel spectrogram pictures. Additionally, the model accurately anticipates fresh data. These models may be used in digital tools that can help dyslexic children, blind people, and people with disabilities.

Table 1. Accuracy of the Deep Learning models

Model	Train Accuracy	Test Accuracy
CNN	98.03%	98%
LSTM	90.1%	93.1%
CNN with Autoencoders	98.96%	99.4%

5. Conclusion

Spoken digits recognition has been implemented using Convolution Neural Network (CNN) and Long Short -Term Memory (LSTM), and the accuracy performance metric has been used to assess each system’s performance. The testing performance of the CNN model for spoken digit identification is good, and its training accuracy is 98%. The LSTM Neural Networks exhibit 90% training accuracy and function pretty accurately during testing. As a result, the CNN model beats the LSTM model in spoken digit recognition. The CNN model's prediction accuracy has also been somewhat enhanced by autoencoders, making it simpler to extract the crucial information from Mel spectrogram pictures. To improve the LSTM model's accuracy, the dataset can be increased.

References

[1] A. Khemani, “Spoken Digit Recognition (Speech Recognition) ",http:// cs230.stanford.edu/projects_fall_2020/ reports/55617928.pdf

[2] P. Sarma, “Automatic Spoken Digit Recognition Using Artificial Neural Network”, International Journal Of Scientific & Technology Research Volume 8, Issue 12, Dec 2019.

[3] Riffat Sharmin, et al., “Bengali Spoken Digit Classification: A Deep Learning Approach Using Convolutional Neural Network”, Elsevier, Science direct, Procedia Computer Science 171 (2020) 1381-1388.

[4] Amirhossein Tavanaei, et al., “Support Vector Data Description for Spoken Digit Recognition”, https:// www.scitepress.org/ papers/2012/37644/37644.pdf

[5] S. Imani, P. Sarma, and K. Samudravijaya. "Automatic Identification of Native Language from Spoken English." Proceedings in FRSM 2019, Kanpur, India, July 6-7, 2019

[6] H. Xie, Li Zhang, C. P. Lim, “Evolving CNN-LSTM Models for Time Series Prediction,” IEEE Access, vol. 8, p. 161519 – 161541, Sep. 2020

[7] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P. Manzagol, “Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion,” Journal of Machine Learning Research, vol. 11, p. 3371-3408, 2010.

[8] Jane Oruh and Serestina Viriri, “Deep Learning-Based Classification of Spoken English Digits”, Journal of Computational Intelligence and Neuroscience, 2022 Sep 28. doi: 10.1155/2022/3364141

[9] G. Hinton, L. Deng, D. Yu et al., “Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups,” IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82–97, 2012.

[10] M. M. Saleem, Deep Learning for Speech Classification and Speaker Recognition, The University of Texas at Dallas, Richardson, TX, USA, 2014.

Author's biography

K. Vaishnavi is a Junior Research Fellow and Research Scholar in the Department of Computer Science and Engineering, PSG College of Technology, Coimbatore. She has 9 years of teaching experience. She handles courses like operating system, Data structures, Database Management system and Data Mining. She has a passion for cognitive computing.

G. Sudha Sadasivam works as a Professor and Head of the department of CSE at PSG College of Technology, Coimbatore. She has 25 years of teaching experience. She handles courses like Compiler Design, Operating Systems, Theory of Computing, Software Engineering, Data Intensive Computing and Distributed Computing. Her research areas include Cloud Computing, Big Data Analytics, Privacy and Security in Big Data Systems. She has completed research projects sponsored by AICTE, UGC, Yahoo, Samsung, Nokia & Cloudera. She is currently carrying out DST-CSRI sponsored research project and SEED/TIDE research projects in the areas of Cognitive Computing. She has published 6 books in the areas of Compiler Design, OOAD, Big Data, Middleware Technologies and Edge Computing. She has published around 50 papers in indexed journals and conferences.