ISSN: 2278-4632
Vol-9 Issue-4 No.01 April 2019

USING MACHINE LEARNING KERAS EXTRACTION OF INFORMATION CLASSIFICATION OF SPEECH

Neelamani Samal[1], Dr. Prakash Chandra Jena[2], Dinesh Manjhi[3]

[1]Asst. Prof. Einstein Academy of Technology and Management, Bhubaneswar, India
[2]Asst. Prof. Einstein Academy of Technology and Management, Bhubaneswar, India
[3]Student, Einstein Academy of Technology and Management, Bhubaneswar, India

Abstract

There is a vast use of Internet in the year 2020 in almost every areas. Every body and every areas has gone online due to lockdown amid COVID. Staring from school education , professional education , research banking even if Kindergarten kids have gone online successfully. [1] Every device and every equipment in the modern world carries forward the interesting area Machine Learning into our lives one step closer. The field of speech and audio recognition is not new but has found widespread usage and general public interest with devices like Alexa. Identification of words, languages and tones is the primary goal of efficient Speech Recognition software, system or application. The accuracy in case of Speech recognition System is of paramount importance as it may lead to misinterpretation of commands and thus may result in unintended function of the system. The most popular tool to implement Machine learning projects in Speech Recognition is Python. Keras is a framework which is primarily used for implementing Artificial Neural Network. The work presented in this thesis is to identify a specific word out of a set of words by training the neural network model in Keras implemented in Python.[2]

Keywords: Machine learning Sound, Speech, Audio, Recognition, Machine Learning, Python, Keras, Neural Network, Artifitial Intelligence

1. Introduction

Speech recognition systems are ubiquitous these days – from Apple's Siri to Google Assistant. These are all new advents though brought about by rapid advancements in technology. The exploration of speech recognition goes way back to the 1950s.These systems have been around for over 50 years.In this paper , I propose to develop an application of audio signal identification and classification based on training and testing my model on datasets of sounds collected over a sample space. The work is proposed to be implemented in Keras which is a Machine Learning implementation platform based on Python and runs on top of Tensor Flow. The IDE shall be Jupyter notebook. In order to perform this task, I used an Anaconda environment (Python 3.7) with the following Python libraries.[3],[4]

2. Results

The following results discussed and some of the results are depicted in the form of screenshots are obtained by comparing the accuracy of Speech Dataset by varying the Keras Loss Models and Optimizers. The dataset included four different words namely “Cat”, “Dog”, “Yes” and “No”. Each word is spoken by approximately 1750 different persons. The dataset is in the form of wave files. I have taken two Optimizers and two Loss Models into consideration while compiling the Model. Both the optimizers and the Loss models are explained in detail in the previous chapter. Optimizers are:

1. Nadam- Nesterov-accelerated Adaptive Moment Estimation
2. SGD- Stochastic Gradient Descent

Two Loss Models namely Probabilistic Loss and Regression Loss are compared for accuracy. In my experiment the Jupyter notebook is running on Ubuntu machine , so it is launched from terminal by running the following command >> jupyter notebook. It opens in default browser, which in my case is brave. The modules are imported in jupyter notebook in the first cell. It is very convenient to work in cells as it makes the code modular. Each cell is executed individually and can be debugged separately thereby making the work flow very streamlined. Data is in the form of wave files.

Table 1 Accuracy Measurement over 15 Epochs(iterations)

Loss Model/ Optimizer	**NADAM**	**SGD**
Categorical Cross Entropy	97.92 %	33.66 %

ISSN: 2278-4632
Vol-9 Issue-4 No.01 April 2019

Mean Squared Error	95.79 %	32.15 %

Table 2 Loss Measurement over 15 Epochs (iterations)

Loss Model/ Optimizer	**NADAM**	**SGD**
Categorical Cross Entropy	.0594	1.3420
Mean Squared Error	.0157	.1835

Analysis of wave files is performed and the audio wave is displayed in the run window with the help of matplotlib module. Audio files are plotted using matplotlib module. The dataset included four different words namely “Cat”, “Dog”, “Yes” and “No”. Each word is spoken by approximately 1750 different persons. Listdir method is used for defining classes. Data Wrangling is the process of cleaning the data and extracting relevant and clean data for testing and validation purpose. More the data is clean and meaningful, more is the accuracy of the prediction made by the model on test data. First, second and third convolution layers are defined for the neural network. Various modules are imported from Keras. Also, training and test data is defined in the 80-20 ratio. Maxpooling and Dropout is carried out on all the layers namely input layer, output layer and hidden layer of neural network. Batch normalization is also performed on all the inputs. Using python's inbuilt modules and methods, various operations are performed. Here modelfit is being done on training data on a batch size of 32 and 15 epochs. Sound file and sound device modules are installed using Python.

The command used for the same is 'pip install'.Whichever module is required to be installed using pip installer, must be registered with python's official website. There are numerous modules available contributed by Python community. After the model is compiled for the different combinations of loss model and optimizer, it is seen that the epochs have started to run. In my paper, I have taken 15 epochs for set of two loss models and two optimizers. Model is compiled taking the validation data for Categorical cross entropy loss and NADAM optimizer. In this I have done total of 15 epochs which took an approximate time of 9 hours. As the epochs progress from 1 to 15, it is observed that loss reduced from 0.1081 to 0.8594. However, the accuracy of the model increased from 96.43% to 97.92% in 15 epochs. Model is compiled taking the validation/test data for Categorical cross entropy loss but here optimizer is changed to SGD. Slight variation in the values of loss and accuracy is seen in this case. Here, a total of 15 epochs have been done which took an approximate time of 9 hours.

As the epochs progress from 1 to 15, it is observed that loss reduced from 1.4018 to 1.3420. However, the accuracy of the model remained very low and increased from 29.47% to 33.66% in 15 epochs. Model is compiled taking the validation/test data for Mean Squared Error along with NADAM. Here, a total of 15 epochs have been done which took an approximate time of 9 hours. As the epochs progress from 1 to 15, it is observed that loss reduced from 0.1668 to 0.0157. For this loss optimizer combination, the model accuracy has drastically increased from 44.41% to 95.79% in 15 epochs.[6] Model is compiled taking the validation/test data for Mean Squared error loss but here optimizer is changed to SGD. Slight variation in the values of loss and accuracy is seen in this case. Here, a total of 15 epochs have been done which took an approximate time of 9 hours. As the epochs progress from 1 to 15, it is observed that loss reduced from 0.1914 to 0.1835. However, the accuracy of the model remained very low and increased from 29.31% to 32.15% in 15 epochs.[7]

3. Conclusion

The results clearly suggest that probabilistic approach towards prediction of the speech data or speech recognition are best suited and recommended. Categorical cross-entropy which is a probabilistic loss model works best for speech data as opposed to Mean sqaured error loss model which is a regression loss model. The reason could be attributed to highly unpredictable and almost infinte set of speech data population. [17],[18] The reason for this the effect of local dialect, gender, pitch, nasal toning and personalized pronunciations.

The second conclusion is the optimizer choice, NADAM is recommended over SGD for the simple plain reason that the gradient descent underperforms in more dynamic and vivid datasets as compared to the adaptive optimizers like ADAM or NADAM.

Future Scope

The work in this paper may be extended or extrapolated to larger datasets with hundreds of words. The results can be directly applied in the decision making for the choice of loss models and optimizer selection while designing such a speech or audio recognition software [19]. Future scope of this thesis includes work on Music Recognition. Music, Vocal,

ISSN: 2278-4632
Vol-9 Issue-4 No.01 April 2019

instrument are areas where prediction and recognition can be performed to enhance the user experience and learning. The simulation results can be obtained much faster with GPU machines and using tensor-flow and keras for GPU. More advanced areas of Human Speech Recognition such as Natural Language Processing, Mood Predictor, Predicting the Emotional State of mind through speech modulation could also be implemented using advanced neural Networks. This time, what we call as a New Normal has now started to become an integral part of our personal, social and professional lives and all of us have now started to camouflage with the machines we have known but never thought could have become our tool, our medium, our connection with everything outside our homes. Artificial Intelligence and Machine Learning which otherwise also was going through the depths of research and development, but has now gained even more attention than ever.

Here, in this paper, I have explored Natural Language Processing application, a subdomain of AI and Machine Learning so that in future we are able to build and create much more interactive and human like experience with machines.

References

[1] Jayashree Padmanabhan Machine Learning in Automatic Speech Recognition: A Survey February 2015 IETE Technical Review 32(4):1-12
[2] Ali Bou Nassif Speech Recognition Using Deep Neural Networks: A Systematic Review February 2019 IEEE Access PP(99):1-1
[3] Y. Xie, L. Le, Y. Zhou and V. V. Raghavan, "Deep learning for natural language processing" in Handbook of Statistics, Amsterdam,The Netherlands:Elsevier, 2018.
[4] J. Padmanabhan and M. J. J. Premkumar, "Machine learning in automatic speech recognition: A survey", IETE Tech.Rev., vol. 32, pp. 240-251, 2015.
[5] H. Singh and A. K. Bathla, "A survey on speech recognition", Int. J. Adv. Res. Comput.Eng.Technol., no. 2, pp. 2186-2189, 2013.
[6] I. Shahin, A. B. Nassif and S. Hamsa,"Novel cascaded Gaussian mixture model- deep neural network classifier for speaker identification in emotional talking environments", Neural Comput.Appl..
[7] Y. Zhang,"Speech recognition using deep learning algorithms", pp. 1-5, 2013, [online] Available: https:// scholar. google. com/ scholar? as q= Speech+ Recognition+ Using+ Deep+ Learning+ Algorithms & as occt= title & hl= en& as sdt= 0% 2C31.
[8] O. Abdel-Hamid, A. Mohamed, H. Jiang, and G. Penn. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP). 2012.
[9] J. Baker, L. Deng, J. Glass, S. Khudanpur, C.-H. Lee, N. Morgan, and D. O'Shaughnessy. Updated MINS report on speech recognition and understanding. IEEE Signal Processing Magazine, 26(4), July 2009.
[10] D. Ciresan, A. Giusti, L. Gambardella, and J. Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In Proceedings of Neural Information Processing Systems (NIPS). 2012.
[11] R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of International Conference on Machine Learning (ICML). 2008.
[12] K. Demuynck and F.Triefenbach. Porting concepts from DNNs back to GMMs. In Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU). 2013.
[13] L. Deng and J. Chen. Sequence classification using the high-level features extracted from deep neural networks. In Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP). 2014.
[14] I. Shahin,"Speaker identification in emotional talking environments based on CSPHMM2s", Eng.Appl.Artif.Intell., vol. 26, no. 7, pp. 1652-1659, Aug. 2013.
[15] P. Domingos,"A few useful things to know about machine learning", Commun. ACM, vol. 55, no. 10, pp. 78-87, 2012.
[16] Santosh K.Gaikwad, Dr.Babasaheb Ambedkar Marathwada, Bharti W.Gawali, 2011, A Review on Speech Recognition Technique.
[17] Dandan Mo,December 4, 2012, A survey on deep learning: one small step toward AI.
[18] Speech Recognition Technique: A Review Sanjib Das Department of Computer Science, Sukanta Mahavidyalaya, (University of North Bengal), India, International Journal of Engineering Research and Applications (IJERA) MayJun 2012.
[19] L. Deng, "A tutorial survey of architectures algorithms and applications for deep learning",APSIPATrans.Signal Inf.Process., vol. 3, no. e2, pp. 1-29, 2014.