STM JOURNALS

Journal of
Artificial Intelligence Research & Advances

ISSN: 2395-6720
Volume 8, Issue 3, 2021
DOI Journal: 10.37591/JoAIRA

http://computers.stmjournals.com/index.php?journal=JoAIRA&page=index

Review JoAIRA

Speech Emotion Recognition using Convolutional Neural Networks

Rajat Mittal*

Abstract

This research work presents an assortment of techniques in Speech Emotion Recognition with the help of spectrograms and handcrafted Deep learning architecture: Convolutional Neural Networks (CNN). Emotional kingdom detection is a crucial part of human-device interplay studies. To make interaction between man and machine more natural, many milestones are conquered in speech emotion recognition, but still this process requires more up-to-the-mark results. To make an attempt for the same, this study represents a three-layers deep, two-dimensional Convolutional Neural Network for the challenging task of emotion detection from spectrograms produced by audio (speech) signals. We teach and examine our version on eight emotions: Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust and Surprised. The mean value of outputs produces a classification of human speech. Our proposed version achieves, on average, a weighted accuracy of 72%.

Keywords: Spectrograms, CNN, emotion, classification, preprocessing, deep learning

INTRODUCTION

When there may be an interplay among a person and a machine, it is essential for a machine to somehow analyze human speech intelligently and understand it semantically and give response likewise; for example, in diagnostic tool for therapists or applications for online tutorial where we observe that the system response relies on the person's emotions, also in robotic interaction or further application is to determine the serious situation in emergency call centers and many more. The inability to understand the emotional state by a machine is itself a significant challenge for scientists. The goal of a speech emotion recognition gadget is to apprehend various emotional situations of the speaker.

So a trouble of speech popularity entails three vital stages: sign pre-processing, the characteristic extraction and the classification. Firstly, the Signal pre-processing where the input signal is divided into various divisions that are used for bringing-out related features. Secondly, Feature extraction, which decreases the size of data and represents them as feature vectors [1]. As the change in the features, we get from prosodic analysis is different according to emotion, the general success rate is transforming. To keep away from this, we studied the impact of spectrograms on emotional recognition. For this, replacing audio traces by their visual representation, techniques for classification of images can be employed for extracting features on speech classification projects. Spectrograms, the maximum broadly used visible illustration of voice, entails the show in their frequency spectrum as they range in time.

The methods of Deep Learning are implemented to do feature extraction and then further classification from spectrogram images.

***Author for Correspondence**
Rajat Mittal
E-mail: rajatmittal0111@gmail.com

Student, Centre for AI in Medicine, Imaging & Forensics, Department of Physics, School of Basic Sciences and Research, Sharda University, Greater Noida, Uttar Pradesh, India

Received Date: November 23, 2021
Accepted Date: December 07, 2021
Published Date: December 28, 2021

Citation: Rajat Mittal. Speech Emotion Recognition using Convolutional Neural Networks. Journal of Artificial Intelligence Research & Advances. 2021; 8(3): 26–32p.

For this, various classifiers have been bought up out there. We see, the speech signals are less frequent, and they generate patterns of distinct representations in spectrogram images.

The task of classification of spectrograms from human speech can be done mainly by two methods, i.e., (i) Traditional classifiers like Self-Organizing Map (SOM), HMM (Hidden Markov Model), Support Vector Machines (SVM), and the Naïve Bayes Algorithm model; and (ii) Deep Learning Algorithms, i.e., Deep Neural Network (DNN), Convolutional Neural Network (CNN), Stack Autoencoder, Deep Boltzmann Machine (DBM), Recurrent Neural Network (RNN) (LSTM) and Deep Belief Network (DBN).

In the sphere of Deep Learning, the Convolution Neural Networks (CNN) enjoyed a massive surge in recognition after AlexNet, a CNN, accomplished today's overall performance in classifying pix with inside the ImageNet challenge. Thus, for audio processing, which is insensitive to the position of patterns on spectrograms and recognized as a perfect technique for classification of features in spectrogram images [2, 3, 4].

RELATED WORK

The traditional methods for this system relied on drawing out the prosodic features from human voice utterances like as, pitch and its harmonics, shimmer etc. and subsequently classifying them using the classifiers like SVM, GMM, HMM etc. Also a unique version of Long Short Term Memory, called Dual-Sequence LSTM (DS-LSTM) that is capable of system sequences of records simultaneously. A novel mechanism for data pre-processing was proposed which uses the Nearest-Neighbor interpolation [5, 6].

A conventional approach proposes a technique to construct a gadget of emotion popularity which changed into elicited from Deep Convolutional Neural Networks (DCNNs). Especially, the log-spectrogram is computed and the Principal Component Analysis technique has been used to lessen the dimensionality and put down the interferences. Then the PCA deep white spectrograms had been spitted into non-overlapping segments. The DCNN was implemented to understand the representation of the emotions from the segments with supervised training speech data [7].

Another manner for the SER gadget with the spectrograms and deep convolutional neural networks has been introduced. The spectrograms constituted of the utterances are entered to the Deep CNN [8]. Another utilization of spectrograms is represented by Li *et al.* [9], where the Neural Network is for detecting multi-type signals and to classify signal varieties manifested in wideband spectrograms. Their network utilizes the fundamental approximation to find the uneven center line in the region of signals and also recognize their class. A relevant work is shown here, where there is a comparison of the results acquired with a Convolutional Neural Network (CNN) with the outcomes received with the aid of using the handmade functions and Support Vector Machine classifiers [10].

Another contributing spectral feature is, Mel-Frequency Cepstral Coefficient (MFCC). Nalini *et al.* developed a system of Speech Emotion Recognition using residual phase and MFCC features with neural network [11].

METHODOLOGY

Proposed Algorithm

Our proposed framework ventures to make use of function mastering techniques for spectrograms which can be made out of audio (speech). For the development of our Deep learning model, we have used the TensorFlow library for fast numerical computing. In addition, for the pre-processing of the model, Scikit Learn was introduced.

We tried to apply a discriminative Convolutional Neural Networks for capacity to study features. Three layers of two-dimensional convolution are stacked with three pooling (2D) layers in among

them. A flattened layer accompanied through a dense layer is appended with the activation 'ReLU' function. The output layer is supported through the SoftMax classifier tool. Our three-layers-deep learning model is being optimized by 'adam'.

We train and compare our version on eight feelings i.e. Neutral, Sad, Calm, Disgust, Happy, Angry, Fearful, and Surprised.

Spectrogram Image Features

Any sound waves generated from any source are built up of low pressure and high-pressure areas propagating throughout a medium. These low-and high-pressure areas form many distinct motifs to each distinguished voice signal. The characteristics like time periods, wavelength and frequencies are used for classification of various types of sounds as we humans possess [1]. Spectrogram is defined as the tridimensional arrangement that reflects the intensity of the audio and frequency distribution with respect to time, as shown in Figure 1.

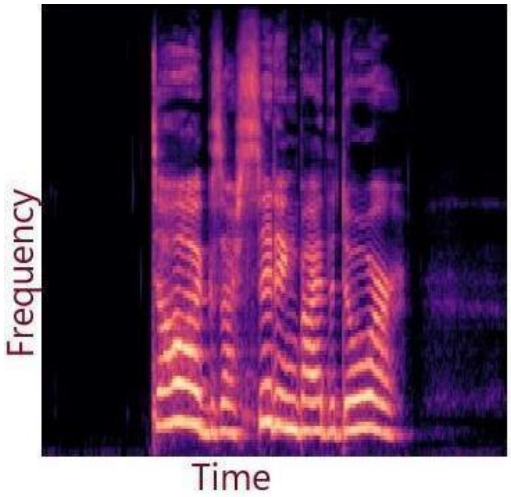

Figure 1. Spectrogram of a sound file '03-01-07-02-02-01-10'.

The aspect features from different playing techniques and effects present in the analyzed records can be easily seen on the spectrograms. The spectrogram's distribution of power is vertically over the frequency-time aircraft at the primary segment of the sound duration that depicts the escalation of sound depth [12]. Once the spectrogram images are generated, they can be used together with numerous ML classifiers after their preprocessing.

Data Preparation and Preprocessing Techniques

Data preprocessing is needed for cleansing the records and making it appropriate for a deep studying version which additionally will increase the accuracy and performance of a version. Now, we have to import all the crucial Python libraries for data preprocessing. The three of them used are, firstly, NumPy, It is the fundamental package to perform scientific calculations or mathematical operations. Secondly, Pandas is an open-supply Python library for statistics analysis. Thirdly, Matplotlib, it is a library that is used for plotting (2D) graphical representation.

For our work, using Scikit-learn we have employed the MultiLabelBinarizer which is required for data with multiple labels. It increases the samples and reduces the class number (here used for eight distinct emotion labels).

The image (input) data has to be split according to the split ratio. To reach the best accuracy of the model, firstly, a lot of training is needed, secondly, a comparatively huge amount of the data should be used just for that purpose.

- Training dataset: This is the element wherein our version is evolved via way of means of schooling algorithms of Deep Learning. The model attempts to memorize the whole dataset and its varied properties.
- Validation dataset: This is portion of the input data which was utilized for validation of our various model fits. Or we are able to say, we have used the validation facts to pick out and refine the model's hyper-parameters.
- TEST data: This input statistics component is utilized for checking out the version hypothesis. It was separated earlier and left untouched until we got to know the model and hyper-parameters, and after that, the model was applicable for test dataset to find an accuracy count of its performance when it will be applied on an actual-world dataset.

Now we have reached the stage where all the required preprocessing of data is completed. Hence the data is ready to be put in a Neural Network which will first recognize the features from each input spectrogram with respect to their corresponding labels (all eight emotions) and further use these extracted features for classification. This may be completed with the use of a set of rules of Deep Learning, Convolutional Neural Network. When we can build a perfect CNN model, then the next step is to train and validate the model.

CONVOLUTION NEURAL NETWORK (CNN)

CNN is a Deep Learning network consisting of a convolutional layer that uses a set of convolutional filters for extracting many atomic elements or patterns at every possible position in the input image and then generating multiple feature maps. Keras was used for the Sequential model. The first convolutional layer in CNN tries to discover the capabilities underlying within the image. After this, we want to layer take hold of the ones characteristic maps from the convolutional layer that is accomplished with the pooling layer which additionally tries to lower dimensionality of the featured maps. So in accordance with the complexity of the input data, as depicted in Figure 2, there could be many sets of the convolutional and pooling layers. The final layer of this CNN is the classifier or output prediction layer.

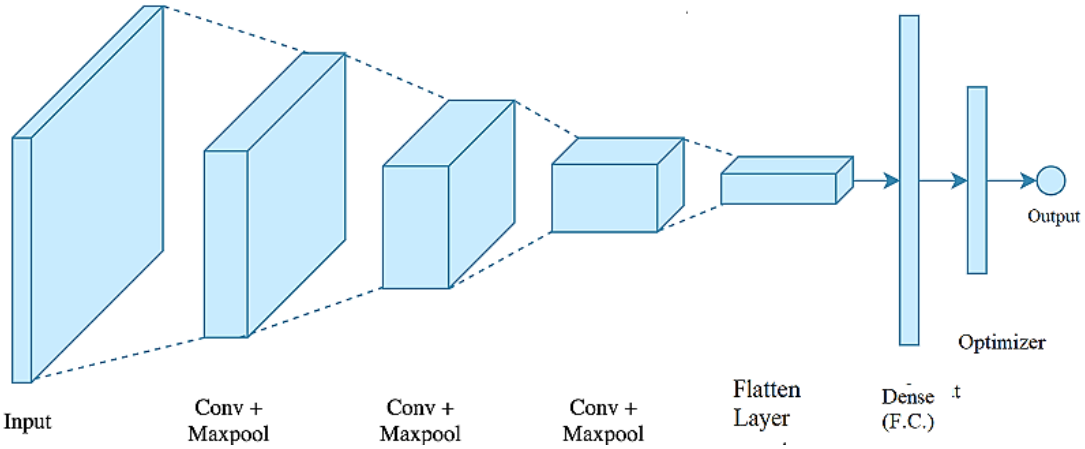

Figure 2. Convolutional Neural Network version used for Speech Emotion Recognition.

For the sake of image compression, all the spectrograms of resolution 270×260 are converted into gray scale as it reduces their size. File compression is commonly performed to boom the processing speed. The CNN sequential version accommodates three convolution layers that are: Conv1 with kernels of size (5×5) and 32 filters have been set, Conv2 with kernels size (5×5) and 64 filters have been set, and Conv3 with kernels of size (5×5) and 32 filters have been set.

Each convolution layer brings up a set of featured maps, these are fed to the top pooling layer, i.e., MaxPooling2D with the pooling of size (5, 5) and value of strides were set as (3, 3) in each pooling

layer. Once the featured maps from three convolutional layers are obtained, the further step is to flatten them i.e. to transform the whole pooled matrix of the feature map into one column (single vector). Hence a flattened layer is added after them.

This step incorporates off the input layer, the fully related or dense layer. The output layer is liable for getting the expected classes. This layer involves an activation function, i.e. 'Softmax', which passes the information through the neural network plus the prediction of error is conceived. This error, inculcated from here, is now back-propagated via the system to enhance the predicted decision.

The fully connected layer is right before the dropout layers which avoids the over-fitting of data. The dropout layer of 25 and 50% dropout probability were deployed to create high coadaptation. The 'ReLU' (Rectifying Linear Unit) activation characteristic escorts all of the three convolutional layers with 0 padding. The proposed CNN model is depicted in Figure 2.

Now, the subsequent step is to bring together this sequential model. The function "compile" expects some parameters which are, loss function, Optimizer and the metrics of model performance. These three are answerable for the compilation of this model. The gradient descent is the set of rules carried out inside the optimizer, and for loss function, we deployed categorical_crossentropy.

To train, the very last step is becoming the CNN model. As mentioned earlier, Keras is used for spectrogram image augmentation which reduces the over-fitting. This is done by flipping (horizontally), rescaling, zooming and shearing the pictures.

So now we have reached the stage where we have to apply experimental techniques. For this, the model was subjected to many different sets of training epochs, filters, kernel sizes and number of strides. This has been done due to some reasons, like over-fitting of the model, high loss, low accuracy etc. After performing many experiments with different values of filters, kernel sizes, and number of strides we can conclude that there is a certain combination of some values which, on training, will show better results in which the model is fit, accuracy is high and it faces low loss value.

MATERIAL

Dataset Description

For the sake of practical collation, this proposed SER device was educated in addition to examined on open supply dataset supplied through the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). Datasets used in this experiment are the speech data of 24 actors (both female and male) [13]. The spectrograms generated are from this dataset. It incorporates 15 sentences for 8 distinct emotions, anger, calm, fear, happiness, disgust, sadness, surprise, and neutral. They were distributed as 60 utterances for every actor. And 24 actors ×60=1440 files. Statement (01= "Kids are speaking via way of means of the door", 02= "Dogs are sitting via way of means of the door"). Using the Scikit tool, the whole set of the input dataset was split into 75% for the training and 25% for the testing [14].

System Specification (Configuration)

The laptop used for the experiment has Processor: *Intel(R) Xeon(R) CPU E5-2673 v4 @ 2.30 GHz 2.29 GHz* with RAM: 16 GB, 1 TB Hard disk and GPU: 4 GB.

EXPERIMENT AND RESULTS

Table 1 depicts the satisfactory overall performance of the experiments finished so far, (numerical) confusion matrix. The schooling became finished with an epoch value, 500. The diagonal numbers (in blue) shown here, are the percentages of every accurate identified class of emotion (or we can say, in which the prediction labels coincide with genuine labels of equal class); the other percentages depict wrongly identified percentages of each emotion class (or we can say, where the prediction labels

coincide with true labels of any other class) [15]. It is clear that *sad, disgusting* and *surprised* predicted execution was over 80% which is quite applaudable, then *happy* and *fearful* emotions also show some good performance and there is need for improvement for the *angry* and *calm* emotions. However, the mean value of accuracy of our model is 72%.

Table 1. Performance: SER System Using CNN model with 500 EPOCHS.

Predicted Labels									
True labels		***Neutral***	***Calm***	***Happy***	***Sad***	***Angry***	***Fearful***	***Disgust***	***Surprised***
	Neutral	50%	34%	1%	1%	0%	6%	4%	5%
	Calm	8%	56%	6%	1%	2%	4%	7%	3%
	Happy	0%	0%	78%	2%	1%	1%	0%	20%
	Sad	3%	0%	2%	83%	3%	4%	3%	2%
	Angry	2%	3%	5%	5%	67%	10%	9%	4%
	Fearful	4%	6%	3%	7%	2%	75%	0%	3%
	Disgust	1%	6%	1%	4%	0%	1%	87%	0%
	Surprised	2%	0%	4%	3%	7%	2%	2%	80%

CONCLUSION

This study concludes a way for speech emotion popularity even as using the tool like spectrograms. Also, they involve the CNN (Convolutional Neural Network). This study represents a three-layers deep, two-dimensional Convolutional Neural Network for the challenging task of emotion detection from spectrograms produced by audio (speech) signals. The signals involved are total eight types and all the signals amplify and achieves on average of 72% accurate result.

Acknowledgment

These studies became supported with the aid of using the Centre for Artificial Intelligence in Medicine, Imaging & Forensics, Sharda University (Dr: Prof. Ashok Kumar). I thank Prof. Ashok Kumar for the assistance of supervision that greatly improved this research project.

REFERENCES

1. Khamparia A, Gupta D, Nguyen NG, Khanna A, Pandey B, Tiwari P. Sound Classification Using Convolutional Neural Network and Tensor Deep Stacking Network. IEEE Access. 2019 Jan; (99): 1–1.
2. Md Zahangir Alom, Taha Tarek M, Christopher Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Van Esesn Brian C, Awwal Abdul AS, Asari Vijayan K. (2018). The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches. [Online]. arXiv:1803.01164 [cs.CV]
3. Vinod Nair, Hinton Geoffrey E. Rectified Linear Units Improve Restricted Boltzmann Machines. Appearing in Proceedings of the 27th International Conference on Machine Learning (ICML), Haifa, Israel. 2010.
4. LeCun Y, Haffner P, Bottou L, Bengio Y. Object Recognition with Gradient-Based Learning. In: Shape, Contour and Grouping in Computer Vision. Lecture Notes in Computer Science. Vol. 1681. Berlin, Heidelberg: Springer; 1999; 319–345. https://doi.org/10.1007/3-540-46805-6_19
5. Jianyou Wang, Xue M, Culhane R, *et al*. (2019). Speech Emotion Recognition with dual-sequence LSTM architecture. [Online]. Available from https://arxiv.org/abs/1910.08874.
6. Taiba Majid Wani, Teddy Surya Gunawan, Syed Asif Ahmad Qadri, Hasmah Mansor, Mira Kartiwi, Nanang Ismail. Speech Emotion Recognition using Convolution Neural Networks and Deep Stride Convolutional Neural Networks. *2020 6th International Conference on Wireless and Telematics (ICWT)*. 2020; 1–6. DOI 10.1109/ICWT50448.2020.9243622.
7. Zheng WQ, Yu JS, Zou YX. An experimental study of speech emotion recognition based on deep

convolutional neural networks. 2015 International Conference on Affective Computing and Intelligent Interaction (ACII). 2015; 827–831. doi: 10.1109/ACII.2015. 7344669.
8. Badshah AM, Ahmad J, Rahim N, Sung Wook Baik. Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network. 2017 International Conference on Platform Technology and Service (PlatCon). 2017.
9. Li W, Wang K, You L. A Deep Convolutional Network for Multi-Type Signal Detection in Spectrogram. Preprints. 2020; 2020050215.
10. Costa Yandre MG, Oliveira Luiz S, Silla Carlos N. An evaluation of Convolutional Neural Networks for music classification using spectrograms. Appl Soft Comput. 2017; 52: 28–38. ISSN 1568-4946.
11. Nalini NJ, Sengottayan Palanivel, Balasubramanian M. Speech Emotion Recognition Using Residual Phase and MFCC Features. Int J Eng Technol. 2013; 5(6): 4515–4527.
12. Krzysztof Czarnecki, Marek Moszyński, Miroslaw Rojewski. Concentrated Spectrogram of audio acoustic signals-a comparative study. HCL Open Science. Proceedings of the Acoustics 2012 Nantes Conference. 2012;23-27. Available from https://hal.archives-ouvertes.fr/hal-00810604/document
13. Wang X, Zhao Y, Pourpanah F. Recent advances in deep learning. Int J Mach Learn Cybern. 2020; 11(4): 747–750. doi: 10.1007/s13042-020-01096-5.
14. Geng Z, Wang Y. Automated design of a convolutional neural network with multi-scale filters for cost-efficient seismic data classification. Nat Commun. 2020; 11(1): 3311. doi: 10.1038/s41467-020-17123-6.
15. Abadi M, *et al*. (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. [Online]. Available: http://arxiv.org/abs/1603.04467.

