Audio Features & Neural Network

Sidhant Gupta[1], Jitesh Sareen[2], Madhur Jain[3]

Information Technology, Bhagwan Parshuram Institute of Technology, Guru Gobind Singh Indraprastha University, India

Abstract: ***The main idea is to talk about the common features used while analyzing audio files. The features discussed are mel frequency cepstral coefficient, zero crossing rate, spectral roll off, spectral centroid and chroma features. Along with the audio features mentioned before, we have also explained about activation function and convolutional neural networks.***

Keywords: Convolutional Neural Network, Audio Features, Activation Functions, Melspectogram, SpectralnCentroid, Zero Crossing Rate

1. Introduction

This paper is most helpful for beginners in the field of deep learning. Those who are new in the field of deep learning and are going to work with audio files for the first time will benefit a lot .We have talked about one of the popular dataset GTZAN which is used in music genre classification. Commonly used activation functions used are also discussed. We have explained about the audio features of a music file along with major processes involved in convolutional neural network like convolution, pooling and dropout.

2. Theory

2.1 Dataset

GTZAN dataset is used in large number of published works.[1] It is one of the most common dataset used while working with music genre classification. Although it is used so much in the field of music detection, it has a lot of missing metadata values. It is not a perfect dataset, it does have a large number of faults. The faults in the dataset include repetitions, mislabeling and distortions. As the dataset possesses a number of faults that does not mean we should not use the dataset, rather we should use the dataset while considering its faults. The dataset has 1000 audio tracks and each of the track is half a minute long. There are 10 genre present in the GTZAN dataset which are **blues, classical, country, disco, hiphop, jazz, reggae, rock, metal and pop.** There are 100 sound clips available for each of these genres. It is not a very large dataset but it still can be used for music genre classification. Like every other real world dataset, it consists of faults like missing metadata values, mislabeling but that does not mean it should not be used at all.

2.2 Activation Function

The main function of the activation function is to perform non linear transformation on the input before sending it to the next layer of neurons. Activation functions are used in neural networks to calculate the weighted sum of input and adds a bias. This value is used to decide whether a neuron should be activated or not. There are various types of activation function linear as well as non linear. Different types of activation functions are used in different scenarios. You might want to use a different activation function for image recognition a different one for speech recognition and so on and so forth. Various types of activation functions are as follows

2.2.1ReLU (Rectified Linear Unit) -It is a widely used activation function in deep learning models. [2] It is quite a fast activation function due to this reason it is successful and is widely used function. It shows high performance and this function is generally used in the hidden layers of the neural networks. It activates only a fewnumber of neurons at a time making it efficient and easy for computations. Mathematically it is max(0,x) which gives x for positive input and 0 otherwise. It is non linear function and it is computationally less expensive as compared to other activation functions.

2.2.2Sigmoid- Sigmoid function[3] is also a widely used function and generally used in feed forward neural networks. It is non linear function which is generally used for binary classification. The result of the sigmoid function lies between 0 and 1 so that is why it is most suited for binary classification. As 1 can be predicted if the result is above 0.5 and 0 otherwise. It is easy to understand and it is used in the output layer of the neural network. Mathematically it is represented as $\mathbf{1/(1+e^{-x})}$

2.2.3 Softmax –It is another activation function [4] used in neural networks. It is used in the output layer of the neural network. It is used when we are dealing with multiclass output data. It is non linear in nature and it is used to find out probabilities to tell the class of each input. It calculates probability of each class and the class with the maximum value is the correct class of the input. It can be said as generalization of the sigmoid function as the sigmoid function is used for binary classification while softmax function is used for multivariate classification i.e. nothing classification involving multiple classes.

2.3 Features

Extracting features from the input audio signal is a very important task. These features will eventually help us find the genre of the audio signal. There are a large nuber of features present in a audio signal but we need to select inly those features which are most relevant to the problem in hand. Features that are important for music genre detection are as follows

2.3.1Mel Frequency Cepstral Coefficient (MFCC): MFCC [5] is a very important feature that needs to be considered while analyzing audio signals. These are small set of features that describe the shape of the spectral envelope. These allow for better representation of sound. It is commonly used in speech recognition systems. It is a very crucial feature while dealing with audio processing.

2.3.2 Spectral Rolloff: It corresponds to the value of frequency below a certain threshold [18] value of the total energy in the spectrum. The threshold value can be set by the user.

2.3.3 Chroma Feature: This feature relates [17] to the 12 different pitch classes. It is a powerful toolused for analyzing music. It has application in audio matching. It is widely used in analyzing audio signals.

2.3.4 Spectral Centroid: The spectral centroid [16] is a measure used in digital signal processing to characterize a spectrum. It indicates where the center of mass of the spectrum is located. Perceptually, it has a robust connection with the impression of brightness of sound. It gives us the point where the entire fre1uency of the signal can be assumed to be concentrated.

2.3.5Zero Crossing Rate: The zero crossing rate [15] is the rate of sign-changes along a signal, i.e., the rate at which the signal changes from positive to negative or back. This feature has been used heavily inboth speech recognition and music information retrieval.

2.3.6Melspectogram: Melspectogram is a combination of two things mel scale and spectogram.Mel scale [19]is basically a non linear transformation of the frequency scale. It is constructed in such a way that sounds of equal distance from one another on the mel scale like 500Hz,1000Hz and 5500Hz,6000Hz sound to humans as if they are equal distance from one another. Spectogram [20] is a used to show how the spectrum of frequencies of a signal vary with time.

2.4 Methodology

We have used deep learning to figure out the genre of the song. Convolution neural networks are widely used for this task

2.4.1 Convolutional Neural Networks (CNN)

Convolutional neural networks [8] are widely used in a large number of domains. These have high popularity as compared to other neural networks. The different areas in which convolutional neural networks can be used are recommendation system, natural language processing, image processing etc. The advantage of convolution neural network is that it can figure out important features without human intervention. Example if there are different classes like mobile, tablet, laptop given to a CNN,it figures out the distinctive features for each class on its own.

2.4.2 Convolution

It is the process in which convolution [14] filter is applied on input to create a feature map. We slide the filter over input at every location to perform element wise matrix multiplication and the sum goes to the feature map. We continue this process till the entire feature map is filled.

2.4.3 Polling

Once the convolution step is over,polling [13] is performed to reduce dimensionality. This is done to reduce the number of parameters which in turn reduces the training time and also helps in dealing with the problem of overfitting. The common type of polling is max polling in which max value is taken from the polling window.

2.5 Neural Network

Artificial Neural Network (ANN) is inspired by biological neural networks that are present in an animal’s brain. This structure helps to perform tasks without being explicitly programmed for a specific task. ANN is composed of one to many separate layers which are basically collection of neurons. These neurons are responsible for processing input and providing output based on their activation function. The outputs of the neurons are fed into the next layer. This whole process repeats except for the output layer which takes input from previous layer and gives the result of the whole computation. One of the neural network widely used for pattern recognition is Convolutional Neural Network (CNN). CNN can be used to identify patterns in audios, images and videos.

2.5.1. Composing factors of CNN

2.5.1(a) Dropout: In dropout, some neurons are ignored by some probability during the training phase [9]. Edges of dropped neurons are not considered for training then. Dropout helps in reducing the network to prevent over fitting.

2.5.1 (b) Max pooling: Through max pooling, dimensions of data can be decreased by taking maximum input from original matrix [10].

2.5.2. Optimizing algorithms for CNN

2.5.2 (a) Adam: Adam optimization has become one of the most popular choice for optimizing image recognition and NLP models. This method computes learning rates for different parameters individually [11]. Adam optimizer starts reducing the cost from the first epoch.

2.5.2 (b) Stochastic Gradient Descent (SGD): SGD uses convergence to reduce the loss and randomly picks features during the training phase [12]. Randomly picking features obviates the need to keep record of already picked examples and features can be processed on the spot.

3. Result

A convolutional neural network (CNN) can be built using the above presented activation functions and optimization algorithms as loss functions having at least 3 layers resulting in no less than 50 % accuracy.

4. Conclusion & Future Scope

Accuracy of prediction depends upon the activation function used and configuring composing factors of CNN with over fitting and under fitting avoidance. CNN are able to achieve at least 50 % accuracy with little training and data. Pattern recognition and recommendation algorithms are in early stages and are yet to be explored further. As the demand for identification and recommendation will increase, need for introducing new features will increase.

References

[1] Sturm, Bob. (2013). The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use.

[2] Nwankpa, C., Ijomah, W., Gachagan, A., & Marshall, S. (2018). Activation Functions: Comparison of trends in Practice and Research for Deep Learning. arXiv preprint arXiv:1811.03378. 163: Ioffe, S., & Szegedy, C.

[3] Han J, Morag C (1995) The influence of sigmoid function parameters on the speed of backpropagation learning. Proceedings Series: Lecture Notes in Computer Science.

[4] Z. Tan, S. Zhou, J. Wan, Z. Lei, S. Z. Li, "Age estimation based on a single network with soft softmax of aging modeling" in ACCV, 2016.

[5] B. Logan, "Mel frequency cepstral coefficients for music modeling", *Proc. Int. Symp. Music Information Retrieval (ISMIR)*, 2000.

[6] Fujishima, Takuya (1999). "Realtime Chord Recognition of Musical Sound: a System Using Common Lisp Music". *Proceedings of the International Computer Music Conference*.

[7] Gouyon F., Pachet F., Delerue O. (2000),On the Use of Zero-crossing Rate for an Application of Classification of Percussive Sounds in *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-00 - DAFX-06), Verona, Italy, December 7–9, 2000*. Accessed 26 April 2011

[8] Hareesh Bahuleyan. Music genre classification using ma-chine learning techniques. CoRR, abs/1804.01149, 2018.

[9] Molchanov, Dmitry, Arsenii Ashukha, and Dmitry Vetrov. "Variational dropout sparsifies deep neural networks." *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017.

[10] Nagi, Jawad, et al. "Max-pooling convolutional neural networks for vision-based hand gesture recognition." *2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*. IEEE, 2011.

[11] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).

[12] Bottou, Léon. "Stochastic gradient descent tricks." *Neural networks: Tricks of the trade*. Springer, Berlin, Heidelberg, 2012. 421-436.

[13] Scherer, D., Müller, A., Behnke, S.: Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition. In: International Conference on Artificial Neural Networks (2010).

[14] Wu, J. (2017). Introduction to Convolutional Neural Networks. National Key Lab for Novel Software Technology Nanjing University, China.

[15] Shete, D., Patil, S., and Patil, P. (2014). " Zero crossing rate and energy of the speech signal of Devanagari script," IOSR-JVSP **4**(1).

[16] Schubert E, Wolfe J, Tarnopolsky A (2004) Spectral centroid and timbre in complex, multiple instrumental textures. In: Proceedings of 8th international conference on music perception & cognition.

[17] J. Urbano, D. Bogdanov, P. Herrera, E. Gomez, and ´ X. Serra. What is the effect of audio quality on the robustness of MFCCs and chroma features. In Proceedings of the International Society for Music Information Retrieval Conference, 2014.

[18] Lippens S, Martens JP, De Mulder T: A comparison of human and automatic musical genre classification. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '04), May 2004, Montreal, Quebec, Canada

[19] Umesh, S. & Cohen, Leon & Nelson, Douglas. (1999). Fitting the Mel scale. 1. 217 - 220 vol.1. 10.1109/ICASSP.1999.758101.

[20] Liu, Xuehao & Delany, Sarah & Mckeever, Susan. (2019). Sound Transformation: Applying Image Neural Style Transfer Networks to Audio Spectograms. 10.1007/978-3-030-29891-3_29.