Investigations on the Performance of Pre-established CNN Model in Music Emotion Detection

Yifan Yu *

Shanghai Pinghe High School, Shanghai, China

* Corresponding author email: yuyifan@shphschool.com

Abstract. Music is a medium for emotional artistic expression. Different people have different understandings of music. Music emotion recognition (MER) has thus become a novel branch in computer music. The goal of this essay is to investigate in the performance of established CNN architectures, such as AlexNet and VGG16, to recognize emotions contained in a song. CAL500 dataset is used as it covers a variety of genres. The dataset is transformed to spectrograms, which can be understood by computers through image recognition. The result of this investigation turned out to be that previous architectures would lead to overfitting within the training of a few batches. Possible explanations for this are that the parameters used in the model are too large for a simple regression task. This research provides some understanding of how CNN works as a network initially designed for image classification. Understanding emotions using spectrograms might require less complex CNN models or new models that are specialized in such tasks.

Keywords: Music Emotion Detection; CNN Model; Spectrogram; Regression.

1. Introduction

Music is a medium through which composers can express their attitudes and emotions. Though emotions in the song can be complex and might be different for everyone, some basic musical elements can still create a general tonality in music. To be specific, a piece of music written in minor would help develop a sorrowful mood even though the composer might want to express something more complicated than pure sadness. As technology develops, especially in the age of artificial intelligence, composing music is no longer a creation that only human beings are able to do. From Rudolf Zaripov's first algorithmic composition on the Ural [1] to the current deep learning model to recreate Bach pieces [2], algorithms have provided us with novel approaches to understanding music. Even though there has been so much work done, emotions in music are still a subjective term. Mankind's understanding of music emotions might be distinct since people might be under different conditions which contribute to different emotional judgments. However, computers can be regarded as unbiased judges, and using algorithms to understand emotions in music might shed some light on how we understand music.

In previous studies, researchers have already realized a lot in extracting some basic features of music. In most cases, previous feature selections were commonly based on components of music, including rhythm patterns, melody lines, harmony, color, and texture. The five features are extracted separately from the past. Rhythm patterns can be determined using beat detection algorithms [3], melody and harmony features are derived from extracting spectrograms of audio, and color and texture are illustrated in the waveform. To combine all the features, previous models widely adopted the use of spectrogram and MFCC features. Spectrograms show the distribution of frequencies in a song against time, which indicates the information of melody and harmony since pitches can be shown in frequencies. In addition, the rhythm pattern is also included indirectly in the model since the change in frequency against time is also shown in the spectrogram. To analyze the spectrogram, researchers have attempted to use the CNN model to recognize features contained in the model. Xin Liu et al. have applied the CNN model to the datasets CAL500 and CAL500exp [4, 5], and their model has hit a high accuracy.

Originally proposed by Fukushima Kunihiko [6] in 1980, CNN models have developed for a long time, with AlexNet by Alex Krizhevsky in 2012 [7], VGG by K Simonyan in 2014 [8], ResNet by K

He in 2015 [9], etc. These CNN models are initially designed to classify images into different categories rather than predicting the degree of each feature. This study would discover some qualities of these models in the music field.

Previous works done did have well-designed models and outstanding accuracy. However, their model only focused on the performance of a simple CNN model on different datasets. As a matter of fact, CNN has a lot more varieties. Different architectures have different accuracy in solving different tasks. Therefore, the correlation between CNN models and spectrogram was not well illustrated in previous models since the model used in too monotonous. Therefore, the essay's goal is to evaluate the use of pre-established CNN model architectures' performance, as well as that of a self-built one, on music emotion detection. This would provide a more in-depth understanding of how CNN recognizes emotions based on the spectrogram. The rest part of the paper is organized as follows. The Sec. 2 would discuss the data and method used, and the Sec. 3 would include results and some discussions of the investigation. Eventually, a brief summary is given in Sec. 4.

2. Data & Method

2.1 Data Set

CAL500 is a widely used dataset which is originally used by Douglas Turnbull et al [5]. This dataset has 502 different songs in 36 genres. Each song, different in length, is split into smaller sections, each section has a length of 10 seconds Setting this length helps regularize the length of each spectrogram. Approximately 8000 spectrograms of each section are extracted in MFCC features, using the FFT (Fast Fourier Transformation), with a sample picture shown below in Fig. 1.

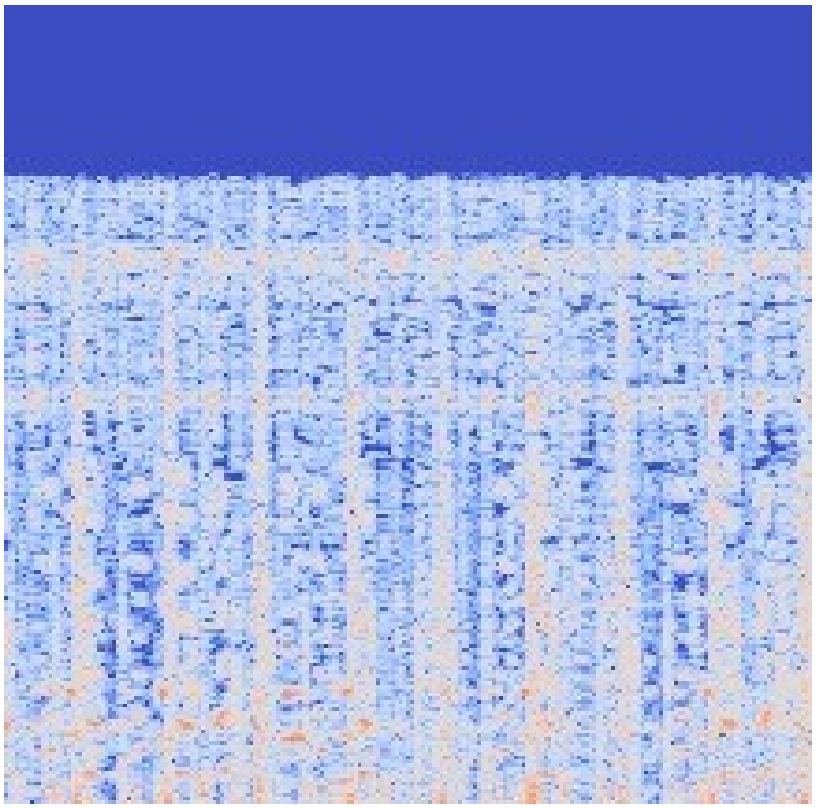

Fig 1. Sample Spectrogram**.**

CAL500 also provides annotations for each song, including different instruments used and emotions contained in the song. Emotions are evaluated in degrees from 1 to 5 following the rules presented Table. 1.

Table 1. The example of values for different labels.

Labels	Values
songName	2pac-trapped.mp3
Emotion-Happy	2
Emotion-Sad	3
Emotion-Calming_/_Soothing	1
Emotion-Arousing_/_Awakening	5
Emotion-Pleasant_/_Comfortable	2
Emotion-Cheerful_/_Festive	1
Emotion-Tender_/_Soft	1
Emotion-Powerful_/_Strong	5
Emotion-Loving_/_Romantic	1
Emotion-Carefree_/_Lighthearted	1
Emotion-Exciting_/_Thrilling	3
Emotion-Emotional_/_Passionate	4
Emotion-Positive_/_Optimistic	1
Emotion-Touching_/_Loving	1
Emotion-Light_/_Playful	1
Emotion-Angry_/_Agressive	5
Emotion-Laid-back_/_Mellow	1
Emotion-Bizarre/_Weird	2

2.2 CNN Model

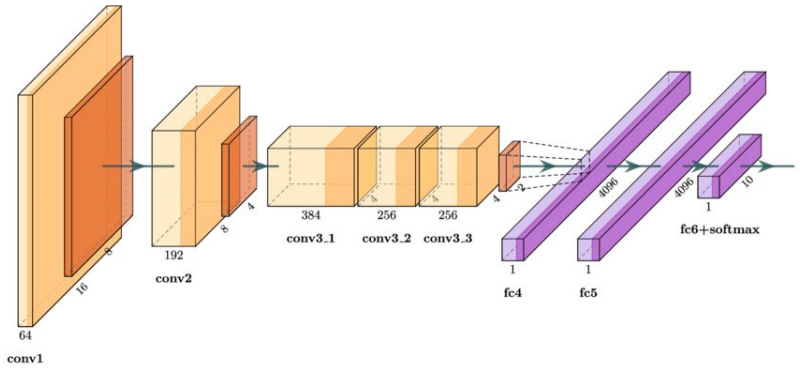

Fig 2. AlexNet Flowchart.

Fig 3. VGG16 Flowchart.

As discussed in the previous section, the CNN model is applied to recognize features in the spectrograms. One typical feature in CNN is the use of convolutional kernels. Convolutional kernels would simplify and combine the information contained in the kernel and thus reduce complexity and improve accuracy. When analyzing the spectrogram, melody lines and harmony features would be

condensed in the kernels and then analyzed to fit the prediction, similar to the human analysis of a music piece. This study mainly uses three different CNN architectures: Alexnet [7], and VGG16-bn [8], ResNet18 [9]. Then, their performance in the task is evaluated. In addition to that, a self-built model is also tested on the dataset. Figs 2-5 are the architectures of the used CNN models [10–12]. The output layers in the flowcharts are left original, but in the study, the layers are changed to regression.

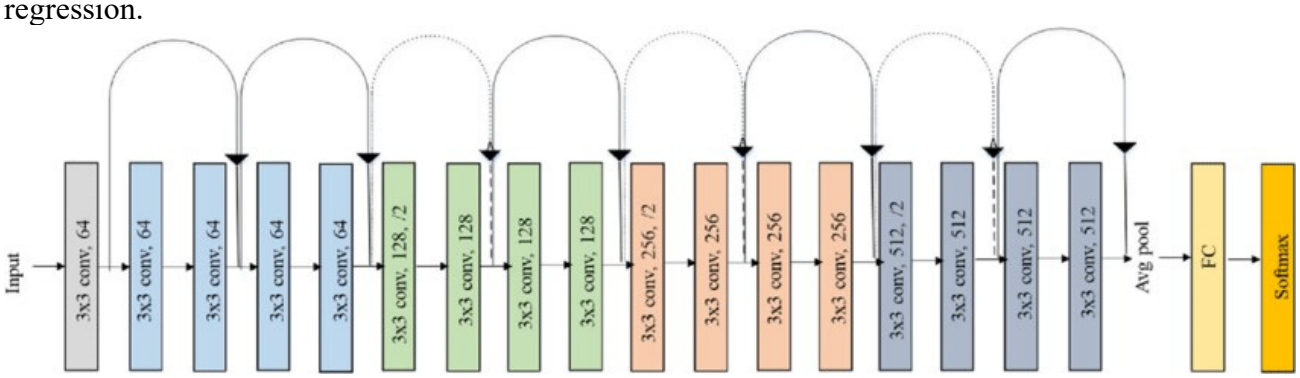

Fig 4. ResNet18 Flowchart.

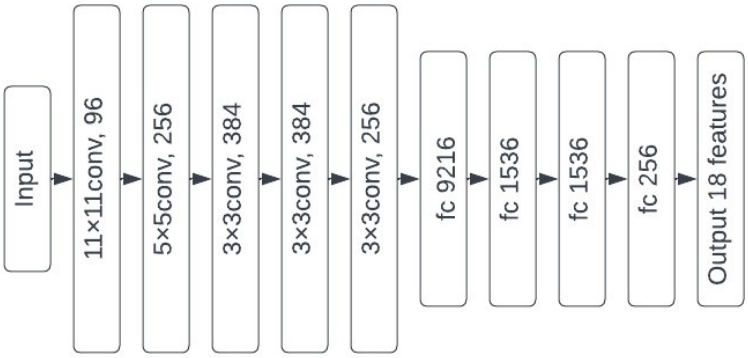

Fig 5. MyModel Flowchart.

2.3 Metrics

For the loss of Alexnet, ResNet, and VGG16 model, Mean Square Error Loss(MSELoss) is used. But for the self built model, L1Loss is used since comparing to MSELoss, L1Loss is linear, very robust and performs well towards the end of training. The formulae are presented as follows:

$$MSELoss: l = L = \{l_1, \dots, l_n\}^T, l_n = (x_n - y_n)^2 \quad (1)$$

$$L1Loss: l = L = \{l_1, \dots, l_n\}^T, l_n = |x_n - y_n| \quad (2)$$

The model would predict the coefficients in each feature with a value between 0 and 1. To calculate the accuracy, the deviation between the predicted value and the annotated value is calculated. Accuracy is then calculated as the ratio of deviation to the sum of annotated values.

2.4 Processing and Training

To simplify the training, data values are normalized into [0,1] span using the following methods: emotion tags values are extracted from each annotation file and divided by 5; each image loaded as RGB values are divided by 255.0. Before data is fed into the model, randomization and data augmentation are employed to make the training more robust. Subsequently, this study splits the dataset into the training set and the validation set with a ratio of 0.8. Then, three different initial models are loaded without being pre-trained. The output layer of the imported PyTorch model has

1000 output classifications. To suit the CAL 500 dataset, the output is then changed to 18 features. This study trains the models for 10 epochs, respectively. Throughout the course of training, the accuracy of the models is shown accordingly.

3. Results & Discussion

3.1 Results & Comparsion

The results of the three model architectures and our CNN model are listed in Table. 2. Seen from the results, our model has the highest accuracy (about 80.2%) which is about 2 times that of other models, indicating the well-performance of the model improvements.

Table 2. Results of Each Model

Model	AlexNet	ResNet	VGG-16	MyModel
Accuracy(percent)	52.3	41.7	51.9	80.2

3.2 Evaluations

These CNN models are initially designed to solve complex image recognition and classification tasks. However, for the music emotion recognition task, the large fully connected layers tend to be overly complex. To elaborate, as loss decreases, the accuracy doesn't improve much. One explanation of this is that the model is overfitting the dataset, leading to both low loss and low accuracy rates. In addition, the accuracy of the model is around 50 percent. This means that the deviation of the prediction is around half of the sum of the coefficients, which implies that the model might be predicting the value that is always around half of the maximum: 0.5 in this case. This prediction is also known as the expected value of the prediction. Just like tossing a dice, the expected value is 3.5, and the expected value of the coefficient is 0.5. Still, this phenomenon can also be explained by the over-fitting theorem.

3.3 Limitations and Strengths

The use of established image recognition neural network architecture seems to be not that successful. The problem of the existed model architecture should be neglected since it's the goal of the essay. Aside from that, other problems and limitations would account for the problem. First is the dataset. CAL500 only contains 500 songs and their annotations. From that dataset, almost 8000 spectrograms are extracted. Still, the dataset might be small for a complex model, e.g., AlexNet and ResNet. In addition, the amplitude, or the loudness, of the songs is not regularized, leading to a considerable distinction in color between songs that have similar features or structures. Secondly, splitting each song by 10 seconds might only contain limited musical information as some songs would repeat a single motive for a longer time, which might affect the accuracy of the prediction. Moreover, in music theory, some songs might change keys, leading to more complex and varying emotions in a single song. Nevertheless, the essay assumes that the emotions are coherent in each song, which is in fact a casual assumption. Finally, the problem of the texture and color of the song is not well considered. Songs that have a similar harmony(chord progression), which can be understood by the model, but different arrangements, faster or slower, brighter or darker, might all contribute to different emotions that can be perceived by human ears but not algorithms through spectrograms.

Though there are so many limitations, the model might shed some light on computer music composing. Previous models are designed to mimic the function of the human optic nerve system, which in nature is separated from the auditory nerve system, and so should the algorithms. On this basis, a more close-to-nature algorithm is perhaps needed for better performance in music emotion recognition. In addition, MyModel has successfully proved that a simpler CNN model would be competent in such task.

3.4 Ethics

When it comes to ethics, the model itself is designed to be unbiased. The predictions is all based on spectrograms, which is a pure analysis of composition, with no personal judgement on the composer's gender or race. In reality, some biased might exist due to the song choices in CAL500 dataset, which is inevitable. To further improve ethics in the study, a better dataset should be used.

4. Conclusion

In summary, this essay investigates the use of established CNN architectures as well as our new proposed CNN architecture in music emotion detection. Specifically, AlexNet, ResNet-18, and VGG-16 are used to predict the components of emotions contained in a song with a value between 0 and 1. The result is not that satisfying since the models are initially designed to categorize images rather than spectrograms used in musical analysis. The problems of the model are explained as the limitations of the investigation, including the deficiencies in the dataset methods used. Overall, these results offer a guideline for future explorations in music emotion recognition using the CNN model. Distinctions between models designed for complex image classification and spectrogram analysis are well explained in the essay.

References

[1] Zaripov R K. An algorithmic description of a process of musical composition. Soviet Physics Doklady. 1960, 5: 479.

[2] Howcroft Jacob. Celebrating Johann Sebastian Bach, March 2019.

[3] Cheng K. Beat This: A Beat Synchronization Project: Beat Detection Algorithm. Rice University, Houston, TX, retrieved from: http://www. owlnet. rice. edc/elec301/Projects01/beat--sync/beatalgo. html, 6.

[4] Liu X, Chen Q, Wu X, et al. CNN based music emotion classification. arXiv preprint arXiv:1704.05665, 2017.

[5] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics, 36(4):193– 202, April 1980.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In Advances in Neural Information Processing Systems, volume 25. Curran Associates, Inc., 2012.

[7] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition, April 2015. arXiv:1409.1556.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, December 2015. arXiv:1512.03385.

[9] Douglas Turnbull, Luke Barrington, David Torres, and Gert Lanckriet. Semantic annotation and retrieval of music and sound effects. IEEE Transactions on Audio, Speech and Language Processing, 16(2):467–476, February 2008.

[10] Nicola Strisciuglio, Manuel Lopez Antequera, and Nicolai Petkov. Enhanced robustness of convolutional networks with a push–pull inhibition layer. Neural Computing and Applications, 32:1–15, 12 2020.

[11] Farheen Ramzan, Muhammad Usman Khan, Asim Rehmat, Sajid Iqbal, Tanzila Saba, Amjad Rehman, and Zahid Mehmood. A deep learning approach for automated diagnosis and multi-class classification of alzheimer's disease stages using resting-state fmri and residual neural networks. Journal of Medical Systems, 44, 12 2019.

[12] Minhaz Ahmed, Yeong Kim, Jin Woo, Rezaul Bashar, and Phill Rhee. Two-person interaction recognition based on effective hybrid learning. KSII Transactions on Internet and Information Systems, 13, 03 2019.