Attention-based Convolutional Neural Network for Music Genre and Mood Classification

Tingyi Li
tingyi@kth.se

Abstract

In recent years, the interest of Music Information Retrieval based on Deep learning has drastically increased while traditional MIR techniques remain difficult, non-universal and sometimes proprietary. This paper aims to further the line of researches by predicting the genre and valence mood of the audio simultaneously by suing a multi-output CNNs to learn the features of mel spectrograms generated from the audio. Baseline models from previous papers have been experimented on and made comparison. In addition, an attention-based CNN model derived from the basic model has been explored. For evaluation, both accuracy and confusion matrices were provided to measure the performance. Results show that attention-based CNN architecture could work pretty well for both music genre and mood classification.

1 Introduction

1.1 Background

Automatic Music Information Retrieval has drawn increasing attention as the ever growing musical data in the world. Traditional MIR techniques such as Support Vector Machines etc. require extracting hand-crafted features from the original songs, which demands a high level of expertise in audio engineering[1,2,3].

In recent years, deep learning methods have become more popular in MIR research. Convolutional neural networks (CNNs) have been actively used for music classification tasks [4,5,7,8]. In the process of extracting patterns in 2D based[6], feature extraction is crucial and CNNs can greatly improve it, which owes to its stacked structure: the output of a convolutional layer is fed into the input of the next layer. Besides, the subsampling layer is another crucial part of CNNs, since it can aggregate features and reduce dimensionality to reduce the computational complexity. Although the use of CNNs has produced satisfactory results in MIR field [1,9,10], it still has some noteworthy problems: in the encoding-decoding structure, the encoder will encode all the input sequences into fixed length vectors. Fortunately, a recent trend in Deep Learning named Attention Mechanisms is very helpful to solve this problem [11]. Attention-based networks have now been applied to a variety of fields, such as handwriting synthesis, machine translation, etc. The way this network works is to select relevant content in each iteration. This method significantly improves the network's ability to process longer input sequences [12]. However, few papers have applied in music genre classification based on audio samples, let alone attention based CNNs in music genre classification.

For mood classification, Liu et al have applied CNN to predict the emotions of songs over eighteen different emotions and reached an accuracy of 71%[13]. However, very few papers have talked about this topic. The lack of studies may be because mood data is hard to collect. Thus this paper uses Spotify's valence metric to investigate this topic.

In this paper, CNNs were applied to build and train a model for music genre classification on FMA dataset. Experiments on various parameters and comparison between current model and baseline models have been carried out. Furthermore, an attention-based convolutional neural network for both genre and mood has been investigated. As a result, the accuracy for genre classification was around 90% and for valence mood classification around 80%. With an additional attention layer, the accuracy was improved to 92% for genre classification while valence classification remains unchanged.

1.2 Related Work

Multiple researchers have done music classification previously. Zhang et al investigated music genre classification using a CNN model with three convolutional layers and three dense layers. In addition, they employed average between max pooling and average pooling to provide more information to the higher level statistical layers. They reached 85% accuracy[17]. While Choi et al used five convolutional layers and one dense layer in their model for genre classification[16]. These are the two baseline models used in this paper.

Also, Choi et al extended their model to convolutional recurrent neural network in future research[10]. They combined CNNs with RNNs, CNNs for feature extraction and RNNs for temporal summariser. The reason why they use the network structure in this way is that RNNs is more flexible in summarizing local features than CNNs. Because different types of music can be influenced by different types of features, such a structure can bring more benefits [4]. Furthermore, Thomas Lidy et al utilizes two different CNNs architectures: a sequential one, and a parallel one, to implement music genre and mood classification. They compares the results come out of these two architectures and found that the latter one achieved higher accuracy, because this structure can record temporal information and timbral relations at the same time. This approach is for MIREX 2016 Train/Test Classification Tasks [6], while Thomas Lidy had implemented spectral convolutional neural network to solve MIREX 2015 music/speech classification problems. This method is mainly used to process input information since the "raw" data from an audio cannot be used as input data directly [14].

As for attention mechanism, there are also many papers on this topic. Jan Chorowski et al extends the attention mechanism to enable it to be used in speech recognition [12]. They utilizes attention-based models to accomplish the TIMIT phoneme recognition task. At every step (time) in generating an output sequence, an attention mechanism is used to select the feature vectors produced by the trained feature extraction mechanism at potentially all of the time steps in the input sequence. The weighted feature vector then helps to generate the next element of the output sequence. Besides, Sanghyun Woo et al [15] conducts convolutional block attention module, which combines CNNs with attention mechanism. What's more, other related work include the work by Alexandros Tsaptsinos, where he uses hierarchical attention network to classify music genre by lyrics [8].

2 Method

In this part, the models used in this paper would be generally introduced in theory as well as some mathematical intuitions.

2.1 CNN

The models applied in this paper to predict genre categories from music are depicted in Figure 2.1.

The baseline models are from Choi et al [16] and Zhang et al [17]. In this paper, the first part focuses on improving the CNN models composed of four convolutional layers, four max pooling layers.

In detail, the convolutional layers were accomplished through applying a number of filters, which is also called kernels, to the input matrix. It convolves the input matrix with each filter, sliding the filter over all spatial locations and computing the dot product of the filter and the convolved input matrix. Since Zhang et al chose 1×1 while Choi et al chose 3×3, 2×2 kernel was chosen in this paper to avoid overfitting. The filter sizes for each layer were 64, 128, 256 and 512 in this paper.

The spectrogram describes the time-frequency characteristic of audio signal and a deep CNN can learn its local features about frequency distribution and variety from spectrogram. Moreover, the stacked convolutional layers could learn combination of different local features. Also, a categorical cross entropy loss was utilized to minimize the distance between the CNN output distribution and the correct label with an Adam optimizer, which is a method for efficient stochastic gradient optimization[20]. The categorical cross entropy loss measures the dissimilarity between the true label distribution y and

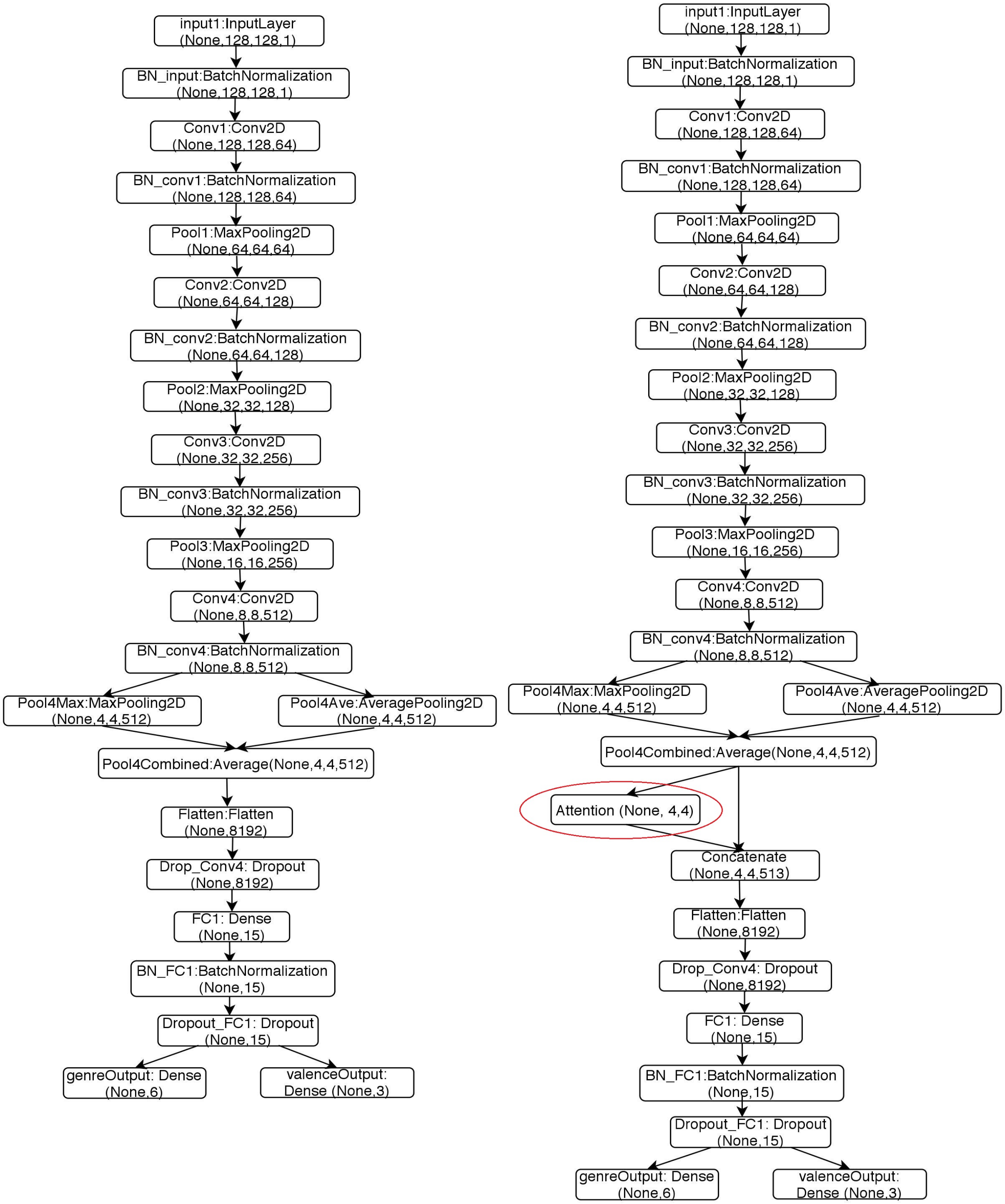
input1:InputLayer (None,128,128,1)
BN_input:BatchNormalization (None,128,128,1)
Conv1:Conv2D (None,128,128,64)
BN_conv1:BatchNormalization (None,128,128,64)
Pool1:MaxPooling2D (None,64,64,64)
Conv2:Conv2D (None,64,64,128)
BN_conv2:BatchNormalization (None,64,64,128)
Pool2:MaxPooling2D (None,32,32,128)
Conv3:Conv2D (None,32,32,256)
BN_conv3:BatchNormalization (None,32,32,256)
Pool3:MaxPooling2D (None,16,16,256)
Conv4:Conv2D (None,8,8,512)
BN_conv4:BatchNormalization (None,8,8,512)
Pool4Max:MaxPooling2D (None,4,4,512)
Pool4Ave:AveragePooling2D (None,4,4,512)
Pool4Combined:Average(None,4,4,512)
Flatten:Flatten (None,8192)
Drop_Conv4: Dropout (None,8192)
FC1: Dense (None,15)
BN_FC1:BatchNormalization (None,15)
Dropout_FC1: Dropout (None,15)
genreOutput: Dense (None,6)
valenceOutput: Dense (None,3)
input1:InputLayer (None,128,128,1)
BN_input:BatchNormalization (None,128,128,1)
Conv1:Conv2D (None,128,128,64)
BN_conv1:BatchNormalization (None,128,128,64)
Pool1:MaxPooling2D (None,64,64,64)
Conv2:Conv2D (None,64,64,128)
BN_conv2:BatchNormalization (None,64,64,128)
Pool2:MaxPooling2D (None,32,32,128)
Conv3:Conv2D (None,32,32,256)
BN_conv3:BatchNormalization (None,32,32,256)
Pool3:MaxPooling2D (None,16,16,256)
Conv4:Conv2D (None,8,8,512)
BN_conv4:BatchNormalization (None,8,8,512)
Pool4Max:MaxPooling2D (None,4,4,512)
Pool4Ave:AveragePooling2D (None,4,4,512)
Pool4Combined:Average(None,4,4,512)
Attention (None, 4,4)
Concatenate (None,4,4,513)
Flatten:Flatten (None,8192)
Drop_Conv4: Dropout (None,8192)
FC1: Dense (None,15)
BN_FC1:BatchNormalization (None,15)
Dropout_FC1: Dropout (None,15)
genreOutput: Dense (None,6)
valenceOutput: Dense (None,3)

Figure 1: CNN architecture

the predicted label distribution $\hat{y}$, and is defined as cross entropy:

$$L(y, \hat{y}) = -\sum_i y_i log(\hat{y}_i)$$

The second part focuses on exploring the effect of attention mechanism on CNN for music genre classification. Thus for the second part, an additional attention layer was added to the previously improved CNN. The CNN learns the representation of the audio signal, while the attention layer computes the weighted sum of all the information extracted from different parts of the input. The output from the max pooling layer and the attention vector are then fed into a fully connected softmax layer together.

2.2 Attention-based model

For each vector x_i in a sequence of inputs x, the attention weights α_i can be computed as follows

$$\alpha_i = \frac{exp(f(x_i))}{\sum_j exp(f(x_i))}$$

where $f(x)$ is the scoring function. In this work, $f(x)$ is the linear function $f(x) = W^T x$, where W is a trainable parameter. The output of the attention layer, attentive x, is the weighted sum of the input sequence.

The reasons to add attention mechanism for music genre classification is that the genre of music is discriminative by its signal spectral characteristics and audio samples from the same genre should share some similar patterns or distributions where different parts of the signal should be given different weights intuitively. In order to improve accuracy, attention mechanism could be applied to the extracted features from convolutional layers, through which the features distribution could be better learned. Therefore, with the additional parallel attention layer, weights to features from different feature maps could be assigned and then for each feature map, weighted sum are then calculated.

Then, the output vector from the max pooling layer and the attention vector are concatenated together and fed further for the final softmax layer. It is worthy to be noted that the reason to concatenate them together is because since the input signals are noisy for the most time, another max pooling layer was added after the convolutional layers to help only select the most salient features and filter noises.

3 Experiments

3.1 Collecting Raw Data

Dataset for this paper is from FreeMusicArchives(FMA). The valence metric is queried from Spotify API. The FMA dataset has 30 second audio samples for a variety of genres, which is a lot less organized than the traditional GTZAN dataset that most MIR practotioners work with. Moreover, FMA data indicates the song name and artist for each song while the GTZAN data do not. In this way, the valence metric could be easily queried from Spotify through their API. The labels this paper focuses on are as follows:

Blues, Classical, Country, Disco, Hip-hop, Jazz, Metal, Pop, Reggae, Rock

These genres were chosen because they were relatively abundant data in FMA dataset. 75 songs from each of the 10 genres were taken randomly to ensure a balanced dataset.

3.2 Data preprocessing

After collecting the raw data, the valence metric could be looked up from Spotify for each song. Valence ranges from 0 to 1 where 0 is the saddest and 1 is the happiest. Measures were taken to ensure that all the song used actually existed in Spotify. The data set had the following valence distribution as in Figure 2.

The valence data was labeled into "Sad" where valence < 0.3, "Happy" where valence ≥ 0.6 and "Neutral" otherwise. This boundary values were chosen because equal number of songs in each subset could be achieved as songs in this data set tends to have lower valence scores.

In this way, a mood label as well as a genre label for each song was assigned in the dataset. The reason why these labels were chosen is because running a multi-ouput CNN where both are categorical cross-entropy is much easier to train than a CNN with one categorical and one regression output. Also the results would be easier to interpret since a categorical accuracy is more intuitive for mood classification rather than a R squared.

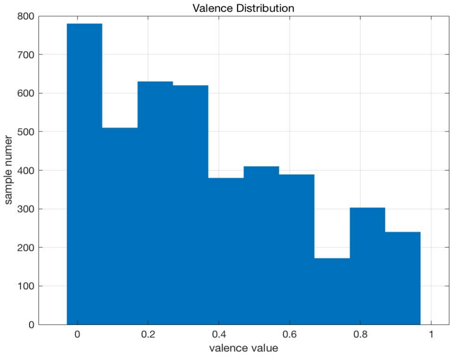

Figure 2: Valence Distribution

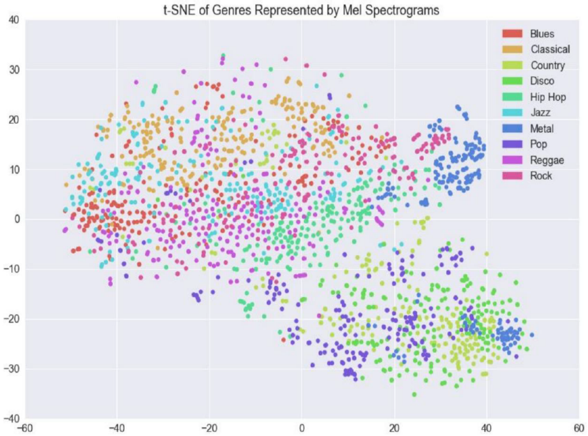

Figure 3: t-SNE representation of the 10 different genres as represented by mel spectrogram

3.3 Feature Extraction

The next step is to extract Mel-Spectrogram as features. Although mel-frequency cepstral coefficients(MFCCs) is one of the audio features usually used in Speech Technology and music genre classification. Most researchers decide to use MFCC features in an attempt to decorrelate the input data to avoid overfitting [18]. However, in this paper, mel-spectrograms were used for the following three reasons.

First, information about the audio signal could be kept as much as possible because the discrete cosine transform eliminates a lot fine details about the original audio. Second, the MFCC features would be beneficial with linear models such as Gaussian Mixture Models(GMMs). But in this paper, CNNs is a strong classifier and the mel-spectrograms are more perceptually intuitive and size efficient. Third, standard regularization techniques such as dropout, L_2 weight decay and maxnorm should suffice. Thus, mel-spectrograms would perform better in this case.

Each audio sample is 30 seconds. Firstly, the whole dataset consist of 750 different songs from 10 genres was splited randomly and exclusively into 90% training set, 5% validation set and 5% test set. Then each song is converted into Mel-spectrograms and divided into 10 slices of 3 seconds, which has the size of $128 \times 128 \times 1$ each. The resulting spectrogram describes the frequency contours as a function of time. Mathematically speaking, the spectrogram represents the time evolution of the 128 amplitudes in the Fourier basis in steps of 20ms with 128 mel frequencies. In other words, each of the mel-spectrogram for a sample has size of 128 by 128 by 1.

3.4 Visualization

Visulization before training is pretty important as it provides intuition on what results should be expected and how well the network works.

In this paper, before training the CNN, mel spectrogram datasets were visualized using the t-Distributed Stochastic Neighbor Embedding(t-SNE) algorithm. t-SNE is a very popular algorithm to reduce dimensionality for high-dimensional data[19]. This algorithm has been widely used in the field of machine learning, because it can effectively transform high-dimensional data into two-dimensional images. In this way, it could reduce the dimensionality of the input data's parameter space while preserving the similarity between examples. The result is shown in Figure 3.

Intuitively from the figure, it is distinct that Pop, Disco and Country form a discriminative group from the others. In this way, t-SNE helps better understand the dataset.

3.5 Experiment Setup

In this paper, Keras was utilized to build CNNs architecture for training. Keras is a deep learning framework based on Tensorflow, Theano and CNTK backends.

In this experiment, the default parameter setting is as follows:$learning\_rate = 1e - 3$, $training\_epochs = 70$, $\lambda = 1e - 3$, $batchsize = 200$. In the first section, experiments on learning rate, batch size, λ ,filter size etc were carried out. For the second part, the default setting is applied.

4 Results

In this section, results of the experiments are presented. Furthermore, tests on different parameters and various model architectures were carried out.

4.1 CNN architecture

Since models from Zhang et al [17] and Choi et al [16] were chosen as baseline models, both the three and five layer CNN were applied on FMA dataset and it turns out zhang's model is underfitting while Choi's is overfitting. It made sense since Zhang et al use three layer CNN while Choi et al use five. Thus, 4 convolutional layers was used in this network. Also, considering both papers, a 2×2 kernel were applied. Only one fully connected layer was used since more would lead to be overfitting. Also, batch normalization was applied to each layer as it drastically accelerates the training speed. Also, a dropout rate of 50% for flattened layer and the dense layer were added in order to regularize the model. Both averaging and max pooling were applied since it would increase accruacy. For the first part of the experiment, a grid search on various parameters was carried out. The results are shown in Table. 1.

Table 1: Paramter testing

test	Learning Rate	Lambda	BatchSize	Genre acc	Valence acc
1	0.01	0.001	100	0.820	0.731
2	0.01	0.002	100	0.744	0.692
3	0.01	0.003	100	0.782	0.683
4	0.01	0.001	200	0.801	0.723
5	0.01	0.002	200	0.756	0.788
6	0.01	0.003	200	0.796	0.676
7	0.001	0.001	100	0.871	0.799
8	0.001	0.002	100	0.853	0.771
9	0.001	0.003	100	0.833	0.790
10	**0.001**	**0.001**	**200**	**0.889**	**0.808**
11	0.001	0.002	200	0.867	0.791
12	0.001	0.003	200	0.852	0.764

From the results, the best performance is test 10 with learning rate of 0.001, batch size of 200 and lambda of 0.001. The test loss function and accuracy and training loss and accuracy are shown in Figure 4 and 5.

The comparison between network in this paper and the baseline models are as follows: From the result in Table 2, the 3 layer network from Zhang et al [17] is obviously underfitting since both training accuracy and test accuracy is less satisfactory while 5 layer network from Choi et al[15]. is overfitting since it achieves a high training accuracy with a low test accuracy. It shows that network in this paper works pretty well.

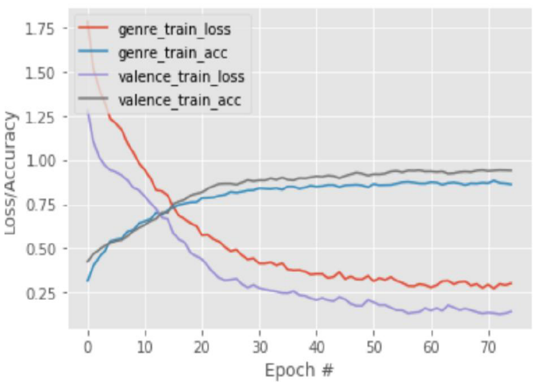

Figure 4: Training loss and accuracy for best performance

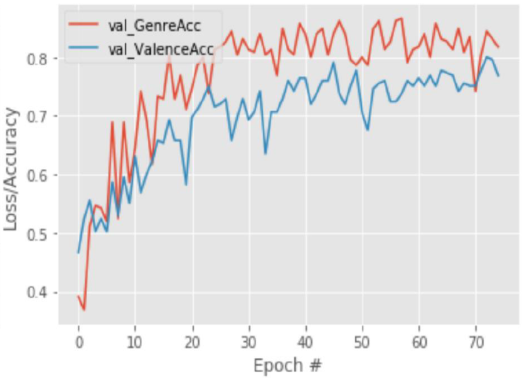

Figure 5: Test accuracy for best performance

Table 2: Comparison with baseline models

Experiment	Genre		Valence	
Model	TrainAcc	TestAcc	TrainAcc	TestAcc
3 layer	0.473	0.384	0.482	0.402
5 layer	0.968	0.820	0.953	0.712
4 layer(this project)	0.908	0.889	0.907	0.808

4.2 Attention-based CNN

In this section, an attention-based model with default parameter setting was trained. The result accuracy is 96.68% for training set and 89.78% for test set in genre classification. Although the improvement on accuracy is not as pronounced as attention mechanism in NLP, the model achieves good results and since the CNN network in this paper is pretty sophisticated, attention mechanism may have less influence than expectation.

For valence classification, the accuracy is almost the same which makes sense because valence classification is less related to spectral characteristics.

The resulting matrices for both valence and genre are shown in Figure 6 and 7

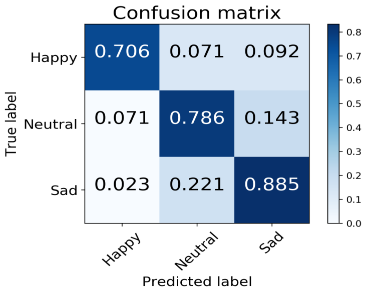

Figure 6: Confusion Matrix for valence mood

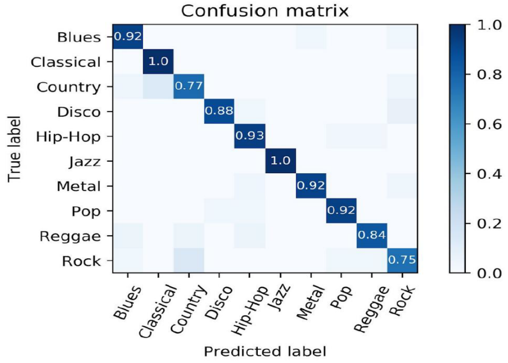

Figure 7: Confusion matrix for genre

From both of the confusion matrices, the results are very satisfactory since there is a distinct diagonal line in each matrix respectively, which means most of the predictions corresponding to each class are correct. Especially for Classical and Jazz, a 100% accuracy was achieved. While Rock and Country seem less easy to classify which may be because they share similar audio patterns during short amount of time(3 seconds). As for valence classification, sad songs are easy to classify while happy songs are relatively hard to distinguish. This may be because less songs were presented in Happy category due to the skewed division of database in order to maintain the balanced distribution.

5 Discussion & Conclusion

In this paper, a sophisticated CNN which could classify music into 10 genres with around 90% accuracy and into 3 valence mood genre with around 80% accuracy. Moreover, further experiment on an attention-based CNN was carried out. The results show that CNNs are good model for genre and mood classification while the results for attention-based model are good although the improvement is not as large as expectation, which may be because the CNN architecture has achieved high accuracy alone.

The possible reasons why this paper reaches a high accuracy are as follows. Firstly, based on previous researches, this paper could gain enough knowledge on how to build a model with satisfactory results from previous papers. Secondly, as stated in the data section, the dataset used in this paper is not large enough and it seemed that the distribution of the songs are skewed a little bit. In this way, there might be some similar samples in both training and test sets, which leads to high accuracy.

In the future, I plan to extend my model to more areas in feature prediction and classification. Also, improvements on datasets coudl be done, such as extending the dataset and preparing an evenly distributed dataset for valence classification. Also, since I only grasp a general idea of attention mechanism, I could carry out different methods to implement the attention layer in order to better extract the weighted features in the future. Since attention mechanism is a strong tool in deep learning, I am very excited about doing more exploration in this direction.

Acknowledgments

In writing this paper, I would avail myself of the opportunity to extend my sincere thanks to my professor, from whose teaching and instruction I have the ability to complete this paper. Besides, I would also like to acknowledge my indebtedness to TAs, who have contributed their time to this report.

Link to Repository:`https://github.com/TingyiLi/dt2119_final_project`

References

[1] Jose Homsi Goulart, Antonio Guido, Rodrigo Maciel, Carlos. (2012). *Exploring different approaches for music genre classification*. Egyptian Informatics Journal.

[2] Pachet, F., Cazaly, D. (2000) *A classification of musical genre*. Content-Based Multimedia Information Access (RIA) Conference, Paris.

[3] Welsh, M., Borisov, N., Hill, J., von Behren, R., and Woo, A(1999) *Querying large collections of music for similarity*. Technical Report UCB/CSD00-1096, U.C Berkeley, Computer Science Division

[4] Keunwoo Choi. & György Fazekas. & Mark Sandler. & kyunghyun Cho (2017) *Convolutional recurrent neural networks for music classification*, IEEE.

[5] Keunwoo Choi. & György Fazekas. & Mark Sandler. (2016) *Explaining deep convolutional neural networks on music classification.* arXiv preprint arXiv:1607.02444, 2016.

[6] Thomas Lidy. & Alexander Schindler. (2016) *Parallel Convolutional Neural Networks for Music Genre and Mood Classification.* MIREX2016.

[7] Cory McKay. & John Ashley Burgoyne. & Jason Hockman. & Jordan B. L. Smith. & Gabriel Vigliensoni. & Ichiro Fujinaga. (2010) *Evaluating the Genre Classification Performance of Lyrical Features Relative to Audio, Symbolic and Cultural Features.* ISMIR 2010, Utrecht, Netherlands.

[8] Alexandros Tsaptsinos. (2017) *Lyrics-Based Music Genre Classification Using a Hierarchical Attention Network.* arXiv:1707.04678.

[9]Sander Dieleman. & Benjamin SchrauIn. (2014) *End-to-end learning for music audio.* IEEE

[10]Keunwoo Choi. & Gyӧrgy Fazekas. & Mark Sandler. (2016) *Automatic tagging using deep convolutional neural networks.* ISMIR2016

[11] Denny Britz. (2016) *Attention and Memory in Deep Learning and NLP* WILDML

[12] Jan Chorowski. & Dzmitry Bahdanau & Dmitriy Serdyuk &Kyunghyun Cho & Yoshua Bengio (2015) *Attention-Based Models for Speech Recognition* NIPS2015

[13] Liu et al. (2018) *CNN based music emotion classification*,http://arxiv.org/abs/1704.05665

[14] Thomas Lidy. (2015) *Spectral Convolutional Neural Network for Music Classification.* MIREX2015.

[15] Sanghyun Woo. & Jongchan Park. & Joon-Young Lee. & In So KIon. (2018). *CBAM: Convolutional Block Attention Module.* arXiv:1807.06521.

[16] Choi, K. & Fazekas, G. & Cho, K. & Sandler, M.B. (2018). *A Tutorial on Deep Learning for Music Information Retrieval.* CoRR, abs/1709.04396.

[17] Zhang, W. & Lei, W. & Xu, X. & Xing, X. (2016). *Improved Music Genre Classification with Convolutional Neural Networks.* INTERSPEECH.

[18] "Mel-Frequency Cepstrum." Wikipedia, Wikimedia Foundation, 12 May 2018, en.wikipedia.org/wiki/Mel-frequency_cepstrum.

[19] Van der Maaten, L. and Hinton, G. (2008) *Visualizing Data Using t-SNE*. Journal of Machine Learning Research, 1, 1-48.

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

6 Modification based on Peer Review

In this section, I presented the changes I've made after peer review. Since the comments are pretty long, only critiques have been presented here and my modification has been stated.

6.1 Literature study 5 / 6 pts

1. **Comments:**
   However, no traditional methods are described. For example, an early overview of the music information retrieval would be a good start and then some methods that use traditional classification methods (e.g. Support Vector Machines), their disadvantages and then Deep Neural Networks that are superior.
   Modification:
   Based on the advice, I added a few papers on early music genre classification. But since the paper is mainly about MIR based on CNNs, I didn't talk much about it.

6.2 Novelty/Originality 3 / 6 pts

1. **Comments:**
   The novelty of the report is not very high since it is not taking aim to extend any of the state of the art papers on music genre classification. In fact, when looking into the purpose of the study it is rather unclear what the authors would like to investigate and perhaps that is why the method feels a bit ad-hoc when it comes to the selection of network architecture.
   Modification:
   The aim is this paper is to try to design a sophisticated model for music genre and mood classification system based on the two baselines models from previous papers. There are two reasons why I would like to investigate this topic: Firstly, music genre classification based on audio samples has always been a hard issue compared to those from NLP. Secondly, attention mechanism has become a hot topic in recent years and most papers combine it with RNN and LSTM. However, recently, researchers have found that attention mechanism could also be combined with CNNs to achieve good results in MIR and NLP etc.

   Based on the advice, I added more papers and explanation on the purpose of the paper in the introduction and abstract parts.
2. **Comments:**
   The author creates a baseline model using a CNN architecture and then adds an attention mechanism to boost the results. These ideas are not new as mentioned in the report and have been implemented before. However, it is interesting, given the dataset, to see how a network can achieve good results and with which hyper-parameter settings. Unfortunately, the report does not state exactly what changes were made to the models when compared to the cited models that the author was inspired, so it makes it hard to evaluate how novel the proposed architecture is.
   Modification:
   This was one mistake that I've made. Based on the advice, I added the detailed implementation of the baseline models and how I chose the parameters of network in this paper.

6.3 Correctness 5 / 6 pts

1. **Comments:**
   However, the report tries to explain the elements that contribute to a Convolutional Neural Network using the baseline and the attention based CNN. This explanation does not contain how the attention layer was integrated into the network and why. It is confusing seeing the diagram of the network that the previous layer (Pool4Combined) is concatenated with the attention layer. Why concatenate the layers ? I think it is a basic concept and should be explained.

   Moreover, the author uses t-SNE to visualize the data which is not introduced when first mentioned and explained after its mention.
   Modification:
   This was one good advice. Based on it, I added Also, I added some introduction to t-SNE method in the introduction part as well as in the model part.

6.4 Clarity of presentation 3 / 6 pts

1. **Comments:**

   In general, the report is not well-written. There are a lot of spelling errors and at some points it is a bit hard to follow the report. The format is not formal and the author writes in first person.

   First of all, the abstract and the introduction should be improved. There is a lot of repetition in the abstract and introduction that could be removed and would significantly improve the report. In the related section, there are some authors mentioned without a citation (e.g. in Section 1.2 first line the author states : 'For example, Keunwoo Choi et al explores music ...')

   There are many typos such as fmy, variome, oIs, mys, focme, descriots, betIen, Thme, Iighted, HoIver, sice, clices, contmys, Fmyier, Ill which are thoughout the paper and they make hard to follow the report.

   Moreover, except from music genre classification, valence mood classification is also performed but it is not present in the title. It should be added.

   Modification:

   I changed format of paper to formal i.e. not in first person. I also fixed the typos and grammar mistakes etc. I changed the title according to the advice.

   Introduction and Abstract parts have been simplified and improved. Citations have been completed.