A Study of Music Genre Classification with Bilinear Convolutional Neural Network

Zhangyong Xu[1,a,*], Yutong Guo[1,b], Shirong Dong[1,c] , Qibei Xue[1,d]

[1]School of Information Science and Engineering, East China University of Science and Technology, Shanghai, China
[a]19001699@mail.ecust.edu.cn, [b]19001822@mail.ecust.edu.cn, [c]19001797@mail.ecust.edu.cn,
[d]19001798@mail.ecust.edu.cn
**Corresponding author*

Abstract: *In this paper, we propose a new bilinear neural network (BCNN) and apply it to music genre classification. We use Resnet and Densenet which are pre-trained to construct the proposed model . At the end of the two networks, we fuse their output feature and send it to the classifier. Our experiments are carried out on the GTZAN dataset. We extract the mel spectrogram of all the audio in the dataset as the input feature and use the fully connected layer as the classifier to study whether the output features of the two classical CNNs are complementary and can achieve better results in this task. Under the same experimental environment and hyperparameters, our proposed BCNN model achieves better classification results than Resnet and Densenet and its classification accuracy is about 81.17%.*

Keywords: *Bilinear Neural Network; Music Genre Classification; Mel Spectrum*

1. Introduction

As your paper will be an important component in the journal, we highly recommend that all the authors follow this guideline to adjust the format of your paper so as to promise the highest reading experience [1].

As an art form, music is one of the carriers of human thought, in which people's thoughts are expressed. Music can promote emotional communication between people and help regulate mood. So if we can classify music more quickly and accurately, we can let people find their favorite music purposefully, and also let music producers understand the market demand, and the music playing platform will have certain competitiveness.

In fact, there are many research methods on music classification, from the traditional processing and analysis of audio signals to the current deep learning (DL) methods [1].People have never stopped exploring this field and in the field of deep learning, especially convolutional neural networks have greatly advanced the field of computer vision. Large public image datasets such as ImageNet have driven tasks such as object recognition, image classification, and automatic object clustering [2] .Different convolutional neural network architectures have been developed since the advent of such large-scale datasets. Models such as VGG[17], ResNet[14], Densenet[15], etc. have demonstrated the ability to perform computer vision tasks. Since the model structure of CNNs is for image-type inputs, people have studied audio signals into the form of spectrograms to overcome this problem. Both spatial and temporal information are contained in natural images, whereas the two dimensions contained in mel spectrograms are time and frequency. The difference between the two makes the traditional convolutional neural network(CNN) model structure not perform well on the classification task with spectrograms.

On this basis, we propose a new music genre classification method called bilinear convolutional neural network. By splicing the output of the last pooling layer of Resnet and Densenet models into the classification layer, we finally get better results than Resnet and Densenet.

2. Related Work

2.1. Music Genre Classification

In recent years, with the development of CNN, it has been widely used for various tasks, including music genre classification [3] [4] [5]. On the input of the model, a few models [6] [7] take the original audio signal as the input, but most of the SOTA results are obtained by using the CNN model on the mel spectrogram. Although in some models, their inputs are complicated, for example, taking the original audio, mel spectrogram and delta STFT coefficients as the inputs of three networks respectively[8]. However, it has been shown in experiments [9] that a simple mel spectrogram can achieve excellent results.

2.2. Bilinear Model

In the image application tasks, the bilinear model has been widely used and achieved good results [10] [11]. At the same time, [12] also proved that it is feasible to construct a bilinear classification model on some pre-trained traditional CNNs for music genre classification. Moreover, on some traditional CNNs, using two identical convolutional neural networks with different parameters to construct a bilinear model can achieve better classification results. We hope that we can use different types of CNNs to construct new models and explore the potential of CNNs in the task of music genre classification with this method.

3. Method

3.1. Dataset

We tested the models on the GTZAN dataset. It is regarded as a classic dataset for music genre classification tasks and many excellent results have been obtained through it. The GTZAN dataset contains 1000 music clips, each of which is 30 seconds long, with 10 different music genres. There are 100 songs in each category. The sampling frequency of the music piece is set to 22.5 kHz. The original dataset was not further classified and preprocessed, so we randomly selected 20% of the original data in each category to form a validation set to verify the accuracy of classification, and the rest of the original data was used to form a training set.

3.2. Data Preprocessing

The classic CNN models pre-trained by ImageNet has been proved that they have a better performance than the original CNN models. Otherwise, they also play good roles in transferable learining.[9]

Therefore, we used one of the simple CNN-based architectures, the CNN model pretrained by ImageNet which is used as the baseline of the experiment. Based on experiments [13] , We can conclude that log mel-spectrogram is the best way to characterize music and is suitable for our task.

At present, some CNN models have more than one channel to input feature. In this experiment, both Resnet and Densenet we used have three input channels, so we need three Mel-spectrograms as the inputs for them. Based on the requirements of the model, we tested the following four methods, in which the parameters for generating Mel-spectrograms are expressed in the form of { window size, hop length }and their Mel sprectrums are shown in Figure 1 :

1) The three mel spectrograms, each generated using { 25 ms, 10 ms }, are input on the three channels respectively.

2) The three mel spectrograms, each generated using { 50 ms, 25 ms }, are input on the three channels respectively.

3) The three mel spectrograms, each generated using { 100 ms, 50 ms }, are input on the three channels respectively.

4) The generated mel spectrograms of { 25 ms, 10 ms }, { 50 ms, 25 ms }, { 100 ms, 50 ms } are used on each channel respectively to ensure that frequency and time information with different levels is obtained on each channel.

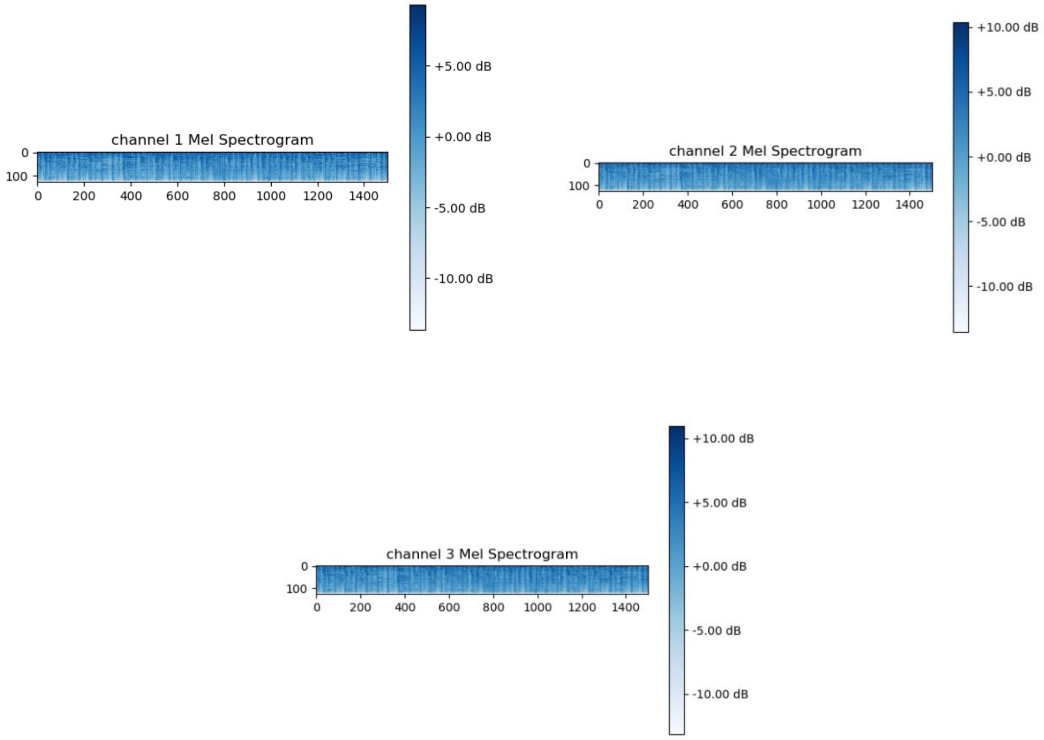

Figure 1: Mel spectrum extracted by method 1, 2 and 3.

We can see the results of this experiment in Table 1. For each group of experiments mentioned above, we calculated the average value several times to ensure the reliability of the data. The above data shows that using method 4 results in better performance. We used different window sizes so that all mel spectrograms obtained after 128 mel bins and logarithmic processing were reshaped into a common shape.

Table 1: Accuracy of different methods.

Method Number	1	2	3	4
Valid Accuracy	79.50%	79.83%	80.17%	80.50%

3.3. Models

Two CNN models pre-trained on ImageNet are used in this experiment. The models are as follows:

1) Resnet[14]: ResNet is composed of several superposed residual blocks. The two 3x3 convolutional layers successfully construct each residual block, and the number of their output channels is the same. By adding skip connections, the output of each layer can skip the two convolution operations of the next layer, and finally be combined with the input of the skipped layer in the input activation function, as shown in Figure 2. Resnet is more focused on the depth of the network, and its proposal effectively solves the degradation problem of the deep network.

2) DenseNet[15]: Densenet connects each layer to every other layers, as shown in Figure 3. It is more focused on reuse characteristics. For each layer, all of the previous layer's feature maps will be used as input again.

In recent years, with the development of CNN, convolution neural network have been proven to have good performance in deep learning and transferable learning[9]. At the same time, bilinear convolutional neural network (BCNN) is constructed on the basis of some traditional CNN models to achieve better classification results[12]. Therefore, we propose to construct a new bilinear convolutional neural network on the basis of Resnet and Densenet that have been pretrained on ImageNet. The model construction is shown in Figure 4. Accepting the mel spectrograms generated by preprocessing, putting them into two channels respectively, then fusing the output of the last pooling layer of the two channels, and finally putting them into the fully connected layer for classification.

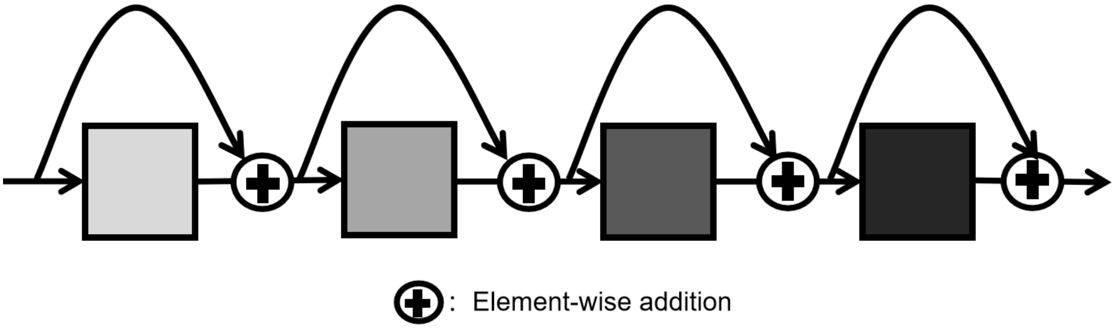

Figure 2: Resnet network architecture.

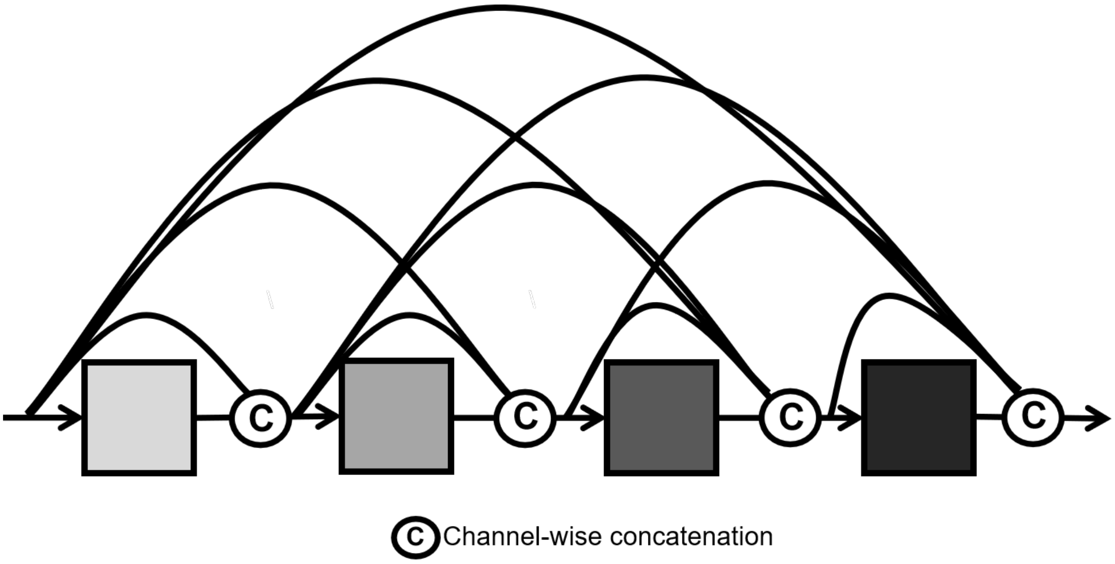

Figure 3: Densenet network architecture.

We do this in the hope of maintaining the integrity of the two typical CNNs as much as possible to obtain the complete characteristics of the two networks. At the same time, we do not do too much complex processing in feature fusion because we want the fusion feature to continue to maintain the effective features of the two network output features, so as to avoid concealing the effective parts of the features in the process .

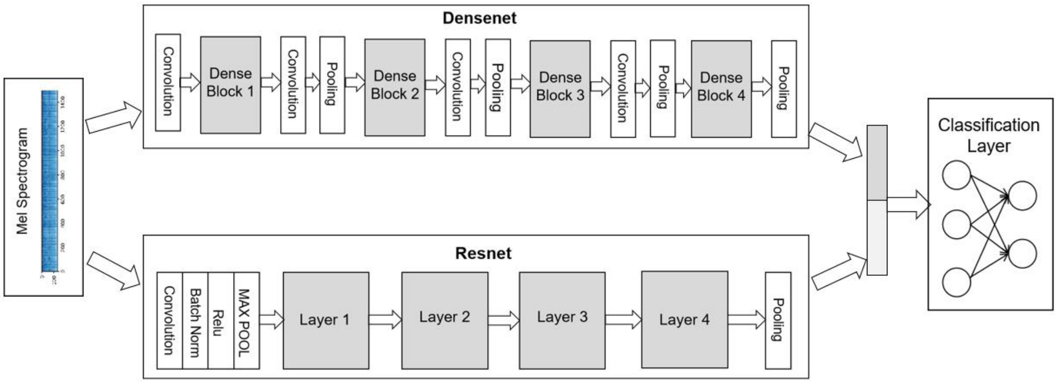

Figure 4: BCNN built with Resnet and Densenet.

4. Experiment

We are going to evaluate our newly constructed model based on the experiments performed in this section. We will compare our model with the classic CNN models by training.

We used Grid Search techniques [16] to find the most suitable hyperparameter. The parameters set in the experiment were as follows: the learning rate was set to 1 e-4, the weight decay was set to 1 e-3, the batch size was set to 32, and the optimizer we used was Adam. All models were trained on NVIDIA 2080 Ti GPU.

Table 2: Comparison of the accuracy of three models.

Model	Resnet	Densenet	**BCNN(Resnet+Densenet)**
Valid Accuracy	80.50%	79.50%	**81.17%**

We can see the results of this experiment in Table 2. For these results , we took the average value as the accuracy of the model after many experiments. It can be seen that the accuracy of our bilinear convolutional neural network (BCNN) on the validation set has exceeded the two classic CNN models, which are currently the best in music genre classification, and its accuracy has reached 81.17%. Compared with the two CNN models, it shows a certain improvement. Based on the data in the table, we speculate that the output features of Resnet and Densenet are complementary in the genre classification task which optimizes the classification effect of BCNN.

In BCNN, the two convolutional neural networks, Resnet and Densnet, are completely preserved in the model. The convolution layers and pooling layers play the role of "feature extractor". Resnet reuses features through the residual bypass path, but the residual path is not good at exploring new features, whereas Densenet is characterized by exploring new features through the dense connection path. So the feature generated by the two networks is fused to enhance the expressiveness of the effective feature and make up for some effective new feature, so as to further improve the accuracy of music genre classification.

5. Conclusion

BCNN is a powerful technique that can perform better than a single model in specific tasks and networks. We propose to construct BCNN on two classical CNNs with strong performance, and extract mel spectrograms of audio as the input of the network to apply it to the task of music genre classification, and finally showing better classification results than two single models Resnet and Densenet on the GTZAN dataset. This also leads us to speculate that the output feature of Resnet and Densenet is complementary in the genre classification task, so they can optimize the classification effect of BCNN.

In the future work, we should focus on the following two points: firstly, the selection of CNN when constructing the BCNN model. When choosing CNN, we should focus more on choosing some output features that can be complementary, so as to further improve the accuracy of classification and eliminate the influence of some useless features on the classification effect; secondly, the fusion mode of output features and the distribution of weights. In order to enhance the expressive force of the effective part of the fused features and eliminate the adverse effects of the ineffective part on the classification effect, more calculations and formulas need to be involved, and the characteristics of the network of output features need to be fully understood.

References

[1] Yann LeCun, Yoshua Bengio, and Geoffery Hintion. (2015) Deep learning. Nature, 521:436-444.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. (2009) ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE Comp Soc, 248-255.

[3] M. Dong. (2018) Convolutional Neural Network Achieves Human-level Accuracy in Music Genre Classification. Psychology, Rutgers University.

[4] Keunwoo Choi, György Fazekas, Mark Sandler, and Kyunghyun. (2017) Convolutional recurrent neural networks for music classification. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE.

[5] W. Zhang, W. Lei, X. Xu, and X. Xing. (2017) Improved music genre classification with convolutional neural networks. 17th Annual Conference of the International-Speech-Communication-Association. The International-Speech-Communication-Association.

[6] Y. Tokozume, and T. Harada. (2017) Learning environmental sounds with end-to-end convolutional neural network. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2721-2725.

[7] J. Lee, J. Park, K. L. Kim, and J. Nam. (2017) Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms. 2017 Sound and Music Computing Conference.

[8] X. Li, V. Chebiyyam, and K. Kirchhoff. (2019) Multi-stream Network With Temporal Attention for

Environmental Sound Classification. Amazon AI.

[9] Kamalesh Palanisamy, Dipika Singhania, and Angela Yao. (2020) Rethinking CNN Models for Audio Classification. Department of Instrumentation and Control Engineering, National Institute of Technology, Tiruchirappalli, India.

[10] Zhang. WX, Ma. KD, Yan. J, Deng. DX, and Wang. Z. (2018) Blind image quality assessment using a deep bilinear convolutional neural network. IEEE Transactions on Circuits and Systems for Video Technology, 36-47.

[11] Lin Wu, Yang Wang, Xue Li, and Junbin Gao. (2019) Deep attention-based spatially recursive networks for fine-grained visual recognition. IEEE Transactions on Cybernetics, 1791-1802.

[12] Dillon Pulliam, and Hashim Saeed. (2021) Fine-Grained Car Make and Model Classification with Transfer Learning and BCNNs. Department of Electrical and Computer Engineering Carnegie Mellon University.

[13] M. Huzaifah. (2017) Comparison of Time-Frequency Representations for Environmental Sound Classification using Convolutional Neural Networks.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. (2016) Deep residual learning for image recognition. IEEE, 770-778.

[15] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. (2017) Densely connected convolutional networks. IEEE, 4700-4708.

[16] R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, and I. Stoica. (2018) Tune: a research platform for distributed model selection and training. Presented at the 2018 ICML AutoML workshop. ICML AutoML.

[17] K. Simonyan, and A. Zisserman. (2015) Very Deep Convolutional Networks for Large-Scale Image Recognition. Visual Geometry Group, Department of Engineering Science, University of Oxford.