Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 10, No. 3, 2024

Comparison Among AlexNet, GoogLeNet and ResNet-18 in Autiomatic Music Genre Classification

Xinyu Hong[1, a]

[1]Maynooth International Engineering College, Fuzhou University, Fuzhou, 350000, China
[a]xinyu.hong.2023@mumail.ie

Abstract: This paper compared the performance among three famous convolutional neural networks in classifying music genres. Being carried out on a prominent dataset named Free Music Archive, the experiments show that ResNet-18 performs much better than AlexNet and GoogLeNet in classifying the music genres in a relatively small dataset. Meanwhile, the classification accuracy of each model for each music genre was also recorded. It indicates that different models could be expert in identifying distinct genres. Several genres, including blues, hip-hop and international, were not closely related to the change of models. In general, ResNet-18 reached the highest average classification accuracy at approximate 80%, while AlexNet did best in finding hip-hop music and GoogLeNet had relatively less difference in recognition rates for every genre. Those findings can serve as a reference in future music genre classification tasks and personalized music recommendation based on big data.

Keywords: Deep learning, music genre classification, convolutional neural networks.

1. Introduction

Listening to digital music has already been one of the common leisure pursuits for individuals.

Alongside the advent of the information revolution, people nowadays can access and download music online much more easily than before. However, due to the lack of standardized metadata information for these streaming media, many music lacks genre information or contains incorrect genre information, which affects the effectiveness of users searching for songs through genre keywords and recommending similar songs through big data. Automatic Music Genre Classification (AMGC) has emerged as a popular technology for mitigating the burden and expenses associated with manually categorizing a vast volume of music genres.[1] It was defined and named by Tzanetakis and Cook, dating back to 2002.[2] During that period, the limited popularity of deep learning resulted in the phenomenon which a significant number of researchers opted for machine learning techniques to classify music genres. For example, Kosina utilized a 3-NN classifier to analyze music features provided by the MARSYAS framework and achieved an average accuracy of 88.35%.[3] Although traditional machine learning methods have been applied in AMGC tasks already, they may not be the optimal choice nowadays. One main reason is that those kinds of technology require a significant amount of human workload to select features of the music for classfiers. In addition, the enhancement of traditional classifiers only has a somewhat restricted influence on the outcome. According to a study conducted by Carlos N. Silla Jr., Alessandro L. Koerich, and Celso A.A. Kaestner, the accuracy of classifiers such as 3-NN, OAA, and FSRR is limited to below 70% when using various music decomposition strategies.[1] With the advent of several models designed based on the LeNet model, the public started to introduce deep learning technology into AMGC tasks. With certainty, the performance of deep learning was general satisfying. In a research carried out by Yan Jiacheng, the convolutional neural network which was designed under the structure of AlexNet reached an average accuracy rate at about 87%, and the network based on VGG16 improved accuracy to approximately 91% in classifying the music dataset GTZAN.[4] What's more, the method that combines RGLU-SE convolutional structure with Bidirectional Long Short-Term Memory (BLSTM) indicated the accuracy to 92.2% about classifying the dataset GTZAN in the context of using attention mechanism for sequence feature aggregation.[5] Recently, the steps of applying convolutional neural network to classify music genres can be concluded generally into three parts: making mel spectrograms, data augmentation and model training.

2. Preparation

2.1. Mel Spectrogram

Mel scale is a set of psychoacoustic pitch units which makes people feel the same difference

between two mel notes adjacent to each other at the same volume. To be more precise, it provides a linear scale for the human auditory system. According to it, we can use a spectrogram as a visual description for the models to learn the features. The frequency of Mel scale(m) can be calculated by such formula below, where f represents Hertz.

2.2. The Dataset

The Free Music Archive (FMA) is an open dataset that contains a wide range of music genres and accurate music metadata. This dataset is valuable for both Music Information Retrieval (MIR) and Automatic Music Genre Classification (AMGC) tasks[6]. I utilized Python to convert music clips from the FMA small dataset, which consists of 8,000 music pieces spanning eight different genres in the Free Music Archive (FMA), into mel spectrograms. Figure 1 presented below shows one of the mel spectrograms.

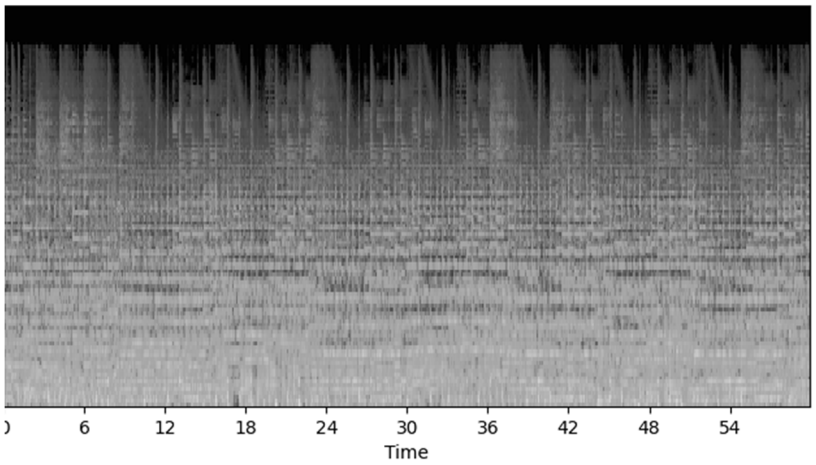


Figure 1. A mel spectrogram

2.3. Data Augmentation

Data augmentation is a technology that expands the dataset and improves the generalization

and robustness of the model. Flipping, rotating, cutting, scaling and translation are several common steps in augmenting data. The inventors of AlexNet, including Alex Krizhevsky, applied two methods in data augmentation. The first one was extracting patches (including their horizontal reflections) from original images, and then training the model with those patches. Another method was performing principal component analysis (PCA) on the RGB pixel values of each image, which reduced the top-1 error rate by up to 1%.[7]

2.4. CNN models

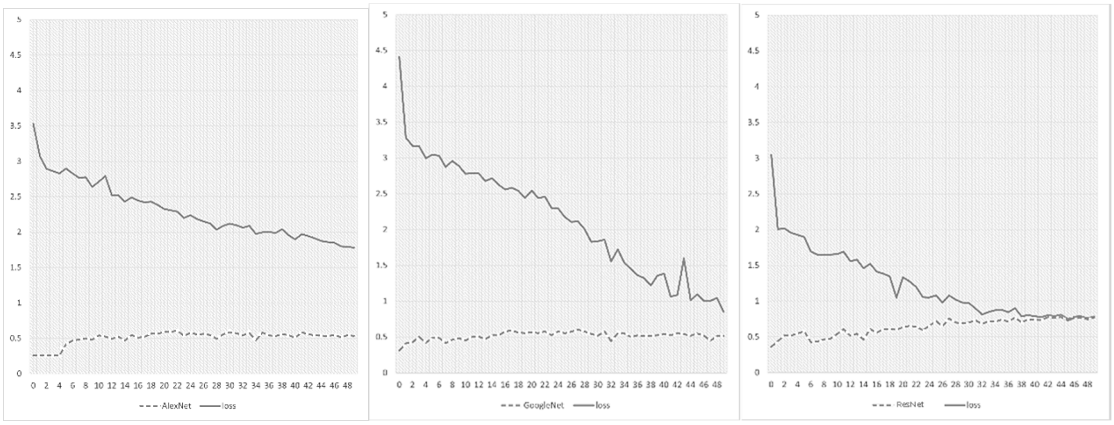
On the basis of the efforts of previous researchers, the performance of convolutional neural networks on computer vision has been improved dramatically, with technology ranging from grouping local connections among pixels to features sharing. In 1998, a group of researchers including LeCun published a convolutional neural model LeNet-5, which has become a source of inspiration for designing new model structures. Although LeNet-5 had been a milestone in deep learning, it still has several shortcomings, like the excessive data dependence. From 2012 to 2023, the establishment of more databases and modifications to neural network architecture have led to the proposal of many famous convolutional neural network models.[8] All models this research includes are derived from LeNet-5 and are published in chronological order. Each of them inherited and updated the previous model in some specific regions. Thus they shared both similarities and differences with each other. Table 1 illustrated their main properties, which can serve as a reference for subsequent analyses. [9-11]

Table 1. Main properties about AlexNet, GoogLeNet and ResNet-18

Model Name	AlexNet	GoogLeNet	ResNet-18
Year Published	2012	2014	2015
Key Innovation	Demonstrated the potential of deep learning in image recognition	Introduced the Inception module to improve computational efficiency	Introduced residual connections to address the problem of gradient vanishing in deep neural networks
Number of Parameters	Relatively high	Lower compared to AlexNet	Lower compared to AlexNet and GoogLeNet
Advantages	Drove the advancement of deep learning in computer vision	Improved computational efficiency and accuracy (through the Inception module)	Alleviated the gradient vanishing problem and enabled deeper network architectures
Disadvantages	A relatively high number of parameters, which can lead to overfitting	May not be as accurate as other models for certain tasks	May not be as accurate as other models for certain tasks

3. Experiment and Results

Using a single processor, the Intel Core i5-12500H, the experiment was conducted based on the Python 3.9 environment and various third-party libraries, including numpy and torch. A total of 8,000 music clips from the FMA-small database were transformed into mel spectrograms and resized to the appropriate input size. And the results of AlexNet, GoogLeNet and ResNet-18 were illustrated in Figure 2.



(a) (b) (c)

Figure 2. The loss and accuracy of each model (AlexNet, GoogLeNet and ResNet-18 from left to right)

It turned out that the accuracy of AlexNet was only around 25% at the beginning, but after the sixth epoch, it significantly improved to around 40%. However, the effect of more epochs on improving accuracy was not very significant, as can be seen from Figure 2(a) that it fluctuated between 50% and 60%. When it came to GoogLeNet, the accuracy in the first epoch was approximately 30% and kept climbing up until the 10th epoch. In comparison with AlexNet, the accuracy of GoogLeNet was also fluctuating from 55% to 60%. Figure 2(b) illustrated the accuracy and loss of GoogLeNet during 50 epochs. Figure 2(c) shows that the accuracy of ResNet-18 maintained the upward trend during 50 epochs. In the first 25 epochs, the accuracy ascended from 30% to 65%. After that, its value maintained above 70%, which proved that it can handle such tasks better.

By comparison, we can intuitively observe that there was little difference in the performance of AMGC tasks between AlexNet and GoogLeNet in the FMA-small dataset. When it came to ResNet-18, the average accuracy significantly improved to 79.58%, which is significantly different from the accuracy of the previous two models, which are 50.89% and 51.53% respectively.

When we look at the trends of the accuracy of the three neural networks changing with the number of training epochs, it is obvious that they all maintained a similar upward trend, and the accuracy differences were not significant in the first 24 epochs. From 25th epoch to 50th epoch, only ResNet-18 still kept its upward trend. By referring to Figure 3, you can more intuitively clarify the above conclusion.

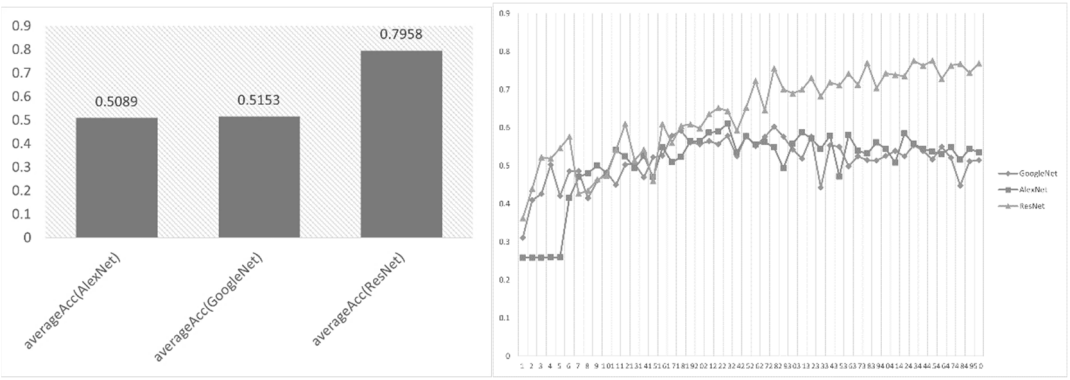


Figure 3. Comparison among three models

4. Analysis

4.1. The Characteristics Of The Experiment

In comparison with datasets such as Million Song Dataset (MSD), Million Musical Tweets Dataset (MMTD), FMA-small contains fewer number of music clips, which made the experiment kind of tricky. To be more precise, the models need to classify the genres while adjusting the limited amount of training samples in order to show a satisfying performance. Meanwhile, we also used the running time of each dataset as a reference when judging the comprehensive performance of those models in this experiment.

4.2. The Similar Performance Between AlexNet and GoogLeNet

As we refer to Figure 3 above, we can easily find that the accuracy of AlexNet was similar to the accuracy of GoogLeNet during every epoch. In specific epochs like epoch 9, GoogLeNet performed even worse than AlexNet. Also, the running time of GoogLeNet was relatively longer.

To comprehend such result above, we could start from the structures of the two models. Both the input of AlexNet and the input of GoogLeNet should be images of pixel size and of three channels R, G and B. Both the two models applied the pretreatment technology like data augmentation and zero mean normalization. Such technology could strengthen the generalization of models, while it may also carry the risk of making the them overly sensitive to certain features of the images (even though some of them are inconsequential). In addition, the use of Local Response Normalization (LRN) could also take responsibility for the deterioration of models. [9-10]

4.3. Better Performance from ResNet-18

The accuracy of ResNet-18 kept an increasing trend after epoch 25. As it reached 70%, the rising rate shrank gradually. Thus, we can consider that the structure of ResNet-18 performed better in such kind of AMGC tasks. Ye J C et al., the authors of ResNet, had tested a model by dataset cifar10. They got a conclusion through the strange result that the model with 56 layers made more mistakes than that one with only 20 layers: adding too much layers in a neural network could increase the risk of neural network degradation (The back features have almost lost the original appearance of the front features). Through the application of inter layer residual hopping, ResNet introduces forward information, alleviates the phenomenon of gradient vanishing, and makes it possible to increase the number of layers in the neural network with a better accuracy.

4.4. The Accuracy of Classifying Specific Genres

Through the experiment, the result told us that AlexNet had the highest accuracy (roughly 75%) classifying the Hip-Hop music than other catagories. The difference in recognition accuracy of GoogLeNet for these specific genres is relatively small, but in comparison, it has a stronger ability to judge Pop music and Folk music. Overall, ResNet-18 was better able to recognize any type of music among those models. And it was better at finding International, Electronic, Pop and Jazz music (with an accuracy rate of over 60%).

If we divide music into two categories based on accuracy, "Strong correlation between accuracy and the model used" and "Weak correlation between accuracy and the model used", then the former would be represented by International music, while the latter are most representative by Instrumental and Blues music.

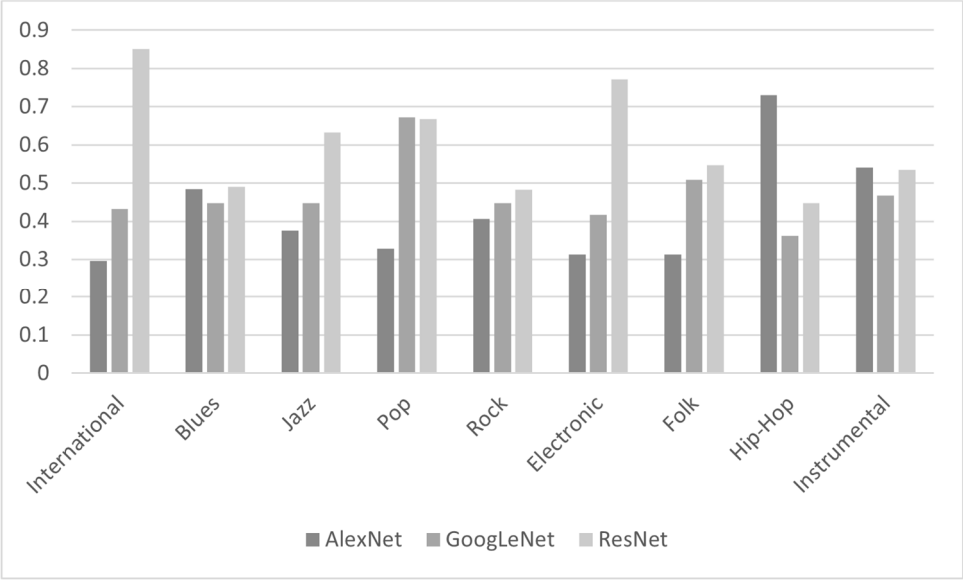


Figure 4. Comparison about accuracy among different models and music genres

5. Conclusions

We are in the developing period of digital audio market. It will undoubtfully incur astonishing labor and time cost if we manually add genre labels to those increasingly new music and old music lacking metadata. And accurate genre information is crucial for various music playback platforms to effectively push big data to users today. In the era of the rise of convolutional neural networks, more and more industry professionals are attempting to use these models for AMGC tasks. The article referred to the relevant theories of AlexNet, GoogLeNet, and ResNet-18 network structures and tested the accuracy of three network classifications on dataset FMA-small. The results were compared and analyzed roughly. They provide a certain reference and guidance direction for model selection during AMGC execution and model optimization for small dataset classification.

Due to limitations in time and my own abilities, although this article has obtained certain experimental data and research results, there are still shortcomings in experimental design and cause analysis. Especially in terms of evaluation indicators, control of the quantity and quality of the dataset used, there is still room for further refinement, which is also the direction for further improvement by the author in the future.

References

[1] Silla Jr, C. N., Kaestner, C. A., & Koerich, A. L. (2007). Automatic music genre classification using ensemble of classifiers. In2007 IEEE International Conference on Systems, Man and Cybernetics, pp. 1687-1692.

[2] Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. IEEE Transactions on speech and audio processing, 10(5): 293-302.

[3] Kosina, K. (2002). Music genre recognition.

[4] Yan, J. (2022) Comparison of Machine Learning and Deep Learning Model Classification of music genres. Information Technology and Informatization, 12:217-220.

[5] Gao, Y. (2020) Research on Music and Audio Classification based on Deep Learning. Thesis of South China University of Technology.

[6] Defferrard, M., Benzi, K., Vandergheynst, P., & Bresson, X. (2016). FMA: A dataset for music analysis. arv preprint arv:1612.01840.

[7] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6): 84-90.

[8] Attri, I., Awasthi, L. K., Sharma, T. P., & Rathee, P. (2023). A review of deep learning techniques used in agriculture. Ecological Informatics, 102217.

[9] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1-9.

[10] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arv preprint arv:1409.1556.

[11] Bae, W., Yoo, J., & Chul Ye, J. (2017). Beyond deep residual learning for image restoration: Persistent homology-guided manifold simplification. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 145-153.