Music Genre Classification via Image Classification Model

Kyeong Joo Jung
Stonybrook University
Incheon, Republic of Korea
kyeongjoo.jung@stonybrook.edu

Doeun Kim
Stonybrook University
Incheon, Republic of Korea
doeun.kim@stonybrook.edu

ABSTRACT

Humans are good at classifying genres of music. It is a simple task for humans. In addition, classifying with the labeled music with a number of dataset is also easy for computing machines. However, there are numerous unlabeled music that people use. Moreover, classifying genres with unlabeled datasets make computers, smartphones, and MP3 players to take much longer time. In this paper, we proposed a method to classify music genres using deep learning with 1000 number of data. We tried various models of image classifiers and found out that Xception is best fit among the models we tried for the music genre classification.

CCS CONCEPTS

• **Computing methodologies** → *Machine learning approaches*;

KEYWORDS

Machine Learning; Deep Learning; Music Genre Classification

ACM Reference Format:
Kyeong Joo Jung and Doeun Kim. 2018. Music Genre Classification via Image Classification Model. In *SPRING'18: 2018 SPRING CSE353/512 Machine Learning Project, 2018, Incheon, Republic of Korea.* ACM, New York, NY, USA, Article 4, 3 pages. https://doi.org/1231242

1 INTRODUCTION

Every song in the universe contains its own music genre. Especially, human can easily distinguish music genres due to their feeling, intuition, and experience. However, computing machine cannot judge music genre of the song by themselves because it doesn't have feeling, intuition, and experience unlike human. In order to let computing machine classify the music genre only by itself, it has to train itself by learning and analyzing the training music dataset. Therefore, music genre classification has been an interesting task that aims to detect and predict the music genre of the song in deep learning area nowadays.

Music can be shown in several ways. Sound is the typical way to recognize the song by human. Unlike human, machine should extract features from song and convert them into other form which is understandable to the machine. Since there are number of features in one song and it is difficult to figure out which features perform the most in the particular task, Deep Neural Network (DNN) has been used recently the most. Previous works have used melspectrograms of music data as the input. However, in this paper, we use image classification to achieve higher accuracy. To use the image as the input, we turned melspectrogram into an image as shown in Figure 1.

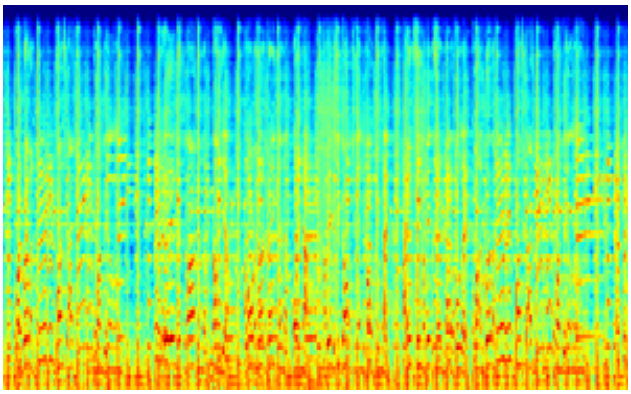

Figure 1: Mel-spectrogram image made by GTZAN .au format music file

In this paper, we compare other state-of-the-art works and some new models with the same number of data. Moreover, unlike other works, we classified music genres into 10 genres. Through the comparison of same number of data, we show which one among the suggested models is better than the previous works. The paper is outlined as follows. In section 2, we introduce related works regarding music genre classification. In section 3, we show which dataset we used and explain our deep neural network models. In section 4, results of the State-of-the-art methods and proposed methods are discussed and show which one of the proposed methods best fit the music genre classification. In section 5, we conclude our paper with contributions, limitations, and future works.

2 RELATED WORKS

There are 3 works that has been done before. In addition, these works use CNN[7], CRNN[3], VGG16[9] models that already exists. These models used AUC metrics to show the performance of the model. AUC (Area under the ROC curve) is also known as the ROC (Receiver Operator Characteristics) which is well used for the multi-class classification. A model is expected to have AUC of higher than 0.5.

First, the model which uses CNN [3] is trained and tested with 214,284 data from Million Song Dataset[2] with 50 tags and 25,856 data from MagnaTagATune[8] with 50 tags. It inputs a mel-spectrogram matrix shaped in 96 x 1366 resulted from each 29.12 seconds long highlight part of the song with 12kHz sampling rate, 96 mel bins, and 256 hop-size. It is structured in two-dimensional convolutional

SPRING'18, Spring 2018, Stonybrook University, SUNY Korea, Incheon, Republic of Korea

ACM ISBN 12412.
https://doi.org/1231242

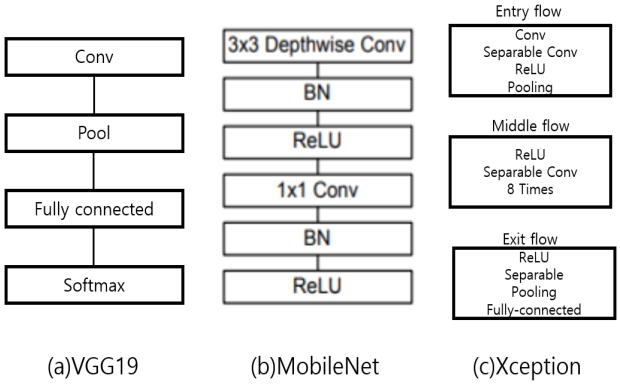

Figure 2: Models that were used for the research - (a) VGG19, (b) MobileNet, (c) Xception

neural networks with 5 convolutional layers of 3x3 filters and 5 max-pooling layers ((2x4)-(2x4)-(2x4)-(3x5)-(4x4)). Its final layer has feature maps size of 1x1. The output layer uses sigmoid functions. Second, the model which uses CRNN is trained and tested with the same dataset as the dataset of the first model and also takes the same mel-spectrogram inputs as the inputs of CNN. However, the CNN model in CRNN is slightly different. It has two-dimensional 4-layer CNN of 3x3 filters and 4 max-pooling layers ((2x2)-(3x3)-(4x4)-(4x4)) in front of the model. Then, CNN connects with 2-layer RNN with Gated Recurrent Units (GRU) [5]. The output layer uses sigmoid functions. Lastly, the model which uses VGG16 [1] extracted 10 seconds sound clips from 2.1 million Youtube videos with 7 different music genres (Pop, Rock, Hip Hop, Techno, Rhythm Blues, Vocal, and Reggae Music) to collect dataset. The clips were converted to .mp4 files, from .mp4 files to .wav files, from .wav files to mel-spectrogram images, and from mel-spectrogram images to 3-channel (RGB) matrix, which is the actual input. The model is constructed with 5 convolutional blocks, 1 fully connected layer, 1 dropout layer, and 1 output layer with softmax.

3 DATASET AND PROPOSED MODEL

It is clear that above three State-of-the-art models commonly used large dataset. In this paper, we used 1000 .au format music files from GTZAN [10] with 10 different music genres (Blues, Classical, Country, Disco, Hip Hop, Jazz, Metal, Pop, Reggae, and Rock). 100 music files are distributed for each music genre.

In order to achieve finding the deep neural network model with the higher accuracy than the state-of-the-art models, we applied 3 additional different image classifiers : Xception[4], VGG19 [9] and MobileNet[6]. The reason of applying image classifiers is that image classification technologies have been highly developed nowadays. Thus, we converted .au format music file to logamplitude mel-spectrogram matrix. Then, we converted the matrix to default color image size of the classifier models such as 299x299, 216x216, 224x224 for Xception, VGG19, and MobileNet model. Therefore, the proposed method takes RGB matrix of the color image as an input.

The difference that the proposed method has compared to the previous models is that it takes the input of images. Prior works transform music data to melspectrograms and use its value for the inputs. This way of approach takes much longer time for data pre-process and cannot be well used with the existing deep neural network classifiers. However, proposed method uses images as input to easily adapt to classifiers. Melspectrograms are turned into images and we use the RGB values as an input.

Above Figure 2 shows how VGG19, MobileNet, and Xception are originally structured. All three deep neural networks have weights pre-trained on ImageNet. Xception has size of 88MB, the highest accuracy of 0.945, and 126 layers. VGG19 has size of 549MB, accuracy of 0.910, and 26 layers. MobileNet has size of 17MB, accuracy of 0.871, and 88 layers.

The proposed method experiments 3 models. The models that were used are VGG19, MobileNet, and Xception. Additionally, as shown in Figure 3, all models flatten output and send it to multilayer perceptron. Then, we add one fully connected layer (dimensionality of output space: 256), followed by one dropout layer, followed by one activation layer with ReLU, and followed by one more fully connected layer with softmax function as an output layer. Moreover, the models are optimized by Nadam. The default layer number of Xception, VGG19, and MobileNet are 126, 26, and 88 each. However, the proposed method did not use the default layers as used models does not fit for image classifiers. Instead, we tried to reduce the number of layers as much as we can because the model will not be able to analyze but memorize the image which will decrease the accuracy when the layers get deeper. The proposed method uses 16, 8, 18 layers for each model.

4 RESULT

For more accurate experiment, we randomly choose and split the dataset of RGB matrices and one-hot-encoded label array into 80 percent for training, 10 percent for validation, and 10 percent for test and shuffle them. We trained and validated the models and measured performance on test set by AUC-ROC. Next, among the models, we calculated average accuracy and average precision from the model that has the highest AUC-ROC to get various evaluation values.

Table 1: Result of the models used for music classification with GTZAN datasets.

Epoch: 30	Model	AUC
State-of-the-art model	CNN	0.6413
State-of-the-art model	CRNN	0.7251
State-of-the-art model	VGG16	0.5101
Proposed Model	VGG19	0.5008
Proposed Model	MobileNet	0.5000
Proposed Model	**Xception**	**0.855**

For each model (MobileNet, VGG19, and Xception), we tried 30 epochs. For VGG19 model, learning rate is 0.000001 and batch size is 8 and MobileNet model learning rate is 0.00001 and batch size is 8. Lastly, Xception model has 0.0001 learning rate and 8 of batch size. As shown in Table 1, result shows that the highest model is Xception which has 0.855 AUC.

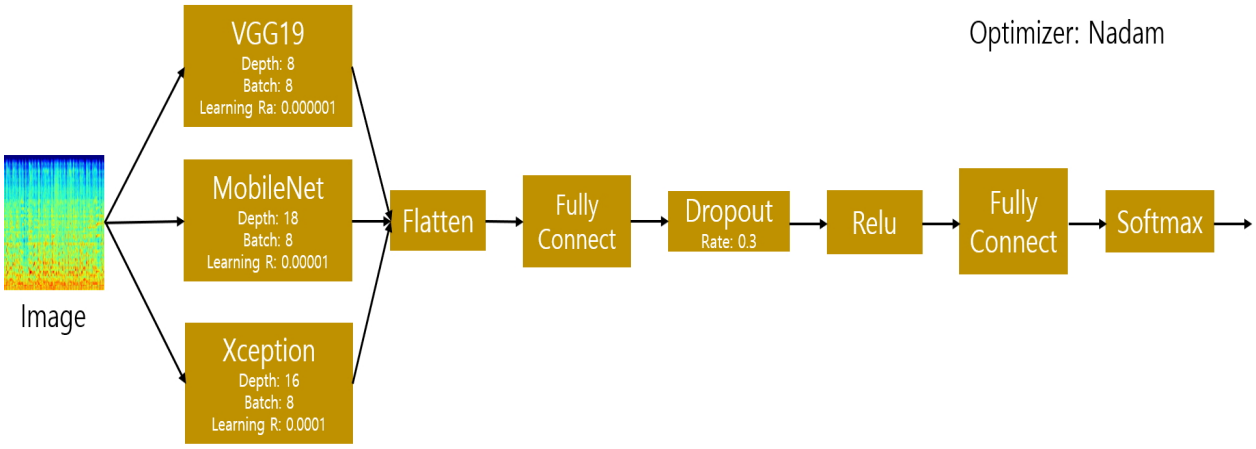

Figure 3: Flowchart of the proposed method - 3 cases (VGG19, MobileNet, and Xception)

CNN, CRNN, and VGG16 models have different dataset from our experiments. However, we replaced their dataset to our dataset which is from GTZAN. The reason to use the same dataset was to correctly compare the AUC of each model. Consequently, results show different AUC with the one recorded in original papers. As shown in Table 1, State-of-the-art models are not as high as the result written in original paper. This is because datasets are not same as the ones used in the original papers and the process of pre-processing is different from the original code. Moreover, models that used VGG19, and MobileNet had extremely low AUC compared to Xception model.

Table 2: AUC, average accuracy, and average precision of the Xception model.

Epoch: 30	AUC	Average Accuracy	Average Precision
Xception	0.855	0.8012	0.563

At last, since Xception model had the highest result, Xception model is the best fit for music genre classification. Moreover, we calculated the value of average accuracy, and average precision as shown in Table 2.

5 CONCLUSION AND FUTURE WORKS

In this paper, we proposed a new model which uses Xception for classifying music genres with higher accuracy than state-of-the-art models. We achieved contributions from this research. Followings are the contribution of this research:

- Compared the accuracy of State-of-the-art models and the proposed models with the same dataset for accurate comparison.
- Experimented various models which have not been used until now and found out which had high or low accuracy.

However, this research still have limitations. Followings are the limitation of this research:

- Difficult to find out the optimal parameters for music as image classification has not been used for music genre classification before.
- Low number of datasets as more number of datasets would be better to achieve accurate and higher accuracy.

Lastly, there are still future works left. With this model, we can also adapt RNN and there might be better parameters for other existing models that did not fit for the classification.

REFERENCES

[1] Hareesh Bahuleyan. 2018. Music Genre Classification using Machine Learning Techniques. *CoRR* abs/1804.01149 (2018). arXiv:1804.01149 http://arxiv.org/abs/1804.01149

[2] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. 2011. The Million Song Dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*.

[3] Keunwoo Choi, George Fazekas, Mark B. Sandler, and Kyunghyun Cho. 2016. Convolutional Recurrent Neural Networks for Music Classification. *CoRR* abs/1609.04243 (2016). arXiv:1609.04243 http://arxiv.org/abs/1609.04243

[4] François Chollet. 2016. Xception: Deep Learning with Depthwise Separable Convolutions. *CoRR* abs/1610.02357 (2016). arXiv:1610.02357 http://arxiv.org/abs/1610.02357

[5] Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *CoRR* abs/1412.3555 (2014). arXiv:1412.3555 http://arxiv.org/abs/1412.3555

[6] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *CoRR* abs/1704.04861 (2017). arXiv:1704.04861 http://arxiv.org/abs/1704.04861

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 1097–1105. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[8] Edith Law, Kris West, Michael Mandel, Mert Bay, and J. Stephen Downie. 2009. Evaluation of algorithms using games: The case of music tagging. In *Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR 2009*. 387–392.

[9] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014). arXiv:1409.1556 http://arxiv.org/abs/1409.1556

[10] G. Tzanetakis and P. Cook. 2002. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing* 10, 5 (Jul 2002), 293–302. https://doi.org/10.1109/TSA.2002.800560