Emotion and Theme Recognition of Music Using Convolutional Neural Networks

Shengzhou Yi, Xueting Wang, and Toshihiko Yamasaki
The University of Tokyo
{yishengzhou,xt_wang,yamasaki}@hal.t.u-tokyo.ac.jp

ABSTRACT

Our team, "YL-UTokyo", participated in the task: Emotion and Theme Recognition in Music Using Jamendo. The goal of this task is to recognize moods and themes conveyed by the audio tracks. We tried several Convolutional Neural Networks with different architectures or mechanisms. As a result, we find that a relatively shallow network achieved better performance on this task.

1 INTRODUCTION

We participated in one of the tasks in MediaEval 2019: Emotion and Theme Recognition in Music Using Jamendo [2]. This task involves the prediction of moods and themes conveyed by a music track. Moods are often defined as feelings conveyed by the music (e.g. happy, sad, dark, melancholy) and themes are associated with events or contexts where the music is suited to be played (e.g. epic, melodic, christmas, love, film, space).

In the task, there are three types of audio representations, including traditional handcrafted audio features, mel-spectrograms, and raw audio inputs. We only used the mel-spectrograms as input to train our model (Figure 1). We tried several Convolutional Neural Networks (CNNs) to find a suitable model for this task. The simplest but effective model we tried is provided by the organizers. It only consists of five convolutional layers and one dense layer at last. We also tried other models with more layers, but they didn't always achieve better results. As a result, the model that achieved the best performance in our experiments is a shallow neural network with only six convolutional layers and one dense layer.

2 RELATED WORK

Image classification performance has improved greatly with the advent of large datasets such as ImageNet [5] using CNN architectures such as VGG [9], Inception [10], and ResNet [6]. There are also many research of music emotion recognition or music classification using CNN architectures [4, 7]. Even though statistical machine learning (e.g. Support Vector Machine [8] and Random Forest [1]) can still achieve good performance in some tasks, deep learning, especially CNN based method, is more popular and can achieve better performance in most tasks. For large-scale datasets, deep learning is much more practicable than statistical machine learning.

MediaEval'19, 27-29 October 2019, Sophia Antipolis, France

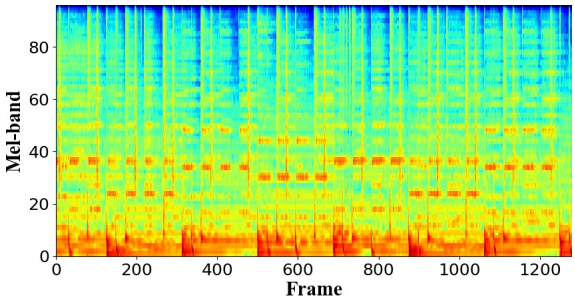


Figure 1: Mel-spectrogram

Table 1: The architecture of 6-layer model

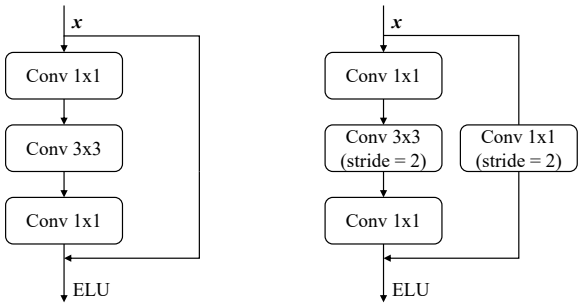
Mel-spectrogram	Input: 96x1280x1
Conv 3x3x32	
MP (2, 2)	Output: 48x640x32
Conv 3x3x64	
MP (2, 4)	Output: 24x160x64
Conv 3x3x128	
MP (2, 2)	Output: 12x80x128
Conv 3x3x256	
MP (2, 4)	Output: 6x20x256
Conv 3x3x512	
MP (3, 5)	Output: 2x4x512
Conv 3x3x256	
MP (2, 4)	Output: 1x1x256
Dense	
Sigmoid	Output: 56x1

3 APPROACH

3.1 Model

We concentrated on finding the most suitable CNN architecture for the task. The baseline is a simple but effective model consisting of five convolutional layers and a final dense layer. We also tried other models with deeper architecture. We tried models with 6, 16, 18 or 25 convolutional layers. In particular, the most shallow model we considered is a fully convolution neural network with ELU activations, six 3x3 convolutional layers, and 32, 64, 128, 256, 512, 256 units for each layer respectively (Table 1).

We also tried some models with the residual architecture [6]. The convolutional block consists of 1x1, 3x3 and 1x1 convolutional layer sequentially. This is the architecture for inputs and outputs with the same size and unit number. For the block that maps inputs



(a) Without downsampling **(b) With downsampling**

Figure 2: Residual architecture

to outputs with smaller size and more units, the stride of 3x3 convolutional layer is two and the shortcut is a 1x1 convolutional layer for downsampling (Figure 2).

3.2 Dataset

The dataset includes 17,982 music tracks with mood and theme annotations. The split for training, validation and test is about $2:1:1$. In total, there are 56 tags, and tracks can possibly have more than one tag. There are three types of audio representations, including traditional handcrafted audio features, mel-spectrograms, and raw audio inputs. The traditional handcrafted audio features are from Essentia [3] using the feature extractor for AcousticBrainz. These features were used in the MediaEval genre recognition tasks. The number of mel-bands of the mel-spectrograms is 96. The raw audio inputs are in MP3 format with 44.1 kHz sampling rate.

3.3 Experiment

We only used the pre-computed mel-spectrograms (Figure 1) as inputs, and we used different data augumentation methods in training, validation and test dataset. Let T be the length of input section [frame]. For training dataset, we randomly cropped a T-frame section from each audio track in every epoch. For validation and test dataset, we respectively cropped 10 and 20 T-frame sections from each audio track at regular intervals. We averaged the predictions over all sections of each audio track. The length of input section T is 1,280 frames. We trained our networking using Adam with the batch size of 64 and the learning rate of 0.001.

4 RESULTS AND ANALYSIS

We compared the performance of the models that have different architectures or mechanisms in Table 2. Suprisingly, the model that achieved the best performance in our experiments was a relatively shallow model that only consists of six convolutional layers, the architecture of which is detailed introduced in Section 3.1. Moreover, the top-5 and bottom-5 tag-wise AUCs of the 6-layer model are showned in Table 3. The performance achieved by the best 6-layer model is in the fifth place among all 29 submissions.

The network with 25 convolutional layers consists of one 7x7, 24 3x3 convolutional layers and five max pooling layers for downsampling. It's commonly believed that deep models can achieve a better performance in image classification task. However, the model with deep architecture didn't always achieve a better performance in this task. We also tried residual architecture that commonly used for improving the performance of neural networks. However, the models with residual architecture didn't have an advantage in performance.

Table 2: Experiment result

Conv Layers	Residual	PR-AUC-macro	ROC-AUC-macro
5 (baseline)	No	0.1161	0.7475
6	**No**	**0.1256**	**0.7532**
16	Yes	0.1125	0.7393
18	Yes	0.1135	0.7460
25	No	0.1009	0.7319

Table 3: Top-5 and bottom-5 tag-wise AUCs

Tag	Rank	PR-AUC-macro	ROC-AUC-macro
summer	1	0.4698	0.9033
deep	2	0.4435	0.9137
corporate	3	0.4017	0.8849
epic	4	0.3886	0.8384
film	5	0.3606	0.7709
etro	52	0.0213	0.7943
holiday	53	0.0186	0.6856
cool	54	0.0185	0.6763
sexy	55	0.0145	0.7327
travel	56	0.0117	0.5990

5 DISCUSSION AND OUTLOOK

The number of samples (18K) in the dataset is relatively smaller than some image datasets (e.g. CIFAR-10: 60K, MS-COCO: 200K, ImageNet: 517K) and the length of audio data (>30s) is relatively longer than some sound datasets (e.g. UrbanSound8K: <4s, ESC-50: 5s, AudioSet: 10s). According to our experience, the generalization ability of models is especially important in this task. Therefore, it is reasonable that relatively shallow VGG-based network with strong generalization ability can achive better performance.

In the future, we plan to use all of the audio representations because we think it is interesting that we treat audio recognition as a multimodal task. Traditional handcrafted audio features and the raw audio inputs may bring great improvement in the performance of our model.

6 CONCLUSION

In our experiments, we applied several convolutional neural networks to recogonize the emotion and theme of music. A shallow VGG-based network that consists of six convolutional layers achieved the best performance with PR-AUC-macro of 0.1256 and ROC-AUC-macro of 0.7532. We think that the generalization ability of the models is very important in this task. The link to our source code: https://github.com/YiShengzhou12330379/Emotion-and-Theme-Recognition-in-Music-Using-Jamendo.

REFERENCES

[1] Miguel Angel Ferrer Ballester. 2018. A Novel Approach to String Instrument Recognition. In *Proceedings of Image and Signal Processing: 8th International Conference*, Vol. 10884. 165–175.

[2] Dmitry Bogdanov, Alastair Porter, Philip Tovstogan, and Minz Won. 2019. *MediaEval 2019: Emotion and Theme Recognition in Music Using Jamendo.* In MediaEval Benchmark Workshop.

[3] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez Gutiérrez, Sankalp Gulati, Herrera Boyer, and others. 2013. Essentia: An audio analysis library for music information retrieval. In *Proceedings of the International Society for Music Information Retrieval.* 493–498.

[4] Keunwoo Choi, George Fazekas, and Mark Sandler. 2016. Automatic tagging using deep convolutional neural networks. *arXiv preprint arXiv:1606.00298* (2016).

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 248–255.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 770–778.

[7] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, and others. 2017. CNN architectures for large-scale audio classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing.* 131–135.

[8] Renato Panda, Ricardo Malheiro, and Rui Pedro Paiva. 2018. Musical Texture and Expressivity Features for Music Emotion Recognition. In *Proceedings of the International Society for Music Information Retrieval.* 383–391.

[9] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[10] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 1–9.