HCMUS at MediaEval 2020: Emotion Classification Using Wavenet Features with SpecAugment and EfficientNet

Tri-Nhan Do[1,3], Minh-Tri Nguyen[1,3],
Hai-Dang Nguyen[1,3], Minh-Triet Tran[1,2,3], Xuan-Nam Cao[1,3]
[1]University of Science, VNU-HCM
[2]John von Neumann Institute, VNU-HCM
[3]Vietnam National University, Ho Chi Minh city, Vietnam
{dtnhan,nmtri17}@apcs.vn,nhdang@selab.hcmus.edu.vn,{tmtriet,cxnam}@fit.hcmus.edu.vn

ABSTRACT

MediaEval 2020 provided a subset of the MTG-Jamendo dataset, aimed to recognize mood and theme in music. Team HCMUS proposes several solutions to build efficient classifiers to solve this problem. In addition to the mel-spectrogram features, new features extracted from the wavenet model is extracted and utilized to train the EfficientNet model. As evaluated by the jury, our best result achieved of 0.142 in PR-AUC and 0.76 in the ROC-AUC measurement. With fast training and lightweight features, our proposed methods are potential to work well with deeper neural networks.

1 INTRODUCTION

Emotions and Themes in Music task in MediaEval [1] is difficult and challenging due to the ambiguity of tags in the real world. Mood is often influenced by human perception, different people will have different feelings, moreover, this is a multi-class classification problem with more than 56 tags. The dataset is pretty unbalanced in the distribution of mood labels, each audio music is composed of multi-labels that there can be many emotions in the same song.

To be able to solve this task, the authors have tried many methods, using many kinds of models, input features or loss functions. Our best result is an ensemble of two kinds of different methods, one using provided mel-spectrogram features with EfficientNet model and the other using waveNet features with MobileNetV2 model [7, 9].

2 RELATED WORK

Data augmentation is important when training neural network model. Traditional audio augmentation methods try to modify the speed of the waveforms or alter the original signal samples with noises, this method need much computational cost. With SpecAugment approach[6], they adjust the spectrogram by warping it in the time direction, masking blocks of consecutive frequency channels, and masking blocks of utterances in time. This approach is more simple, cost less time and resources.

WaveNet model is applicable in many problem of signal processing, time series forecasting and music generation[4]. Therefore, the authors also try following this approach by using a pre-trained WaveNet model to extract feature vectors from raw audio and then, using those features as inputs for convolutional neural networks.

3 APPROACH

We follow many approaches which include two main inputs: mel-spectrogram features and wavenet features.

3.1 Data analysis

As in the figure, the green part shows the audio with only one label mood/theme, the yellow part shows the audio with 2 to 3 moods, the red part shows the audio with more than 3 moods. Number of sample audio for training is 9949 with a total of 17885 moods. On average, each class will have 319 audio with a standard deviation of 202.75. The maximum number of moods of an audio is 8. Mood / theme that appears most is happy with 927 audios.

We can see that the data is extremely unbalanced, and some classes have no audio representing it entirely. Therefore, it is necessary to have a way to reduce the complexity of the data.

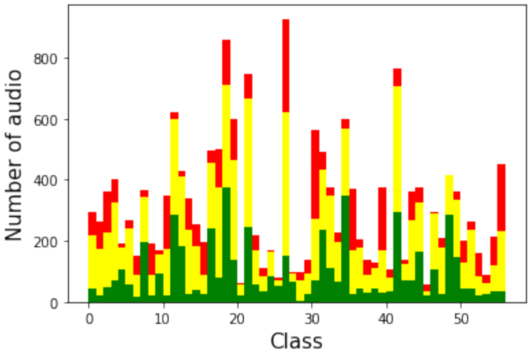

Figure 1: Histogram of mood and theme of training set

3.2 Data preprocessing

3.2.1 Data balance: To reduce the ambiguity of the data, the authors try to change each audio's label from multi-label to single label, keeping the most significant tag of each audio, reduce standard deviation, give preference to moods with little data.

3.2.2 Features preprocessing: **Wavenet feature**: Based on the idea of using wavenet as classifier for raw waveform music audio [5, 10], the authors use WaveNet-style autoencoder model that conditions an autoregressive decoder on temporal codes learned from the raw audio waveform, this model was pretrained from high-quality dataset of musical notes Nsynth [2].

Based on the dataset's statistic, the minimum length of audio is 30 seconds and based on the limitation of the authors' training machine, sound samples greater than 400 seconds in length will be

MediaEval'20, December 14-15 2020, Online

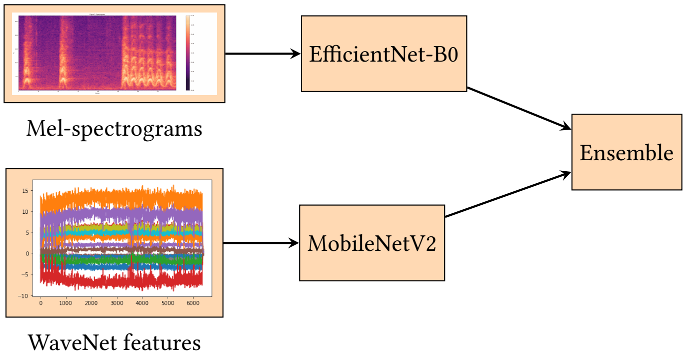

Figure 2: Overview of submission 1.

trimmed to take the middle part. Each sample is again randomly cut for 30 seconds and then extract features from them. This approach is quite subjective and causes loss of input data, we planned to experiment with random cutting from 400 seconds of audios after each epoch. The output of a 30 seconds audio is 16 frames multiply with 937-time steps.

Mel-spectrogram: Each sample feature has 96 channels and time frames are randomly cropped to 6950 after each epoch.

3.3 Data augmentation

SpecAugment: To train models more efficiently, the authors include segmentation method SpecAugment which was introduced by Google. This method masks blocks of consecutive time steps and channels in each mel-spectrogram. The result when using this method is increased significantly, PR-AUC-macro is improved from 0.134 to 0.139.

Each input have 70% chance to be augmented by using SpecAugment, each mel-spectrogram will have two blocks of time masking and two blocks of channel masking.

3.4 Deep Neural Network model

Since both mel-spectrogram features and wavenet feature can be expressed as images, the authors use convolutional models such as MobileNet and EfficientNet. The mel-spectrogram features are passed to EfficientNet-B0, on the other hand, the waveNet features are passed to MobileNetV2 and EfficientNet-B7. Because waveNet features are not large enough to fit EfficientNet-B7, the authors duplicate the number of channels so that these kinds of features can be used.

In addition, we also tested the SVM model, InceptionNet, Resnet, and to capture the long-term temporal characteristics, self-attention was added as in the method of AMLAG 2019[8], but this method produce a slight improvement in the result.

3.5 Loss function

For the loss function, binary cross entropy loss (BCE) is applied for both MobileNet V2 and EfficientNet. The authors also try to apply Focal Loss[3] since the dataset is pretty unbalanced, however it does not gain better results on our dataset after the balance step.

4 EXPERIMENTS AND RESULTS

Our experiments are done on a computer server with Nvidia Quadro k6000 graphic card. Method A,B and D are not submitted to the challenge. We realize that data balancing method leads to a better result comparing to the original dataset with default labels. Based on the experiments on the validation set, our ensemble models are calculated with factors of 0.7 and 0.3 for mel-spectrogram features and wavenet features to gain the best results.

Method	Features and Model	PR-AUC-macro
A	Mel-spectrogram EfficientNet-B0	0.127
B	Mel-spectrogram EfficientNet-B0 with data processing	0.134
C (run2)	Mel-spectrogram EfficientNet-B0 using augmentation	0.139
D	WaveNet MobileNetV2	0.102
E (run3)	WaveNet EfficientNet-B7	0.105
F (run1)	Ensemble C and D	0.1413
G (run4)	Ensemble C and E	0.1414

Table 1: Experiment results

5 CONCLUSION AND FUTURE WORKS

The EfficientNet model was shown to be more efficient than previous models in the mood and theme classification problem. The results can be improved by training mel-spectrogram features on other more complex EfficientNet models.

Although the result when training on wavenet features is not higher than mel-spectrogram features, but when assembling two models using these features, the results are improved, it shows that wavenet can extract other aspects of the dataset. Because the wavenet features were extracted by using a pretrained model, the augmentation methods have not been fully applied, for the future work, there are still more improvements to come when training WaveNet-style autoencoder models on the Jamendo dataset.

ACKNOWLEDGMENTS

Research is supported with computing infrastructure by SELAB and AILAB, University of Science, Vietnam National University - Ho Chi Minh City.

REFERENCES

[1] Philip Tovstogan Minz Won Dmitry Bogdanov, Alastair Porter. 2020. MediaEval 2020: Emotion and theme recognition in music using Jamendo. In *MediaEval 2020 Workshop*.

[2] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. 2017. Neural audio synthesis of musical notes with wavenet autoencoders. In *International Conference on Machine Learning*. PMLR, 1068–1077.

[3] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.

[4] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).

[5] Sandeep Kumar Pandey, HS Shekhawat, and SRM Prasanna. 2019. Emotion recognition from raw speech using wavenet. In *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*. IEEE, 1292–1297.

[6] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779* (2019).

[7] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4510–4520.

[8] Manoj Sukhavasi and Sainath Adapa. 2019. Music theme recognition using CNN and self-attention. *arXiv preprint arXiv:1911.07041* (2019).

[9] Mingxing Tan and Quoc V Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946* (2019).

[10] Xulong Zhang, Yongwei Gao, Yi Yu, and Wei Li. 2020. Music Artist Classification with WaveNet Classifier for Raw Waveform Audio Data. *arXiv preprint arXiv:2004.04371* (2020).