# SELAB-HCMUS at MediaEval 2021: Music Theme and Emotion Classification with Co-teaching Training Strategy

Phu-Thinh Pham[1,3], Minh-Hieu Huynh[1,3], Hai-Dang Nguyen[1,3], Minh-Triet Tran[1,2,3]
[1]University of Science, VNU-HCM
[2]John von Neumann Institute, VNU-HCM
[3]Vietnam National University, Ho Chi Minh city, Vietnam
{phpthinh18,hmhieu18}@apcs.fitus.edu.vn,nhdang@selab.hcmus.edu.vn,tmtriet@fit.hcmus.edu.vn

## ABSTRACT

MediaEval 2021 offers the third challenge motivating studies on automatically recognizing the emotions and themes conveyed in a music recording. Team SELAB-HCMUS proposes various methods to deal with this problem. In this competition, we have applied an efficient training strategy to solve the task. In addition, with short segments of input representations, the model achieves better results and a reduction in training time. From the official evaluation, the best result achieves 0.1435 PR-AUC and 0.7599 ROC-AUC measurements.

## 1 INTRODUCTION

The third edition of Emotions and Themes in Music task [12] is introduced in MediaEval workshop 2021. The aim of the task is to predict the mood and themes of given raw audio, with 56 tags in total, and audio can be labeled multiple tags, which could be considered as a multi-label classification problem.

Recognizing themes and emotions in music tracks is really important and has a wide range of applications, such as music recommendation systems or music analysis. However, this task is considered to be extremely difficult. Determining the emotional perspectives of a song can be quite ambiguous because a song's emotion or mood heavily depends on the sentiments of the one listening to it. Moreover, the length of each audio file is quite long, which can lead to an exponential growth in the number of parameters for the deep learning models, while the emotions and themes in music recordings could be determined by short segment audio.

In order to tackle these problems, we tried to find some alternative solutions to preprocess the data instead of training on the whole audio. Furthermore, we utilized some pre-trained CNN (Convolutional Neural Network) models to build the ensemble model, which achieves 0.1435 PR-AUC and 0.7599 ROC-AUC. Along with the mentioned methods, we also applied a model training approach called co-teaching to improve the efficiency of our models.

## 2 APPROACH

We have approached this problem in several different ways, which can be divided into 3 subsections, namely data pre-processing, model architecture, and co-teaching.

### 2.1 Data pre-processing

*2.1.1 Data shortening.* At first, we attempted to train the whole length of each track; however, it takes too much time to train, approximately 20 minutes/epoch. Through data analysis, we recognize that each music track can be counted as a sequence of repeated rhythms. Thus, instead of training the whole track, we perform a randomly cut on each track. During the training stage, each mel-spectrogram instance will be randomly cut to have the dimension of $96 \times 960$. In validating and testing stages, each sample is divided into 16 segments, and the final prediction would be the average score. This method has facilitated the training stage, which reduces the training time remarkably from 20 minutes to 15 minutes per epoch while preserving the models' accuracy.

*2.1.2 Mixup.* We adapt the learning principle Mixup [13] by training on convex combinations of pairs of data and their labels. The previous work [6, 7] has shown the essentialness of this method to enhance the performance and generalization of the models.

*2.1.3 SpecAugment.* We use a simple data augmentation method SpecAugment [8], originally used for speech recognition, to train models more effectively. This technique masks blocks of frequency channels and consecutive time steps in mel-spectrogram features. However, in validation and testing, we do not apply SpecAugment.

*2.1.4 Data balancing.* Following the study [1], we attempt to reduce the ambiguity in the data by changing the labels from multi-tags to single-tags. This has been proved to be effective in giving better results in comparison to default labels.

### 2.2 Model Architecture

Since the mel-spectrogram feature can be considered images, we can apply CNN models, which are frequently used in image and video processing, to the audio-based problem. The models used in our experiment are EfficientNet-B0 [10], ReXNet-100 [3], MixNet-S [11], ResNet [4], RegNet [9]. After conducting several experiments and evaluations on those models, EfficientNet and RexNet are the most suitable architectures for the task.

We tried to improve these two models more by applying the co-teaching paradigm, which will be explained in Section 2.3.

### 2.3 Co-teaching

Co-teaching [2] is an effective learning paradigm that trains 2 networks simultaneously, allowing them to teach each other. First, in each mini-batch, each network ignores a small part of data, considering only useful knowledge. Therefore, for each epoch, each network selects small-loss instances in the data and uses them to

*MediaEval'21, December 13-15 2021, Online*

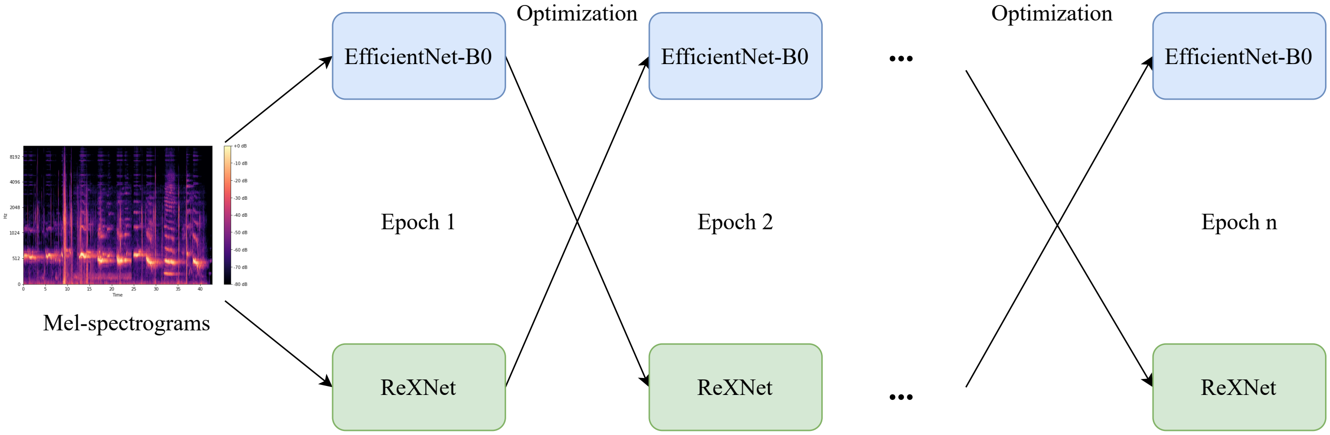

**Figure 1: Overview of the system trained with Co-teaching strategy**

optimize its peer network. This allows them to communicate with each other to distinguish helpful data to be used.

From the observation that this approach is potential, we have applied this method for the task. Since the system consists of 2 models, one model is EfficientNet [10]; for the other one, we choose ReXNet [3]. This is a combination of two networks discussed in Section 2.2. The illustration of the system flow is shown in Figure 1. In our experiments, we set the forget rate $\alpha = 0.05$, which decides the amount of data to be ignored.

## 3 SUBMISSIONS AND RESULTS

### 3.1 Experimental setup

To conduct experiments, we use Adam [5] optimizer for 100 epochs. We start training with a learning rate of $1 \times 10^{-3}$. The learning rate will be decreased 10 times after 5 consecutive epochs without improvement. The loss function used for development is Binary Cross Entropy (BCE). The experiments are carried out on Google Colab Pro, with the GPU NVIDIA Tesla P100.

### 3.2 Submissions

We submitted 4 runs, corresponding to 4 models having the highest PR-AUC measurement, to the MediaEval 2021 organizers.

- Run 1: Model EfficientNet-B0 trained by Co-teaching strategy with the peer network ReXNet in parallel.
- Run 2: Model ReXNet trained by Co-teaching strategy with the peer network EfficientNet-B0 in parallel.
- Run 3: Model ReXNet using data processing.
- Run 4: Ensemble model from Run 1 and Run 2, with the factor of 0.65 and 0.35 for Run 1 and Run 2, respectively.

### 3.3 Results

After experimenting and evaluating the selected models, the results are shown in Table 1. In comparison, the EfficientNet and ReXNet models trained with the co-teaching method have better accuracy than the traditional-trained models. Intended to increase the overall accuracy, from these two models, we created the ensemble model. To maximize the result, we apply linear optimization to find the optimal coefficient of each model based on the result from the validation set. Finally, we get the best result of 0.1435 in PR-AUC, which is slightly higher than working on an individual model.

| Model | Description | PR-AUC-macro |
|---|---|---|
| Ensemble<br>(Run 4) | Run 1 + 2 | 0.1435 |
| EfficientNet-B0<br>(Run 1) | Co-teaching<br>w/ ReXNet | 0.1415 |
| ReXNet<br>(Run 2) | Co-teaching<br>w/ EfficientNet-B0 | 0.1343 |
| ReXNet<br>(Run 3) | Using proposed<br>data processing | 0.1262 |
| EfficientNet-B0 [1] | Using data augmentation | 0.139 |

**Table 1: Model performance evaluation on the test set.**

## 4 CONCLUSION AND FUTURE WORKS

This paper introduces a method to process data before training, which is data shortening. We apply CNN, an approach commonly used in image-based classification problems, to the audio-based classification problem. Besides, by realizing the problem with a large number of tags, we make use of the co-teaching paradigm, which is designed to deal with the problem of the noisy labels. With all of those efforts, we achieve our highest PR-AUC of 0.1435.

For future work, we first aim to investigate the factors that could improve the model's accuracy. Then, we want to design more efficient models specified for the music emotion recognition task.

## ACKNOWLEDGMENTS

This work was funded by Gia Lam Urban Development and Investment Company Limited, Vingroup and supported by Vingroup Innovation Foundation (VINIF) under project code VINIF.2019.DA19

## REFERENCES

[1] Tri-Nhan Do, Minh-Tri Nguyen, Hai-Dang Nguyen, Minh-Triet Tran, and Xuan-Nam Cao. 2020. HCMUS at MediaEval 2020: Emotion Classification Using Wavenet Features with SpecAugment and EfficientNet. In *MediaEval 2020 Workshop*.

[2] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *arXiv preprint arXiv:1804.06872* (2018).

[3] Dongyoon Han, Sangdoo Yun, Byeongho Heo, and YoungJoon Yoo. 2021. Rethinking Channel Dimensions for Efficient Model Design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 732–741.

[4] 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.

[5] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[6] Khaled Koutini, Shreyan Chowdhury, Verena Haunschmid, Hamid Eghbal-Zadeh, and Gerhard Widmer. 2019. Emotion and theme recognition in music with Frequency-Aware RF-Regularized CNNs, In MediaEval 2019 Workshop. *arXiv preprint arXiv:1911.05833*.

[7] Khaled Koutini, Hamid Eghbal-Zadeh, Matthias Dorfer, and Gerhard Widmer. 2019. The receptive field as a regularizer in deep convolutional neural networks for acoustic scene classification. In *2019 27th European signal processing conference (EUSIPCO)*. IEEE, 1–5.

[8] 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).

[9] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. 2020. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10428–10436.

[10] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*. PMLR, 6105–6114.

[11] Mingxing Tan and Quoc V Le. 2019. Mixconv: Mixed depthwise convolutional kernels. *arXiv preprint arXiv:1907.09595* (2019).

[12] Philip Tovstogan, Dmitry Bogdanov, and Alastair Porter. 2021. MediaEval 2021: Emotion and Theme Recognition in Music Using Jamendo. In *MediaEval 2021 Workshop*.

[13] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412* (2017).