Emotion and Theme Recognition in Music with Frequency-Aware RF-Regularized CNNs

Khaled Koutini, Shreyan Chowdhury, Verena Haunschmid, Hamid Eghbal-zadeh, Gerhard Widmer
Johannes Kepler University Linz
firstname.lastname@jku.at

ABSTRACT

We present CP-JKU submission to MediaEval 2019; a Receptive Field-(RF)-regularized and Frequency-Aware CNN approach for tagging music with emotion/mood labels. We perform an investigation regarding the impact of the RF of the CNNs on their performance on this dataset. We observe that ResNets with smaller receptive fields – originally adapted for acoustic scene classification – also perform well in the emotion tagging task. We improve the performance of such architectures using techniques such as Frequency Awareness and Shake-Shake regularization, which were used in previous work on general acoustic recognition tasks.

1 INTRODUCTION

Content based emotion recognition in music is a challenging task in part because of noisy datasets and unavailability of royalty-free audio of consistent quality. The recently released MTG-Jamendo dataset [2] is aimed at addressing these issues.

The Emotion and Theme Recognition Task of MediaEval 2019 uses a subset of this dataset containing relevant emotion tags, and the task objective is to predict scores and decisions for these tags from audio (or spectrograms). The details of this specific data subset, task description, data splits, and evaluation strategy can be found in the overview paper [1].

Convolutional Neural Networks (CNNs) achieve state-of-the-art results in many tasks such as image classification [8, 10], acoustic scene classification [4, 16] and audio tagging [5]. These models can learn their own features and classifiers in an end-to-end fashion, which as a result reduces the need for task-specific feature engineering. Although CNNs are capable of learning high-level concepts given very simple and low-level information, the careful design of the network architectures in CNNs is a crucial step in achieving good results.

In a recent study [14, 16], Koutini et al. showed that the *receptive field (RF)* of CNN architectures is a very important factor when it comes to processing audio signals. Based on these findings, a regularization technique was proposed, that can significantly boost the performance of CNNs when used with spectrogram features. Further, in [17] a drawback of CNNs in the audio domain is highlighted, which is caused by the lack of spatial ordering in convolutional layers. As a solution, *Frequency-Aware (FA) Convolutional Layers* were introduced, to be used in CNNs with the commonly-used spectrogram input.

The proposed RF-regularization and FA-CNNs have shown great promise in several tasks in the field of Computational Auditory Scene Analysis (CASA), and achieved top ranks in international challenges [16]. In this report, we extend the previous work to Music Information Retrieval (MIR) and demonstrate that these models can be used to recognize emotion in music, and achieve new state-of-the-art results.

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

2 SETUP

2.1 Data Preparation

We used a sampling rate of 44.1 kHz to extract the input features. We apply a Short Time Fourier Transform (STFT). The window size for the STFT is 2048 samples and the overlap between windows is 75% for submissions 1, 2 and 3, and 25% for submissions 4 and 5. We use perceptually-weighted Mel-scaled spectrograms similar to [4, 14, 16], which results in an input having 256 Mel bins in the frequency dimension.

2.2 Optimization

In a setup similar to [14, 16, 17], we use Adam [13] for 200 epochs. We start with 10 epochs warm-up learning rate, we train with a constant learning rate of 1×10^{-4} for 60 epochs. After that, we use a linear learning rate scheduler for 50 epochs, dropping the learning rate to 1×10^{-6}. We finally train for 80 more epochs using the final learning rate.

2.3 Data Augmentation

Mix-up [21] has proven essential in our experiments to boost the perfomance and the generalization of our models. These results are consistent with experience from our previous work [14, 16, 17].

3 ADAPTING CNNS

Convolutional Neural Networks (CNNs) have shown great success in many acoustic tasks [4–6, 9, 11, 14–20]. In our submissions, we build on this success and investigate their performance on tasks more specific to music. We use mainly adapted versions of ResNet [8]. We adapt the architectures to the task using the guidelines proposed in Koutini et al.[14][1]. We use the CNN variants introduced in [17].

3.1 Receptive Field Regularization

Limiting the receptive field (RF) has been shown to have a great impact on the performance of a CNN in a number of acoustic recognition and detection tasks [14, 16]. We investigated the influence of the receptive field in this task in a setup similar to [14].

Figure 1 shows the PR-AUC on both the the validation (val) and testing (test) sets, for ResNet models with different receptive fields

[1]The source code is published at https://github.com/kkoutini/cpjku_dcase19

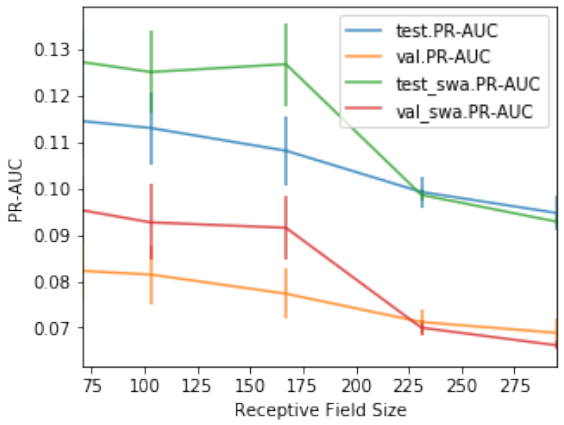

Figure 1: PR-AUC for ResNets with different RFs

and their SWA (see Section 3.4 below) variants. The results show the larger receptive field causes performance drops in accordance to the findings of [14]. Moreover, further experiments showed that size of the receptive field over the time dimension has lower significance on performance.

3.2 Frequency-Awareness and FA-ResNet

Figure 1 shows that smaller-RF ResNets perform better. As shown in [17], Frequency-Awareness can compensate for the lack of freuqency information caused by the smaller RF. We use Frequency-Aware ResNet (FA-ResNet) introduced in [17].

3.3 Shake-Shake Regularization

The Shake-Shake regularization [7] is proposed for improved stability and robustness. As shown in [16] and [17], although Shake-Shake ResNets do not perform well in the original acoustic scene classification problem, it performed really well in this task.

3.4 Model Averaging

Stochastic Weight Averaging: Similar to [16, 17], we use Stochastic Weight Averaging (SWA) [12]. We add networks weights to the average every 3 epochs. The averaged networks turned out to out-perform each of the single networks.
Snapshot Averaging: When computing the final prediction we also average the predictions of 5 snapshots of the networks during training. Specifically, we average the model with the highest PR-AUC on the validation set with the last 4 SWA models' predictions during training.
Multi-model Averaging: We average different models that have different architectures, initialization and/or receptive fields over time.

4 SUBMISSIONS AND RESULTS

4.1 Submitted Models

Overall, we submitted five models to the challenge: the first three are variations of the approach described above; the other two were models tested during our experiments, and were submitted as additional baselines against which to compare our modified CNNs.
ShakeFAResNet We average the prediction of 5 Shake-Shake regularized FA-ResNets with different initlizations. Their frequency RF is regularized as explained in Section 3.1. They have however different RF over the time dimension.
FAResNet similar to Shakefaresnet, but without Shake-Shake regularization.
Avg_ensemble We average the prediction of all the models included in both Shakefaresnet and Faresnet. In addition, we add a RF-regularized ResNet and DenseNet as introduced in [14].
ResNet34 In our preliminary experiments, Vanilla Resnet-34 outperformed Resnet-18 and Resnet-50 on the validation set, so we picked this architecture as an additional baseline.
CRNN The CRNN network was motivated by the notion that global structure of musical features could affect the perception of certain aspects of music (like mood), as mentioned by Choi et al [3]. We use an architecture similar to the one used by Choi et al, where the CNN part acts as the feature extractor and the RNN part acts as a temporal aggregator. This approach increased the performance from the baseline CNN and the Resnet-34.
CP_ResNet (not submitted to the challenge) We also show the results of a single model RF-regularized ResNet.

Table 1: PR-AUC results

Submission	Validation PR-AUC	Testing PR-AUC
ShakeFAResNet*	.1132	.1480
FAResNet*	.1149	.1463
Avg_ensemble*	**.1189**	**.1546**
ResNet34	.0924	.1021
CRNN	.0924	.1172
CP_ResNet	**.1097**	**.1325**
VGG-ish-baseline	-	.1077
popular baseline	-	.0319

*: indicates an ensemble.

4.2 Results

Table 1 shows the results of our submitted systems and compares them with the baselines. We can see that our RF-regularized and Frequency-Aware CNNs outperform the baselines by a significant margin, resulting in ranking as the top 3 submissions in the challenge. The systems that are marked with a star *, are ensembles of multiple models and snapshots (Section3.4). Table 1 also shows a single RF-regularized ResNet (CP_ResNet) can perform very well compared to the baselines.

ACKNOWLEDGMENTS

This work has been supported by the LCM – K2 Center within the framework of the Austrian COMET-K2 program, and the European Research Council (ERC) under the EU's Horizon 2020 research and innovation programme, under grant agreement No 670035 (project "Con Espressione").

REFERENCES

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

[2] Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. 2019. The MTG-Jamendo dataset for automatic music tagging.

[3] Keunwoo Choi, György Fazekas, Mark Sandler, and Kyunghyun Cho. 2017. Convolutional recurrent neural networks for music classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2392–2396.

[4] Matthias Dorfer, Bernhard Lehner, Hamid Eghbal-zadeh, Christop Heindl, Fabian Paischer, and Gerhard Widmer. 2018. Acoustic scene classification with fully convolutional neural networks and I-vectors. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Challenge (DCASE2018)*.

[5] Matthias Dorfer and Gerhard Widmer. 2018. Training general-purpose audio tagging networks with noisy labels and iterative self-verification. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*. 178–182.

[6] Hamid Eghbal-Zadeh, Bernhard Lehner, Matthias Dorfer, and Gerhard Widmer. 2016. CP-JKU Submissions for DCASE-2016: A Hybrid Approach Using Binaural i-Vectors and Deep Convolutional Neural Networks. In *DCASE 2016-challenge on Detection and Classification of Acoustic Scenes and Events*. DCASE2016 Challenge.

[7] Xavier Gastaldi. 2017. Shake-shake regularization. *arXiv preprint arXiv:1705.07485* (2017).

[8] 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.

[9] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson. 2017. CNN Architectures for Large-Scale Audio Classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2017-03). 131–135. https://doi.org/10.1109/ICASSP.2017.7952132

[10] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4700–4708.

[11] Turab Iqbal, Qiuqiang Kong, Mark Plumbley, and Wenwu Wang. Stacked convolutional neural networks for general-purpose audio tagging. DCASE2018 Challenge.

[12] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. 2018. Averaging Weights Leads to Wider Optima and Better Generalization. *arXiv preprint arXiv:1803.05407* (2018).

[13] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

[14] 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 *Proceedings of the European Signal Processing Conference (EUSIPCO)*. A Coruña, Spain.

[15] Khaled Koutini, Hamid Eghbal-zadeh, and Gerhard Widmer. 2018. Iterative Knowledge Distillation in R-CNNs for Weakly-Labeled Semi-Supervised Sound Event Detection. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)* (2018-11). 173–177.

[16] Khaled Koutini, Hamid Eghbal-zadeh, and Gerhard Widmer. 2019. *CP-JKU submissions to DCASE'19: Acoustic Scene Classification and Audio Tagging with Receptive-Field-Regularized CNNs*. Technical Report. DCASE2019 Challenge.

[17] Khaled Koutini, Hamid Eghbal-zadeh, and Gerhard Widmer. 2019. Receptive-field-regularized CNN variants for acoustic scene classification. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*.

[18] Donmoon Lee, Subin Lee, Yoonchang Han, and Kyogu Lee. Ensemble of Convolutional Neural Networks for Weakly-Supervised Sound Event Detection Using Multiple Scale Input. DCASE2017 Challenge.

[19] Bernhard Lehner, Hamid Eghbal-Zadeh, Matthias Dorfer, Filip Korzeniowski, Khaled Koutini, and Gerhard Widmer. 2017. Classifying Short Acoustic Scenes with I-Vectors and CNNs: Challenges and Optimisations for the 2017 DCASE ASC Task. In *DCASE 2017-challenge on Detection and Classification of Acoustic Scenes and Events*. DCASE2017 Challenge.

[20] Yuma Sakashita and Masaki Aono. 2018. Acoustic scene classification by ensemble of spectrograms based on adaptive temporal divisions. DCASE2018 Challenge.

[21] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond Empirical Risk Minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.