Music theme recognition using CNN and self-attention

Manoj Sukhavasi, Sainath Adapa
manoj.sukhavasi1@gmail.com,adapasainath@gmail.com

ABSTRACT

We present an efficient architecture to detect mood/themes in music tracks on `autotagging-moodtheme` subset of the MTG-Jamendo dataset. Our approach consists of two blocks, a CNN block based on MobileNetV2 architecture and a self-attention block from Transformer architecture to capture long term temporal characteristics. We show that our proposed model produces a significant improvement over the baseline model. Our model (team name: AMLAG) achieves 4th place on PR-AUC-macro Leaderboard in `MediaEval 2019: Emotion and Theme Recognition in Music Using Jamendo`.

1 INTRODUCTION

Automatic music tagging is a multi-label classification task to predict the music tags corresponding to the audio contents. Tagging music with themes (action, documentary) and mood (sad, upbeat) can be useful in music discovery and recommendation. `MediaEval 2019: Emotion and Theme Recognition in Music Using Jamendo` aims to improve the machine learning algorithms to automatically recognize the emotions and themes conveyed in a music recording [3]. This task involves the prediction of moods and themes conveyed by a music track, given the raw audio on the `autotagging-moodtheme` subset of the MTG-Jamendo dataset [4]. The overview paper [3] describes the task in more detail, and also introduces us to a baseline solution based on VGG-ish features. In this paper, we describe our Fourth place submission on PR-AUC-macro Leaderboard [1] which improves the results significantly on the baseline solution.

2 RELATED WORK

Conventionally feature extraction from audio relied on signal processing to compute relevant features from time or frequency domain representation. As an alternative to these solutions, architectures based on Convolutional Neural Networks(CNN) [6] have become more popular recently following their success in CV, speech processing. Extensions to CNNs have also been proposed to capture the long term temporal information in the form of CRNN [7]. Recently [20] has shown that self-attention applied to music tagging captures temporal information. This architecture was based on the transformer architecture which was very successful in Natural Language Processing (NLP)[19]. In this paper, we propose two methods MobileNetV2 and MobileNetV2 with self-attention which are based mainly on these two previous works [1, 20].

[1]https://multimediaeval.github.io/2019-Emotion-and-Theme-Recognition-in-Music-Task/results

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

3 APPROACH

We used the pre-computed Mel-spectrograms made available by the organizers of the challenge[2]. No additional pre-processing steps were undertaken other than the normalization of the input Mel-spectrogram features.

As image-based data augmentation techniques have been shown to be effective in audio tagging [1, 2], we used transformations such as Random crop and Random Scale. Additionally, we also employed SpecAugment and Mixup. SpecAugment[14] proposed initially for speech recognition, masks blocks of frequency channels or time steps of a log Mel-spectrogram. Mixup [22] samples two training examples randomly and linearly mixes them (both the feature space and the labels).

We propose two methods: MobilenetV2 architecture, and MobileNetV2 architecture combined with a self-attention block to capture long term temporal characteristics. We describe both of these methods in detail below.

3.1 MobileNetV2

It has been shown previously that using pre-trained ImageNet models helps in the case of audio tagging [1, 13]. Hence, we employed MobileNetV2 [17] for the current task. Since Mel-spectrograms are single channel, the input data is transformed into a three-channel tensor by passing it through two convolution layers. This tensor is then sent to the MobileNetV2 unit. As the number of labels is different here, the linear layer at the very end is replaced. No other modifications were performed to the original MobileNetV2 architecture.

3.2 MobileNetV2 with Self-attention

The architecture described in sub-section 3.1 might not be able to capture the long-term temporal characteristics. The dataset consists of tracks with varying lengths with a majority longer than 200s. Self-attention has been shown to capture long-range temporal characteristics in the context of music tagging [20]. Hence self-attention mechanism can be helpful in the current task. In this section, we describe our extended MobileNetV2 architecture with self-attention.

The architecture consists of 2 main blocks: modified MobileNetV2 (identical to the architecture described in [1]) to capture freq-time characteristics, and the self-attention block to capture long term temporal characteristics.

Similar to the transformer model [19], multi-head self-attention with positional encoding was implemented for the current architecture. Since our task consists only of classification we use only the encoder part of it similar to BERT [9]. Our implementation is based on the architecture described in [20]. We use 4 attention heads and 2 attention layers. The input sequence length is 16 and has embedding size of 256.

[2]https://github.com/MTG/mtg-jamendo-dataset

The control flow within this architecture is as follows:

- Input is a Mel-spectrogram tensor of length 4096 (number of bands being 96). This input is divided length-wise into 16 segments, with each segment's length being 256.
- Each of the 16 slices is sent through the modified MobileNetV2 block to extract the features.
- The feature maps are then fed into the Self-attention block. At the end of this block, two dense layers are put to use to generate the predictions.
- Additionally, the feature maps from the MobileNetV2 block are also used to generate predictions. With each segment, we have a set of predictions. All the sixteen predictions are averaged to obtain the final prediction.

As described above, the architecture generates two predictions: one solely using the MobileNetV2, and the other using the MobileNetV2 and the Self-attention blocks. While training, combined loss from both the predictions are used for back-propagation.

4 TRAINING AND RESULTS

We made two submissions under the team name AMLAG[3], one each using the two architectures described in sections 3.1 and 3.2. Both the submissions employ the same Mel-spectrogram inputs and Binary Cross-entropy loss as the optimization metric. PyTorch [15] was used for training the model in both cases.

For *submission 1*, the AMSGrad variant of the Adam algorithm [12, 16] with a learning rate of 1e-3 was utilized for optimization. Whenever the overall loss on the validation set stopped improving for five epochs, the learning rate was reduced by a factor of 10. For this training we use input Mel-spectrogram of length 6590, padding is used to make all the inputs of constant length. We observed that not all classes benefited from being trained together (see Figure 1). Hence, following the approach taken in [5], early stopping was done separately for each class based on the loss value for that particular class. Additionally, an attempt was made to find subsets of classes that train well together, but ultimately the overall performance had been lower than when all the classes were jointly trained. This is one avenue for future research with this dataset.

To prepare *submission 2*, we use input Mel-spectrogram of length 4096, padding is used to make all the inputs of constant length. We train the model for 120 epochs while utilizing Adam as the initial optimizer. We then employ an optimization technique proposed in [10, 20]: the optimizer is switched from Adam to Stochastic gradient descent (with Nesterov momentum [18]) after 60 epochs for better generalization of the model. Early stopping was done jointly for all classes based on the macro-averaged AUC-ROC on the validation set.

We present the results for both the submissions in Table 1. Also, results from the baseline approach that uses VGG-ish architecture are shown for comparison purposes. In all the metrics, the MobileNetV2 with a self-attention block exhibits an improvement over solely using the MobileNetV2. With respect to the baseline model, *submission 2* proved to be an improvement over all but the micro-averaged F-score and Precision metrics. On the task leaderboard, our model achieved 4th position in case of PR-AUC-macro, and 5th position in case of F-score-macro.

[3]https://github.com/sainathadapa/mediaeval-2019-moodtheme-detection

	Baseline (vggish)	Submission 1	Submission 2
PR-AUC-macro	0.107734	0.118306	0.125896
ROC-AUC-macro	0.725821	0.732416	0.752886
F-score-macro	0.165694	0.151891	0.182957
precision-macro	0.138216	0.135673	0.145545
recall-macro	0.30865	0.306015	0.39164
PR-AUC-micro	0.140913	0.150605	0.151706
ROC-AUC-micro	0.775029	0.784128	0.797624
F-score-micro	0.177133	0.152349	0.164375
precision-micro	0.116097	0.098133	0.10135
recall-micro	0.37348	0.340428	0.434691

Table 1: Performance on the test dataset

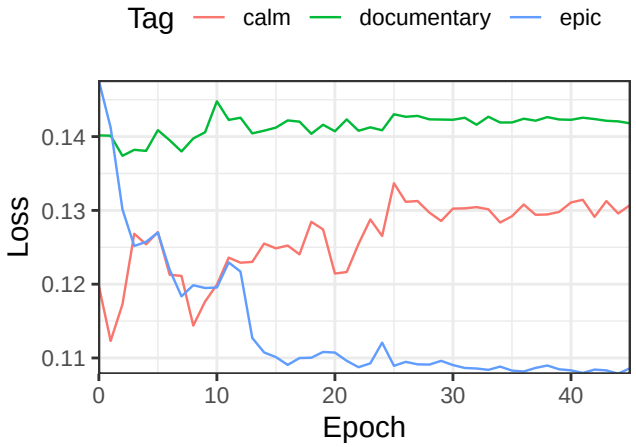

Figure 1: Trend in loss values for three sample classes while training the MobileNetV2 model. The plot illustrates the detail that not all classes were benefiting from joint training. In this case, the loss for *epic* class is decreasing while the loss for *calm* is increasing, *documentary* loss is almost stagnant.

5 OTHER APPROACHES

Some of the approaches that we have tried, but haven't observed better performance are listed below:

- A dense layer architecture that uses OpenL3 embeddings [8]
- A dense layer architecture that uses the pre-computed statistical features from Essentia using the feature extractor for AcousticBrainz. This data was made available by the organizers, along with the raw audio and Mel-spectrogram data.
- CNN architecture that directly uses the raw audio representation, as described in [11]
- Similar to using the MobileNetV2 in Section 3.1, we tested another ImageNet pre-trained architecture - ResNeXt model [21].

REFERENCES

[1] Sainath Adapa. 2019. Urban Sound Tagging using Convolutional Neural Networks. (2019). arXiv:cs.SD/1909.12699

[2] Ruslan Baikulov. 2019. Argus solution Freesound Audio Tagging 2019. (2019). https://github.com/lRomul/argus-freesound Accessed: 2019-10-01.

[3] Dmitry Bogdanov, Alastair Porter, Philip Tovstogan, and Minz Won. 2019. MediaEval 2019: Emotion and Theme Recognition in Music Using Jamendo. In *2019 Working Notes Proceedings of the MediaEval Workshop, MediaEval 2019*. 1–3.

[4] Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. 2019. The MTG-Jamendo Dataset for Automatic Music Tagging. In *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML 2019)*. Long Beach, CA, United States. http://hdl.handle.net/10230/42015

[5] Rich Caruana. 1998. A dozen tricks with multitask learning. In *Neural networks: tricks of the trade*. Springer, 165–191.

[6] Keunwoo Choi, George Fazekas, and Mark Sandler. 2016. Automatic tagging using deep convolutional neural networks. *arXiv e-prints*, Article arXiv:1606.00298 (Jun 2016), arXiv:1606.00298 pages. arXiv:cs.SD/1606.00298

[7] Keunwoo Choi, György 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

[8] Jason Cramer, Ho-Hsiang Wu, Justin Salamon, and Juan Pablo Bello. 2019. Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3852–3856.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv e-prints*, Article arXiv:1810.04805 (Oct 2018), arXiv:1810.04805 pages. arXiv:cs.CL/1810.04805

[10] Nitish Shirish Keskar and Richard Socher. 2017. Improving Generalization Performance by Switching from Adam to SGD. *CoRR* abs/1712.07628 (2017). arXiv:1712.07628 http://arxiv.org/abs/1712.07628

[11] Taejun Kim, Jongpil Lee, and Juhan Nam. 2019. Comparison and Analysis of SampleCNN Architectures for Audio Classification. *IEEE Journal of Selected Topics in Signal Processing* 13, 2 (2019), 285–297.

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

[13] Mario Lasseck. 2018. Acoustic Bird Detection With Deep Convolutional Neural Networks. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop*. Tampere University of Technology.

[14] 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).

[15] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. (2017).

[16] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. 2019. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237* (2019).

[17] 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.

[18] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. 2013. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*. 1139–1147.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv e-prints*, Article arXiv:1706.03762 (Jun 2017), arXiv:1706.03762 pages. arXiv:cs.CL/1706.03762

[20] Minz Won, Sanghyuk Chun, and Xavier Serra. 2019. Toward Interpretable Music Tagging with Self-Attention. *arXiv e-prints*, Article arXiv:1906.04972 (Jun 2019), arXiv:1906.04972 pages. arXiv:cs.SD/1906.04972

[21] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1492–1500.

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