Frequency Dependent Convolutions for Music Tagging

Vincent Bour
lileonardo, Paris, France
vincent.bour@lileonardo.com

ABSTRACT

We present a deep convolutional neural network approach for emotions and themes recognition in music using the MTG-Jamendo dataset. The model takes mel spectrograms as input and tries to leverage translation invariance in the time dimension while allowing convolution filters to depend on the frequency. It has led the lileonardo team to achieve the highest score of the 2021 MediaEval Multimedia Evaluation benchmark for this task[1].

1 INTRODUCTION

Emotions and Themes Recognition in Music using Jamendo is a multi-label classification task of the 2021 MediaEval Multimedia Evaluation benchmark. Its goal is to automatically recognize the emotions and themes conveyed in a music recording by means of audio analysis. We refer to [9] for more details on the task.

2 RELATED WORK

In the wake of its successes in computer vision, the use of convolution neural networks has become a very common approach for audio and music tagging and often leads to state-of-the-art results (see [10] for a comparison of different CNN approaches in music tagging). Among other things, [10] shows that a convolutional neural network trained on small excerpts of music constitutes a simple but efficient method for music tagging.

A number of previous works highlight the importance of the choice of filters with respect to the frequency dimension. In [7], the authors use domain knowledge to design vertical filters aimed to capture different spectral features. In [4], the authors add a channel in order to make the convolution filters frequency aware. They also study the influence of the receptive field and show that models with limited receptive field in the frequency dimension perform better.

As it is often the case with multi-label tagging tasks, there is a pronounced imbalance between positive and negative classes. The proportion of positive examples for each label in the training set ranges from 0.6% to 9.3%. To address this issue, the authors of [3] tried to adapt the loss function and used focal loss [6], which increases the weight of the wrongly classified examples.

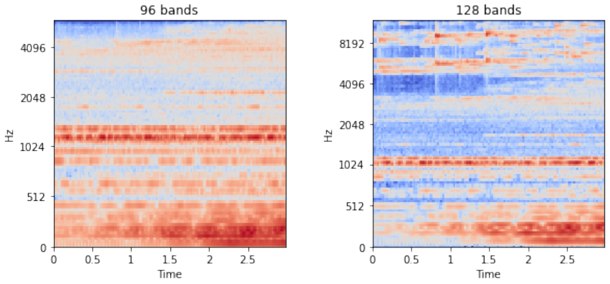
3 APPROACH

According to the results of [5], deep convolutional neural networks seem to be able to perform comparably well for music tagging tasks when used on raw waveforms or on mel spectrograms. After having experimented on a smaller dataset, and in order to save computation time, we have decided to restrict ourselves to using mel spectrograms as input.

The task of recognizing emotions and themes seems particularly well adapted to training on small excerpts of music. Indeed, while an instrument tag can be attributed to a song when it is only present on a small part of it, emotions and themes can often be recognized on most parts of the song. Moreover, by reducing the input time to lengths as small as 1 second, we observed that the ability of the network to perform its task did not radically diminish.

We have computed 128 frequency bands mel spectrograms from the original audio files by keeping the sample rate of 44100 Hz, using a fft window length of 2048 with 50% hop length and a low-pass filter with maximum frequency set to 16 kHz (see Figure 1 for a comparison between these mel spectrograms and the original 96 bands mel spectrograms provided by the MTG-Jamendo dataset).

Figure 1: An example of 96 and 128 bands mel spectrograms



We found that a simple CNN model overfitted rather quickly when trained on small random chunks of the tracks, unless a carefully engineered learning rate scheduling scheme was used. When we tried to train it on the 128 bands spectrograms, it performed significantly worse and overfitted even more quickly.

In order to try and overcome the class imbalance issue, we used a weighted loss, where for each label, the weight of the positive class is inversely proportionate to its frequency in the training set:

$$l(x,y) = -\frac{1}{c}\sum_{i=1}^{c}\left(\frac{2}{1+p_i}y_i\log(x_i) + \frac{2p_i}{1+p_i}(1-y_i)\log(1-x_i)\right),$$

where p_i is the frequency of the positive class in the train set for label i and $c = 56$ is the number of labels.

We also adopted an input stem constituted of three convolutional layers followed by a max pooling layer, inspired by [8] and now widely used in ResNets [1]. It significantly improved the results, probably because it resulted in a better low level analysis from the initial part of the model. It divides the time and frequency dimensions by a factor four while outputting 128 channels. Four more convolutional layers are aimed at performing a higher level analysis of the spectrogram. The dimension is then reduced to 1 × 1 by using a dense layer on the remaining frequency dimensions and by averaging along the time dimension.

[1] https://multimediaeval.github.io/2021-Emotion-and-Theme-Recognition-in-Music-Task/results

MediaEval'21, December 13-15 2021, Online

Table 1: Description of the layers for 128 frequency input dimension and 224 time input dimension

Layer	Input	Kernel / stride	Channels	Dropout
conv1	128 × 224	3 × 3 / 2	64	0
conv2	64 × 112	3 × 3 / 1	64	0
conv3	64 × 112	3 × 3 / 1	128	0
maxpool	64 × 112	3 × 3 / 2		
filter1	32 × 56	3 × 3 / 2	128	0.2
filter2	16 × 28	3 × 3 / 2	256	0.2
filter3	16 × 28	3 × 3 / 1	256	0.2
filter4	8 × 14	3 × 3 / 2	256	0.2
collapse	4 × 7	4 × 1 / 1	512	0.2
avgpool	1 × 7	1 × 7 / 1		
fc	1 × 1	1 × 1	1024	0.5
output	1 × 1	1 × 1	56	0.5

In order to give the model the possibility to detect different features at different frequencies, we tried and let the convolution filters of the higher level layers vary with the frequency. Despite the six fold increase in parameters with this approach, it gave better results on the 128 bands mel spectrograms.

The choice of this architecture comes from the hypothesis that the problem is invariant by translation in the time dimension, that the first layers compute low level visual features, and that at a higher level, the characteristic features depend on the frequency and are not similar on the low, middle or high frequency parts of the spectrogram.

We trained the network on randomly chosen small excerpts of each track. Final predictions are obtained by averaging the outputs for every small size segment in the track.

4 EXPERIMENTAL RESULTS

We evaluated the model described in Table 1, first with normal 3 by 3 convolutions, then with frequency dependent convolutions in layers `filter1-4`.

We trained these two models both on the original 96 bands mel spectrograms provided by the MTG-Jamendo dataset and on the 128 bands ones described in the previous section. We also tried input time lengths of 128 and 224, corresponding to excerpts of approximately 3s and 5.2s. Finally, we trained the models with a classic binary cross-entropy loss, with the weighted loss described in section 3 and with the focal loss with parameters $\alpha = 0.25$ and $\gamma = 2$. We then averaged the predictions of the trained models for any given choice of these hyperparameters to obtain the results in Table 2.

All models were trained for 200 epochs using SGD with Nesterov momentum of 0.9, a learning rate of 0.5 and weight decay of 2e-5. We used a cosine learning rate decay, mixup with $\alpha = 0.2$ [11] and stochastic weight averaging on the last 40 epochs [2].

A PyTorch implementation is available online[2].

Table 2: Experimental results

	Validation		Test	
	PR	ROC	PR	ROC
convs	**0.1169**	0.7434	0.1483	0.7715
freq-dep	0.1155	**0.7452**	**0.1504**	**0.7744**
mels-96	**0.1173**	**0.7471**	0.1479	**0.7723**
mels-128	0.1140	0.7389	**0.1483**	0.7710
input-128	0.1176	0.7448	0.1500	0.7743
input-224	**0.1180**	**0.7468**	**0.1507**	**0.7753**
bce	**0.1184**	0.7411	0.1472	0.7702
weighted	0.1143	**0.7453**	**0.1488**	**0.7744**
focal	0.1145	0.7420	0.1469	0.7678
ensemble	0.1179	0.7461	0.1506	0.7752

5 DISCUSSION AND OUTLOOK

The same training process was used for all models both for the sake of simplicity and in order to provide better comparison possibilities. However, the models with frequency dependent convolutions have a much higher number of parameters and may benefit from more regularization. It seems that with appropriate regularization, no overfitting occurs and the network could be trained for a longer number of epochs.

The results obtained on the test set, when broken down by label, are very different from those obtained on the validation set, in a way that is very consistent across all the models we have trained. On average, the score is significantly higher on the test set than on the validation set. Some labels (deep, summer, powerful...) are always better predicted on the test set whereas other labels (movie, action, groovy...) are always better predicted on the validation set. The fact that this is consistent across the models seem to show an inherent difference in the data rather than to be indicative of a high variance.

An increase in the input time length results in a small improvement. It is not clear whether this improvement is due to the additional input data seen by the network resulting in a virtual increase in the number of epochs, if it comes from a better management of side effects introduced by padding, or from better averages seen by the fully connected layer. Since the receptive field at the average pooling layer is independent of the input length, the model cannot actually use features present at a longer time scale.

The actual influence of the number of bands in the spectrogram and of the frequency of the low pass filter applied before the fft is not clear and would need further study.

The introduction of residual blocks in the place of the four filter layers seemed to provide a small but limited improvement. A resnet version of the model would be a good candidate to further improve the results.

ACKNOWLEDGMENTS

The author is grateful to François Malabre for very helpful discussions.

[2] https://github.com/vibour/emotion-theme-recognition

REFERENCES

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity Mappings in Deep Residual Networks. In *Computer Vision – ECCV 2016*. 630–645.

[2] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. 2018. Averaging weights leads to wider optima and better generalization. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*. 876–885.

[3] Dillon Knox, Timothy Greer, Benjamin Ma, Emily Kuo, Krishna Somandepalli, and Shrikanth Narayanan. 2020. MediaEval 2020 Emotion and Theme Recognition in Music Task: Loss Function Approaches for Multi-label Music Tagging. In *Proc. of the MediaEval 2020 Workshop, Online, 13–15 December 2020*.

[4] Khaled Koutini, Hamid Eghbal-zadeh, and Gerhard Widmer. 2019. Receptive-Field-Regularized CNN Variants for Acoustic Scene Classification. In *Acoustic Scenes and Events 2019 Workshop (DCASE2019)*. 124–128.

[5] Jongpil Lee, Jiyoung Park, Luke Kim, and Juhan Nam. 2017. Sample-level Deep Convolutional Neural Networks for Music auto-tagging Using Raw Waveforms. In *Proceedings of the 14th Sound and Music Computing Conference, July 5-8, Espoo, Finland*.

[6] 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.

[7] Jordi Pons, Oriol Nieto, Matthew Prockup, Erik M. Schmidt, Andreas F. Ehmann, and Xavier Serra. 2018. End-to-end Learning for Music Audio Tagging at Scale. In *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23–27, 2018*. 637–644.

[8] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[9] Philip Tovstogan, Dmitry Bogdanov, and Alastair Porter. 2021. MediaEval 2021: Emotion and Theme Recognition in Music Using Jamendo. In *Proc. of the MediaEval 2021 Workshop, Online, 13–15 December 2021*.

[10] Minz Won, Andres Ferraro, Dmitry Bogdanov, and Xavier Serra. 2020. Evaluation of CNN-based Automatic Music Tagging Models. In *17th Sound and Music Computing Conference (SMC2020)*.

[11] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*.