MediaEval 2019: Emotion and Theme Recognition in Music Using Jamendo

Dmitry Bogdanov, Alastair Porter, Philip Tovstogan, Minz Won
Universitat Pompeu Fabra, Spain
name.surname@upf.edu

ABSTRACT

This paper provides an overview of the Emotion and Theme recognition in Music task organized as part of the MediaEval 2019 Benchmarking Initiative for Multimedia Evaluation. The goal of this task is to automatically recognize the emotions and themes conveyed in a music recording by means of audio analysis. We provide a large dataset of audio and labels that the participants can use to train and evaluate their systems. We also provide a baseline solution that utilizes VGG-ish architecture. This overview paper presents the task challenges, the employed ground-truth information and dataset, and the evaluation methodology.

1 INTRODUCTION

Emotion and theme recognition is a popular task in music information retrieval that is relevant for music search and recommendation systems. We invite participants to try their skills at recognizing moods and themes conveyed by the audio tracks.

The last emotion recognition task in MediaEval [1] was in 2014, and there has been decline of interest since then. We bring the task back with openly available good quality audio data and labels from Jamendo.[1] Jamendo includes both mood and theme annotations in their database.

While there is a difference between emotions and moods, for this task we use the mood annotations as a proxy to understanding the emotions conveyed by the music. Themes are more ambiguous, but they usually describe well the concept or meaning that the artist is trying to convey with the music, or set the appropriate context for the music to be listened in.

Target audience: Researchers in areas of music information retrieval, music psychology, machine learning a generally music and technology enthusiasts.

2 TASK DESCRIPTION

This task involves the prediction of moods and themes conveyed by a music track, given an audio signal. Moods are often feelings conveyed by the music (e.g. happy, sad, dark, melancholy) and themes are associations with events or contexts where the music is suited to be played (e.g. epic, melodic, christmas, love, film, space). We do not make a distinction between moods and themes for the purpose of this task. Each track is tagged with at least one tag that serves as a ground-truth.

Participants are expected to train a model that takes raw audio as an input and outputs the predicted tags. To solve the task, participants can use any audio input representation they desire, be it traditional handcrafted audio features, spectrograms, or raw audio inputs for deep learning approaches. We also provide a handcrafted feature set extracted by the Essentia [2] audio analysis library as a reference. We allow the use of third-party datasets for model development and training, but this should be mentioned explicitly by participants if they do this.

We provide a dataset that is split into training, validation and testing subsets with mood and theme labels properly balanced between subsets. The generated outputs for the test dataset will be evaluated according to typical performance metrics.

3 DATA

The dataset used for this task is the *autotagging-moodtheme* subset of the MTG-Jamendo Dataset [3], built using audio data from Jamendo and made available under Creative Commons licenses. In contrast to other open music archives Jamendo targets its business on royalty free music for commercial use, including music streaming for venues. It ensures a basic technical quality assessment for their collection, thus the audio quality level is significantly more consistent with commercial music streaming services.

This subset includes 18,486 audio tracks with mood and theme annotations. There are 56 distinct tags in the dataset. All tracks have at least one tag, but many have more than one. The top 40 tags are shown in the Figure 1.

As part of the pre-processing of the dataset, some tags were merged to consolidate variant spellings and tags with the same meaning, (e.g., "dreamy" to "dream", "emotion" to "emotional"). The exact mapping is available in the dataset repository.[2] In addition, tracks shorter than 30 seconds were removed and tags used by less than 50 unique artists were discarded. Some tags were discarded while generating training, validation, and testing splits to ensure the absence of an artist and album effect [5] resulting in 56 tags after all pre-processing steps.

We provide audio files in 320kbps MP3 format (152 GB) as well as compressed .npy files with pre-computed mel-spectrograms (68 GB). Scripts and instructions to download the data are provided in the dataset repository.

3.1 Training, validation and test data

The MTG-Jamendo dataset provides multiple random data splits for training, validation and testing (60-20-20%). For this challenge we use *split-0*. Participants are expected to develop their systems using the provided training and validation splits.

[1] https://jamendo.com

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

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

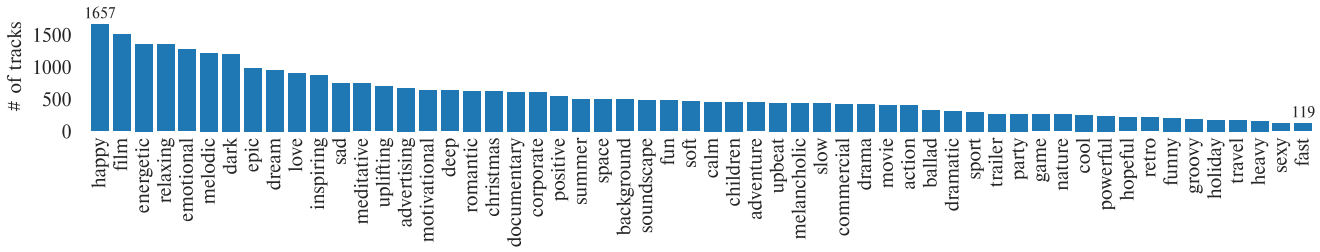

Figure 1: All tags

The validation set should be used for tuning hyperparameters of the models and regularization against overfitting by early stopping. These optimizations should not be done using the test set, which should be only used to estimate the performance of the final submissions.

We place no restrictions on the use of third party datasets for the development of the systems. In this case, we ask that participants also provide a baseline system using only data from the official training/validation set. Similarly, if one wants to append the validation set to the training data to build a model using more data for the final submission, a baseline using only training set for training should be provided.

4 SUBMISSIONS AND EVALUATION

Participants should generate predictions for the test split and submit those to the task organizers as well as self-computed metrics. We provide evaluation scripts in the GitHub repository.[3]

To have a better understanding of the behavior of the proposed systems, we ask participants to submit both *prediction* scores (probabilities or activation values) and binary classifications *decisions* for each tag for each track in the test set. We provide a script to calculate activation thresholds and generate decisions from predictions by maximizing macro F-score. See the documentation in the evaluation scripts directory in the dataset repository for instructions on how to do this.

We will use the following metrics, both types commonly used in the evaluation of auto-tagging systems:

- Macro ROC-AUC and PR-AUC on tag prediction scores.
- Micro- and macro-averaged precision, recall and F-score for binary decisions.

Participants should report the obtained metric scores on the validation split and test split if they have run such a test on their own. Participants should also report whether they used the whole development dataset or only a part for each submission.

We will generate rankings of the submissions by ROC-AUC, PR-AUC and micro and macro F-score. For leaderboard purposes we will use PR-AUC as the main metric, however we encourage comprehensive evaluation of the systems by using all metrics with the goal of generating more valuable insights on the proposed models when reporting evaluation results in the working notes. A maximum of five evaluation runs per participating team are allowed.

[3]https://github.com/MTG/mtg-jamendo-dataset/tree/master/scripts/mediaeval2019

Table 1: Baseline results

Metric	VGG-ish	Popular
ROC-AUC	0.725	0.500
PR-AUC	0.107	0.031
precision-macro	0.138	0.001
recall-macro	0.308	0.017
F-score-macro	0.165	0.002
precision-micro	0.116	0.079
recall-micro	0.373	0.044
F-score-micro	0.177	0.057

Note that we rely on the fairness of submissions and do not hide the ground truth for the test split. It is publicly available for benchmarking as a part of the MTG-Jamendo Dataset outside this challenge. For transparency and reproducibility, we encourage the participants to publicly release their code under an open-source/free software license on GitHub or another platform.

5 BASELINES

5.1 VGG-ish baseline approach

We used a broadly used VGG-ish architecture [4] as our main baseline. It consists of five 2D convolutional layers followed by a dense connection. The implementation is available in the MTG-Jamendo Dataset repository. We trained our model for 1000 epochs and used the validation set to choose the best model. We found optimal decision thresholds for the activation values individually for each tag, maximizing macro F-score. The evaluation results on the test set are presented in Table 1.

5.2 Popularity baseline

The popularity baseline always predicts the most frequent tag in the training set (Table 1). For the training set of *split-0* this is "happy".

6 CONCLUSIONS

By bringing Emotion and Theme recognition in Music to MediaEval we hope to benefit from contributions and expertise of a broader machine learning and multimedia retrieval community. We refer to the MediaEval 2019 proceedings for further details on the methods and results of teams participating in the task.

ACKNOWLEDGMENTS

We are thankful to Jamendo for providing us the data and labels.

This work has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 765068.

REFERENCES

[1] Anna Aljanaki, Yi-Hsuan Yang, and Mohammad Soleymani. 2014. Emotion in Music Task at MediaEval 2014.. In *MediaEval*.

[2] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J.R. Zapata, and X. Serra. 2013. Essentia: An Audio Analysis Library for Music Information Retrieval. In *International Society for Music Information Retrieval Conference*. Curitiba, Brazil.

[3] Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. 2019. The MTG-Jamendo Dataset for Automatic Music Tagging. In *Proceedings of the Machine Learning for Music Discovery Workshop, 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. http://mtg.upf.edu/node/3957

[4] Keunwoo Choi, George Fazekas, and Mark Sandler. 2016. Automatic tagging using deep convolutional neural networks. *arXiv preprint arXiv:1606.00298* (2016).

[5] Arthur Flexer and Dominik Schnitzer. 2009. Album and Artist Effects for Audio Similarity at the Scale of the Web. In *Sound and Music Computing Conference*.