The MTG-Jamendo Dataset for Automatic Music Tagging

Dmitry Bogdanov [1] **Minz Won** [1] **Philip Tovstogan** [1] **Alastair Porter** [1] **Xavier Serra** [1]

Abstract

We present the MTG-Jamendo Dataset, a new open dataset for music auto-tagging. It is built using music available at Jamendo under Creative Commons licenses and tags provided by content uploaders. The dataset contains over 55,000 full audio tracks with 195 tags from *genre*, *instrument*, and *mood/theme* categories. We provide elaborated data splits for researchers and report the performance of a simple baseline approach on five different sets of tags: genre, instrument, mood/theme, top-50, and overall.

1. Introduction

Music auto-tagging and related tasks of music genre, emotion and instrument recognition are common research topics in Music Information Retrieval that require annotated datasets of music audio. So far, only a few public datasets are available for researchers to prototype their systems for these tasks. In the context of auto-tagging, recent studies commonly use three datasets (Law et al., 2009; Bertin-Mahieux et al., 2011; Defferrard et al., 2017) summarized in Table 1 and known for their limitations. Researchers typically prototype on the smaller MTAT dataset and then validate their models on the larger MSD or FMA.

These datasets suffer from different problems such as the availability of only short audio segments instead of full tracks, low and inconsistent audio encoding quality, low coverage in terms of number of unique tracks and artists (MTAT), and noisy annotations (Choi et al., 2017). In the case of FMA, the origin of audio recordings is not well-moderated, with many recordings being of a low technical quality inconsistent with requirements for many industrial applications that are typically ensured by mastering and quality control teams in music distribution. In addition, there is no standard data split for benchmarking on MTAT and MSD, which leads to confusion in comparisons of the reported results.

Dataset	MTAT	MSD	FMA	MTG-Jamendo
Tracks	5,405[1]	505,216[2]	106,574[3]	55,609
Artists	230	N/A	16,341	3,565
Tags	188	522,366	161	195
Tag groups	No[4]	No[4]	No[5]	Yes[6]
Full tracks	No	No	Yes	Yes
CC-licensed	No	No	Yes	Yes
Bitrate	32	104	263	320
Sample rate	16	22-44.1	44.1	44.1

[1] Tracks are split into 25,877 individually annotated segments. [2] 7-Digital previews have been available. [3] Smaller subsets are also provided. [4] Mixed tags without categorization. [5] Only genre tags. [6] Genres, instruments, and moods/themes.

Table 1. Popular music auto-tagging datasets, compared to the proposed MTG-Jamendo Dataset. All audio is available as MP3 with different average bitrate (Kbps) and sample rate (KHz).

Considering these limitations, we propose a new dataset that consists of curated music with high quality full-length audio, which is closer to commercial music collections. This dataset may help developments in music auto-tagging as well as related tasks such as genre, instrument, and mood recognition. We provide an elaborated split and a benchmark for these tasks using a simple baseline.

2. Dataset

To create the MTG-Jamendo Dataset dataset we employed music publicly available on the Jamendo platform[1] under Creative Commons licenses. In contrast to other open music archives (such as the one used for FMA) 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, and we may expect its audio quality level to be significantly more consistent with commercial music streaming services.

We gathered a subset of music tracks from Jamendo annotated by tags provided by content uploaders and available via an API. In total we provide 55,701 full music tracks (with at least 30s duration) encoded as 320kbps MP3 (509 GB of audio). These tracks are annotated by 692 tags. The median

[1] https://jamendo.com

*Equal contribution [1]Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain. Correspondence to: Dmitry Bogdanov <dmitry.bogdanov@upf.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

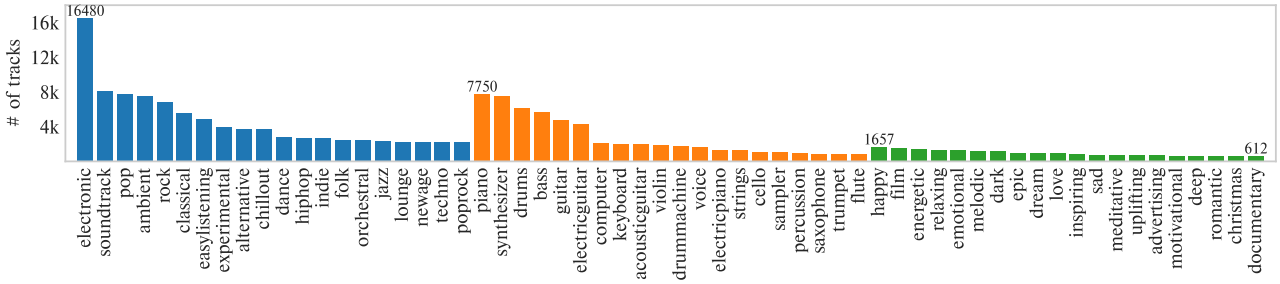

Figure 1. Top 20 tags per category (genre, instrument, mood/theme).

duration of a track is 224s and overall there are 3,777 hours of audio. Starting from this core dataset, we merged some tags to consolidate variant spellings and tags with the same meaning,[2] re-mapping 99 tags (less than 15%).

For the auto-tagging task, we propose to use a version of the dataset with tag filtering that only includes tags that have at least 50 unique artists. The statistics for this final dataset are presented in Table 2 and Figure 1. All tags are annotated by category and researchers can either work on all tags or subsets of tags they are interested in (genres, instruments, moods/themes, and top-50). We provide five random dataset splits (training, validation, and testing sets) ensuring that:

- No track appears in more than one set and no tracks in any set are from the same artist present in other sets;
- The same tags are present in all three sets across all splits;
- Tags are represented by at least 40 and 20 tracks from 10 and 5 artists in training and validation/testing sets, respectively.

In contrast to exiting datasets, we propose to use multiple splits for repeated evaluations to avoid possible biases of a particular random split.[3] Importantly, our split also avoids "artist and album effects" (Flexer & Schnitzer, 2009) which lead to possible overestimation of performance when a testing or validation sets contain tracks from the same artists/albums as the training set. The approximate split ratios are 60%, 20% and 20% for training, validation and testing. Partitioning scripts are provided to create alternative splits ensuring these characteristics in the data. We include track, artist and album identifiers in the dataset metadata.

The dataset, detailed statistics, pre-processing scripts and the implementation of the baseline are available online.[4] The metadata is released under the CC BY-NC-SA 4.0 license, while the audio files are available under their original Creative Commons licenses.

Subset	Tags	Tracks	Artists	ROC-AUC	PR-AUC
Genre	95	55,215	3,547	78.14	11.01
Instrument	41	25,135	2,006	67.47	12.74
Mood/Theme	59	18,486	1,533	67.19	8.19
All	195	55,609	3,565	71.97	7.36
Top-50	50	54,380	3,517	75.49	19.24

Table 2. The MTG-Jamendo Dataset statistics and baseline results.

3. Baselines for auto-tagging

As a simple baseline experiment, we tested a well-known music auto-tagging model proposed by Choi et al. (2016). It is a 5-layer convolutional neural network using 3×3 square filters. We followed all the original parameter settings. For each track we only used a centered 29.1s audio segment. All models were trained for 100 epochs. For more stable learning, we adopted an optimization trick introduced by Won et al. (2019). It starts with ADAM with a $1e{-}4$ learning rate and then switches to SGD with a learning rate of $1e{-}3$ and $1e{-}4$ at 60th and 80th epochs, respectively. At every switch point, it loads a saved model with the best Area Under the Receiver Operating Characteristic curve (ROC-AUC).

We report the obtained results in Table 2. Along with ROC-AUC, we also report Area Under the Precision Recall curve (PR-AUC) because ROC-AUC can give over-optimistic scores when the data is unbalanced (Davis & Goadrich, 2006), which is our case. The reported results in Table 2 are averages of tag-wise ROC-AUC and PR-AUC.[5] Sparse tags report extremely poor PR-AUC, and therefore the average PR-AUC is low. For comparison, we also report results using top-50 tags, which is a common evaluation setup in previous works and for which our baseline was originally optimized (Choi et al., 2016).

[2] E.g., mapping "synth" to "synthesizer", "guitarra" to "guitar". The exact mapping is available online.

[3] A few tags are discarded in the splits to guarantee the same list of tags across all five splits. This results in 87 genre tags, 40 instrument tags, and 56 mood/theme tags available for the evaluation.

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

[5] Per-tag ROC-AUC and PR-AUC are reported online.

Acknowledgements

This work was funded by the predoctoral grant MDM-2015-0502-17-2 from the Spanish Ministry of Economy and Competitiveness linked to the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502).

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

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688382 "AudioCommons".

References

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., and Lamere, P. The million song dataset. In *International Society for Music Information Retrieval Conference*, 2011.

Choi, K., Fazekas, G., and Sandler, M. Automatic tagging using deep convolutional neural networks. *arXiv preprint arXiv:1606.00298*, 2016.

Choi, K., Fazekas, G., Cho, K., and Sandler, M. The effects of noisy labels on deep convolutional neural networks for music classification. *arXiv preprint arXiv:1706.02361*, 2017.

Davis, J. and Goadrich, M. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240. ACM, 2006.

Defferrard, M., Benzi, K., Vandergheynst, P., and Bresson, X. FMA: A dataset for music analysis. In *International Society for Music Information Retrieval Conference*, 2017.

Flexer, A. and Schnitzer, D. Album and artist effects for audio similarity at the scale of the Web. In *Sound and Music Computing Conference*, 2009.

Law, E., West, K., Mandel, M. I., Bay, M., and Downie, J. S. Evaluation of algorithms using games: The case of music tagging. In *International Society for Music Information Retrieval Conference*, pp. 387–392, 2009.

Won, M., Chun, S., and Serra, X. Toward interpretable music tagging with self-attention. *arXiv preprint arXiv:1906.04972*, 2019.