IMPROVING MUSIC TAGGING FROM AUDIO WITH USER-TRACK INTERACTIONS

Andres Ferraro, Dmitry Bogdanov, Xavier Serra
Music Technology Group - UPF
{first.lastname}@upf.edu

Jay Ho Jeon, Jason Yoon
Kakao Corp
{jay.jh},{jason.yoon}@kakaocorp.com

EXTENDED ABSTRACT

Automatic tagging of audio is an important research topic that can serve for organizing large collections of music. By tagging music, we can facilitate recommendations according to user interests and generate playlists [3]. Current state-of-the-art systems for music auto-tagging using audio are based on deep learning, in particular, convolutional neural networks (CNNs) [2]. Previous works have shown that it is possible to improve the tagging by combining other sources of information like biographies of the artists, images from the covers of the albums or reviews from users [4].

In this preliminary work, we study the possibility of improving the tagging of tracks using audio and collaborative filtering (CF) information. We use Matrix Factorization (MF) to obtain a representation of the tracks from the user-track interaction matrix. Visualizing the space of MF representations using t-SNE we can see how the corresponding tags tend to group together. For this reason, we propose to apply the tag prediction directly from the MF representations of tracks and combine these predictions with the ones obtained from the audio.

In our study we use a common dataset split [5] of the Million Song Dataset (MSD) [1]. This split includes 201,680 tracks for training, 11,774 for validation and 28,435 for testing. We include the audio retrieved from 7digital.com and user-track interactions from the Echo Nest Taste Profile Subset [1]. Additionally, the Last.fm Dataset [1] is related to MSD and provides tags in the song level, which are extracted from last.fm. The tags were crowdsourced and cover genre, instrumentation, moods and eras.

We train a 200-dimension latent representation of tracks from the user-track interaction matrix based on WARP [6]. WARP is based on stochastic gradient descent and its novelty is that it approximates the ranking of the relevant items efficiently by doing a sampling trick. In our case, WARP samples tracks for a user and updates the representations only when the prediction is wrong, that is the sampled track is disliked by the user while being predicted higher than the liked tracks.

To validate our idea, we used t-SNE to visualize the track representations by reducing it to two dimensions. From this visualization, we can see that some tags are grouped together. In Figure 1 we show a small part of this latent space where is clear that tracks of the same tag are grouped together.

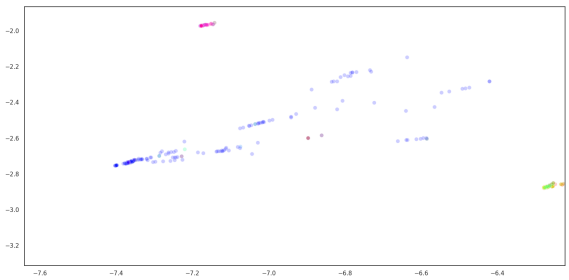

Figure 1: Small section of the latent space where tracks with the same tags (Hip-Hop) are grouped.

In this work, we use one of the state-of-the-art music auto-tagging models based on CNN to predict the tags for a large collection of audio and refine it by track-to-tag similarities in the MF latent space. We used the audio of 201,680 tracks for training and reserved 28,435 tracks for testing. We trained a model to predict the tags from the audio based on the architecture described by Choi [2] and used this as a baseline (*VGG*).

For each tag, we selected the tracks with the highest predicted probability by the VGG model and computed the mean of their MF representations. We used the resulting tag centroids to compute cosine distances to all music tracks, and use those as tag probabilities (Figure 2). This way we obtain both tag probabilities from VGG and distance-based probabilities comparing tracks to tag centroids in the MF space (*MF-centroids*).

In Table 1 we compare the performance of both models together with a hybrid method combining both predictions by averaging the obtained tag probabilities for each track (*Hybrid*). We see that the performance the VGG baseline is much higher then using *MF-centroids*, but combining both in a hybrid performance increases.

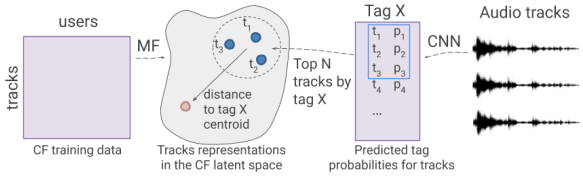

Figure 2: Overview of the proposed *MF-centroids* method.

Name	ROC AUC	PR AUC
VGG baseline	0.8672	0.2774
MF-centroids	0.5879	0.0953
Hybrid	**0.8700**	**0.2830**

Table 1: Performance of the tagging models.

We propose further directions to improve tagging results. One option is to use a clustering method to identify outliers for each tag and reduce the probability of potentially wrong predictions. Another option is to train a model to predict the tags directly from the MF representation learning other relations between tags and tracks which might not necessarily be captured by audio. Furthermore, we plan to explore ways to use the MF tag representations to generate track recommendations and generate playlists for users.

ACKNOWLEDGMENTS

This research has been supported by Kakao Corp.

REFERENCES

[1] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, 2011.

[2] Keunwoo Choi, György Fazekas, Kyunghyun Cho, and Mark Sandler. A tutorial on deep learning for music information retrieval. *arXiv preprint arXiv:1709.04396*, 2017.

[3] Andres Ferraro, Dmitry Bogdanov, Jisang Yoon, KwangSeob Kim, and Xavier Serra. Automatic playlist continuation using a hybrid recommender system combining features from text and audio. In *Proceedings of the ACM Recommender Systems Challenge 2018*, page 2, 2018.

[4] Sergio Oramas, Oriol Nieto, Francesco Barbieri, and Xavier Serra. Multi-label music genre classification from audio, text, and images using deep features. *arXiv preprint arXiv:1707.04916*, 2017.

[5] Jordi Pons, Oriol Nieto, Matthew Prockup, Erik M. Schmidt, Andreas F. Ehmann, and Xavier Serra. End-to-end learning for music audio tagging at scale. In *Proceedings of the 19th International Society for Music Information Retrieval*, pages 637–644, 2018.

[6] Jason Weston, Samy Bengio, and Nicolas Usunier. Large scale image annotation: learning to rank with joint word-image embeddings. *Machine learning*, 81(1):21–35, 2010.