MULTIMODAL METRIC LEARNING FOR TAG-BASED MUSIC RETRIEVAL

Minz Won[⋆,†] *Sergio Oramas*[⋆⋆] *Oriol Nieto*[⋆⋆] *Fabien Gouyon*[⋆⋆] *Xavier Serra*[⋆]

[⋆] Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
[⋆⋆] Pandora, Oakland, CA, United States of America

ABSTRACT

Tag-based music retrieval is crucial to browse large-scale music libraries efficiently. Hence, automatic music tagging has been actively explored, mostly as a classification task, which has an inherent limitation: a fixed vocabulary. On the other hand, metric learning enables flexible vocabularies by using pretrained word embeddings as side information. Also, metric learning has proven its suitability for cross-modal retrieval tasks in other domains (e.g., text-to-image) by jointly learning a multimodal embedding space. In this paper, we investigate three ideas to successfully introduce multimodal metric learning for tag-based music retrieval: elaborate triplet sampling, *acoustic* and *cultural* music information, and domain-specific word embeddings. Our experimental results show that the proposed ideas enhance the retrieval system quantitatively and qualitatively. Furthermore, we release the *MSD500*: a subset of the Million Song Dataset (MSD) containing 500 cleaned tags, 7 manually annotated tag categories, and user taste profiles.

Index Terms— Metric learning, Music retrieval, Multimodality, Auto-tagging

1. INTRODUCTION

Text-based search is one of the most common ways of browsing the internet. This information behavior is also prevalent when exploring music libraries: from querying editorial metadata (e.g., title, artist, album) to high-level music semantics (e.g., genre, mood). However, the annotation of music tags is demanding and time-consuming, especially when large music collections are available. To scale such annotation process, audio-based automatic music tagging has been actively explored by music information retrieval (MIR) researchers [1]. However, this categorical classification has an intrinsic limitation: it can only use a fixed vocabulary. When an out-of-category tag is queried, music tagging models tend to not properly generalize since new tags are not considered during training. In a real world scenario, users query a virtually unlimited amount of music tags. Hence, the music retrieval system needs to be more flexible beyond categorical models.

[†]Work performed during an internship with Pandora in 2019.

As opposed to categorical classification models, metric learning aims to construct distance metrics for establishing similarity of data [2, 3]. It can form a similarity metric between two instances from the same modality using shared weights (e.g., Siamese networks [4]) and this can be also easily expanded towards multiple modalities [5, 6]. By jointly learning a multimodal embedding space, metric learning has already demonstrated its suitability for cross-modal retrieval such as image-to-text [7, 5] and video-to-audio [8]. Metric learning facilitates the nearest neighbor search in the embedding space directly, while classification models require a two-step retrieval (i.e., tagging and ranking). Also, metric learning enables abundant vocabulary when pretrained word embeddings are used to represent tags as side information [7, 9].

Recent work in MIR showed the advantage of using metric learning with pretrained word embeddings for audio-based music tagging and classification [9]. Based on the proposed model, we investigate several ideas to successfully introduce metric learning for tag-based music retrieval.

Contribution. Our contribution is four-fold: (*i*) we show the importance of elaborate triplet sampling, (*ii*) we explore *cultural* and *acoustic* information to represent music, (*iii*) we examine domain-specific word embeddings, and (*iv*) we present a manually cleaned dataset for reproducibility.

2. MODEL

2.1. Related work

A triplet network [10] is a type of metric learning that uses a triplet loss to fit a metric embedding, where a positive example x_p belongs to the same class as an anchor x_a, and a negative example x_n is a member of a different one. The triplet network is optimized to satisfy $Sim(x_a, x_p) > Sim(x_a, x_n)$, where $Sim(.)$ is a learned similarity metric. As it learns by comparisons, instead of using direct labels, the triplet approach is expandable to leverage various data sources that are not explicitly labeled. Thanks to its flexibility, deep metric learning with the triplet loss has been actively used to solve a set of diverse MIR problems [6, 9].

Choi et al. [9] proposed a triplet network that learns a multimodal embedding of audio and word semantics. To handle unseen labels, the authors used pretrained GloVe embed-

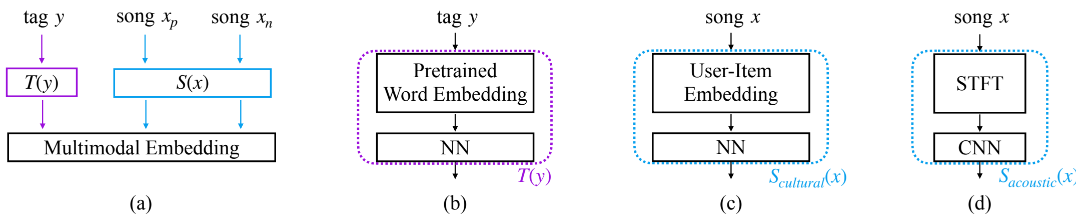

Fig. 1: (a) Overall architecture of the tag-based music retrieval model. (b) Tag embedding branch. (c) Song embedding branch with cultural information. (d) Song embedding branch with acoustic information.

dings [11] as side information. An audio embedding branch learns the mapping of the audio input to the multimodal embedding space. And another branch maps pretrained word embeddings to the shared multimodal embedding space. This metric learning model with side information demonstrated its versatility in multi-label zero-shot annotation and retrieval tasks. Since it can perform cross-modal retrieval (i.e., text-to-music), we adopt this architecture design as the backbone of our tag-based music retrieval model.

2.2. Model description

2.2.1. Architecture overview

Similar to previous work [9], our model is based on two branches. One branch $T(y)$ learns the mapping of tag semantics y to the embedding space, and another branch $S(x)$ learns the mapping of song information x to the shared embedding space — see Figure 1-(a). The model is trained to minimize the following loss function L:

$$L = [D(E_a, E_p) - D(E_a, E_n) + \delta]_+, \tag{1}$$

where D is a cosine distance function, δ is a margin, and E_a, E_p, E_n are mapped embeddings of anchor tag, positive song, and negative song, respectively. $[\cdot]_+$ is a rectified linear unit. The margin δ prevents the network from mapping all the embeddings to be the same (i.e., $L = 0$ for any inputs). With learnable transformations $T(y)$ and $S(x)$, the equation can be rewritten as:

$$L = [D(T(y_a), S(x_p)) - D(T(y_a), S(x_n)) + \delta]_+, \tag{2}$$

where y_a is the anchor tag input, and x_p and x_n are positive and negative song inputs, respectively. The following subsections depict the details of each branch $T(y)$ and $S(x)$.

2.2.2. Tag embedding

Figure 1-(b) shows the tag embedding branch $T(y)$. A given tag y passes through the pretrained word embedding model which results in a 300-dimensional vector. By using the pretrained word embeddings, the system can handle richer vocabulary than categorical models. For example, one can expect the system to handle plural forms (*guitar* and *guitars*), synonyms (*happy* and *cheerful*), acronyms (*edm* and *electronic dance music*), and dialectal forms (*brazil* and *brasil*). As our baseline, we use Word2Vec [12] embeddings pretrained with Google News dataset. The tag embedding is input to a neural network which is fully connected to a 512-dimensional hidden layer followed by a 256-dimensional output layer.

2.2.3. Song embedding

Pachet et al. [13] outlined three main types of music information: *editorial*, *cultural*, and *acoustic*. Most of the previous works in music tagging [1] and multimodal metric learning [6, 9], focused mainly on acoustic information to represent music. In our work, we attempt to operate on not only acoustic information but also cultural information in music retrieval. Cultural information is produced by the environment or culture. One of the most common methods to obtain cultural information is collaborative filtering [14].

The song embedding branch with cultural information $S_{cultural}(x)$ consists of a user-item embedding and a neural network — Figure 1-(c). The user-item embedding is obtained by factorizing a user-song interaction matrix. Weighted matrix factorization with the alternating least squares [15] is used, yielding both user and song embeddings of 200 dimensions each. User embeddings are discarded and song embeddings are used as our input. The input of the neural network is fully connected to a 512-dimensional hidden layer followed by a 256-dimensional output layer.

The song embedding branch with acoustic information $S_{acoustic}(x)$ learns audio-based music representation using a convolutional neural network (CNN) — Figure 1-(d). According to previous research [1], a simple 2D CNN with 3×3 filters could achieve competitive results to state-of-the-art in music tagging when it uses a short chunk of audio inputs (≈4s). For simplicity, we adopt the short-chunk CNN [1] to train our acoustic embedding.

The model is optimized using ADAM [16] with 10^{-4} learning rate, and 10^{-4} weight decay. The model is trained for 200 epochs where 1 epoch includes 10,000 triplets. For input preprocessing, audio files are downsampled to 22.5kHz then converted to Mel spectrograms using 1024-point FFT with 50% overlap and 128 Mel bands.

3. DATASET

The Million Song Dataset (MSD) [17] is a collection of metadata and precomputed audio features for 1 million songs. Along with this dataset, the Echo Nest Taste Profile Subset [18] provides play counts of 1 million users on more than 380,000 songs from the MSD, and the Last.fm Subset provides tag annotations to more than 500,000 songs from the MSD. We take advantage of these two subsets of the MSD to build our own dataset. Tags in the Last.fm Subset are very noisy, including 522,366 distinct tags. We performed a cleanup process of the dataset (e.g., merge synonyms or acronyms, fix misspelling) in order to have fewer tags while supported with a reasonable number of annotations. The detailed cleanup process is described in our online repository.

The final dataset contains 500 tag groups (from now on we call these groups "tags"), which yields 1,352 distinct Last.fm tags. These 500 tags are then manually classified in a lightweight taxonomy of 7 classes (genre, mood, location, language, instrument, activity, and decade). 158,323 distinct tracks are tagged with these 500 tags with an average of 3.1 tags per track and each track has user play counts. We release the final dataset as the *MSD500*.

In this paper, we use two different subsets of the proposed dataset which are *MSD100* and *MSD50*. Music tags are highly skewed towards few popular tags and handling this skewness is another big topic in data-driven approaches. Models are optimized to predict more popular tags in the training set while evaluation metrics are averaged over tags. To avoid the undesired effect of the high skewness, we only use the top 100 tags in our experiments which results in 115k songs (*MSD100*).

Although we have user information in our dataset, the interaction counts are not scalable compared to industry standards [19]. This may underrepresent the predictive power of cultural information. Hence, we build another subset which includes 39,402 songs with Last.fm tags and user-item embeddings from more than 100B in-house user explicit feedback. In this case we only use the top 50 tags (*MSD50*) because the size of the dataset became smaller during the mapping process. As the in-house user feedback includes sensitive information, we only release the song IDs and their tags of the *MSD50*. All data splits have been done at an artist level to avoid unintentional information leakage.

4. EXPERIMENTS

In this section we introduce three experiments which can be critical to enhance our metric learning approach for tag-based music retrieval. All models are evaluated with mean average precision (MAP) over the labels and precision at 10 (P@10). Reproducible code and dataset are available online. [1]

[1]https://github.com/minzwon/tag-based-music-retrieval

Metrics	Random	Balanced	Balanced-weighted
MAP	0.1658	0.1675	**0.1852**
P@10	0.2990	0.3160	**0.3500**

Table 1: Performance of different samplings (MSD100).

4.1. Sampling matters

The number of possible triplets grows cubically as the number of observations grows. Thus, triplet sampling is crucial in deep metric learning [20], as it matters equally or more than the choice of loss functions. In this subsection, we explore three different sampling methods: random sampling, balanced sampling, and balanced-weighted sampling.

Random sampling randomly chooses one song to generate an anchor-positive pair. Then a negative example is randomly sampled from a set of songs without the anchor tag. With this method, more popular tags are more likely to be sampled as the anchor tag. Also, songs with less popular tags are less likely to be sampled as negative examples due to their small numbers.

To alleviate this problem, the balanced sampling method uniformly samples an anchor tag first and then select a positive song. Minor tags may have equal possibilities to popular tags to be sampled as an anchor tag. By sampling negative examples from the batch of the positive songs, we can also expect more balanced tag distribution of negative examples.

For more efficient training, various triplet sampling methods have been proposed such as hard negative mining [21], semi-hard negative mining [22], and distance weighted sampling [20]. We combine the distance weighted sampling [20] with the aforementioned tag balancing method (i.e., balanced-weighted sampling). As in balanced sampling, we select an anchor tag and a positive song. From the batch of positive songs, we sample negative examples. Sampling weights are inversely proportional to their cosine distances from the anchor tags in the embedding space. Thus, more informative (harder) negative examples are more likely to be sampled while not loosing semi-hard and soft negative examples.

As shown in Table 1, balanced-weighted sampling outperforms other sampling methods. This proves that sampling matters for training our tag-based music retrieval model. Note that here we only used acoustic information for the song embedding to control the experiment. From now on, the following experiments use the balanced-weighted sampling method.

4.2. Acoustic and cultural music representation

We believe certain groups of tags are more related to acoustic information while others may be more culturally relevant. A tag *piano*, for example, can be predicted using the user-item matrix if there is a specific group of users who heavily listened to songs with piano. However, originally, the tag *piano* is associated with acoustic information. When there is a

Metrics	MSD100			MSD50		
	Cul-E	Acoustic	Concat	Cul-E	Cul-I	Acoustic
MAP	0.1155	**0.1852**	0.1775	0.2163	**0.4719**	0.3062
P@10	0.3200	**0.3500**	0.3120	0.4500	**0.6380**	0.4680

Table 2: Performance of cultural and acoustic models.

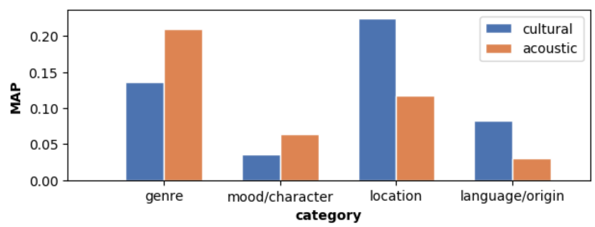

Fig. 2: Category-wise MAP on *MSD100*.

song beloved by the aforementioned user group, if we only use cultural information, the song can be regarded as piano music even when no piano can be acoustically perceived in the song. As another example, a tag *K-pop* can be predicted based on acoustic information since there are common acoustic characteristics of *K-pop*. However, if the song is not from Korea and is not being consumed in Korea, it should not be tagged as *K-pop*. To investigate the capability of two different information sources, we train our metric learning model with cultural information only and acoustic information only: $S_{cultural}$ and $S_{acoustic}$, respectively.

As shown in Table 2, the acoustic model outperforms the cultural model on *MSD100*. However, if we take a closer look at category-wise scores, the cultural model shows its strength in *location* and *language/origin* tags (Figure 2). This supports our hypothesis that the modality selection has to be associated with its original source of information. But a more important factor than the information source is the size and quality of available data. In Table 2 (*MSD50*), we have two different cultural models Cul-E and Cul-I, which use the EchoNest Taste Profile and our in-house user explicit feedback, respectively. Since our in-house data are of industry scale and explicit, they are richer than the publicly available data. As cultural information becomes richer (Cul-I), the cultural model outperforms the acoustic model. In addition, we observed that the cultural model with richer information (Cul-I) is superior in every tag category including *genre* and *mood*. As observed, acoustic and cultural models show different strengths, but the foremost important factor of the modality selection is the size and quality of available user-item interactions and audio data. We also experimented with a hybrid model with simple concatenation of cultural and acoustic embeddings but it did not improve results (Table 2-Concat).

[0]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).

Tag	GoogleNews	Domain-specific
Jungle	jungles, dense_jungle, dense_jungles, rainforest, thick_jungles, Amazon_jungle, Amazonian_jungle, steamy_jungles, hilly_jungle, swamps	**breakbeat, dub, drum_n_bass, drum'n'bass, grime, deep_house, ragga, dubstep, acid, acid_house**

Table 3: Nearest words in GoogleNews and domain-specific word embeddings. Music-related words are emboldened.

4.3. Domain-specific word embeddings

We use pretrained Word2Vec [12] embeddings as a part of our tag branch $T(y)$. Since they are trained with Google News, the embeddings are not expected to have musical context.

We pretrain our own word embeddings with musical text data. We use the corpus of text from the subtask 2B of the SemEval-2018 Hypernym Discovery Task [2]. It contains an already tokenized 100M-word corpus including Amazon reviews, music biographies, and Wikipedia pages about theory and music genres. We train a Word2Vec model on this corpus with a window of 10 words yielding word embeddings for unigrams, frequent bigrams and trigrams of 300 dimensions.

We could not discover any quantitative performance gain by using our domain-specific word embeddings. However, as shown in Table 3, the domain-specific word embeddings may include more musical context. For example, for the unseen query *jungle*, a model with domain-specific embeddings could successfully retrieve relevant items while conventional embeddings could not. Also, domain-specific music corpora include frequent bigrams and trigrams, such as *deep house* or *smooth jazz*, which are not typically captured in word embeddings trained on general text corpora. More qualitative examples are included in our online repository.

5. CONCLUSION

In this paper, we explored three different ideas to enhance the quality of metric learning for tag-based music retrieval. Balanced-weighted sampling could successfully improve the evaluation metrics. Cultural and acoustic models showed different strengths based on the information source of the given tag but the foremost important factor is the size and quality of available data. Finally, domain-specific word embeddings showed their suitability for music retrieval by including more musical context.

As future work, in-depth comparison of acoustic and cultural models is necessary to better understand how the size and the quality of data affect the results. Also, a hybrid method of fusing acoustic and cultural information should be explored. Finally, to meet real-world expectations, multi-tag retrieval systems have to be considered.

[2]https://competitions.codalab.org/competitions/17119#learn_the_details-terms_and_conditions

6. REFERENCES

[1] Minz Won, Andres Ferraro, Dmitry Bogdanov, and Xavier Serra, "Evaluation of cnn-based automatic music tagging models," in *Proc. of the 17th Sound and Music Computing*, 2020.

[2] Eric P Xing, Michael I Jordan, Stuart J Russell, and Andrew Y Ng, "Distance metric learning with application to clustering with side-information," in *Advances in neural information processing systems*, 2003.

[3] Kilian Q Weinberger and Lawrence K Saul, "Distance metric learning for large margin nearest neighbor classification.," *Journal of Machine Learning Research*, vol. 10, no. 2, 2009.

[4] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, 2015.

[5] Liwei Wang, Yin Li, and Svetlana Lazebnik, "Learning deep structure-preserving image-text embeddings," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[6] Sergio Oramas, Francesco Barbieri, Oriol Nieto, and Xavier Serra, "Multimodal deep learning for music genre classification," *Transactions of the International Society for Music Information Retrieval. 2018; 1 (1): 4-21.*, 2018.

[7] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov, "Devise: A deep visual-semantic embedding model," in *Advances in neural information processing systems*, 2013.

[8] Didac Surís, Amanda Duarte, Amaia Salvador, Jordi Torres, and Xavier Giró-i Nieto, "Cross-modal embeddings for video and audio retrieval," in *Proc. of the European Conference on Computer Vision*, 2018.

[9] Jeong Choi, Jongpil Lee, Jiyoung Park, and Juhan Nam, "Zero-shot learning for audio-based music classification and tagging," *In Proc. of the 20th International Society for Music Information Retrieval Conference*, 2019.

[10] Elad Hoffer and Nir Ailon, "Deep metric learning using triplet network," in *International Workshop on Similarity-Based Pattern Recognition*. Springer, 2015.

[11] Jeffrey Pennington, Richard Socher, and Christopher D Manning, "Glove: Global vectors for word representation," in *Proc. of the 2014 conference on empirical methods in natural language processing*, 2014.

[12] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," *International Conference on Learning Representations*, 2013.

[13] Francois Pachet, "Knowledge management and musical metadata," *Idea Group*, vol. 12, 2005.

[14] Jonathan L Herlocker, Joseph A Konstan, and John Riedl, "Explaining collaborative filtering recommendations," in *Proc. of ACM conference on Computer supported cooperative work*, 2000.

[15] Yifan Hu, Yehuda Koren, and Chris Volinsky, "Collaborative Filtering for Implicit Feedback Datasets," in *Proc. of the 8th IEEE International Conference on Data Mining*, 2008.

[16] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 2015.

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

[18] Brian McFee, Thierry Bertin-Mahieux, Daniel P.W. Ellis, and Gert R.G. Lanckriet, "The million song dataset challenge," *Proc. of the 21st international conference companion on World Wide Web*, 2012.

[19] Filip Korzeniowski, Oriol Nieto, Matthew McCallum, Minz Won, Sergio Oramas, and Erik Schmidt, "Mood classification using listening data," *In Proc. of the 19th International Society for Music Information Retrieval Conference*, 2018.

[20] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl, "Sampling matters in deep embedding learning," in *Proc. of the IEEE International Conference on Computer Vision*, 2017.

[21] Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, Pascal Fua, and Francesc Moreno-Noguer, "Discriminative learning of deep convolutional feature point descriptors," in *Proc. of the IEEE International Conference on Computer Vision*, 2015.

[22] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. of the IEEE conference on computer vision and pattern recognition*, 2015.