TOWARD UNIVERSAL TEXT-TO-MUSIC RETRIEVAL

SeungHeon Doh, Minz Won, Keunwoo Choi, Juhan Nam

Graduate School of Culture Technology, KAIST, South Korea

ByteDance, USA

Gaudio Lab, South Korea

ABSTRACT

This paper introduces effective design choices for text-to-music retrieval systems. An ideal text-based retrieval system would support various input queries such as pre-defined tags, unseen tags, and sentence-level descriptions. In reality, most previous works mainly focused on a single query type (tag or sentence) which may not generalize to another input type. Hence, we review recent text-based music retrieval systems using our proposed benchmark in two main aspects: input text representation and training objectives. Our findings enable a universal text-to-music retrieval system that achieves comparable retrieval performances in both tag- and sentence-level inputs. Furthermore, the proposed multimodal representation generalizes to 9 different downstream music classification tasks. We present the code and demo online.

Index Terms—Cross-modal retrieval, Text-based retrieval, Music retrieval

1. INTRODUCTION

The demand for efficient music retrieval has been increasing as massive music libraries become easily accessible. While various methods have been proposed for efficient retrieval [1, 2], text-based retrieval remains the most prevalent [3]. Text-based retrieval is challenging because it needs to handle not only editorial metadata (e.g., title, artist, release year) but also semantic information (e.g., genre, mood, theme). Furthermore, modern retrieval systems, such as voice assistants [4], need to generalize to sentence-level natural language inputs beyond fixed tag vocabularies.

While much research has addressed text-based retrieval, there are two dominant approaches: classification and metric learning. Classification models [5, 6] are trained with a set of fixed tag labels, and then the predicted tags are utilized in retrieval. Despite its successful classification performance, this approach is limited to a fixed vocabulary. In contrast, metric learning models are more flexible by using pre-trained word embeddings [7, 8] or language models [9, 10, 11]. Especially pre-trained language models enable free-form text inputs for music retrieval by representing sentence-level semantics. There are multiple loss functions (e.g., triplet loss, contrastive loss) for metric learning based on its training objective.

An ideal text-based retrieval system needs to be flexible to allow various input types (e.g. word, sentence) and abundant vocabularies. For example, one can use broadly used tags, such as genre, to explore various input types (e.g. word, sentence) and abundant vocabularies. Sometimes the input queries may include unseen types of music tags. Also, another can use more detailed sentence-level descriptions to discover music. However, to the best of our knowledge, previous works mainly focused on improving a single type of input queries. Also, they are using respective datasets and evaluation metrics which makes it difficult to choose the appropriate solution for universal music retrieval.

To address this issue, we perform the holistic evaluation of recently proposed text-to-music retrieval approaches. We first review the training objectives and modality encoding of the previous works, then propose a novel stochastic sampling of text inputs to enable a generalizable text encoder in Section 2. We introduce a text-music paired dataset and an evaluation benchmark to assess the system’s generalizability in Section 3. Section 4 describes experimental results. Finally, Section 5 proposes reusable insights for designing universal text-to-music retrieval systems.

2. MUSIC AND TEXT REPRESENTATION LEARNING

This section introduces recent text-music representation learning models (illustrated in Figure 1). We carefully review three training objectives and their modality encoders. In the following descriptions, \( x^a \) denotes a music audio example, \( x^t \) its paired text data, \( f(\cdot) \) an audio encoder, and \( g(\cdot) \) a text encoder. Each modality input is processed by the corresponding encoder \( f \) or \( g \) (described in 2.2). Each encoder consists of a backbone model, a linear projection, and an \( l_2 \) normalization layer. We denote the two output embeddings of audio and text as \( z^a = f(x^a) \) and \( z^t = g(x^t) \), respectively. The
The goal of classification model is to learn multi-label classification of the tags. The classification model does not have a text encoder since we directly perform multi-label classification of the tags.

2.1. Training Objective

**Classification Model** The goal of classification model is to learn a linearly discriminate embedding space. This can be also interpreted from the similarity-based metric learning perspective as introduced in [2]. The prediction score of model for each class follows:

\[ y = \text{sigmoid}(z^a \cdot c_y), \]

where \( c_y \) is a centroid vector for each class (parameters of the last dense layer). To maximize the similarity between the audio embedding \( z^a \) and the positive class \( c_y \), the objective function is formulated as follows:

\[ L_{ce} = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})) \]  \hspace{2cm} (1)

where \( y \) is a multi-hot label from the ground truth. Since the prediction score of a track is utilized as a similarity score with the centroid vector (the tag label), the classification-based model serves as the baseline system for tag-based retrieval. The classification model is limited to a fixed vocabulary since it cannot take advantage of text embeddings in a zero-shot retrieval scenario.

**Triplet-Loss Model** The goal of triplet-loss models is to learn an embedding space where relevant input pairs are mapped closer than irrelevant pairs in the latent space. The objective function is formulated as follows:

\[ L_{a\rightarrow t} = [0, \delta - z^a \cdot z^t_{pos} + z^a \cdot z^a_{neg}]_+ \]  \hspace{2cm} (2)

where \( \delta \) is the margin, \( z^t_{pos} \) denotes the paired text embedding for the music audio, and \( z^a_{neg} \) denotes irrelevant text embedding. \([ \cdot ]_+\) indicates a rectified linear unit. In practice, an efficient negative sampling is crucial in triplet-based metric learning. We applied the distance-weighted sampling method used in [8]. Using structure-preserving constraints [13, 14], we utilize a symmetric loss function: \( L_{a\rightarrow t} = (L_{a\rightarrow t} + L_{t\rightarrow a})/2 \).

**Contrastive-Loss Model** The core idea of contrastive-loss models is to reduce the distance between positive sample pairs while increasing the distance between negative sample pairs. Unlike triplet-loss models, contrastive-loss models can utilize a large number of negative samples that exist in a mini batch \( N \). During training, the audio and text encoders are jointly trained to maximize the similarity between \( N \) positive pairs of (music, text) associations while minimizing the similarity for \( N \times (N - 1) \) negative pairs. This is known as InfoNCE loss [15, 16] and formulated as follows:

\[ L_{a\rightarrow t} = -\log \frac{\exp(z^a_i \cdot z^t_j/\tau)}{\sum_1^N \exp(z^a_i \cdot z^t_j/\tau)} \]  \hspace{2cm} (3)

where \( \tau \) is a learnable parameter. The loss function is designed as follows: \( L_{a\rightarrow t} = (L_{a\rightarrow t} + L_{t\rightarrow a})/2 \).

2.2. Audio Encoding

For all experiments, we utilize a modified version of Music Tagging Transformer [12] as our audio encoder. The first four convolution layers capture local acoustic features and the following four transformer layers summarize the sequence. The output of convolutional layers (audio sequence) attaches [CLS] token at the first position, and the output of the last layer of the transformer at the [CLS] token is treated as the feature that represents the whole audio. It is, finally, linearly projected into an embedding space. We use mel spectrograms as input without any augmentation.

2.3. Text Encoding

We use tag and sentence text representation for input of the text encoders. For this, we use a pre-trained word embedding GloVe [17] and a pre-trained Bidirectional Encoder Transformer (BERT) [18] with a base-uncased architecture. In using both text encoders, tag and sentence representations are processed differently. The tag representation uniformly samples one tag among multi-label texts. The sentence representation uses the entire multi-label text by concatenating multi-label text. In the case of the GloVe model, the sentence text is tokenized by white space, and projected to joint embedding space, then average the sentence embedding sequence.\(^3\) In the case of BERT model, the input text sequence is tokenized by wordpiece tokenizer, and the max sequence length is 64. Similar to audio feature embeddings, the text sequence is attached with a [SOS] token at first position and the output of the last layer of the transformer at the [SOS] token are treated as the feature representation of the text, after layer normalization and a linear projection to the embedding space.

2.4. Stochastic Text Representation

In a preliminary study, we found that there is a strong association between text representation (train stage) and text query types (test stage). As somewhat obviously, the model works better when the input forms during the training phase and test phase are homogeneous, there are no references studying the relationship between text representation and retrieval performance. To use the advantages of both, we propose a stochastic text representation. During the training stage, we select \( K \) words from a text sentence length of \( L \). \( K \) is uniformly randomly sampled among integer numbers from 1 (word length) to \( L \) (sentence length). Unlike the dropout method, which determines the length by probability value, stochastic sampling has a dynamic input length.

\(^3\)We tested early fusion and late fusion of word embedding, but there was no significant difference in the results.

| MSD Subset | # of Track | # of Artist | # of Album | # of Tag | # of Caption | Avg.Tag | A/S | Genre | Style | Inst | Vocal | Mood | Theme | Culture |
|------------|------------|-------------|-------------|----------|--------------|--------|-----|-------|-------|------|-------|-------|--------|---------|
| Top50s [5, 6] | 241,889    | 25,239      | 67,495      | 50       | 11,418       | 1.72   | No  | ✓     | ✓     | ✓    | ✓     | ✓     | ✓      | ✓       |
| CALS [12]  | 233,147*   | 24,569      | 63,349      | 50       | 4,408        | 1.31   | Yes | ✓     | ✓     | ✓    | ✓     | ✓     | ✓      | ✓       |
| ECALS (Ours) | 513,977    | 32,650      | 89,920      | 1054     | 139,541      | 10.18  | Yes | ✓     | ✓     | ✓    | ✓     | ✓     | ✓      | ✓       |

Table 1. Comparison of the existing MSD-subset and the proposed ECALS subset. A/S stands for artist stratified. (*) CALS includes additional un-annotated tracks for semi-supervised learning. This table only shows tag annotated dataset.
3. EXPERIMENT

3.1. Music-Text Pair Dataset (ECALS)

With the growing interest in sentence-level retrieval tasks [10, 11], it is desirable to have a music-caption paired dataset. However, no dataset is available for re-implementation. To address this problem, we concatenate the tag from different annotation sources. Based on Million Song Dataset (MSD) [19], we propose the ECALS (Extended Clean tag and Artist-Level Stratified) subset by merging the CALS subset [12] with 50 Last.fm tags [8] and 1,402 AllMusic [20] tag annotation. As a result, the ECALS subset has 0.51 million 30-second clips and 140k unique tag captions, including genre, style, instrument/vocal, mood/theme, culture categories. Table 1 shows the size and statistics of the MSD subset. The test track of the ECALS subset is the same as the CALS subset, and only the train, validation track, and annotation tags have been increased. Using the ECALS dataset, we evaluate tag-level and sentence-level retrieval tasks.

3.2. Evaluation Dataset

For unseen-query retrieval and downstream evaluation, we select various datasets related to music semantic understanding. The selection criteria are as follows: if a dataset has 1) commercial music for retrieval, 2) publicly assessed (at least upon request) and 3) categorical single or multi-label annotations for supporting text-based retrieval scenarios. We summarize all the datasets and tasks in Table 3. MagnaTagATune (MTAT) [21] consists of 26k music clips from 5,223 unique songs. Following a previous work [22, 10], we use their published splits and top 50 tags. We do not compare result with previous works using different split [11, 23]. MTG-Jamendo (MTG) [24] contains 55,094 full audio tracks with 183 tags about genre, instrument, and mood/theme. We use the official splits (split-0) in each category for tagging, genre, instrument, and mood/theme tasks. For single-label genre classification, we use the fault-filtered version of GTZAN (GZ) [25] and the ‘small’ version of Free Music Archive [26] (FMA-Small). For the vocal attribute recognition task, we use K-pop Vocal Tag (KVT) dataset [27]. It consists of 6,787 vocal segments from K-pop music tracks. All the segments are annotated with 42 semantic tags describing various vocal styles including pitch range, timbre, playing techniques, and gender. For the categorical mood recognition task, we use Emotify dataset [28]. It consists of 400 excerpts in 4 genres with 9 emotional categories.

3.3. Evaluation

Text-based Retrieval Depending on the type of input query, text-based music retrieval is divided into tag-level and sentence-level. Since the evaluation of tag-level retrieval is the same as label-wise evaluation of the auto-tagging task, we use the conventional macro version of ROCAUC and PRAUC metrics [7, 8]. We report both evaluation results on the top 50 vocabularies of CALS [12] and the 1054 large vocabularies of ECALS.

For evaluation of sentence-level retrieval, we build an audio-sentence subset by randomly sampling 1000 (audio, sentence) pairs from our testing split. Following the previous work [23], the sentence-level retrieval performance is evaluated by measuring Recall at K (K=1, 5, 10), mean average Precision at 10 (mAP10), and Median Rank (MedR). In case of classification model, we annotate multi-label tags on the music items with the best f1 score thresholds. And we perform sentence-level retrieval on the frequency of words overlapping with the sentence query.

Zero-shot Transfer and Probing For evaluation of unseen query retrieval and generalization ability, we measure the zero-shot transfer and probing performance, respectively. The zero-shot transfer measures the prediction score as the cosine similarity between the audio embedding of music and the text embedding of unseen tag [29]. For the probing task, we trained two shallow classifiers (linear models and one-layer MLPs) with the average pooled embedding from the frozen audio encoder. For rigorous comparison, we follow the probing protocol of previous studies [23, 22].

3.4. Training Details

The input to the audio encoder is a 9.91-second audio signal at 16 kHz sampling rate. It is converted to a log-scaled mel spectrogram with 128 mel bins, 1024-point FFT with a hann window, and a hop size of 10 ms. During training, we randomly sample an audio chunk from a 30-second waveform. All models are optimized using Adam and with a batch size of 64. We use different learning rates for

---

Table 2. Tag based, and Sentence based Retrieval result. Used In refers to previous studies using the same method.

| Dataset | Task         | # of Track | # of Tag | Avg.Tag | Metric          |
|---------|--------------|------------|----------|---------|-----------------|
| MTAT    | Tagging      | 25,860     | 50       | 2.70    | ROC/PR          |
| MTG-top50s | Tagging       | 54,380     | 50       | 3.07    | ROC/PR          |
| GTZAN   | Genre        | 55,094     | 87       | 2.44    | ROC/PR          |
| FMA-Small | Genre        | 8,000      | 8        | 1.00    | ACC             |
| KVT     | Vocal        | 6,787      | 42       | 22.78   | F1              |
| MTG-MT  | Mood/Theme   | 17,982     | 56       | 1.77    | ROC/PR          |
| Emotify | Mood         | 400        | 40       | 1.00    | ACC             |

Table 3. Downstream tasks/datasets for music semantic

| Dataset | Task      | # of Track | # of Tag | Avg.Tag | Metric          |
|---------|-----------|------------|----------|---------|-----------------|
| MTG     | Binary    | 90.2 / 39.5| 86.4 / 8.8| 4.0 | R@1 13.8 20.1 8.3 86 |
| MTG     | Genre     | 89.2 / 36.0| 82.6 / 6.1| 2.8 | R@5 11.2 18.6 6.6 51.5 |
| MTG     | Instrument| 88.6 / 37.1| 76.8 / 5.3| 5.4 | R@10 22.1 35.0 13.0 17 |
| MTG     | Mood/Theme| 89.2 / 37.6| 81.6 / 6.2| 6.4| R@10 21.8 32.7 12.8 19.5 |
| MTG     | Sentence  | 86.9 / 30.2| 81.7 / 5.1| 1.6 | R@10 6.2 12.0 3.9 68 |
| MTG     | Tag       | 87.7 / 35.0| 78.8 / 5.4| 6.7 | R@10 23.6 36.6 14.1 16 |
| MTG     | Stochastic| 88.4 / 35.0| 83.6 / 6.3| 6.6 | R@10 25.1 39.4 14.6 16 |
| MTG     | Ours      | 90.6 / 40.2| 86.4 / 8.8| 2.5 | R@10 13.7 22.5 7.4 47 |
| MTG     | Sentence  | 87.0 / 32.5| 77.6 / 5.1| 6.8 | R@10 25.4 38.4 15.3 17 |
| MTG     | Stochastic| 89.8 / 38.0| 84.8 / 7.7| 10.2| R@10 29.8 42.8 18.7 13 |
Contrastive-loss models were trained with a 0.2 initialized learning rate. The models with the BERT text encoder were with a learning rate of 5e-5. Contrastive-loss models were trained with a 0.2 initialized learning rate.

A similar to the retrieval task, the contrastive-loss model in the zero-shot transfer setup, although we used elaborated negative sampling. We interpret that larger negative sampling from a batch is more suitable than simple averaging. Another possible explanation is that the language model can summarize the sequence better than simple averaging. This is because the language model is versatile enough to handle sentence-level inputs when abundant music tags are available for training. However, the model cannot generalize to unseen tags (even if they are synonyms or acronyms), the classification model is a reliable solution.

5. CONCLUSION

In this paper, we introduced effective design choices for universal text-to-music retrieval by evaluating recent frameworks on a carefully designed dataset and downstream tasks. We mainly focused on training objectives and text representation. Experimental results revealed that retrieval performance heavily depends on text representation. Contrastive learning of text-music representation using 44 million data [10] significantly outperforms other approaches trained with 0.5 to 3.3 million dataset [22, 30, 31] in MTAT tagging. Unfortunately, MTAT was the only dataset with reported downstream evaluation. Contrastive learning of text-music representation using a pre-trained language model and stochastic sampling achieved the best retrieval performance.

In summary, contrastive learning of text-music representation using 44 million data [10] significantly outperforms other approaches trained with 0.5 to 3.3 million dataset [22, 30, 31] in MTAT tagging. Unfortunately, MTAT was the only dataset with reported downstream evaluation. Contrastive learning of text-music representation using a pre-trained language model and stochastic sampling achieved the best retrieval performance. The pre-trained text encoder was with a learning rate of 5e-5. Contrastive-loss models were trained with a 0.2 initialized temperature $\tau$, and triplet-loss models with a 0.4 margin $\delta$.

4. RESULTS

Table 2 shows the retrieval performances of different models using tag-level and sentence-level inputs. The classification model is a competitive baseline for tag-based retrieval (Table 2-left). Although the model cannot generalize to unseen tags (even if they are synonyms or acronyms), the classification model is a reliable solution when abundant music tags are available for training. However, the classification model could not handle sentence-level inputs because it is only trained with tag-level queries due to its inherent design.

The pre-trained language model is versatile enough to handle both tag-level and sentence-level inputs. The pre-trained word embedding could also take sentence-level inputs by averaging the word embeddings, but the performance is not comparable. One possible reason is that the language model can summarize the sequence better than simple averaging. Another possible explanation is that the language model (BERT) was trained with larger data than the word embedding. Our proposed stochastic sampling approach further improves the performance when it is applied to the text encoder.

Contrastive learning consistently showed better retrieval performance than triplet approaches in tag-level and sentence-level inputs, although we used elaborated negative sampling. We interpret that larger negative sampling from a batch is more suitable than triplet sampling in retrieval tasks. In summary, contrastive learning of text-music representation using a pre-trained language model and stochastic sampling achieved the best retrieval performance.

We report the zero-shot transfer and probing results in Table 4. A similar to the retrieval task, the contrastive-loss model in the zero-shot transfer task showed robust performance in almost all datasets. Compared to recent text-music representation learning approaches [9, 11, 10], we see that contrastive-loss models achieve competitive results and show significant improvements on the MTAT and GTZAN dataset. All probing results of contrastive-loss models are close to state-of-the-art performance and achieve state-of-the-art performance on GTZAN and KVT datasets.

We also believe the inclusion of large-scale data can improve the performance. Recent multimodal representation learning approaches [16] have shown breakthrough in many domains by taking advantage of enormous data from the web. Similar trend is found in our downstream evaluation. Contrastive learning of text-music representation using 44 million data [10] significantly outperforms other approaches trained with 0.5 to 3.3 million dataset [22, 30, 31] in MTAT tagging. Unfortunately, MTAT was the only dataset with reported performance from various previous works.

5. CONCLUSION

In this paper, we introduced effective design choices for universal text-to-music retrieval by evaluating recent frameworks on a carefully designed dataset and downstream tasks. We mainly focused on training objectives and text representation. Experimental results revealed that retrieval performance heavily depends on text representation. Contrastive models achieve better performance than triplet models in both retrieval and downstream tasks. Furthermore, our proposed stochastic text representation achieved robust performance in tag-level, caption-level, and zero-shot query retrieval cases. However, our current dataset is limited to music tags, such as genre, mood, and instrument. A more generalizable music retrieval system needs to cover other musical attribute, such as the tempo, key, chord progression, melody, artist, etc. To overcome the limitations of annotated labels, multi-task learning of multiple datasets or a teacher-student model can be an alternative. Reproducible code, pre-trained models are available online for future research.

https://github.com/seungheondoh/music-text-representation
6. REFERENCES

[1] Meinard Müller, Frank Kurth, David Damm, Christian Fremer, and Michael Clausen, “Lyrics-based audio retrieval and multimodal navigation in music collections,” in International conference on theory and practice of digital libraries, 2007.

[2] Jongpil Lee, Nicholas J Bryan, Justin Salamon, Zeyu Jin, and Juhan Nam, “Metric learning vs classification for disentangled music representation learning,” in ISMIR, 2020.

[3] Douglas Turnbull, Luke Barrington, David Torres, and Gert Lanckriet, “Semantic annotation and retrieval of music and sound effects,” IEEE Transactions on Audio, Speech, and Language Processing, 2008.

[4] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I Hong, “hey alexa, what’s up?” a mixed-methods studies of in-home conversational agent usage,” in Proceedings of the designing interactive systems conference, 2018.

[5] Keunwoo Choi, George Fazekas, and Mark Sandler, “Automatic tagging using deep convolutional neural networks,” in ISMIR, 2016.

[6] Juhan Nam, Keunwoo Choi, Jongpil Lee, Zu-Yu Chou, and Yi-Hsuan Yang, “Deep learning for audio-based music classification and tagging: Teaching computers to distinguish rock from bach,” IEEE signal processing magazine, 2018.

[7] Jeong Choi, Jongpil Lee, Jiyoung Park, and Juhan Nam, “Zero-shot learning for audio-based music classification and tagging,” in Proc. ISMIR, 2019.

[8] Minz Won, Sergio Oramas, Oriol Nieto, Fabien Gouyon, and Xavier Serra, “Multimodal metric learning for tag-based music retrieval,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020.

[9] Minz. Won, Justin. Salamon, Nicholas J. Bryan, Gautham J. Mysore, and Xavier. Serra, “Emotion embedding spaces for matching music to stories;” in ISMIR, 2021.

[10] Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel PW Ellis, “Mulan: A joint embedding of music audio and natural language;” in ISMIR, 2022.

[11] Ilaria Manco, Emmanouil Benetos, Elio Quinton, and György Fazekas, “Contrastive audio-language learning for music;” in ISMIR, 2022.

[12] Minz Won, Keunwoo Choi, and Xavier Serra, “Semi-supervised music tagging transformer;” in ISMIR, 2021.

[13] Liwei Wang, Yin Li, and Svetlana Lazebnik, “Learning deep structure-preserving image-text embeddings;” in Proceedings of the computer vision and pattern recognition, CVPR, 2016.

[14] Keunhyoung Kim, Jongpil Lee, Sangeun Kum, and Juhan Nam, “Learning a cross-domain embedding space of vocal and mixed audio with a structure-preserving triplet loss;” in ISMIR, 2021.

[15] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, “Representation learning with contrastive predictive coding;” arXiv preprint arXiv:1807.03748, 2018.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., “Learning transferable visual models from natural language supervision;” in International Conference on Machine Learning (ICML), 2021.

[17] Jeffrey Pennington, Richard Socher, and Christopher D Manning, “Glove: Global vectors for word representation;” in Proceedings of the conference on empirical methods in natural language processing, EMNLP, 2014.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding;” in Proceedings of NAACL-HLT, 2018.

[19] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere, “The million song dataset;” in Proc. ISMIR, 2011.

[20] Alexander Schindler and Peter Knees, “Multi-task music representation learning from multi-label embeddings;” in International Conference on Content-Based Multimedia Indexing (CBMI), 2019.

[21] Edith Law, Kris West, Michael I Mandel, Mert Bay, and J Stephen Downie, “Evaluation of algorithms using games: The case of music tagging;” in ISMIR, 2009.

[22] Rodrigo Castellon, Chris Donahue, and Percy Liang, “Codi-fied audio language modeling learns useful representations for music information retrieval;” in ISMIR, 2021.

[23] Ilaria Manco, Emmanouil Benetos, Elio Quinton, and György Fazekas, “Learning music audio representations via weak language supervision,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022.

[24] Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Bresson, “The mtg-jamendo dataset for automatic music tagging;” 2019.

[25] Bob L Sturm, “The gtzan dataset: Its contents, its faults, their effects on evaluation, and its future use,” arXiv preprint arXiv:1306.1461, 2013.

[26] Michâel Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson, “Fma: A dataset for music analysis;” in ISMIR, 2016.

[27] Keunhyoung Luke Kim, Jongpil Lee, Sangeun Kum, Chae Lin Park, and Juhan Nam, “Semantic tagging of singing voices in popular music recordings;” IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020.

[28] Anna Aljanaki, Frans Wiering, and Remco C Veltkamp, “Studying emotion induced by music through a crowdsourcing game,” Information Processing & Management, 2016.

[29] Jeong Choi, Jongpil Lee, Jiyoung Park, and Juhan Nam, “Zero-shot learning and knowledge transfer in music classification and tagging,” arXiv preprint arXiv:1906.08615, 2019.

[30] Pablo Alonso-Jiménez, Xavier Serra, and Dmitry Bogdanov, “Music representation learning based on editorial metadata from discogs;” in ISMIR, 2022.

[31] Fabien Gouyon Andreas F. Ehmann Matthew C. McCallum Filip Korzeniowski Sergio Oramas, “Supervised and unsupervised learning of audio representations for music understanding;” in ISMIR, 2022.