# Multimodal music information processing and retrieval: survey and future challenges

we can recognize music played on different instruments, with different timings, intensity changes, tonalities, tunings, background noises and so on.

Better results through multimodal approaches.

In this paper, we review the existing literature about Music Information Retrieval.

Ways to digitalize music information: audio, lyrics, album covers, symbolic scores.

Multimodal music information processing as an MIR approach which takes as input multiple modalities of the same piece of music.

Multimodal music processing tasks:

some tasks have been extensively studied with a multimodal approach, such as audioto-score alignment, score-informed source separation, music segmentation, emotion or mood recognition; other tasks, instead, have been little explored and are worth of more attention.

classification of music, synchronization of different representations, similarity computation between two or more modalities, and time-dependent representation.

- Similarity:

Often, this task has the purpose of retrieving documents from a collection through a query.

A very common example of explicit queries for retrieving another modality is queryby- humming or query-by example, in which the query is represented by an audio recording and the system retrieves the correct song.

An example of implicit query systems, instead, are recommender systems and playlist generators.

Most of the recent research in this field tries to exploit multimodal approaches – also called hybrid – involving metadata, user context, audio features.

An emerging field in the retrieval context is the so-called multimodal queries [16], [24].

- Classification:

The classification process consists in taking as input a music document and returning one or more labels.

Examples: mood or emotion recognition, genre classification, artist identification, derivative works classification, instrument classification, tonic identification, expressive musical description.

Feature extraction in multimodal approaches

Various types of features can be extracted from each modality. In this section, we provide a general description for audio, video, textual and symbolic score features.

- text features [70]: The most common text representations are based on tf-idf. In order to make BoW and tf-idf models effective, a few preliminary steps are usually performed, such as punctuation and stop-words removal and stemming. More sophisticated methods are also available, allowing for topic- or semantics-based analysis, such as Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), Explicit Semantic Analysis (ESA) and CNN feature extraction.

Conversion to common space

The conversion of the extracted features to a common space is often a mandatory step in early fusion approaches. The conversion to a common space consists in the mapping of the features coming from different modalities to a new space where they are comparable.

many papers describe techniques which include a mapping of the features to a common space, both in the pre-processing and in the processing stages.

Common methods include: normalization, conversion from one modality to another, machine learning algorithms and dimensionality reduction algorithms which usually search for a new space where data samples are representable with a fewer number of dimensions without losing the ability to separate them

Future Directions

First of all, we note the unavailability of datasets of suitable size. Although a few datasets have been recently created [37], [76], [78], a great effort should still be carried out in this direction. Indeed, existing multimodal music datasets are usually characterized by limited sizes and only rarely include a wide range of modalities.

We argue that this limit is due to two main reasons: first, the precise alignment of

various modalities is a hard computational task and should be controlled by human supervision; second, no largely adopted standard exists for multimodal music representation.

# Audio Retrieval with Natural Language Queries: A Benchmark Study

The objectives of this work are cross-modal textaudio and audio-text retrieval, in which the goal is to retrieve the audio content from a pool of candidates that best matches a given written description and vice versa. Text-audio retrieval enables users to search large databases through an intuitive interface: they simply issue free-form natural language descriptions of the sound. … these tasks have received limited attention in the existing literature.

The vast and unabated growth of user-generated content in recent years has introduced a pressing need to search ever-growing databases of multimedia. Free-form natural language sentences (i.e. sequences of text that are written as they would be spoken) form an intuitive and powerful interface

for composing search queries for these databases since they allow for expressing virtually any concept. Spanning multiple modalities, different retrieval strategies were developed for content as diverse as text (including web pages and books), images [1], and videos [2], [3]. Surprisingly, while search

engines currently exist for these modalities (e.g. Google, Flickr and YouTube, respectively), unstructured audio is not accessible in the same way.

It is important to distinguish content-based retrieval from that based on metadata, such as the title of a video or song or an audio tag. Metadata retrieval is feasible for manuallycurated databases such as song or movie catalogues. However, content-based retrieval is more important in user-generated

data, which often has little structure or metadata. There are methods to search for audio which matches an audio query [4], [5], but satisfying the requirement to input an example audio

query can be difficult for a human (e.g. making convincing frog sounds is difficult). We, on the other hand, propose a framework which enables the searching of a sound database using detailed free-form natural language queries of the desired sound .

Furthermore, natural language queries are a familiar user interface widely used in current search engines. Therefore, our proposed audio retrieval with free-form text queries could be a first step towards more natural and flexible audio-only search.

Text-based audio retrieval could also be beneficial for video

retrieval. The majority of recent works that address the textbased video retrieval task focus heavily on the visual and text domains [1], [2], [6], [7]. Since audio and visual information inherently have natural semantic alignment for a significant portion of video data, text-based audio retrieval could also be used for querying video databases by only considering the audio stream of the video data.

Learning to retrieve audio, given natural language queries, requires data with paired text and sound. Audio captioning <<definition of captioning>> datasets naturally lend themselves to this task, since they contain audio and a matching text description for the sound.

However, existing captioning datasets are limited in size and in the diversity of their audio content.

Audio captioning consists of generating a natural language description for a sound [27]. This requires a more detailed understanding of the sound than simply mapping the sound to a set of labels (sound event recognition).

Audio-based retrieval. Multiple content-based audio retrieval frameworks, in particular query by example methods, leverage the similarity of sound features that represent different aspects of sounds (e.g. pitch, or loudness) [37], [38], [39], [5]. More recently, [4] use a twin neural network framework to learn to encode semantically similar audio close together in the embedding space. [40] address multimedia event detection using only audio data, while [41] tackle near-duplicate video retrieval by

audio retrieval.

Text-based video retrieval. More closely related to our work, a number of methods showed that embedding video and text jointly into a shared space (such that their similarity can be computed efficiently) is an effective approach [1], [3], [2], [6], [43], [7], [44] (though other formulations, such as

computing similarities directly in visual space have also been explored [45]). One particular trend has been to combine cues from several “experts”—pre-trained models that specialise in different tasks (such as object recognition, action classification etc.) to inform the joint embedding. Recently, transformerbased architectures have demonstrated impressive results for text-based video retrieval [7], [46], [47]. In this work, we propose to adapt three expert-based methods: the Mixture of Embedded Experts method of [2], the Collaborative Experts model of [6], and the Multi-Modal Transformer [7] by repurposing them for the task of audio retrieval (described in more detail in Sec. IV).

Cross-domain audio retrieval. Methods that retrieve audio by matching associated text, such as metadata or sound event labels, have the implicit assumption that the text is relevant [48]. 2In contrast, [49] is an early work that proposes to link audio and text representations in hierarchical semantic and acoustic spaces. [50] builds on this using mixture-of-probability-expert models for each of the modalities. ……

Other works have explored using images [55] or video data [56], [57], [58], [59], [60] as queries for retrieving audio. More recently, [61] use a twin network to learn a shared latent text and sound space for cross-modal retrieval.

Problem formulation. Given a natural language query (i.e. a written description of an audio event to be retrieved) and a pool of audio samples, the objective of text-audio (abbreviated to

t2a) retrieval is to rank the audio samples according to their similarity to the query.

Methods. To tackle the problem of text-audio retrieval, we propose to learn cross-modal embeddings. we aim to learn embedding functions, that project each audio sample and text sample into a shared space, such that ai  and ti in said space are close when the text describes the audio, and far apart

otherwise.

Experiments

Retrieval performance metrics are recall at rank k and median and mean rank.

In this section, we first compare text-audio retrieval (t2a) and audio-text retrieval (a2t) performance on audio-centric and visual-centric datasets. Next, we perform an ablation study on the contributions of different experts and present our baselines for the proposed AUDIOCAPS, CLOTHO, and SOUNDDESCS benchmarks. Finally we perform experiments to assess the influence of pre-training, audio segment duration and training dataset size, and give qualitative examples of retrieval results. Throughout the section, we use the standard retrieval metrics: recall at rank k (R@k) which measures the percentage of targets retrieved within the top k ranked results (higher is better), along with the median (medR) and mean (meanR) rank. For all metrics, we report the mean and standard deviation of three different randomly seeded runs.

# TOWARD UNIVERSAL TEXT-TO-MUSIC RETRIEVAL

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.

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, 3, 4], text-based retrieval remains the most prevalent [5, 6, 7]. 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 [8], 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 [9, 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 [10, 11] or language models [12, 13, 14, 15]. Especially pre-trained language models enable free-form text inputs for music retrieval by representing sentence-level semantics.

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 the music library. Sometimes the input queries may include unseen types of music tags. Also, another can use more detailed sentence-level descriptions to discover music.

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 [21] and a pre-trained Bidirectional Encoder Transformer (BERT) [22]

with a base-uncased architecture.

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 embedding, the text sequence attaches [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 which is layer normalized and then linearly projected embedding space.

# MULAN: A JOINT EMBEDDING OF MUSIC AUDIO AND NATURAL LANGUAGE

Music tagging and content-based retrieval systems have traditionally been constructed using pre-defined ontologies covering a rigid set of music attributes or text queries. This paper presents MuLan: a first attempt at a new generation of acoustic models that link music audio directly to unconstrained natural language music descriptions. MuLan takes the form of a two-tower, joint audio-text embedding model trained using 44 million music recordings (370K hours) and weakly-associated, free-form text annotations. Through its compatibility with a wide range of music genres and text styles (including conventional music tags), the resulting audio-text representation subsumes existing ontologies while graduating to true zero-shot functionalities. We demonstrate the versatility of the MuLan embeddings with a range of experiments including transfer learning, zero-shot music tagging, language understanding in the music domain, and cross-modal retrieval applications.

Classifiers are generally trained to label examples with predefined and fixed class inventories, which are often manually specified as a structured ontology indicating interclass relationships. Empowered by recent advances in neural language modeling and their demonstrated transfer learning competence, researchers have begun exploring less restrictive natural language interfaces to access

the categorical information underlying raw content signals. The majority of this work has been in the visual and audio event domain, where a recent series of studies have demonstrated the utility of jointly embedding media content with natural language captions [1–5]. These joint embeddings have demonstrated strong capabilities in a range of applications, including transfer learning, cross-modal retrieval, automatic captioning, and zero-shot classification.

The success of these efforts strongly depends on large-scale training resources and hefty neural network architectures that are flexible enough to model the complex, non-monotonic relationship between language and other modalities. In particular, the visual domain has greatly benefited from the availability of large amounts of captioned images available across the web [1]. However, in

the general environmental audio domain, such large-scale audio-caption pairs are less readily available and related efforts have relied on small captioned datasets [6, 7]. Critically, these datasets do not span the diversity of sound-descriptive language and their success in the more difficult zero-shot setting has been lacking [3, 8, 9].

our strategy is to assemble a collection of textual annotations extracted from metadata, comments, and playlist data and map them to a training set of over 44 million internet music videos. As was the case with image-text model training in [1], our text data only truly refers to the musical content in a fraction of cases. Therefore, we also explore text pre-filtering using a text classifier separately trained to identify music descriptions.

Music text joint embedding models

Content-based music information retrieval requires linking the rich semantics expressible to free-form text with both broad and finegrained musical properties. One approach is to consider a large number of text label classes and try to ground the semantics in music with a multi-label classification task …..

our two-tower parallel encoder approach results in a joint embedding space that provides a natural language interface to arbitrary music audio. This opens up downstream opportunities for cross-modal retrieval, zero-shot tagging, and language understanding.

Evaluation:

Music Retrieval from Text Queries

Given a music search collection and a text query, MuLan provides the ability to retrieve the music clips that are closest to the query in the embedding space. This evaluation is relevant to music retrieval applications, where content features can offer finer-grained and more complete similarity information when compared with metadata-based methods [41].

Results and Discussion

Music Retrieval from Text Queries

Even though we start with a BERT checkpoint pretrained with massive language resources, training MuLan with only AudioSet clips and label annotations provides very limited ability to ground indomain natural language to music. Such limited crossmodal supervision does not generalize to the rich semantics that appear in the playlist titles and descriptions, which are more in line with the complex queries that are presented to real-world music search engines. We observe significant gain after including the large-scale short-form tags mined from the internet, which helps the model learn to ground more fine-grained music concepts. There is additional gain when including comments and playlist data, where the complete sentences are helpful for grounding the more complex queries, including multi-term queries (e.g.‘instrumental action movie soundtrack’), compositional queries (e.g. ‘classical music with middle eastern influence’), and even queries with negation (e.g. ‘hard rock without vocals’). Again, we find that training is surprisingly robust to annotation noise, achieving similar performance using unfiltered training text.

# Contrastive audio-language learning for music

As one of the most intuitive interfaces known to humans, natural language has the potential to mediate many tasks that involve human-computer interaction, especially in application-focused fields like Music Information Retrieval.

Our approach consists of a dual-encoder architecture that learns the alignment between pairs of music audio and descriptive sentences, producing multimodal embeddings that can be used for text-to-audio and audio-to-text retrieval out-of-the-box.

# Improving information retrieval by semantic embedding

### Vec4IR Word embeddings for information retrieval.

<https://github.com/lgalke/vec4ir>

### Information Extraction from Natural Language on the Web using LLMs and Iterative Set Expansion

<https://github.com/yaminivibha/LLM_InformationRetrieval>

### gensim – Topic Modelling in Python

<https://github.com/piskvorky/gensim>

Gensim is a Python library for topic modelling, document indexing and similarity retrieval with large corpora. Target audience is the natural language processing (NLP) and information retrieval (IR) community.

### txtai is an all-in-one embeddings database for semantic search, LLM orchestration and language model workflows

<https://github.com/neuml/txtai>

### LLM Embedder

LLM Embedder is fine-tuned based on the feedback from LLMs. It can support the retrieval augmentation needs of large language models, including knowledge retrieval, memory retrieval, examplar retrieval, and tool retrieval. It is fine-tuned over 6 tasks: Question Answering, Conversational Search, Long Conversation, Long-Range Language Modeling, In-Context Learning, and Tool Learning. For more details please refer to https://github.com/FlagOpen/FlagEmbedding/llm\_embedder/README.md

<https://github.com/FlagOpen/FlagEmbedding/tree/master/FlagEmbedding/llm_embedder>

### BGE Embedding

BGE embedding is a general Embedding Model. We pre-train the models using retromae and train them on large-scale pair data using contrastive learning. You can fine-tune the embedding model on your data following our examples. We also provide a pre-train example. Note that the goal of pre-training is to reconstruct the text, and the pre-trained model cannot be used for similarity calculation directly, it needs to be fine-tuned. For more training details for bge see <https://github.com/FlagOpen/FlagEmbedding/blob/master/FlagEmbedding/baai_general_embedding/README.md> .

<https://github.com/FlagOpen/FlagEmbedding/tree/master/FlagEmbedding/baai_general_embedding>

### One Embedder, Any Task: Instruction-Finetuned Text Embeddings

We introduce Instructor, an instruction-finetuned text embedding model that can generate text embeddings tailored to any task (e.g., classification, retrieval, clustering, text evaluation, etc.) and domains (e.g., science, finance, etc.) by simply providing the task instruction, without any finetuning. Instructor achieves sota on 70 diverse embedding tasks!

<https://github.com/xlang-ai/instructor-embedding>

### COVID19-document-retrieval-with-BERT

This project is about developing a document retrieval system to return titles and the context of scientific papers containing the answer to a given user question. We will be using CORD-19 CORD-19 dataset for the dataset which is the biggest corpus of academic papers about COVID-19 and related coronavirus research. In this project, we will use and evaluate 4 different methods for creating contextual sentence embeddings. These embeddings then will be used to find the most semantically meaningful document that relates to our question.

<https://github.com/AGiannoutsos/COVID19-document-retrieval-with-BERT>

### Document-Retrieval-and-Question-Answering-with-BERT

In first part of this repository we developed, with 2 different techniques, a Document Retrieval System, which aims to return titles of scientific papers containing the answer to a given user question. To achieve the goal of this exercise, we first read Sentence Embeddings using Siamese BERT-Networks, in order to understand how to create sentence embeddings. We decided to transform sentences to embeddings, with the help of SBERT Library. So the main concept behind our retrieval system was to compare sentence embeddings (question's and possible answer's embedding) by cosine similarity.

<https://github.com/spyros-briakos/Document-Retrieval-and-Question-Answering-with-BERT>