Master thesis on Sound and Music Computing

Universitat Pompeu Fabra

Learning Sound Representations Using Triplet-loss

Kohki Mametani

Supervisor: Xavier Favory

Co-Supervisor: Frederic Font

September 2020

Master thesis on Sound and Music Computing

Universitat Pompeu Fabra

Learning Sound Representations Using Triplet-loss

Kohki Mametani

Supervisor: Xavier Favory

Co-Supervisor: Frederic Font

September 2020

Contents

Dedication

(Optional, if used placed on a right page next to an empty left page)

I would like to dedicate this work to...

Acknowledgement

(Optional, if used placed on a right page next to an empty left page)

I would like to express my sincere gratitude to:

- My supervisor
- My co-supervisor
- My family

Abstract

The abstract should have at least 200 but not more than 600 words. Placed on a right page next to a blank left page. A list of keywords (approximately 3 to 5) should be just below the abstract, preceded by the word "Keywords". Keywords should be separated by ";".

Keywords: Imaging techniques; Cloud computing; Alzheimer

Chapter 1

Introduction <in progress>

As an introduction to our work, we begin with a historical context of our work, briefly going back to the prehistoric times, describing the emergence of need for information technology to glance at how inherently important to humanity such technology is. From that point, the exponential growth of information in terms of both quality and quantity resulted in the establishment of today's society, sometimes referred as *Information Society*, in which the usage, creation, distribution of information is a significant activity. Instead of looking through each of relevant technologies, the center of our attention always lie at multimedia, especially at audio. The aim of this chapter is to provide a perspective that navigate readers in the course of the technological development from earliest pioneering work to more complex state-of-the-art methods which we will discuss in the next chapter.

Human's need for information has relatively gradually evolved as they expanded their intellectual activity and population. At least by 40,000-60,000 years ago, in hunter-gathering age, people mostly relied on immediate and environmental clues for decision making and most lived day to day without keeping much information for the future. With the emergence of agriculture around 10,000 years ago, however, people started recording and sharing information such as weather to maximize the outcome (their crops) in the future. The activity of storing, retrieving, and distributing information saw the first acceleration since its origin when the Sumerian

in Mesopotamia developed writing in about 3000 BC.

Despite the constant growth of need for information over the history, the availability and the accessibility of information has greatly improved in the recent decades. ...

But the term *information technology* in its modern sense first appeared in 1958, around the time when the early digital computers were moving into mass production.

The amount of online data is growing exponentially. According to IBM Marketing Cloud study, 90% of this data on the internet has been created since 2016. To give an idea of how much and rapidly data is shared, shared, and stored on the internet today, the following activities are recognized as popular contributors of the data growth: users on YouTube uploads 500 hours of new video each minute of every day and Instragram users upload over 100 million photos and videos every day.

The internet differs from most of the mass media it replaces in an obvious and very important way: it's bidirectional. We can send messages through the network as well as receive them. The ability to exchange information online, to upload as well as download, is one of the main factors that immensely accelerated the amount of information we can access today.

Regarding the use of online data and its applications, users not only upload the data but also search in a massive multimedia databases including audio, speech, text, video, image, and their combinations. Users typically submit search queries that express a broad intent, which makes the system often return large result sets. A query from a user can be a single word, multiple words or a sentence. The search results are generally presented in a list of results and users examine a few samples of the results until they find what they are looking for. This process is often designed on a basis of various technical consideration, for example, how to organize data efficiently, how to precisely respond to diverse user requests with inconsistent set of vocabulary, or how to present the search results to users.

Methods used in content-based retrieval of multimedia data on the internet usually consider a number of common problems. User generated data on multimedia sharing

websites such as image tags on Flickr, video descriptions on YouTube are often noisy, unstructured and produced (or recorded) under different conditions.

Chapter 2

State of the Art

In the last chapter, we saw information technology has a decisive role in shaping our society as well as some breakthroughs of the technology that altered how we manage, retrieve and distribute information and critically changed our way of thinking and living. The advancements in information technology is an ongoing central force of digital innovation in today's society. Yet we think there are a plenty of opportunities left behind. This chapter encompasses this long-running technological advances from the past, making readers certain of the state-of-the-art studies that are relevant to our work, and shows how seamlessly our methodology will progress from the current point in the next chapter.

As seen in the introduction, information technology can mean any kind of systems for storing, retrieving, and sending information. In this work, we focus exclusively on information technology which deal with information that are processed and structured from audio data. With our subsequent methodology, more specifically, we aim to improve an content-based audio search system which allows people to access sounds efficiently and intuitively by looking inside a cluster of sounds grouped by content similarity instead of examining a linear list of sounds. To achieve this, we think of two independent issues which are equally important: the way of representing sounds (production of audio features) and the way of presenting sounds based on the given representation (clustering approach). We discuss these two issues in

the context of an information retrieval process which begins when a user enters a query for sounds into the system and finds matching objects. There is no doubt that such a system greatly helps people's search for digital sounds on the internet which is increasingly daunting because of the information overload we discussed in the last chapter.

The following sections offer a collection of the most up-to-date work and ideas regarding audio retrieval and not be filled with a mere assortment of details about related new techniques. We give a special importance to the context of why an effort of research or development was made and what problem did it try to address. In this way, we might have to give up some trivial facts and details of the topic, but we believe it is inevitable so that readers can focus on a contiguous progression of previous development toward our proposed method in the next chapter. Each of the following sections is dedicated to a fundamental technique of our methodology. Therefore, readers can get to know the state-of-the-art of each technique and, by the end of chapter, will be all ready for the discussion of our experiments and analysis.

The rest of this chapter is organized as follows: in the next section, we discuss the most general topic of our research: information retrieval, focusing on its connection to sound and music computing instead of going into detail of applications with non-audio information. Then, in Section 2, a class of traditional audio features, the building blocks of audio retrieval, are introduced. In Section 3, we cover the audio features that are recently developed with the help of deep learning. Then, among those new kind of features, we focus on those learned from similarity learning and discuss its practical use for sound retrieval in Section 4. In Section 5, a review of clustering methods is presented with the focus on two popular approaches, the partitional clustering and the graph-based clustering. Due to the large volume and many topics to be covered, Figure 1 visualizes the thematic structure of the following sections and allows readers to focus on specific topics.

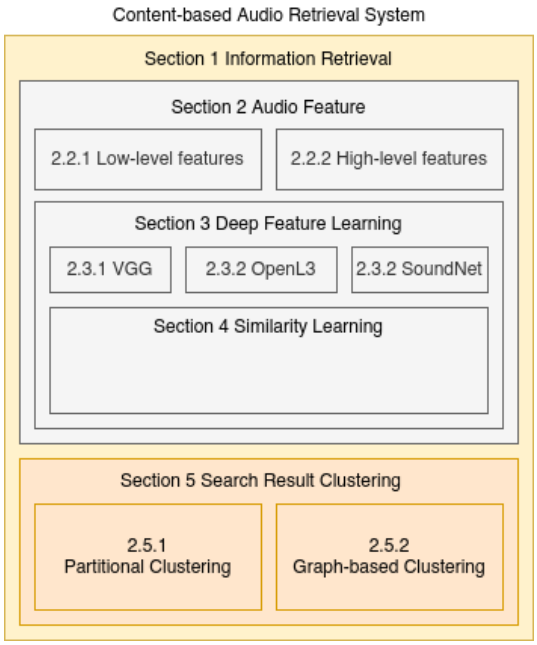

Figure 1: Overview of the sections in this chapter. Each section in the same group shares a topic in common.

2.1 Information Retrieval and Audio

The purpose of this section is to emphasize that our work lies at an intersection of information retrieval (IR) and audio engineering problems, pointing out a close relationship between IR systems and audio engineering with examples. We introduce fundamental technologies that are essential to modern IR process such as database technology. Also a number of terminologies used in the IR research, which appear thoroughly in the following sections, are carefully introduced since those may not be clear for readers from audio engineering fields.

2.1.1 Information Retrieval

Information retrieval is the science of finding information that satisfies an information need from a large collection of those resources [1]. Most IR systems start off by taking as input a text, a search query which describes user's information needs formally or informally, for example search keywords for web search engines. Then, given a query, the system ranks the data objects which are represented by information in the database. The rank of relevance between the query and each object is

measured using a numeric score on how well the object matches with the query [2]. Then, it shows the top ranking objects to the user. While traditionally IR is used to mean searching for documents (an unstructured text data) from a collection of other documents in computers [1], IR is becoming the dominant form of any kind of information access. Today, the objects that IR systems deal with are not only texts but a wide variety of multimedia data are also supported such as images [3] and sound [4], and the retrieval process can be performed using either full-text query or content-based query [5].

Techniques and algorithms which enable efficient retrieval process vary with the kind of information to search for. For example, Music Information Retrieval (MIR), which is a branch of IR and deeply related to our work, focuses on retrieving information from music. Since our work focuses on the research and development of content-based audio retrieval which aim to deal with a various types of sounds, including artificially-created sounds such as music loops, we will come back to this topic and discuss advances in techniques and knowledge of the MIR research.

2.1.2 Database System

As already mentioned, database system is one of the underlying technologies in IR. It is employed in any kind of web-based applications [6] and makes possible an satisfying IR experience on the internet . However, database searching itself is not considered as IR because it lacks the ranking process of the search results [7]. This means an IR system returns the results which may or may not match the query in the order of relevance while typically a database system returns nothing when no matching object is available in its data collection. Furthermore, an IR system does not limit the user to a specific query language while such languages like SQL or MongoDB are commonly used in database systems [8]. Despite the key differences, it is worth taking a close look at the database technology, specifically in regard to web services, because it is an integral part of many popular web-based IR systems, for example image retrieval on Flickr and audio retrieval on Freesound, and importantly, some of those applications will appear as examples in the following

sections.

Database on the web

Considering different concerns, a website is usually separated into two independent components: one that considers the activity of clients, helping them interact with data on a website, and the other that takes command of the web servers which essentially provide service to one or many clients. Many users of the internet, even non-technical users, know fairly well about the first part, especially the graphical user interfaces (GUI). The development of such interfaces is known as front-end web development. The interfaces are designed using programming languages such as HTML, CSS, and JavaScript, that can be executed on client devices with a web browser. Modern front-end development is accustomed to support intuitive GUIs and often built on the top of commonly-shared templates, among which the most popular one is Twitter's Bootstrap[1]. As such, most of users who already got used to modern web experience will easily guide themselves when they visit a new website.

What's often left invisible and unknown to the public is a large number of server-side processes running in the back end. Through front end interfaces, users can communicate with the server of a website and properly command them to return the data they look for. Efficiency of the back end system highly affects the performance of the front end activity. Indeed, even basic user experience such as authentication and permission, requires communicating and retrieving data stored in server-side database systems. As such, the database for the website is placed and accessed in the back end.

Database in practice `<in progress>`

A database system and a file system have some functions in common and the similarity of two systems may confuse some readers. Practically, both system store the data in files: a database system stores data with predefined data formats in a table and a file system stores data (the pointers which indicate where and what data is

[1] https://getbootstrap.com/

stored in a hardware storage).

About SQL and NoSQL for querying

About RAID6 for secure database

About how data is typically organized, not nested much

2.1.3 Multimedia Information Retrieval <in progress>

2.1.4 Freesound <in progress>

Freesound[2] is a non-profit organization which hosts a collaborative web-based repository of Creative Commons Licensed sounds. It has been maintained at the Music Technology Group (MTG) in Universitat Pompeu Fabra, Barcelona. The repository was launched in 2005 with an initial goal to provide a large royalty-free sound databases with sound researchers to test their algorithms [9]. Since then, it has widely adopted by sound artists, application developers and all sorts of content creators and became a standard website for sharing Creative Commons sounds.

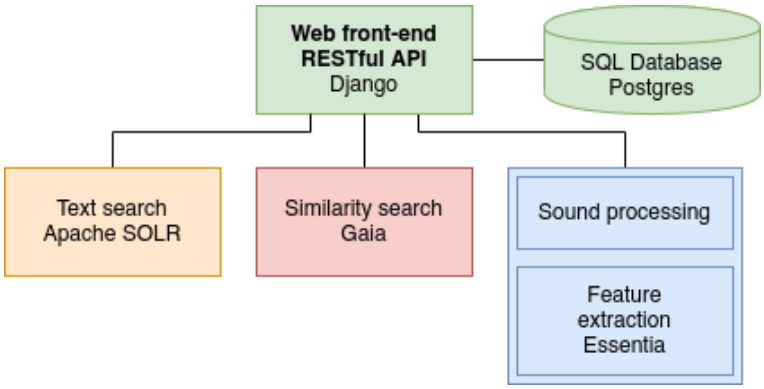

Figure 2: Overview of system architecture of Freesound [9]. The architecture is a good example of the use of database system in web-based IR systems as discussed in the Section 2.1.2

2.1.5 Content-based and context-based search <in progress>

Two underpinnings for audio retrieval are how audio data is represented (audio feature) and how data will presented to users. Before breaking down the two topics

[2]https://freesound.org/

in the following sections, this section aims to provide readers with a context for the sound retrieval research, covering from historical examples to an example of popular real-world application, explaining emerging technological advancements which affect its future research directions.

2.1.6 Toward intelligent audio retrieval <in progress>

So far, we have seen various efforts made in the research and development of efficient IR for multimedia information, focusing on audio. Speaking of audio retrieval, humans are inherently good at remembering/retrieving audio since we are born with a set of abilities to discriminate a wide range of sounds, especially in the context of speech. Moreover, we are also able to relate and match similar sounds [10]. In fact, we have the capability to detect and relate sound events or "acoustic objects" which we have never encountered before, based on how that phenomenon stands out against the background [11]. To computers, however, a raw audio signal data is essentially a featureless collection of bytes with most rudimentary information embedded such as file name, format, and sampling rate. This does not readily allow efficient management and searches for sounds. To accomplish the task, we first need to transform sounds into a small set of parameters (audio features) and, in the next section, we will go into detail of various analysis techniques to achieve this goal.

2.2 Audio Features

In machine learning and pattern recognition, a feature is most briefly defined as an individual measurable property or characteristic of a phenomenon being observed [12]. Features that appear in this work are always real-valued numeric while they may vary in different contexts such as binary (e.g. "on" or "off"); categorical (e.g. "A", "B", "AB" or "O", for blood type); ordinal (e.g. "large", "medium" or "small").

Feature extraction is also an active research topic in multimedia IR. In case of image retrieval system, IR systems traditionally relied on manually created textual annotation to measure the relevance between user's query and information in resources

(keyword annotation paradigm) [13]. With the rise of the amount of information, however, the annotation approach had to be replaced with a faster and automatic feature extraction. Researchers found that the use of low-level visual features such as shapes and colors can be superior to traditional annotation-based approach in terms of robustness because automation removed the human subjectivity [14]. Furthermore, features are used to map from a large, complex data space into a small, simple feature space so that the system looks for information in a smaller search space efficiently [15].

In regard to audio retrieval, audio features lie at the heart of our work. Within a scheme of typical audio retrieval applications, audio features are obtained in a preliminary process, which is referred as a feature extraction process, since the raw signal is too large, noisy and redundant for most of the interesting analysis [16]. In this context, feature extraction is a process that convert the raw audio waveform (time series data points) into a set of numerical vectors.

This section aims to offer a review about two general classes of audio features, acoustic features and symbolic features, that researchers and developers commonly find useful for a range of audio applications. Audio features discussed in this section are crafted in traditional approaches and based on from low-level features such as fundamental frequency and harmonicity of signals to high-level features which are designed to provide semanticity of sounds such as "sharpness".

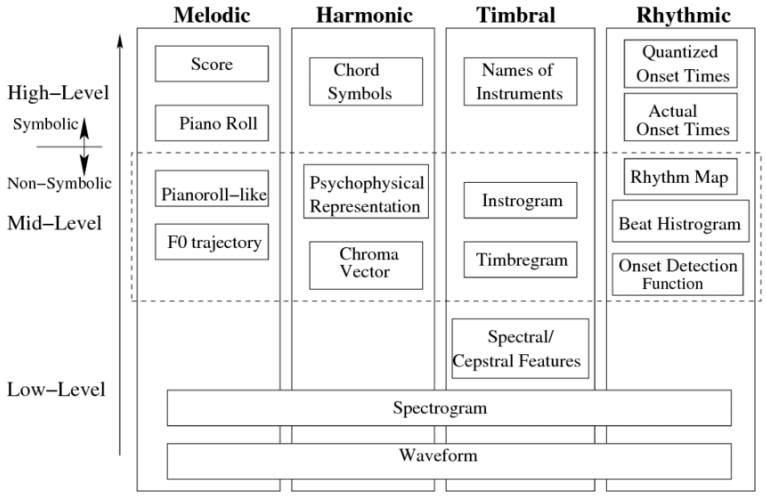

Figure 3:

2.2.1 Acoustic features

Pitch, loudness, and timbre are the acoustic features which are traditionally used to describe sounds and commonly used in many audio retrieval applications [17, 18, 19]. Those features are closely related to our audio perception and sensory stimuli we receive when we hear sounds. Audio perception is itself a complicated discipline. Interested readers are referred to [20, 21] to find a complete coverage of audio features and their effects on perception. Normally, acoustic feature extraction is applied to a short frame of the original signal, computing features roughly once every 10 ms over a window of 20 or 30 ms. It is important to note that this windowing operation, assuming the acoustic feature between a frame is constant, lose information from the original signal. The information loss is often perceptually trivial as long as the parameters are properly tuned for the specific application.

Pitch

Pitch is a perceptual frequency-related scale estimated by taking a series of short-time Fourier spectra. For each of these frames, the frequencies and amplitudes of the peaks are measured and an approximate greatest common divisor algorithm is used to calculate an estimate of the pitch. In the following mathematical formulation, pitch is stored as a log frequency. A perfect young human ear can hear frequencies in the 20-Hz to 20-kHz range and most of the application are configured to identify pitch within this range as accurately as possible.

Loudness

Loudness is approximated by the signal's root-mean-square (RMS) level in decibels, which is calculated by taking a series of windowed frames of the sound and computing the square root of the sum of the squares of the windowed sample values.

Timbre

2.2.2 Symbolic features <in progress>

Unlike acoustic features which are commonly used to many different purposes, symbolic audio features are designed for a specific purpose and deeply reflected by application specifics. Figure 3 shows a class of popular audio features used in n Music Information Retrieval research, including both acoustic and symbolic features. Take an example of the high-level features, *Score* and *Piano Roll* are essential in applications such as melody analysis but not always required for other applications such as beat detection. This is why those are classified as high-level features from the *Melodic* category.

2.2.3 Practical Feature Extraction

In the context of the typical audio engineering development, it is usual that a developer uses open-source libraries for feature extraction tasks instead of coding them from scratch. This is because feature extraction processes demand a very careful optimization to the extent that only a few experts can attain and, more importantly, there are a number of reputable open-source libraries available [22, 23, 24]. Among them, Essentia [25] is an open-source audio analysis tool which powers the similarity search functionality of Freesound. This library supports an extensive collection of reusable algorithms for audio feature extractions as shown in the complete list[3].

As for low-level audio features, Essentia supports a multitude of features such as loudness, dynamic range, zero crossing rates, as well as many spectral-based features including MFCCs, spectral energy, and pitch salience. Furthermore, many other algorithms are distinctly categorized for specific applications, such as mood detection, key detection, melody extraction and audio fingerprinting, and each algorithm is regarded as a high-level feature extractor.

[3]https://essentia.upf.edu/algorithms_reference.html

2.3 Deep Feature Learning

As introduced in the last section, traditional approaches which more or less require audio-specific domain knowledge were dominant in this field. However, recently, the use of deep learning has seen growing popularity and success because the new approach replaces laborious feature engineering with automated feature learning. With deep learning approaches, audio features are often a byproduct of automated feature learning unlike predefined features as we have seen in the traditional approaches. In this section, we go into detail of recent advancements in machine learning research in the context of audio feature learning. We discuss how it can be used to produce more expressive audio representations for audio information retrieval.

2.3.1 VGG

VGG is a Convolutional Neural Network architecture proposed by Karen Simonyan and Andrew Zisserman of Oxford Robotics Institute in 2014 [26]. It was submitted to Large Scale Visual Recognition Challenge 2014 (ILSVRC2014) and the model achieves 92.7% top-5 test accuracy in ImageNet. Originally, the VGG model takes an RGB image as input and passe the image through a stack of convolutional layers, where the filters are used with a very small receptive field: 3x3 (which is the smallest size to capture the notion of left/right, up/down, and center). At each layer, the results are followed by non-linear activation and max-pool. Max-pooling is performed over a 2x2 pixel window, with stride 2. The objective of max-pooling is to down-sample an image representation, reducing its dimensionality and allowing for genelization of the input representation in a lower dimension. With its relative simplicity of implementation and superior performance, it has became one of the standard baseline models in the computer vision research.

The architecture of VGG model is specifically optimized for image classification tasks, however, the architecture has also gained popularity in the audio research field. Since spectrogram can be essentially treated as if a 2 dimensional vector like a grayscale image, a simple adaptation to an audio task is possible relatively easily

by changing the input layer to a spectrogram input.

2.3.2 Case: AudioSet <in progress>

Sound representations from AudioSet [27] use a spectrogram-based CNN architecture trained on a classification task.

2.3.3 Case: OpenL3 <in progress>

OpenL3 [28] also uses a spectrogram-based CNN architecture but trained through self-supervised learning of audio-visual correspondence in videos.

2.3.4 Case: SoundNet <in progress>

About SoundNet and how it extracts features

2.4 Similarity Learning

Classification in machine learning is a supervised learning in which the computer program identifies to which of a set of categories a new observed data belongs, on the basis of a training set of examples consisting of pairs of instances and its category. Classification models have many applications and a variety of the trained models is already used in our everyday life, such as in image recognition [ref], semantic segmentation [ref], and e-mail spam filter [ref] which identifies an incoming e-mail to your inbox to be a spam or not. A lot of the audio features we discussed in the previous section were indeed all by-products of audio classification tasks such as environmental sound classification

While classification learning enables a classifier to identify an observed data with a probability distribution over a set of categories (probabilistic classification), the task is performed without a measure of similarity, explaining how probable the new input belongs to a category but ignoring how similar it is to other data from the existing category. This missing particulars are crucial to feature representations. This is because, with the learned feature representations, what we want to do in

this work is not to use as an input for classification but to make use of it for better organization of information that audio features represent, grouping them in a way that looks coherent to human, dealing with data that are not observed in the training set of data.

Similarity learning is also an area of supervised machine learning which is related closely related to classification but it considers similarity. The goal is to learn to estimate how similar a feature is to other features by optimizing a similarity function that measures a distance in the feature space. However, in spite of being specialized in similarity, a rich line of work has focused solely on features obtained as a by-product of classification learning and similarity learning did not enjoy much attention. For example, the idea of using neural networks to extract features that respect certain relationships dates back to the 90s. Siamese Networks [29] find an embedding space such that similar examples have similar embeddings and vice versa. However, it was not feasible to train such a neural network given the limited compute power at the time and their non-convex nature [30].

With sufficient data, computational power, recent advancement in machine learning allowed for the training of Siamese architecture using triplet losses. Sampling strategy, appropriate distance metric, and the structure of the network are the challenging factors for researchers to improve the performance of the network model. In this section, while several ideas and implementations regarding similarity learning are introduced, our focus lies at the details of the architecture based on triplet loss.

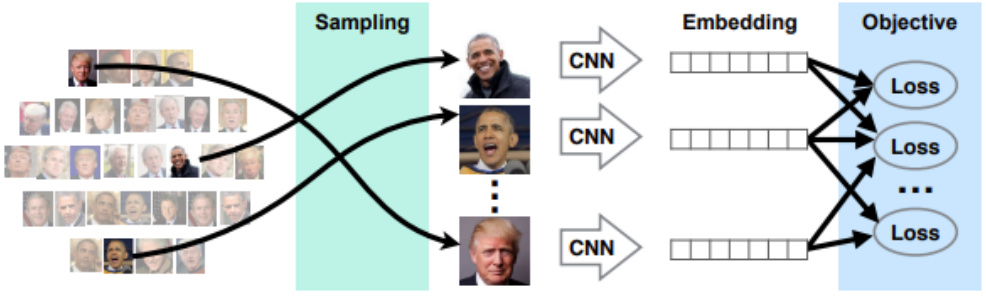

Figure 4: `Similar architecture for audio application will be shown`

2.4.1 Learning Approach <in progress>

Supervised Learning <in progress>

Weakly Supervised Learning <in progress>

2.4.2 Triplet <in progress>

2.4.3 Mining Strategy

It is known that choosing which triplets to use are essential for achieving good generalization in similarity learning [**?**]. For example, a network will not be trained well when it is fed with many easy triplets (a pair of highly distinct categories). Also, when triplet loss is used with a large amount of data, the number of possible combination of triplets easily explodes. To solve this problem, we need to formulate triplet mining strategy which ensures the difficulty of triplets while keeping semantics and train a deep neural network to learn diverse sound representations.

Hard Mining <in progress>

Random Mining <in progress>

2.4.4 Case: FaceNet <in progress>

In face recognition, triplet loss is used to learn good embeddings of faces.

2.4.5 Triplets using tags <in progress>

2.5 Clustering

Our work strives for shift from the traditional audio retrieval to a more knowledgeable approach which provides users with content-based relationships between sounds to access what they look for. The new intelligent way of audio retrieval is analogous to the process of knowledge acquisition. Essentially, knowledge acquisition is the process of storing new information in memory in a way that it can be efficiently

retrieved later [31]. The efficiency of this process depends heavily on the representation quality of the information and also how they are organized in the storage. Take for instance, language acquisition is not only about memorizing a large vocabulary but what's also important is the organization of such information, grouping vocabulary 0by semantic similarity (synonyms) or putting them in a context with sentences.

So far, we have seen the recent advancements in the studies that aim to convert featureless audio signal into a expressive but compact representation of the acoustic information. We concluded in the last section that audio features learned with similarity learning are the most suitable for our purposes. At this point, what's left is choosing an approach for organization which is appropriate for our tasks and the nature of the information we use. With the method of organization, we additionally need to design an intuitive interface through which one can browse sounds and retrieve what he or she looks for.

In our work, we are interested in audio features which are the informational representation of audio data and constructed exclusively on a basis of similarity among sound samples. Therefore, it is reasonable to organize such information in a way that puts similar objects closer and distant ones more distant. In this way, the presented group of audio features resembles what is knowledge in aforementioned knowledge acquisition process. Mathematically speaking, such algorithm belongs to a class of the cluster analysis which requires a distance function so that it measures the distance between objects (audio features).

In this section, we introduce a number of clustering analysis techniques, giving a special importance to the partitional approaches. Before breaking into a specific clustering algorithms, we review the use of clustering algorithms in the context of information retrieval. Specifically, we discuss search result clustering which is an integral part of our research and provide an several review of previous applications in the field.

2.5.1 Search Result Clustering

Search Result Clustering (SRC) is a challenging algorithmic task that requires partitioning the results returned by search engine into classes. This process of partitioning is called Clustering. More specifically, it is a process of organizing similar search results into the same cluster so that each result belongs to a coherent and thematic cluster. Usually, when a traditional search engine is given an ambiguous query, the search engine tries to satisfy this query with high recall, meaning it returns many results which are not necessarily relevant to the user's information need [32]. SRC assists users in such a situation, allowing them to look through the clusters of relevant topics without refomulating the query.

Yippy[4] is a sophisticated example of search engine with an extensive SRC support which receives over 400,000 visitors a month (based on the analysis by SimilarWeb[5]). Yippy itself does not parses and indexes the webpages on the Internet as an usual search engine does. Instead, it works as what's called meta search engine [33], giving users efficient ways of examining a large volume of search results that it obtains from multiple search engines. With the search query *barcelona*, for example, the SRC process on Yippy resulted in 632 clusters from a large clusters labeled *Hotel* and *Lionel Messi* to smaller ones such as *Airport*.

SRC is also a commonly-researched topic in the context of online sound collections where a wide variety of sounds are freely shared, such as the aforementioned Freesound. Searches for audio sounds are particularly time-consuming because users cannot skim across the results as they do in text searches and have to spend at least a few seconds to listen to the sound. Furthermore, in these platforms, the collection is generated by its users instead of coming from professional studios. This results in non-uniformly annotated content compared to professional libraries, which involve experts for annotating and organizing the collections [9], and finding a desired sound in online sound collection by a search query is more difficult and less consistent [34]. Therefore, the technique to identify useful subsets in their results is

[4]https://yippy.com/
[5]https://www.similarweb.com/website/yippy.com

immensely valuable for the search in sound collections.

Many approaches to obtain audio features were already introduced in the last section. In relation to SRC, audio features which requires a large taxonomy such as AudioSet [27] is not suitable. This is because a large size of taxonomies is not applicable in the context of everyday sounds and online collections where the content to describe is too diverse and involves many different types of concepts. This increases the cost of labeling and also introduces technical difficulties such as unbalanced dataset and inability to cluster sounds properly which are unknown in the existing taxonomies. On the other hand, since similarity learning does not require a predefined classes and train audio features directly using a distance metric, sounds which belong to none of existing clusters will ideally be put aside distantly from other clusters.

The interesting research topic is how the difference in audio features influence the performance of the clustering tasks. This is because clustering approaches we will discuss in the following sections, makes full use of given audio features for clustering unlike a classification task in which a model learns to get rid of irrelevant information from given representations. This also indicates the quality of the representation is crucial to the performance of the clustering process.

2.5.2 Partitional Clustering

Partitional clustering is a class of clustering methods that require the number of clusters to be defined before starting the process. K-means clustering is the most popular implementation of the partitional approach and it aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest means (cluster centers or cluster centroid).

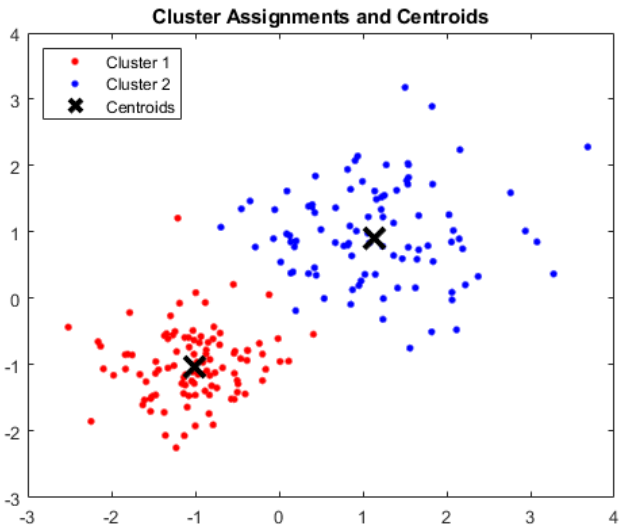

Figure 5:

2.5.3 Graph-based Clustering

2.5.4 Self-organizing Map

2.5.5 Evaluation of Clustering Quality

About mAP, MRR, MR, Top1, Top10, metric

List of Figures

List of Tables

Bibliography

[1] Manning, C. D., Raghavan, P. & Schütze, H. *Introduction to Information Retrieval* (Cambridge University Press, Cambridge, UK, 2008). URL `http://nlp.stanford.edu/IR-book/information-retrieval-book.html`.

[2] Xu, G., Zhang, Y. & Li, L. *Web Mining and Social Networking: Techniques and Applications* (Springer-Verlag, Berlin, Heidelberg, 2010), 1st edn.

[3] Goodrum, A. A. Image information retrieval: An overview of current research **3**, 63–67 (2000).

[4] Foote, J. An overview of audio information retrieval. *Multimedia Syst.* **7**, 2–10 (1999). URL `https://doi.org/10.1007/s005300050106`.

[5] Lew, M. S., Sebe, N., Djeraba, C. & Jain, R. Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl.* **2**, 1–19 (2006). URL `https://doi.org/10.1145/1126004.1126005`.

[6] Williams, H. E. & Lane, D. *Web Database Applications with PHP & MySQL, 2nd Edition* (O'Reilly & Associates, Inc., USA, 2004).

[7] Jansen, B. J. & Rieh, S. Y. The seventeen theoretical constructs of information searching and information retrieval. *Journal of the American Society for Information Science and Technology* **61**, 1517–1534 (2010). URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21358`. `https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.21358`.

[8] Vombatkere, K. & Webberman, A. Information retrieval: Web based analysis and searching. *CSC 461: Database Systems Term Paper* (2017). URL `https://www.cs.rochester.edu/courses/261/fall2017/termpaper/submissions/14/Paper.pdf`.

[9] Font, F., Roma, G. & Serra, X. Freesound technical demo. In *ACM International Conference on Multimedia (MM'13)*, 411–412. ACM (ACM, Barcelona, Spain, 2013).

[10] Mandel, D. R., Jusczyk, P. W. & Pisoni, D. B. Infants' recognition of the sound patterns of their own names. *Psychological Science* **6**, 314–317 (1995). URL `https://doi.org/10.1111/j.1467-9280.1995.tb00517.x`. PMID: 25152566, `https://doi.org/10.1111/j.1467-9280.1995.tb00517.x`.

[11] Kumar, A., Singh, R. & Raj, B. Detecting sound objects in audio recordings. In *2014 22nd European Signal Processing Conference (EUSIPCO)*, 905–909 (2014).

[12] Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)* (Springer-Verlag, Berlin, Heidelberg, 2006).

[13] Rui, Y., Huang, T., Ortega, M. & Mehrotra, S. Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology* **8**, 644–655 (1998).

[14] Mujawar, J., Shukla, S., Dayma, Y. & Thakare, P. R. Feature extraction techniques for multimedia information retrieval system. *Imperial journal of interdisciplinary research* **2** (2016).

[15] Manolescu, D. Feature extraction - a pattern for information retrieval (2000). URL `https://micro-workflow.com/PDF/plop98.pdf`.

[16] Bello, J. P. Low-level features and timbre (2016). URL `http://www.nyu.edu/classes/bello/MIR_files/timbre.pdf`.

[17] Wold, E., Blum, T., Keislar, D. & Wheaton, J. Content-based classification, search, and retrieval of audio. *IEEE MultiMedia* **3**, 27–36 (1996). URL `https://doi.org/10.1109/93.556537`.

[18] Tzanetakis, G., Ermolinskyi, A. & Cook, P. Pitch histograms in audio and symbolic music information retrieval. *Journal of New Music Research* **32** (2002).

[19] Kumar, R. & Chandy, A. Audio retrieval using timbral feature. 222–226 (2013).

[20] Moore, B. C. J. *An Introduction to the Psychology of Hearing (Academic Press) (Academic Press)* (2003).

[21] Bregman, A. *Auditory Scene Analysis: The Perceptual Organization of Sound*, vol. 95 (1990).

[22] McFee, B. *et al.* librosa: Audio and music signal analysis in python. 18–24 (2015).

[23] Giannakopoulos, T. pyaudioanalysis: An open-source python library for audio signal analysis. *PloS one* **10** (2015).

[24] Bullock, J. Libxtract: A lightweight library for audio feature extraction (2007).

[25] Bogdanov, D. *et al.* Essentia: An open-source library for sound and music analysis. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, 855–858 (Association for Computing Machinery, New York, NY, USA, 2013). URL `https://doi.org/10.1145/2502081.2502229`.

[26] Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations* (2015).

[27] Gemmeke, J. F. *et al.* Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017* (New Orleans, LA, 2017).

[28] Cramer, J., Wu, H.-H., Salamon, J. & Bello, J. P. Look, listen and learn more: Design choices for deep audio embeddings. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 3852–3856 (Brighton, UK, 2019).

[29] Bromley, J. *et al.* Signature verification using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence* **7**, 25 (1993).

[30] Wu, C., Manmatha, R., Smola, A. J. & Krähenbühl, P. Sampling matters in deep embedding learning. *CoRR* **abs/1706.07567** (2017). URL `http://arxiv.org/abs/1706.07567`. `1706.07567`.

[31] McNamara, D. S. Learning: Knowledge representation, organization, and acquisition (2002).

[32] Scaiella, U., Ferragina, P., Marino, A. & Ciaramita, M. Topical clustering of search results. In *Proceedings of the fifth ACM international conference on Web search and data mining*, 223–232 (New York, NY, USA, 2012).

[33] Glover, E., Lawrence, S., Birmingham, W. & Giles, C. Architecture of a metasearch engine that supports user information needs (2000).

[34] Xavier Favory, X. S., Frederic Font. Comparing audio features for unsupervised sound classification. *unpublished preprint* (2020).