Master thesis on Sound and Music Computing

Universitat Pompeu Fabra

Learn Sound Representations Using Triplet-loss

Kohki Mametani

Supervisor: Xavier Favory

Co-Supervisor: Frederic Font

September 2019

Master thesis on Sound and Music Computing
Universitat Pompeu Fabra

Learn Sound Representations Using Triplet-loss

Kohki Mametani

Supervisor: Xavier Favory

Co-Supervisor: Frederic Font

September 2019

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

This is an example paragraph. As you can see, the main text uses a font size of 12 pt and a line spacing of 1.5. Neither the paragraphs nor the first lines of paragraphs should be indented.

There is no very strict page limit. Your number of pages will be strongly influenced by the size and total number of your figures and tables. It is recommended staying within 30-50 pages. Do not try to fill as many pages as you can. Longer theses are not necessarily of higher quality and of more non-redundant content than shorter theses. Certainly, a master thesis of 15 pages is too short, and a master thesis of 100 pages is too long. Also, here you can see a sample reference [**?**]

1.1 Motivation

1.2 Objectives

1.3 Structure of the Report

Chapter 2

State of the Art

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 created, shared, and stored on the internet today, the following activities are recognized as popular contributors of the data growth: users on YouTube uploads 400 hours of new video each minute of every day and Instragram users upload over 100 milion photos and videos every day.

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 the user examines a few samples of the results until she finds a desired data. This process is often designed on a basis of variaous technical consideration, for example, how to organize data efficiently, how to precisely respond to dirverse user requests with inconsistent set of vocabulary, or how to present the search results to users.

2.1 Sound Retrieval

Among the variation of multimedia, we focus on audio data, more specifically on the general sound including artificial and natural sound. This section aims to offer a contextual review about sound retrieval, specifically how to obtain a desired data from a large online audio databases.

2.1.1 Sound and Music Computing

Sound and music computing (SMC) is a research field that studies the whole sound and music communication chain from a multidisciplinary point of view. By combining scientific, technological and artistic methodologies it aims at understanding, modeling and generating sound and music through computational approaches.

2.1.2 Large Online Databases

2.1.3 Freesound

2.1.4 In search of ideal sound

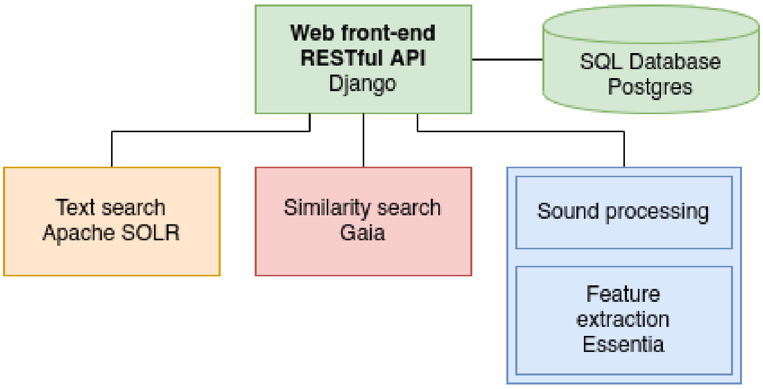

Figure 1:

The rest of this chapter is organized as follows: in the next section, a traditional class of audio feature, the building blocks of sound 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, we provide a review of clustering methods with the focus on two popular approaches, which are the partitional clustering and the graph-based clustering.

2.2 Audio Features

In machine learning and pattern recognition, a feature is an individual measurable property or characteristic of a phenomenon being observed [Bishop]. It also known in statistics as an explanatory variable (or independent variable, although features may or may not be statistically independent). While features may variously be binary (e.g. "on" or "off"); categorical (e.g. "A", "B", "AB" or "O", for blood type); ordinal (e.g. "large", "medium" or "small"); integer-valued (e.g. the number of occurrences of a particular word in an email); or real-valued (e.g. a measurement of blood pressure), in our work, we focus on features that are numeric.

Audio features that allows for extracting properties from audio signal, detecting if any pattern is present in the signal, and how a particular signal is correlated to another similar signals lie at the heart of our work. This section aims to offer an review about some general classes of audio 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 audio, such as "brightness" and "sharpness".

Within a typical scheme of audio applications, audio features are obtained in a preliminary process, which is referred as a feature extraction process. Feature extraction is a process that convert the raw audio waveform (time series data points) into a set of numerical vectors. When analyzing digital audio signals, the raw signal is too large, redundant, and noisy.

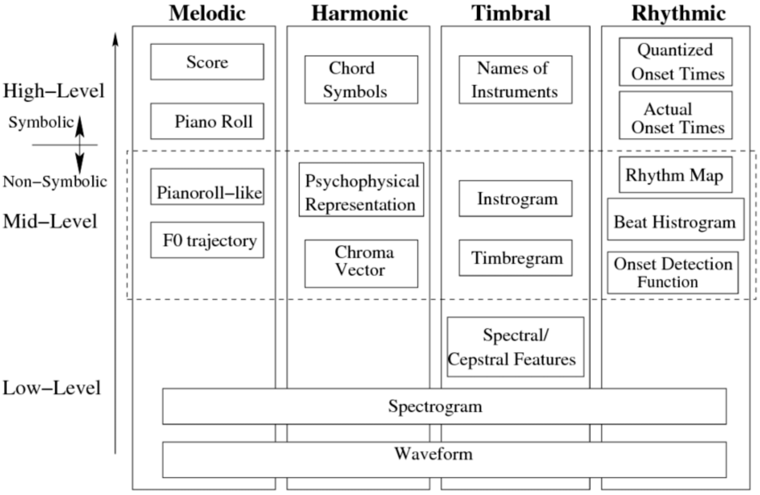

Figure 2:

2.2.1 Low-level features

Low-level audio features are obtainable instantaneously using several signal processing techniques. Normally, low-level 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.

2.2.2 High-level features

2.3 Deep Feature Learning

Traditional approaches which require domain knowledge of human perception were dominant in this field. However, recently, the use of deep learning has seen growing popularity and success since the approach replaced laborious feature engineering with automated feature learning. With deep learning approaches, sound representations (features that represent sounds) are often a byproduct of automated feature learning instead of fixed high-level features, such as MFCCs, which were used in traditional approaches.

2.3.1 VGG

VGG is a Convolutional Neural Network architcture, It was proposed by Karen Simonyan and Andrew Zisserman of Oxford Robotics Institute in the the year 2014. 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 a RGB images 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 2Œ2 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 dimesion.

The VGG model design is based on a computer vision task, specifically image classification, however, it is also a common architecture in the audio field. Since spectrogram is a 2 dimensional vector like a grayscale image, a simple adaptation to an audio task can be done relatively easily by changing the input layer to a spectrogram input.

2.3.2 Case: AudioSet

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

2.3.3 Case: OpenL3

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

2.3.4 Case: SoundNet

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, semantic segmentation, and e-mail spam filter which identifies an incoming e-mail to your inbox to be a spam or not. A lot of the features we discussed in the previous section were indeed all by-products of 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 is 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 data, grouping the data 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 data is to other data points by optimizing a similarity function that measures a distance between data points. However, despite 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 [4] 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.

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.

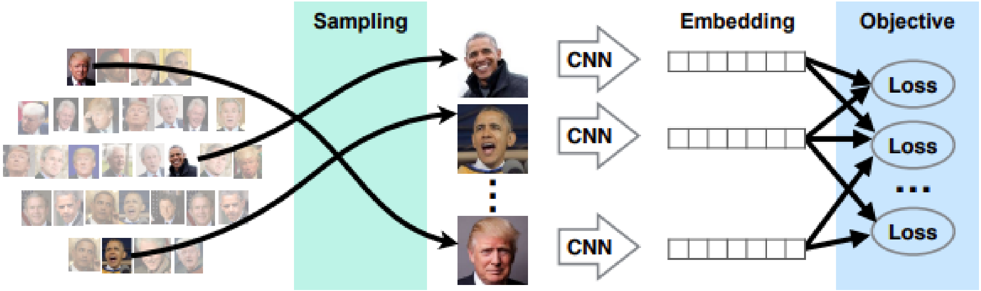

Figure 3:

2.4.1 Learning Approach

Supervised Learning

Weakly Supervised Learning

2.4.2 Benefits

2.4.3 Problems

2.4.4 Triplet

2.4.5 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 ensure the difficulty of triplets while keeping semantics and train a deep neural network to learn diverse sound representations.

Hard Mining

Random Mining

2.4.6 Case: FaceNet

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

2.4.7 Triplets using tags

2.5 Search Result Clustering

Search Result Clustering is commonly used in online sound collections where a wide variety of content are freely shared, such as Freesound.org. The technique allows users to identify useful subsets in their results. 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 [?]. Therefore, finding a desired sound in online sound collection by a search query is more difficult and less consistent [?].

For Search Result Clustering in such a condition, one approach is to obtain sound representation using large hierarchical structures called taxonomy [?]. However, a large size of taxonomies is not applicable because, in the context of everyday sounds and online collections, 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 which are unknown in the existing taxonomies.

We are interested in how those sound representations influence the performance of sound clustering tasks. This is because a clustering model (e.g. K-means clustering model) makes full use of given sound representations for their task unlike a classification task in which a model learns to get rid of irrelevant information from given representations. Therefore, the quality of the representation is crucial to the quality of sound clustering. Specifically, we focus on Search Result Clustering [**?**] in the context of online audio collections which use sound representations to organize search-result audio content into coherent groups in the context.

2.5.1 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).

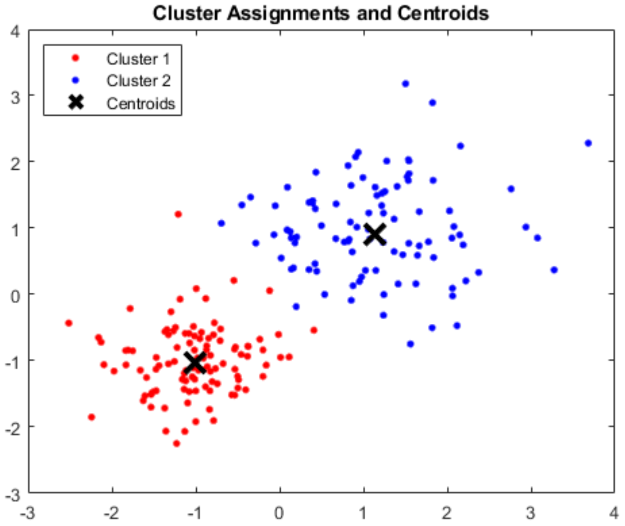

Figure 4:

2.5.2 Graph-based Clustering

Chapter 3

Methods

This is an example paragraph. As you can see, the main text uses a font size of 12 pt and a line spacing of 1.5. Neither the paragraphs nor the first lines of paragraphs should be indented.

There is no very strict page limit. Your number of pages will be strongly influenced by the size and total number of your figures and tables. It is recommended staying within 30-50 pages. Do not try to fill as many pages as you can. Longer theses are not necessarily of higher quality and of more non-redundant content than shorter theses. Certainly, a master thesis of 15 pages is too short, and a master thesis of 100 pages is too long.

3.1 Materials

Chapter 4

Results

This is an example paragraph. As you can see, the main text uses a font size of 12 pt and a line spacing of 1.5. Neither the paragraphs nor the first lines of paragraphs should be indented.

There is no very strict page limit. Your number of pages will be strongly influenced by the size and total number of your figures and tables. It is recommended staying within 30-50 pages. Do not try to fill as many pages as you can. Longer theses are not necessarily of higher quality and of more non-redundant content than shorter theses. Certainly, a master thesis of 15 pages is too short, and a master thesis of 100 pages is too long.

4.1 Tables and graphics

This is an example paragraph. As you can see, the main text uses a font size of 12 pt and a line spacing of 1.5. Neither the paragraphs nor the first lines of paragraphs should be indented.

There is no very strict page limit. Your number of pages will be strongly influenced by the size and total number of your figures and tables. It is recommended staying within 30-50 pages. Do not try to fill as many pages as you can. Longer theses are not necessarily of higher quality and of more non-redundant content than shorter theses. Certainly, a master thesis of 15 pages is too short, and a master thesis of 100 pages is too long.

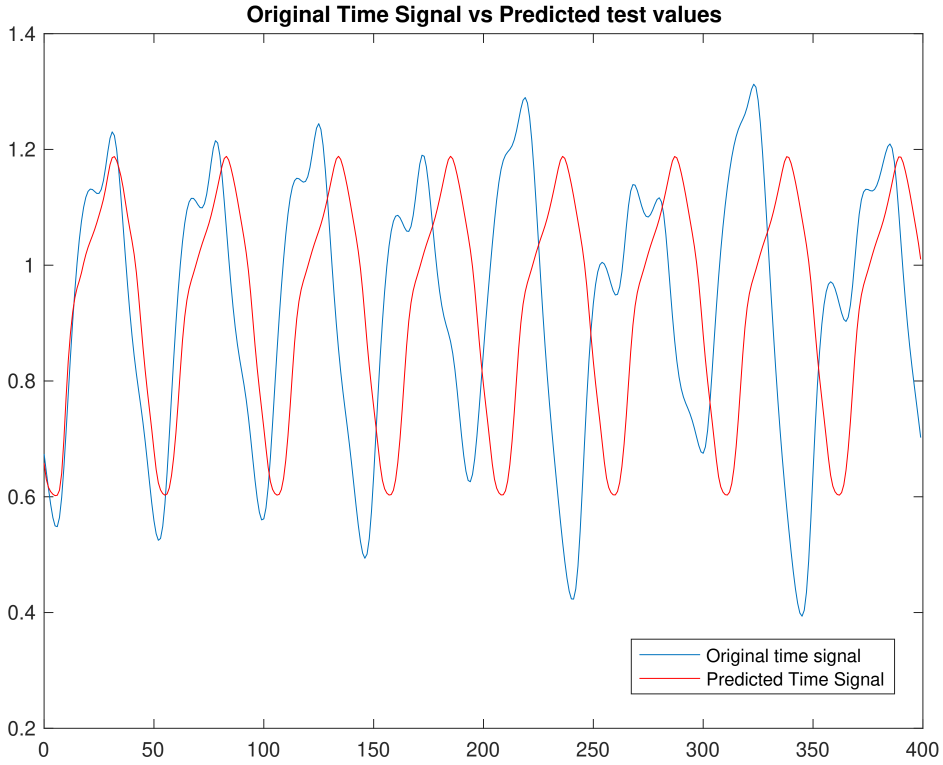

Figure 5: This is an example of a figure and its caption.

Table 1: This is an example of a table and its caption.

PCA	**Residual mean (in absolute values)**
Original PCA	0.1267
PCA on Centroid 1	0.1249
PCA on Centroid 2	0.1214

Chapter 5

Discussion

This is an example paragraph. As you can see, the main text uses a font size of 12 pt and a line spacing of 1.5. Neither the paragraphs nor the first lines of paragraphs should be indented.

There is no very strict page limit. Your number of pages will be strongly influenced by the size and total number of your figures and tables. It is recommended staying within 30-50 pages. Do not try to fill as many pages as you can. Longer theses are not necessarily of higher quality and of more non-redundant content than shorter theses. Certainly, a master thesis of 15 pages is too short, and a master thesis of 100 pages is too long.

5.1 Discussion

5.2 Conclusions

List of Figures

List of Tables

Appendix A

First Appendix

This is an example paragraph. As you can see, the main text uses a font size of 12 pt and a line spacing of 1.5. Neither the paragraphs nor the first lines of paragraphs should be indented.

There is no very strict page limit. Your number of pages will be strongly influenced by the size and total number of your figures and tables. It is recommended staying within 30-50 pages. Do not try to fill as many pages as you can. Longer theses are not necessarily of higher quality and of more non-redundant content than shorter theses. Certainly, a master thesis of 15 pages is too short, and a master thesis of 100 pages is too long.

Appendix B

Second Appendix