Audio Feature Learning with Triplet-Based Embedding Network

Xiaoyu Qi, Deshun Yang, and **Xiaoou Chen**
Institute of Computer Science and Technology, Peking University
128th, ZhongGuanCun North Street, Haidian District, Beijing, 100080, P.R. China
{xyqi, yangdeshun, chenxiaoou}@pku.edu.cn

Abstract

We propose a triplet-based network for audio feature learning for version identification. Existing methods use hand-crafted features for a music as a whole while we learn features by a triplet-based neural network on segment-level, focusing on the most similar parts between music versions. We conduct extensive experiments and demonstrate our merits.

Introduction

A version of a music is a new performance of a previous work. Given a query audio piece, version identification aims to retrieve all the versions in a database. Music version identification has wide application, especially in music copyright monitoring, while it remains challenging due to the variances in tempo, key, etc. Previous work mainly focus on extracting features from pitch class profiles or chroma, which well captures the tonal content of a music, and has been proved to be more effective than original audio spectrums descriptors such as Mel-frequency cepstrum coefficients.

Most version identification algorithms use some form of dynamic time-warping (DTW) (Serra 2011). Balen et al.(Bertin-Mahieux and Ellis 2011) extracted hash code called 'jump code' from beat-aligned chroma. Balen et al. (van Balen et al. 2014) proposed cognition-inspired descriptors as audio features. From then on, more methods such as information theoretic measures have been proposed with little performance improvement. Having seen the glass-ceiling for hand-crafted features, we thus turn to machine learning algorithms for higher learning capability to bridge the semantic gap in the long term.

Version identification is basically a problem of similarity metric learning. To this end, our goal is to embed the data to a feature space where versions of the same music are close to each other and music from different versions stay relatively far away. We propose a triplet-based neural network, which learns audio similarity from triplet input containing a query audio piece, a same version and a different version. In addition, in order to match music accurately, we conduct the learning process on segment-level instead of song-level. The experimental results demonstrate that our method outperforms hand-crafted ones. Our contribution can be summarized into two-folds. 1) To our knowledge, this is the first time that a triplet deep learning method has been applied to audio metric learning. 2) We propose an end-to-end framework and the triplet-based network shall be applicable for large-scale commercial use due to its high efficiency and great potential as the amount of audio data grows.

Method Overview

Given a training set $\mathcal{D} = \{(a_i, v_i)\}_{i=1}^{N}$, where $a_i = \{e_{j,k}\}_{l \times d}$ is an audio track, and v_i denotes the version it belongs to. $e_{j,k} \in (0, 1)$ indicates audio energy of different frequency bands. l is the audio duration and d is the number of frequency bands. Our goal is to learn a similarity metric $S(\cdot, \cdot)$ with $\mathcal{D}$ in order to make a prediction of the similarity degree given a new music pair (a_j, a_k).

It is impractical to train a classifier for each kind of version due to the large amount of music tracks and small number of versions. To solve this problem, we learn by distance comparison within triplets. Furthermore, we experiment on segment-level for more precise learning.

Our framework includes two parts. First, we construct triplets set $\tilde{\mathcal{D}}$ from $\mathcal{D}$ on segment-level. $\tilde{\mathcal{D}} = \{(x_i, x_i^+, x_i^-)\}_{i=1}^{N}$, where x_i is a music segment, x_i^+ denotes a same version of x_i, while x_i^- represents a non-version one. Then, we embed audio with a triplet-based network and form the similarity metric on the embedded feature.

Audio segmentation and triplets generation

In this section, we conduct audio segmentation and construct triplets set. We experiment on segment-level for two reasons, focusing on the most significant parts and solving the problem that audio versions of different genres usually vary in length while triplets should have equal length. We encode audio tracks and build an index for efficient segment retrieval to construct triplets. The concrete algorithm is as follows.

A song is originally represented as a sequence of d-dimensional vectors e_k. In the first step, each vector is encoded into a 64-bit vector with two parts. One part concerns the audio vector itself, consisting of four 12-bit sub-codes. In each sub-code, the bits corresponding to the top m largest dimensions of the vector are set to 1 and others 0, with m being 6, 5, 3 and 1. The other part of a code represents the

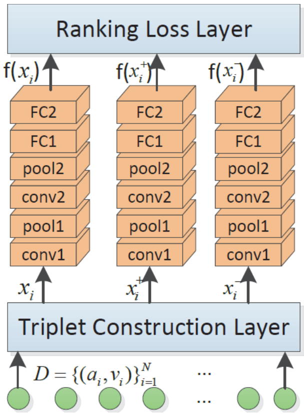

Figure 1: Triplet network

average of the neighbors of the vector. Based on the first parts of codes, we build a code index which indicates the occurrences of each unique code in all the songs in $\mathcal{D}$.

Then, we divide a_i into segments $x_i, i = 1, 2...n$ of length l, with step size d. For each code in a segment x_i, find out through the index all occurrences of the code in a_j, calculate the Hamming distance between the x_i segment and the occurrence-aligned x_i^+ segment. Generate pairs of similar music segments between x_i and x_i^+.

For each similar segment pair (x_i, x_i^+), we randomly select a music segment from a different song and denote it as x_i^- to generate a triplet (x_i, x_i^+, x_i^-).

Triplet-based network feature learning

We propose a triplet-based embedding neural network for audio metric learning. Given an audio triplet (x_i, x_i^+, x_i^-), triplet-based neural network embeds them to a vector space, generating $f(x_i), f(x_i^+)$ and $f(x_i^-)$ respectively. Specifically, as we show in Figure 1, $f(\cdot)$ consists of two convolution layers, with Relu as activation function, followed by two max-pooling layers and two full connection layers. With the audio embeddings, we adopt the online distance metric learning algorithm with cosine similarity proposed in (Wu et al. 2013). The similarity function is denoted as follows:

$$S(x_i, x_j) = f(x_i)^T f(x_j)/(||f(x_i)|| \times ||f(x_j)||) \qquad (1)$$

The hinge loss (Wan et al. 2014) is defined as:

$$L(D;\theta) = \sum_{i=1}^{N} max(1 - S(x_i, x_i^+) + S(x_i, x_i^-), 0) \qquad (2)$$

where θ represents all of the parameters in our model, and N denotes the number of training instances. θ is updated by back-propagation and each base network in the triplet network share the same architecture and parameter weights.

Experiment

We use the SecondHandSongs dataset (SHSD) which is an official list of cover songs within the Million Song Dataset (MSD) released in 2011. SHSD is the largest public cover song dataset up to now, including 18196 tracks in 5854 cliques. To build our model, we use the training split of SHSD dataset with 10150 tracks and test on the standard Query1500 testing set (including 500 triplets).

We clip the training audios into segments and experiment on length from 100 to 700 with a step of 50. We randomly sampled 10^6 triplets for training and 10^5 for testing, 10 to 100 times of our network parameter number complexity which is about 10^4.

We experiment on various parameter weights and set the convolution kernel size to 3, pooling size and stride both to 2 for best performance. The first full connection layer has 100 nodes and the second has 50 nodes as output. The triplet network is implemented with Caffe.

We test the model on the Query1500 data set which is given in the form of (x_i, x_i^+, x_i^-), first two of the same version and the third different from the first. We calculate querying accuracy by comparing the similarity function results $S(x_i, x_i^+)$ and $S(x_i, x_i^-)$.

$$Acc = \frac{1}{T} \sum_{i=1}^{T} g(S(x_i, x_i^+) - S(x_i, x_i^-)) \qquad (3)$$

where $g(c) = 1$ if $c > 0$ and $g(c) = 0$ otherwise. T denotes the testing instance numbers. The results are in Table 1.

Method	Jump codes	DTW	Cognition	Triplet
Acc	77.4%	80.0%	73.2%	82.2%

Table 1: Performances on Query1500

Acknowledgments

This work was supported by National Hi-Tech Research and Development Program (863 Program) of China under Grant2014AA015102 and the Natural Science Foundation of China(No.61370116).

References

Bertin-Mahieux, T., and Ellis, D. P. 2011. Large-scale cover song recognition using hashed chroma landmarks. In *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 117–120. IEEE.

Serra, J. 2011. Identification of versions of the same musical composition by processing audio descriptions. *Department of Information and Communication Technologies*.

van Balen, J.; Bountouridis, D.; Wiering, F.; Veltkamp, R. C.; et al. 2014. Cognition-inspired descriptors for scalable cover song retrieval. In *proceedings of the 15th international conference on Music Information Retrieval*.

Wan, J.; Wang, D.; Hoi, S. C. H.; Wu, P.; Zhu, J.; Zhang, Y.; and Li, J. 2014. Deep learning for content-based image retrieval: A comprehensive study. In *Proceedings of the 22nd ACM international conference on Multimedia*, 157–166. ACM.

Wu, P.; Hoi, S. C.; Xia, H.; Zhao, P.; Wang, D.; and Miao, C. 2013. Online multimodal deep similarity learning with application to image retrieval. In *Proceedings of the 21st ACM international conference on Multimedia*, 153–162. ACM.