MIREX 2020: COVER SONG IDENTIFICATION WITH AN END-TO-END METRIC LEARNING BASED ON CONVOLUTION NEURAL NETWORK

Jian Zhao

NetEase Cloud Music

zhaojian01@corp.netease.com

Huaping Liu

NetEase Cloud Music

liuhuaping@corp.netease.com

ABSTRACT

This document describes our submission to audio cover song identification task of MIREX 2020.we propose an end-to-end neural network architecture that is trained to represent each audio track as a single fixed length embedding vector.We extract each audio track's embedding directly out of its original audio representation, rather than from other indirect audio representation, such as dominant melody,beat&[1]. We use a trained end-to-end metric learning based on convolution neural network to the task of extracting each audio track's embedding.Further more, the task's training goal reduces to a simple Euclidean distance computation.This algorithm we submitted improves state-of-the-art accuracy.

1. INTRODUCTION

Cover songs are different versions of the same original music. They usually share similar melody, but differ in one or several other dimensions, such as lyrics, structure, tempo, instrumentation, genre, etc. Cover song identification is an important task because it can prevent music copyright infringement and apply to similar music retrieval systems.

2. IMPLEMENTATION DETAILS

The end-to-end neural network architecture mapping each audio track to a single fixed length embedding, and trained to minimize cover song pairs Euclidean distance in the embedding space, while maximizing it for non cover song pairs,allowing us to use the triplet loss function like in FaceNet[2].We define the pairs and non pairs of songs as follows, which will be the input data of the end-to-end neural network(Figure 1):

Query set (Q): The query set includes 120 songs with fifty covers each.

Reference set: The reference set includes P of the remaining covers for each of the 120 query songs, and N covers for each other songs not included in the query set.

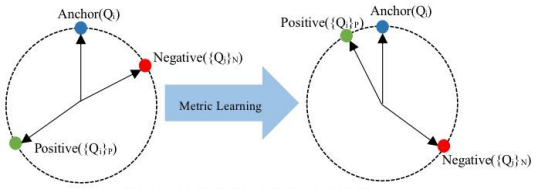

Figure 1: The Triplet loss in Euclidean similarity

We use softmax and cross entropy loss to pre-train our model. It uses classification layer to replace the fixed length embedding and triplet loss layers in Figure 1. Using softmax pre-training to initialize the weights of the network has two main benefits. First, we notice that the cross entropy loss produces stabler convergence than triplet loss. Secondly, while triplet selection is faster with larger minibatches, smaller mini-batches usually yield better generalization in Stochastic Gradient Descent (SGD).

Proved by our experiment,with softmax pre-trained neural network can achieve lower EER and higher ACC than neural networks without pre-training.

3. REFERENCES

[1] Guillaume Doras, Geoffroy Peeters. Cover detection using dominant melody embeddings. ISMIR 2019, Nov 2019, Delft, Netherlands. ffhal-02457735.

[2] Li Chao, et al., "Deep Speaker: an End-to-End Neural Speaker Embedding System." arXiv preprint arXiv:1705.02304 (2017).