CASPERNET FOR COVER SONG IDENTIFICATION

Xingjian Du, Zhesong Yu, Bilei Zhu
ByteDance AI Lab
{duxingjian.real, yuzhesong, zhubilei}@bytedance.com
Xiaoou Chen
Peking University
chenxiaoou@pku.edu.cn

ABSTRACT

This paper describes the algorithm for MIREX 2020 cover song identification task. We utilize ResNet50-IBN to learn a descriptor toward cover song identification.

1. INTRODUCTION

Cover song identification has long been an interesting topic for researchers in Music Information Retrieval as its potential applications in music license management, music retrieval, etc. Over the past ten years, the researchers initially attempt to employ dynamic programming toward this task, such as chroma [4, 5]. And some attempted to model music for cover song identification, such as 2DFM [1]. Furthermore, a few researchers used Deep Learning for this task recently [3, 6]. We designed a more powerful network structure to extract compact representations from music.

2. METHOD

We have a training dataset $D = \{(x_n, t_n)\}$, where x_n is a recording and t_n is a one-hot vector denoting to which song (or class) the recording belongs. Different versions of the same song are viewed as the samples from the same class, and different songs are regarded as the different classes. We aim to train a classification network model f_θ parameterized by θ from D. Then, this model could be used for cover song retrieval. More specifically, after the training, given a query q and references r_n in the dataset, we extract latent features $f_\theta(Q), f_\theta(R_n)$ using the network, which we call music representations, and use a metric s to measure their similarity.

2.1 Feature

CQT, mapping frequency energy into musical notes [2], is extracted by *Essentia* for our experiment. The audio is resampled to 22050 Hz, the number of bins per octave is set as 12 and Hann window is used for extraction.

2.2 Training Scheme

We design spectrum sitiching augmentation when training the model. The training loss includes cross entropy loss, triplet loss and center loss.

Second Hand Songs 100K (SHS100K), which is collected from *Second Hand Songs website* and *Youtube* [6], contains 8858 songs with various covers and 108523 recordings in total. In our experiments, we split this dataset into three subsets – *SHS100K-TRAIN*, *SHS100K-VAL* and *SHS100K-TEST* with a ratio of $8 : 1 : 1$ for training, validation and testing respectively.

2.3 Retrieval

the network is used to extract music representation. As shown in Figure **??**, given a query q and a reference r, we first extract their CQT descriptors Q and R respectively, which are fed into the network to obtain music representations $f_\theta(Q)$ and $f_\theta(R)$), and then the similarity s is defined as their cosine similarity:

$$s(f_\theta(Q), f_\theta(R)) = \frac{f_\theta(Q)^T f_\theta(R)}{|f_\theta(Q)||f_\theta(R)|} \quad (1)$$

After compute pair-wise similarity between query and references in dataset, a ranking list is returned for evaluation.

3. REFERENCES

[1] Thierry Bertin-Mahieux and Daniel PW Ellis. Large-scale cover song recognition using the 2d fourier transform magnitude. In *International Society for Music Information Retrieval Conference*, 2012.

[2] Judith C Brown. Calculation of a constant q spectral transform. *The Journal of the Acoustical Society of America*, 89(1):425–434, 1991.

[3] Sungkyun Chang, Juheon Lee, Sang Keun Choe, and Kyogu Lee. Audio cover song identification using convolutional neural network. In *Workshop Machine Learning for Audio Signal Processing at NIPS*, 2017.

[4] Joan Serrà, Emilia Gómez, Perfecto Herrera, and Xavier Serrà. Chroma binary similarity and local alignment applied to cover song identification. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(6):1138–1151, 2008.

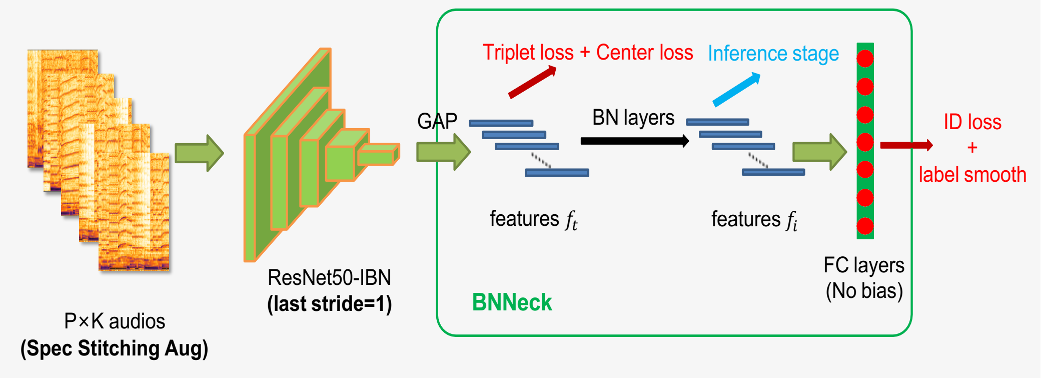

Figure 1. The pipeline of our feature extraction model

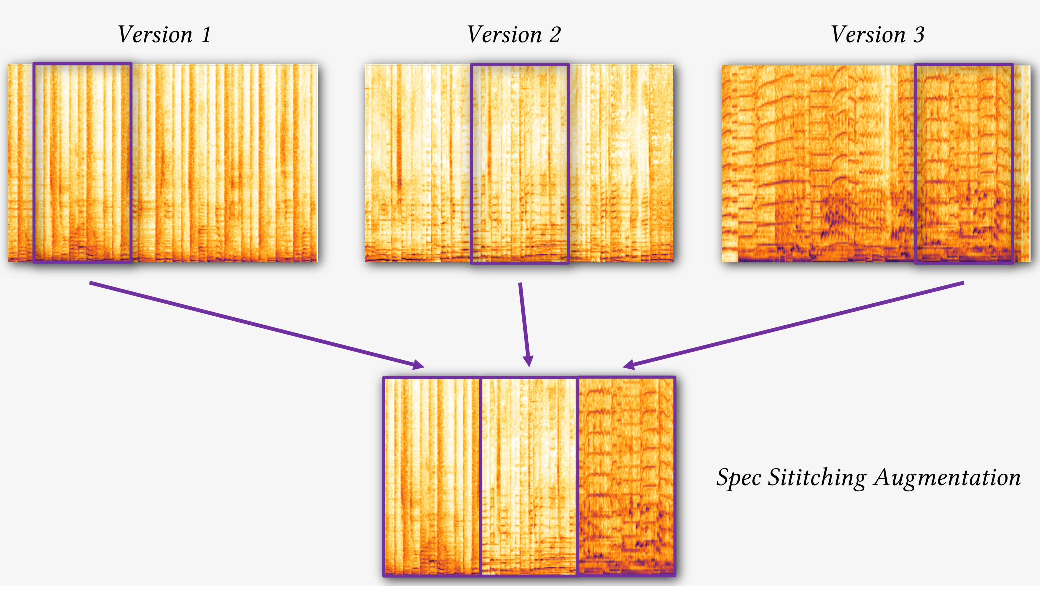

Figure 2. The segments of spectrogram from different versions of the same song are stitched along the time dimension.

[5] Joan Serrà, Xavier Serrà, and Ralph G Andrzejak. Cross recurrence quantification for cover song identification. *New Journal of Physics*, 11(9):093017, 2009.

[6] Xiaoshuo Xu, Xiaoou Chen, and Deshun Yang. Key-invariant convolutional neural network toward efficient cover song identification. In *2018 IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2018.