CQTNET: CONSTANT-Q TRANSFORM NETWORK FOR COVER SONG IDENTIFICATION

Zhesong Yu, Xiaoshuo Xu, Xiaoou Chen, Deshun Yang
Wangxuan Institute of Computer Technology, Peking University
{yzs, xsxu, chenxiaoou, yangdeshun}@pku.edu.cn

ABSTRACT

This paper describes the algorithm for MIREX 2019 cover song identification task. We utilize Convolutional Neural Networks (CNNs) to learn a descriptor toward cover song identification. Viewing different songs as different classes and versions as samples, we train the network through classification criteria. After the training, the network is used to extract music representation for 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 [7,8]. 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 [3,9]. Based on TPPNet [10], we designed a more powerful network structure to extract compact representations from music.

2. METHOD

As shown in Figure 1, 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 *Librosa* [4] for our experiment. The audio

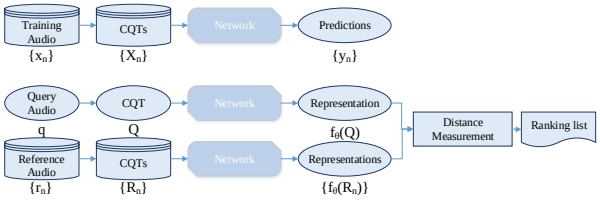

Figure 1. Training procedure and retrieval procedure.

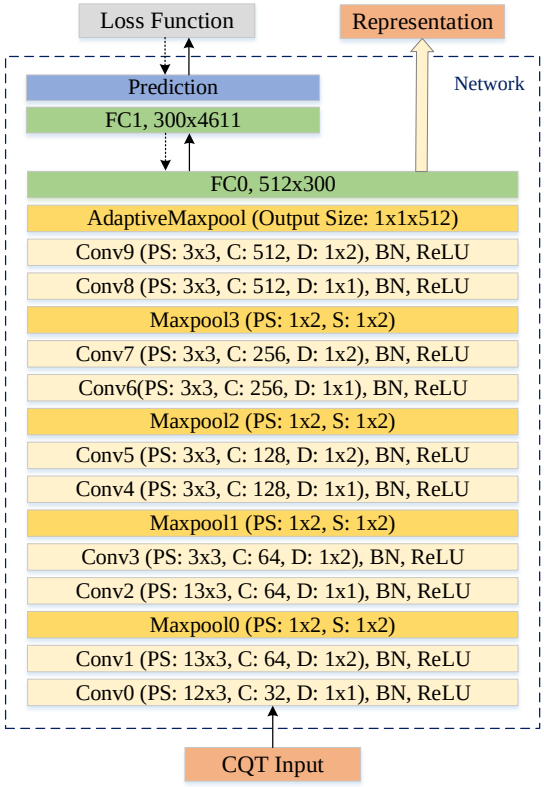

Figure 2. Network structure. KS: kernel size, C: channel number, D: dilation and S: stride. The stride is set to 1×1 for the convolutional layers, and pooling layers has a dilation of 1×1. The output dimension is 4611, the number of classes in the training set.

is resampled to 22050 Hz, the number of bins per octave is set as 12 and Hann window is used for extraction with a hop size of 512. Finally, a 20-point mean filter is applied to CQT, and the resulting feature is a sequence with a feature rate of about 2 Hz. It could also be viewed as a $84 \times T$ matrix where T depends on the duration of input audio.

2.2 Network Structure

We stack small filters following with max pooling operations, except that in initial layers, we design the height of filter to be 12 or 13 (see Figure 2). This setting results in that the units of the third layer have a receptive field with a height of 36; it spans across three octaves or thirty-six semitones. Besides, we also introduce dilated convolution into the model. This structure helps enlarge the receptive field and works well in speech synthesis and speech denoising [5, 6]. More importantly, our model does not involve any downsample pooling operation in the frequency dimension; in other words, the vertical stride is always set to 1. Furthermore, after several convolutional and pooling layers, we apply an adaptive pooling layer to the feature map, whose length varies depending on the input audio.

2.3 Training Scheme

We design a multi-length training strategy to resolve this problem. For each batch, we sample some recordings from the training set and extract CQTs from them. For each CQT, we randomly crop three subsequences with a length of $200, 300$ and 400 for training, corresponding to roughly 100s, 150s and 200s respectively.

We make some data augmentation when training the model. We sample a changing factor from $(0.7, 1.3)$ for each recording in the batch following uniform distribution and simulate tempo changes using *Librosa* [4] on the recording before cropping subsequences.

Second Hand Songs 100K (SHS100K), which is collected from *Second Hand Songs website* and *Youtube* [9], 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.

2.4 Retrieval

the network is used to extract music representation. As shown in Figure 1, 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] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, pages 18–25, 2015.

[5] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[6] Dario Rethage, Jordi Pons, and Xavier Serra. A wavenet for speech denoising. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5069–5073. IEEE, 2018.

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

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

[9] 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.

[10] Zhesong Yu, Xiaoshuo Xu, Xiaoou Chen, and Deshun Yang. Temporal pyramid pooling convolutional neural network for cover song identification. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4846–4852, 2019.