TEMPORAL FEEDBACK CONVOLUTIONAL RECURRENT NEURAL NETWORKS FOR MUSIC GENRE CLASSIFICATION

Taejun Kim, Juhan Nam
Graduate School of Culture Technology
Korea Advanced Institute of Science and Technology (KAIST)
{taejun,juhannam}@kaist.ac.kr

ABSTRACT

This paper describes models used for audio classification task of the Music Information Retrieval Evaluation eXchange (MIREX) 2019. We recently proposed a novel convolutional recurrent neural networks (CRNNs) architecture named temporal feedback CRNNs (TF-CRNNs) for keyword spotting [4]. The architecture is inspired by efferent connections in the human auditory system, which is the feedback pathway from the brain to ears. Experimental results show that the models are effective for keyword spotting. In this paper, to further explore the capability of TF-CRNNs, we also evaluate the TF-CRNNs on music genre classification in this study.

1. INTRODUCTION

One of goals in deep learning is to reduce domain knowledge required for a task. However, when a neural network architecture is designed, domain knowledge is often leveraged. A method often used as domain knowledge is utilizing the brain mechanism. In our previous study, we also proposed a novel convolutional recurrent neural networks (CRNNs) architecture inspired by efferent connections in the human auditory system, which is the feedback pathway from the brain to ears [4].

2. ARCHITECTURE

The proposed architecture named temporal feedback CRNN (TF-CRNN) has temporal feedback connections that are conceptually analogous to the mechanism of the outer-hair cells [5]. More specifically, a temporal feedback connection is a connection from a hidden state of a recurrent neural network (RNN) at the previous time step h^{t-1} to a convolutional block at the current time step h^t as illustrated in Figure 1. The feedback signal is used to scale features extracted from the convolutional block in a channel-wise manner. Figure 2 describes a convolutional block in the TF-CRNN. The channel-wise feature scaling weights are computed by a fully connected (FC) layer using the temporal feedback.

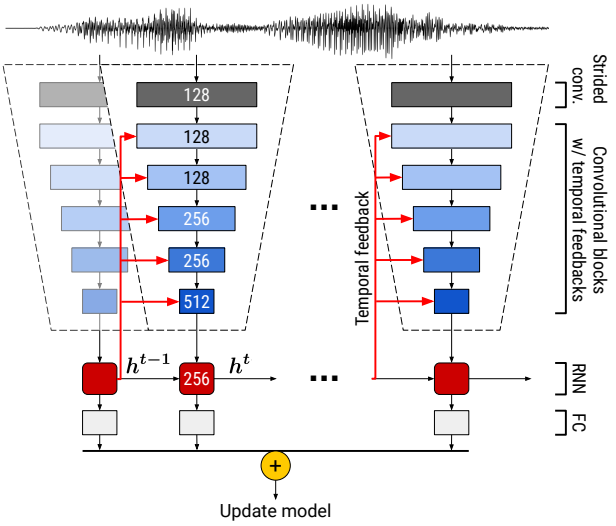

Figure 1. The architecture of temporal feedback CRNNs (TF-CRNNs). The number of convolutional filters are denoted in the boxes.

3. EXPERIMENTAL SETUP

3.1 Dataset

A medium size split of FMA dataset [2] is used for music genre classification, which contains 84,337 / 10,957 / 11,262 (training / validation / test split) tracks in total 25,000 tracks of a length of 30 seconds and 16 genres which are used as target classes.

3.2 Implementation details

Raw waveforms are directly used as inputs for the models. An audio track is divided into 2.5 second segments for training the models. In the test phase, the model predictions are summarized over the track length. All filters and max pooling layers have a size of 3. Gated Recurrent Units (GRUs) [1] are used to implement RNNs of which weights are initialized with zeros. The models are trained using stochastic gradient descent with Nesterov momentum of 0.9 with a batch size of 23.

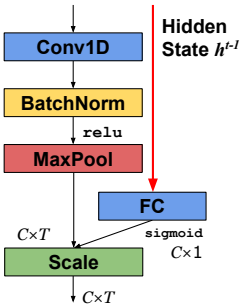

Figure 2. A convolutional block of TF-CRNNs taking the temporal feedback as an input. The feedback signal is used for channel-wise feature scaling. C and T indicate channel and time dimensionality, respectively.

Table 1. Performances of models on music genre classification. The results are averages of 3 experiments. Standard deviations are denoted in parentheses.

Model	Accuracy
CRNN	0.6078 (0.0101)
TF-CRNN	0.6117 (0.0060)
SampleCNN [3]	**0.6314 (0.0119)**
HC-features+SVM [2]	0.6300 (–)

4. RESULTS

We evaluate CRNNs without temporal feedbacks and TF-CRNNs on music genre classification and compare them with other approaches. Table 1 summarizes the results. The temporal feebacks improve the performance of CRNNs on music genre classification as same as on keyword spotting. However, TF-CRNNs could achieve the state-of-the-art performances on keyword spotting in our previous study, but not on music genre classification.

5. ACKNOWLEDGMENTS

This research was supported by BK21 Plus Postgraduate Organization for Content Science (or BK21 Plus Program) of Korea.

6. REFERENCES

[1] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[2] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. Fma: A dataset for music analysis. In *International Society of Music Information Retrieval (ISMIR)*, 2017.

[3] Taejun Kim, Jongpil Lee, and Juhan Nam. Comparison and analysis of SampleCNN architectures for audio classification. *Journal of Selected Topics in Signal Processing*, 13(2):285–297, 2019.

[4] Taejun Kim and Juhan Nam. Temporal feedback convolutional recurrent neural networks for keyword spotting. *arXiv preprint (to be uploaded)*, 2019.

[5] Richard F. Lyon. *Human and Machine Hearing: Extracting Meaning from Sound*. Cambridge University Press, New York, NY, USA, 1st edition, 2017.