AUDIO CLASSIFICATION TASKS USING RECURRENT NEURAL NETWORK

Guangxiao Song, Shenyi Ding, Zhijie Wang
College of Information Science and Technology, Donghua University
guangxiaosong@hotmail.com

ABSTRACT

In our submission of MIREX 2018, we propose a model using Recurrent Neural Network (RNN) and scattering transform to tackle with music genre/mood classification tasks. In order to get a balance between information integrity and feature extraction in preprocessing phase, we employ the scattering transform. Then, multi-layer RNN is used to extract higher-level features from the scattering coefficients. Softmax classifier is used for the final classification.

1. INTRODUCTION

As a powerful and popular learning method, deep learning has successfully been applied in computer vision, speech recognition and natural language processing (NLP) in recent years. The primary reason of these successes is that deep learning related algorithms can automatically extract high-level features relevant to certain tasks from raw data or processed data. Researchers also have applied deep learning to music classification task with different preprocessing methods. Nam [1] uses unsupervised learning on bag-of-features to initialize a generative stochastic neural network (restricted Boltzmann machine, RBM), then finetune the neural network with musical tags. Convolutional Neural Network (CNN) gains a lot of success in recognition tasks, such as image classification and speech recognition. Base on inspiration of its outstanding performance in feature extraction, Choi [2] uses deep full convolutional neural networks (FCNs) with mel-spectrogram inputs to deal with music auto-tagging task. Contrast to CNN, which learns high-level features layer by layer from static data, Recurrent Neural Network (RNN) can learn correlations through different time steps well, especially from sequential data. Musical data are sequential and different kinds of tags need various time scales. Specifically, instrument (guitar, strings, piano) is at scale of milliseconds and the rhythm, genre (classic, rock, pop) of music is at the scale of seconds and musical mood (slow, quiet, soft) needs longer. Therefore, learning long term correlations through time in musical data is important for music genre/mood classification task. This suggests that RNN architecture is a suitable for music genre/mood classification potentially [3, 4].

However, straightforwardly feeding raw musical data to RNN is impracticable because of the limitation of current hardware. To exploit the advantages of deep learning algorithms, musical data have to be shrunk by preliminary feature extraction. This preprocessing should be moderate, and retain useful information as much as possible in order that deep learning algorithms can extract features further contrast to traditional machine learning algorithms, whose performance depends on extracted features heavily. For more compatibly combining the preprocessing with deep learning algorithms, we use scattering transform to reduce the size of musical data. This method not only retains the stability but also recovers the information lost by a mel-frequency averaging with modulus operators and wavelet decompositions [5]. Furthermore, Long Short Term Memory (LSTM), a structure to manipulate the hidden states of RNN, can deal with long-term relationships which is necessary for classification task.

2. PROPOSED ARCHITECTURE

Figure 1 shows the overall structure of the proposed method to tackle music genre/mood classification tasks. Our architecture consists of two parts. One is machine learning part using multilayer RNNs with LSTM, because RNN is suitable and powerful algorithm for sequential data, meanwhile, LSTM is a gated unit based structure which can learn long-term relationships by training process. And multilayer structure can improve the feature extraction and learning capacity further. In order to develop the ability of feature extraction as much as possible, we use scattering transform as preprocessing stage. This transform can extract useful features from raw data and recovery the information loss in feature extraction operation at the same time. It makes the raw data decrease to an appropriate size and retain abundant information for deep neural networks.

3. RECURRENT NEURAL NETWORK

The concept of gated units in RNNs is introduced firstly in [6] to overcome the vanishing gradients problems, named Long-Short Term Memory. The vanishing gradients are alleviated by allowing the network to conserve its memory over quantity of time steps during both forward

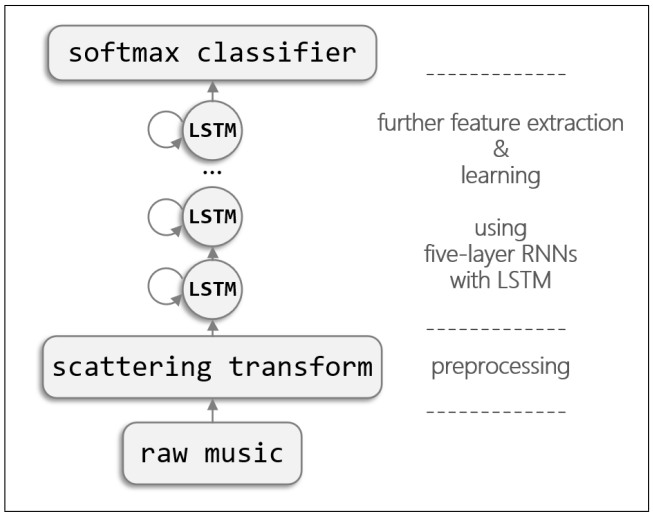

Figure 1. Overall structure of the proposed architecture.

and backward phase. LSTM can be formatted by following equations:

$$\begin{pmatrix} i_t \\ f_t \\ o_t \\ g_t \end{pmatrix} = \begin{pmatrix} \sigma(W_i[x_t, h_{t-1}] + b_i) \\ \sigma(W_f[x_t, h_{t-1}] + b_f) \\ \sigma(W_o[x_t, h_{t-1}] + b_o) \\ f(W_g[x_t, h_{t-1}] + b_g) \end{pmatrix} \quad (1)$$
$$c_t = f_t * c_{t-1} + i_t * g_t$$
$$h_t = o_t * f(c_t)$$

where i_t, o_t and f_t are input, output and forget gates at time step t respectively. σ and f are sigmoid and hyperbolic tangent activation function. g_t is cell updates vector and cell vector c_t is used to update the hidden state h_t. "$*$" represents the elementwise multiplication.

We use 5-layer LSTM neural network which is constructed by stacking each hidden layer on the top of previous layer, in order to improve the ability of representation of our architecture in this paper. Additionally, generalization of the proposed deep RNN is improved by applying dropout between each layer.

4. SCATTERING TRANSFORM

Scattering transform is chosen as the preprocessing method. It produces the preliminary or low-level features that beneficial to classification. The main reason of the benefits is that scattering transform yields local translation invariant and stable representation while avoiding information loss as much as possible when the time scale is large ($> 500ms$) [7].

A time average operation $S_0x(t) = x * \phi(t)$ is applied to signal x (x is the data of one frame extracted from the entire music signal) that brings out locally translation invariant but high frequency contents lost. Fortunately, this information loss can be recovered by a wavelet modulus transform $|x * \psi_\lambda(t)|$, where the index λ is an integer that represents the scale of the wavelet, guaranteed by the number of wavelets per octave Q, and Λ is the maximum value of λ. The first order of scattering transform is obtained by convoluting $|x * \psi_{\lambda_1}|$ with the low-pass filter $\phi(t)$, $S_1x(t, \lambda_1) = |x * \psi_{\lambda_1}| * \phi(t)$.

Iteratively averaging and recovering the results of last order in a cascade way composes deep scattering transform. In the proposed model, we use two orders of scattering transform, and the second is defined as $S_2x(t, \lambda_1, \lambda_2) = ||x * \psi_{\lambda_1}| * \psi_{\lambda_2}| * \phi(t)$.

5. SETTINGS

We use 512 hidden states in each layer. Dropout is set as 0.75. Learning rate is 0.00001. ScatNet matlab toolbox and tensorflow are used to implement scattering transform and RNN in our submission.

6. REFERENCES

[1] J. Nam, J. Herrera, K. Lee, A deep bag-of-features model for music auto-tagging, arXiv preprint arXiv:1508.04999.

[2] K. Choi, G. Fazekas, M. Sandler, Automatic tagging using deep convolutional neural networks, arXiv preprint arXiv:1606.00298.

[3] Song G, Wang Z, Han F, et al. Transfer Learning for Music Genre Classification. International Conference on Intelligence Science 2017:183-190.

[4] Song G, Wang Z, Han F, et al. Music Auto-Tagging Using Deep Recurrent Neural Networks. Neurocomputing, 2018, 292.

[5] S. Mallat, Group invariant scattering, Communications on Pure and Applied Mathematics 65 (10) (2012) 1331–1398.

[6] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780.

[7] J. Andén, S. Mallat, Deep scattering spectrum, IEEE Transactions on Signal Processing 62 (16) (2014) 4114–4128.