Neurocomputing 000 (2018) 1–7

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

# Music auto-tagging using deep Recurrent Neural Networks

Guangxiao Song, Zhijie Wang*, Fang Han*, Shenyi Ding, Muhammad Ather Iqbal

*College of Information Science and Technology, Donghua University, Shanghai 201620, China*

ARTICLE INFO

*Article history:*
Received 14 June 2017
Revised 10 February 2018
Accepted 15 February 2018
Available online xxx

Communicated by Dr Zhiyong Wang

*Keywords:*
Music auto-tagging
Deep learning
Music information retrieval
Recurrent Neural Network

ABSTRACT

Musical tags are used to describe music and are cruxes of music information retrieval. Existing methods for music auto-tagging usually consist of preprocessing phase (feature extraction) and machine learning phase. However, the preprocessing phase of most existing method is suffered either information loss or non-sufficient features, while the machine learning phase depends on heavily the feature extracted in the preprocessing phase, lacking the ability to make use of information. To solve this problem, we propose a content-based automatic tagging algorithm using deep Recurrent Neural Network (RNN) with scattering transformed inputs in this paper. Acting as the first phase, scattering transform extracts features from the raw data, meanwhile retains much more information than traditional methods such as mel-frequency cepstral coefficient (MFCC) and mel-frequency spectrogram. Five-layer RNNs with Gated Recurrent Unit (GRU) and sigmoid output layer are used as the second phase of our algorithm, which are extremely powerful machine learning tools capable of making full use of data fed to them. To evaluate the performance of the architecture, we experiment on Magnatagatune dataset using the measurement of the area under the ROC-curve (AUC-ROC). Experimental results show that the tagging performance can be boosted by the proposed method compared with the state-of-the-art models. Additionally, our architecture results in faster training speed and less memory usage.



## 1. Introduction

In music information retrieval (MIR) area, musical tags are important for artist identification, genre classification or other purposes. Tags represent the high-level information of each music clip, such as instrument (piano, guitar, strings), mood (quiet, soft, weird), genre (classical, rock, jazz) and so on. In the past, they are collected manually by musicians or some music fans. In order to save both time and labor, automatic tagging techniques, namely content-based MIR techniques, have been researched and developed [1].

MIR techniques consist of two phases, preprocessing (feature extraction) and machine learning phase. It is desired that preprocessing phase maintains a delicate balance between feature extraction and information integrity, and that machine learning phase makes use of more information as possible. Most of these existing MIR techniques use traditional machine learning methods such as support vector machine, random forest and decision tree [2–4] to gain complicated relationship from original musical signals to abstract tags. But these machine learning algorithms do not have the capacity of feature extraction. Their performances heavily depend on the performance of the time-consuming features carried out in the preprocessing phase. For example, mel-frequency cepstral coefficient (MFCC) which is widely used in audio processing, is efficient extracting way when using short time scales. When applied to traditional machine learning algorithms, features of short time scales do not perform well in musical tagging task. So, it is necessary to enlarge the scale to make it more suitable for music tagging applications. Another popular transformation way of musical data is mel-frequency spectrogram. It can achieve the purpose of larger scale with stability to time-warping deformation but losing information which has significant influence on the machine learning phase [5–8].

As a powerful and popular learning method, deep learning has successfully been applied in computer vision [9–13], speech recognition [14–16], audio recognition [17] and natural language processing (NLP) [18–20] 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 auto-tagging task with different preprocessing methods. In [21], a feedforward artificial neural network (multilayer perceptron, MLP) is used for classification. Nam et al. [22] uses unsupervised learning on bag-of-features to initialize a generative stochastic neural network (restricted Boltzmann machine, RBM), then fine-tune the neural network with musical tags. Convolutional

* Corresponding authors.
*E-mail addresses:* wangzj@dhu.edu.cn (Z. Wang), yadiahan@163.com (F. Han).

https://doi.org/10.1016/j.neucom.2018.02.076
0925-2312/

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 et al. [23] 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 [24,25], especially from sequential data. Musical data are sequential and different kinds of tags need various time scales. Specifically, instrument (guitar, strings, piano) is at the 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 short and long term correlations through time in musical data is important for auto-tagging task. This suggests that RNN architecture is a suitable for musical tagging task potentially.

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 in this paper. This method not only retains the stability but also recovers the information lost by a mel-frequency averaging with modulus operators and wavelet decompositions [26]. We believe these advantages can maximize the capacity of deep learning algorithms, albeit the combination of scattering transform and deep learning is rare up to date. Furthermore, Gated Recurrent Unit (GRU), a structure to manipulate the hidden states of RNN, can deal with long-term relationships which is necessary for auto-tagging task.

Base on the discussions above, we propose an architecture combines scattering transform with five-layer RNNs using GRU [27] in this paper. We use sigmoid output layer for the last RNN layer, and binary cross-entropy loss to compute the objective of the training. The test result achieves a competitive score of the area under the ROC-curve (AUC-ROC) on Magnatagatune dataset, which is higher than the state-of-the-art models.

During the experiments, with the intention of finding the best RNN unit and the numbers of layers and hidden states, Long Short-Term Memory (LSTM) [28], Batch-Normalized LSTM (BN-LSTM) [29] units using different hyperparameters have been evaluated as well. For better comparison, CNN with scattering coefficients has been tested. And the proposed algorithm converges to a quite high accuracy rapidly at early epoch and has more efficient training process because of fewer trainable parameters, relative to CNN models.

## 2. Proposed architecture

Fig. 1 shows the overall structure of the proposed method to tackle music auto-tagging task. Deep learning techniques are often used for automatically feature extraction. But in MIR, raw music data are too huge to existent processors if used as inputs for deep neural networks directly. Therefore, our architecture consists of two parts. One is machine learning part using multilayer RNNs with GRU, because RNN is suitable and powerful algorithm for sequential data, meanwhile, GRU is a variation of gated unit structure which can learn long-term relationships by training process. And multilayer structure can improve the feature extraction and learning capacity further. Considering the situation of the existing hardware, musical data should be reduced. 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 retains abundant information for deep neural networks. In the next two sections, we describe the scattering transform and multilayer RNN architecture in detail.

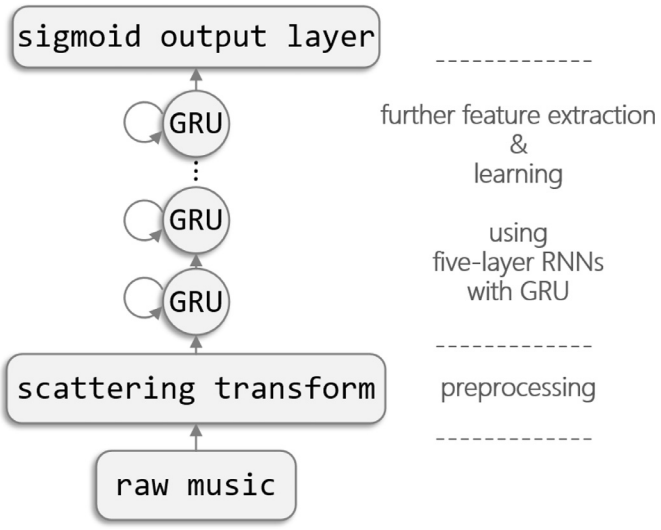


**Fig. 1.** Overall structure of the proposed architecture.

## 3. Scattering transform

Choosing scattering transform as preprocessing part is because it can enlarge time scale and recover the information loss, which commonly used preprocessing methods in MIR, such as the mel-frequency spectrogram and MFCC lack. We first describe the calculation process of two-order scattering transform used in our architecture. Then introduce the scattering transform normalizer, in order to reduce redundancy and increase the invariance of scattering coefficients.

The multiplicity of musical information at different time scales is the major difficulty in auto-tagging task. Instrument is at the scale of milliseconds, genre is at the scale of seconds and musical mood needs longer scale. Existing effective and popular used methods in MIR are MFCC and mel-spectrogram. MFCC is efficient at time scales up to 25 ms. And mel-spectrogram can enlarge the scale but cause information loss. The lost information are quite important for many audio applications. So mel-spectrogram is often calculated over small time windows about 25 ms. To enlarge the time scale without too much information loss, we adopt scattering transform as our preprocessing method.

For an audio signal $x$, Andén and Mallat [5] show that mel-spectrogram coefficients are approximately equal to averaged squared wavelet coefficients $|x \star \psi_{\lambda_1}|^2 \star |\phi|^2(t)$, where $\psi_{\lambda_1}$ is a bandpass filter of wavelets to handle with high bandwidth frequencies, $\phi(t)$ is a lowpass filter to handle with low frequencies, and $\star$ is convolution operation as it in discrete wavelet transform (DWT). To avoid amplifying the outliers among these coefficients by the square operator, the square is removed and $|x \star \psi_{\lambda_1}| \star \phi(t)$ is computed instead. The information loss emerges here due to the application of the time averaging filter $\phi(t)$. To solve this information loss problem, we recover the lost information by using a modulus of wavelet transform $\psi_{\lambda_2}$, formalized as $||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|$. The previous wavelet modulus coefficients averaged by the lowpass fil-

Please cite this article as: G. Song et al., Music auto-tagging using deep Recurrent Neural Networks, Neurocomputing (2018), https://doi.org/10.1016/j.neucom.2018.02.076