Transfer Learning for Music Genre Classification

Guangxiao Song, Zhijie Wang(✉), Fang Han(✉), and Shenyi Ding

College of Information Science and Technology, Donghua University, Shanghai 201620, China
wangzj@dhu.edu.cn, yadiahan@163.com

Abstract. Modern music information retrieval system provides high-level features (genre, instrument, mood and so on) for searching and recommending conveniently. Among these music tags, genre is the most widely used in practice. Machine learning technique has the ability of cataloguing different genres from raw music. A disadvantage of it is that the final performance heavily depends on the used features. As a powerful learning algorithm, deep neural network can extract useful features automatically and effectively instead of time-consuming feature engineering. But deeper architecture means larger data are needed to train the neural network. In many cases, we may not have enough data to train a deep network. Transfer learning solves the problem by pre-training the network in a similar task which has enough data, then fine-tuning the parameters of the pre-trained network using the target dataset. Magnatagatune dataset is used for pre-training the proposed five-layer Recurrent Neural Network (RNN) with Gated Recurrent Unit (GRU). And in order to reduce the input of the network, scattering transform is used in this paper. Then GTZAN dataset is used as the target dataset of genre classification. Experimental results show the transfer learning way can achieve a higher average classification accuracy (95.8%) than the same deep RNN which initials the parameters randomly (93.5%). In addition, the deep RNN using transfer learning converges to the final accuracy faster than using random initialization.

Keywords: Music genre classification · Transfer learning · Deep learning

1 Introduction

Music genre is important to many applications, such as music recommender system and information retrieval. Automatic genre classification system has been developed using machine learning technique recent years. Most of these systems have the ability of cataloguing different music genres from raw music contents [1–3].

Mel-frequency cepstral coefficient (MFCC) and Mel-spectrogram are widely used in genre classification task. Because they can extract variant features from raw data for the learning process. But the performance of genre classification

Z. Shi et al. (Eds.): ICIS 2017, IFIP AICT 510, pp. 183–190, 2017.
DOI: 10.1007/978-3-319-68121-4_19

benefits from features over long-time scale (>500 ms) while MFCC is efficient around time scale of 25 ms, and enlarging the time scale leads the information loss when using mel-spectrogram [4,5]. Differently, scattering transform can recover the information loss by wavelet decompositions, meanwhile, extract long-time scale features by lowpass filters [6,7].

Deep learning makes massive of success in different areas, for instance, computer vision [8–10], speech recognition [11,12], and natural language processing [13,14]. These algorithms can extract high-level features automatically layer by layer, different from traditional machine learning classifiers, such as Support Vector Machine (SVM), Nearest Neighbors, and Decision Trees, which are heavily dependent on the result of feature extraction. Among its several typical models, Recurrent Neural Network (RNN) is widely used for sequential data. And RNN is good at learning the relationship through time [15]. But in purpose of achieving good performance, deep neural network needs large amount of data. In condition of the target dataset need to be classified is not enough, we can use a large data, which is the same or similar to the target dataset, to pre-train the deep neural network, then replace the connections to classifier according to the target classification number and fine-tune the parameters of the pre-trained network. This process is called transfer learning [16]. In this paper, we use Magnatagatune dataset [17] and GTZAN dataset [18] as the large and the target dataset respectively. 5-layer RNN using Gated Recurrent Unit (GRU) [19] and softmax classifier are used. Additionally, for reducing the input of deep RNN, we use scattering transform as its preprocessing.

The results of the experiment show that the proposed 5-layer RNN reaches a high accuracy when using transfer learning, and the same architecture using random initialization converges more slowly to a lower accuracy.

2 Transfer Learning Process

The architecture of the proposed method is shown in Fig. 1. The overall process consists of two parts. One part is deep RNN training on a large musical dataset

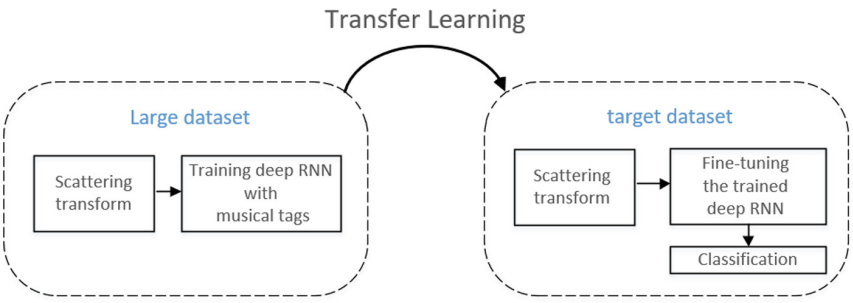

Fig. 1. The architecture of the proposed transfer learning process

(Magnatagatune dataset is used in this paper). The other part is genre classification process after fine-tuning the previous trained deep RNN by target dataset (GTZAN dataset is used in this paper). Specifically, scattering transform is applied at the beginning of each part, in order to reduce the raw music data and to extract features preliminarily for the next process of neural network training. 5-layer RNN with GRU and softmax classifier are trained with tagged music clips as the deep RNN we mentioned. At last, we use the target genre classification dataset (GTZAN) to fine-tune the trained parameters of RNN.

2.1 Scattering Transform

In genre classification task, large time scale (>500 ms) invariant signal representation is important. As widely used methods in audio processing, mel-spectrogram can enlarge the time scale but remove information which is crucial to genre classification. And MFCC is efficient at time scales up to 25 ms. Unlike the previous methods. Scattering transform can provide invariants over large time scales without too much information loss.

For an audio signal x, scattering transform defined as $S_n x$, where n represent the order. $S_0 x = x \star \phi(t)$ has locally invariant property because of the time averaging operation, but it leads to high frequency information loss which can be retrieved by the wavelet modulus coefficients $|x \star \psi_{\lambda_1}(t)|$. To make the wavelet modulus coefficients invariant to translation, a time averaging unit is applied. The first layer of scattering transform defined as:

$$S_1 x(t, \lambda_1) = |x \star \psi_{\lambda_1}| \star \phi(t) \tag{1}$$

Andén [7] indicates that if wavelets filter-bank ψ_{λ_1} have the same frequency resolution as the mel-windows, then $S_1 x$ coefficients can be approximate to the mel-filter-banks coefficients. The difference is that applying a bank of higher frequency wavelet filters ψ_{λ_2} with a modulus to the wavelet modulus coefficients can recover the lost information. The same as previous operation, adding a low-pass filter $\phi(t)$ make the coefficients translation invariant. Then the second layer of scattering transform defined as:

$$S_2 x(t, \lambda_1, \lambda_2) = ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi(t) \tag{2}$$

2.2 Deep Recurrent Neural Network

RNNs have an aptitude for handling sequential information, such as speech recognition and NLP. RNN structure can be described as transitions from previous to current states. For classical RNN, this transition is formulized as:

$$h_t = f(W_h[x_t, h_{t-1}] + b_h) \tag{3}$$

In order to solve the problem of vanishing gradients of RNN. Gated structure named LSTM introduced by Hochreiter [15]. The LSTM unit allows that information of more timesteps can be memorized. And the memories are stored by memory cells. Then the LSTM can decide to forget, output, or change the saved memories. As a popular variant of LSTM, GRU is simpler and effective as well. It uses gate Zt and gate Rt to update the hidden state. Theses gates are given by:

$$\begin{aligned} \begin{pmatrix} z_t \\ r_t \end{pmatrix} &= \begin{pmatrix} \sigma(W_z[x_t, h_{t-1}] + b_z) \\ \sigma(W_r[x_t, h_{t-1}] + b_r) \end{pmatrix} \\ g_t &= f(W_g[x_t, r_t * h_{t-1}] + b_g) \\ h_t &= (1 - z_t) * h_{t-1} + z_t * g_t \end{aligned} \tag{4}$$

We use 5-layer GRU 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 [20].

3 Datasets and Experiment Setup

Magnatagatune and GTZAN dataset are used as the large and target dataset respectively. All the clips are transformed to mono and sampled by 16 kHZ. Magnatagatune has 25863 clips and each clip is annotated with 188 different musical tags such as genre, mood, and instrument. We use the last 2105 clips (distributed in folder 'f') for validation, others for training. We use 512 hidden states in each layer. Dropout is set as 0.7. Learning rate is 0.00001. And we use AUC-ROC score [21] to evaluate the performance of our model to avoid imbalance of the dataset. When the AUC-ROC score is stable, we stop the training and save the model. GTZAN dataset has 1000 clips of 10 genres and each genre contains 100 clips evenly. As the target dataset, it is randomly shuffled and the mean accuracy of 10 times of 10-fold cross validation is used for the final test accuracy. Among the 10 folds in total, we use 1 fold for testing, and the others for training. Each time of 10-fold cross validation, we change the output number of the softmax classifier to 10 (the genre number of GTZAN dataset), then fine-tune the parameters of pre-trained model from Magnatagatune dataset.

4 Experiment Results and Analysis

As shown in Fig. 3, both random initialization and transfer learning models (pre-training process is shown in Fig. 2) of 5-layer RNN with GRU using scattering transform preprocessing converge to quite high accuracy in training. And the models using transfer learning need about 100 epochs to be stable. But the

random initialed models need more. This phenomenon not only appears in the three random picked training processes, but also in the unpicked to be shown. It indicates that the transfer learning initials the model better, and improves the speed of convergence.

Comparing with other works of recent years in Table 1, our approach shows a competitive accuracy (95.8%) in genre classification task on GTZAN dataset. Even the model using random initialization can also reach a high accuracy (93.5%) relatively. The combination of scattering transform and deep RNN has been evaluated, and by using this architecture, it performs well in music genre classification.

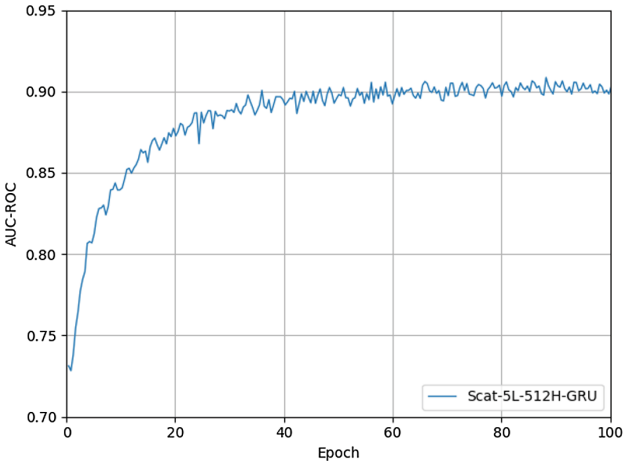

Fig. 2. Validation AUC-ROC score of 5-layer GRU neural network using scattering transformed input

Table 1. Average test accuracy of different models on GTZAN dataset

Model	Average test accuracy
Panagakis and Kotropoulos [22]	92.4%
Andén and Mallat [7]	91.6%
Lee et al. [5]	90.6%
Dai and Liu [23]	93.4%
Random initialization	93.5%
Transfer learning	**95.8%**

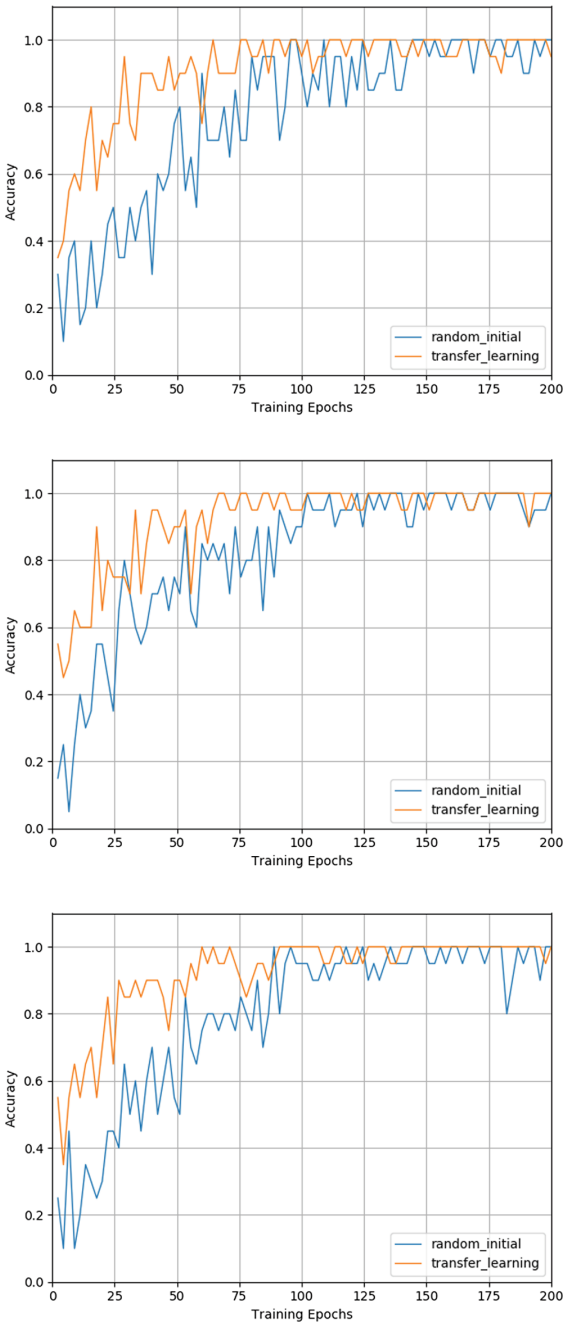

Fig. 3. Three random picked training processes of 10-cross validation, Blue lines represent the RNN using random initialization, and the orange lines represent the RNN using transfer learning. And the accuracy is tested by a random batch of training data (Color figure online)

5 Conclusion

In this paper, we use transfer learning in music genre classification by using 5-layer RNN with GRU and scattering coefficients as its input. When applying the transfer learning from a large music dataset (Magnatagatune is used in this paper), our model shows a faster convergence and higher average accuracy than the same model of random initialization on the target dataset (GTZAN is used in this paper). And the accuracy of transfer learning approach is competitive comparing with the state-of-the-art models as well. The effectiveness of deep RNN combined with scattering transform and transfer learning has been verified in music genre classification task.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (Grants nos. 11572084, 11472061, 71371046), the Fundamental Research Funds for the Central Universities and DHU Distinguished Young Professor Program (No. 16D210404).

References

1. Song, Y., Zhang, C.: Content-based information fusion for semi-supervised music genre classification. IEEE Trans. Multimedia **10**(1), 145–152 (2008). doi:10.1109/tmm.2007.911305
2. Meng, A., Ahrendt, P., Larsen, J., Hansen, L.K.: Temporal feature integration for music genre classification. IEEE Trans. Audio Speech Lang. Process. **15**(5), 1654–1664 (2007). doi:10.1109/tasl.2007.899293
3. Lampropoulos, A.S., Lampropoulou, P.S., Tsihrintzis, G.A.: Music genre classification based on ensemble of signals produced by source separation methods. Intell. Decis. Technol. **4**(3), 229–237 (2010). doi:10.3233/idt-2010-0083
4. McDermott, J.H., Simoncelli, E.P.: Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. Neuron **71**(5), 926–940 (2011). doi:10.1016/j.neuron.2011.06.032
5. Lee, C.H., Shih, J.L., Yu, K.M., Lin, H.S.: Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features. IEEE Trans. Multimedia **11**(4), 670–682 (2009)
6. Mallat, S.: Group invariant scattering. Commun. Pure Appl. Math. **65**(10), 1331–1398 (2012). doi:10.1002/cpa.21413
7. Andén, J., Mallat, S.: Deep scattering spectrum. IEEE Trans. Signal Process. **62**(16), 4114–4128 (2014). doi:10.1109/tsp.2014.2326991
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012). doi:10.1145/3065386
9. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. doi:10.1109/cvpr.2016.90 (2016)
11. Mohamed, A., Dahl, G.E., Hinton, G.: Acoustic modeling using deep belief networks. IEEE Trans. Audio Speech Lang. Process. **20**(1), 14–22 (2012). doi:10.1109/tasl.2011.2109382

12. Feng, X., Zhang, Y., Glass, J.: Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1759–1763. IEEE (2014). doi:10.1109/icassp.2014.6853900
13. Sutskever, I., Martens, J., Hinton, G.E.: Generating text with recurrent neural networks. In: Proceedings of the 28th International Conference on Machine Learning (ICML 2011), pp. 1017–1024 (2011)
14. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
15. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997). doi:10.1162/neco.1997.9.8.1735
16. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2010). doi:10.1109/tkde.2009.191
17. Berenzweig, A., Logan, B., Ellis, D.P., Whitman, B.: A large-scale evaluation of acoustic and subjective music-similarity measures. Comput. Music J. **28**(2), 63–76 (2004). doi:10.1162/014892604323112257
18. Tzanetakis, G., Cook, P.: Musical genre classification of audio signals. IEEE Trans. Speech Audio Process. **10**(5), 293–302 (2002). doi:10.1109/tsa.2002.800560
19. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
20. Zaremba, W., Sutskever, I., Vinyals, O.: Recurrent neural network regularization. arXiv preprint arXiv:1409.2329 (2014)
21. Davis, J., Goadrich, M.: The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 233–240. ACM (2006). doi:10.1145/1143844.1143874
22. Panagakis, Y., Kotropoulos, C., Arce, G.R.: Music genre classification using locality preserving non-negative tensor factorization and sparse representations. In: ISMIR, pp. 249–254 (2009)
23. Dai, J., Liu, W., Ni, C., Dong, L., Yang, H.: "Multilingual" deep neural network for music genre classification. In: Sixteenth Annual Conference of the International Speech Communication Association (2015)