Music Genre Classification Using Transfer Learning

Beici Liang
QQ Music, Tencent Music Entertainment
Shenzhen, China
beiciliang@tencent.com

Minwei Gu
QQ Music, Tencent Music Entertainment
Shenzhen, China
torogu@tencent.com

***Abstract*—We proposed a transfer learning approach for audio-based classification of 11 western music genres, including Rock, Pop, Rap, Country, Folk, Metal, Jazz, Blues, R&B, Electronic Music and Classical Music. Multiple models were investigated. The best one can achieve 0.9799 ROC-AUC and 0.8938 PR-AUC on a private dataset of 1100 songs.**

***Keywords*-music genre classification; transfer learning;**

I. Introduction

Music tags can present high-level information for music contents. They are commonly related to genre, mood, instrumentation and so on. Music auto-tagging is therefore considered as a combination of multiple tasks such as genre classification and instrument recognition. Convolutional Neural Networks (CNN) have been widely used to learn the relationship between tags and the audio content. An effective trained-model can automatically provide content-based annotations and be useful for applications in online streaming services or music management tools.

In this paper, our focus are tags associated to western music genres. Given that state-of-the-art auto-tagging models focus on top-50 tags of a dataset, they are not able to detect music with tags that are out of the top-50, for instance, "classical music" in the Million Song Dataset[1] (MSD), and "blues" in the MagnaTagATune Dataset[2] (MTT). To solve this, we propose a transfer learning approach. The knowledge gained from the pre-trained music auto-tagging models can be applied to the target task of music genre classification. We demonstrate the effectiveness of our method in 1100 audio recordings consisting of 11 genres: Rock, Pop, Rap, Country, Folk, Metal, Jazz, Blues, R&B, Electronic Music and Classical Music[3].

II. Methods

Transfer learning consists of a source task and a target task. The neural network trained in the source task can be reused in the target task after adapting the network to a more specific dataset. This has been gaining more attentions in music informatics research for alleviating the data sparsity problem. Transfer learning also has the ability to be used for different classification and regression tasks [1]. For instance, Liang et al. [2] obtained features from pre-trained CNNs, which were trained using a large synthesized dataset in the source task. These features were used in the target task of pedal-on/off classification from acoustic piano recordings.

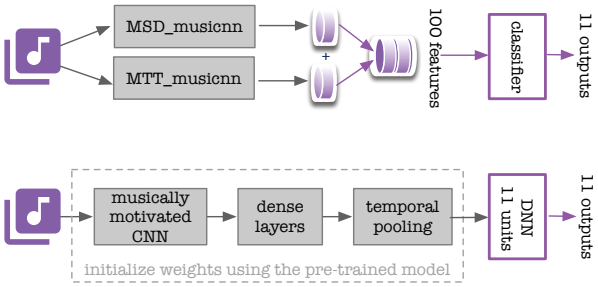

Figure 1: Schematic of the two proposed methods in the target task of transfer learning.

A. Source Task

Music auto-tagging was used as the source task. The *muscinn* library [3] contains a set of pre-trained musically motivated CNN for audio tagging. We used the following available models: `MTT_musicnn`, `MSD_musicnn`, and `MSD_musicnn_big`. The `MTT_musicnn` was trained with the MagnaTagATune dataset, while the `MSD` models were trained with the Million Song Dataset. It is noted that `MSD_musicnn_big` is a larger model with more capacity[4].

The above three models have similar architectures. However, two different 50-tags vocabularies were considered because two different datasets were used for training. These models achieve state-of-the-art performance on both datasets. More details on the training process of these models can be found in [4].

B. Target Task

To adapt the pre-trained models in the target task, we propose two methods as illustrated in Figure 1. One is to use two pre-trained models (`MTT_musicnn` and `MSD_musicnn`)

[1]http://millionsongdataset.com/

[2]http://mirg.city.ac.uk/codeapps/the-magnatagatune-dataset

[3]Due to the copyright issue, we are not able to make this private dataset public online.

[4]By the time we conducted our experiments, this model obtained the best performance in music auto-tagging task.

978-1-7281-4272-2/20/$31.00 ©2020 IEEE
DOI 10.1109/MIPR49039.2020.00085

Authorized licensed use limited to: PES University Bengaluru. Downloaded on June 23,2023 at 06:05:14 UTC from IEEE Xplore. Restrictions apply.

Table I: Performance of different models in the target task.

Model	ROC-AUC	PR-AUC
Features+KNN	0.6385	0.1878
CNN_MSD	0.9782	0.8831
CNN_MSD_big	**0.9799**	**0.8938**

as feature extractors. After concatenating the corresponding 50-dimensional output from each model, a feature vector of 100 dimensions is generated and classified into 11 genres using a machine learning model. This is considered as the baseline method in the following experiments.

The other proposed method is to change the final dense layer of a pre-trained model (MSD_musicnn or MSD_musicnn_big) in order to output probabilities of 11 genres instead of 50 tags. Retraining this model uses weights from the selected pre-trained model during initialization.

III. Experiments

A. Dataset and Setup

For each genre, there are 100 commercial songs in mp3 format, 75% of which were used for training and 25% for evaluation. To efficiently compare different models in the target task, we randomly sampled four segments from every song for each epoch. Here each segment has a duration of 3 seconds. There are 150 epochs in total. Only the model with the minimum evaluation cost was saved as the best model.

For the baseline method, k-nearest neighbors algorithm (KNN) was selected as the machine learning model for classification. Its parameters were optimized using grid-search. Other configurations related to the CNN models were kept the same as the ones used in [4].

B. Experimental Results

We denoted the baseline method as Features+KNN, and the two models in the other method as CNN_MSD and CNN_MSD_big, respectively. To compare the proposed methods, two sets of performance measurements are used: ROC-AUC and PR-AUC[5]. Best results based on the evaluation set are presented in Table I, where 0.9799 ROC-AUC and 0.8938 PR-AUC are obtained using CNN_MSD_big.

Using the model with the highest performance measurements, i.e., CNN_MSD_big, we also conducted a song-level evaluation. To this end, several predictions are computed for a song (by a moving window of 3s). Final genre of the song is decided by majority voting. Confusion matrix from the song-level results is presented in Figure 2. We can obtain an overall accuracy of 0.8836.

IV. Conclusions

In this paper, we demonstrated the effectiveness of the proposed transfer learning method in western music genre classification. Pre-trained models were gained from the source task of music auto-tagging. They can be successfully adapted to the target task and achieve high performance measurements.

[5]ROC: Receiver Operating Characteristic. PR: Precision Recall. AUC: Area Under the Curve.

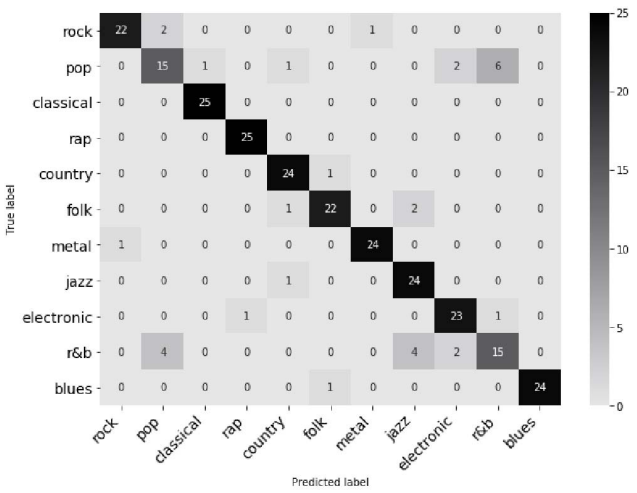

Figure 2: Confusion matrix from the song-level results.

However, we observe that Pop and R&B are easily misclassified. Pop is a broad genre that describes all music that is popular. Thus it is difficult to separate Pop from other genres. Likewise, R&B encompasses elements from Pop, Rap, Electronic Music and so on. It is possible to be misclassified into these genres. We assume that using a thresholding method instead of majority voting could alleviate the song-level misclassification problems.

Acknowledgment

We would like to thank our colleagues, Meiying Zhang, Weilong Wang and Yundong Sun for their contributions in the dataset preparation.

References

[1] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Transfer learning for music classification and regression tasks," in *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017, pp. 141–149.

[2] B. Liang, G. Fazekas, and M. Sandler, "Transfer learning for piano sustain-pedal detection," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019.

[3] J. Pons and X. Serra, "musicnn: Pre-trained convolutional neural networks for music audio tagging," *arXiv preprint arXiv:1909.06654*, 2019.

[4] J. Pons Puig, O. Nieto, M. Prockup, E. M. Schmidt, A. F. Ehmann, and X. Serra, "End-to-end learning for music audio tagging at scale," in *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, 2018, pp. 637–644.

Authorized licensed use limited to: PES University Bengaluru. Downloaded on June 23,2023 at 06:05:14 UTC from IEEE Xplore. Restrictions apply.