Improving CNN-based Music Tagging with Query2Label Transformer

Chengxi Yang
Shanghai Jiao Tong University
arcadia-y@sjtu.edu.cn

Abstract

Automatic music tagging, a multi-label classification task, aims to assign semantic tags to music pieces, encompassing genres, moods, and instruments. This task faces challenges such as label imbalance, subjectivity in tagging, and the multiple instance problem. Current convolutional neural network (CNN)-based approaches, though effective, suffer from slow convergence and suboptimal precision. To address these issues, we propose an enhanced music tagging model that combines a modified ResCNN backbone with a Query2Label (Q2L) Transformer framework. By utilizing learnable label embeddings as queries within the Transformer architecture, our model achieves superior performance on the MagnaTagATune dataset. Experimental results demonstrate that our approach significantly improves mean average precision and exhibits faster convergence compared to ResCNN with acceptable additional computational overhead. The source code will soon be available at `https://github.com/Arcadia-Y/q2l-music-tagging`.

1 Introduction

Automatic music tagging is a multi-label classification task, which aims to assign appropriate tags to a given piece of music. Tags cover a wide range of semantic information of music, including genre, mood, and instruments, and can be used for applications such as recommendation, curation, playlist generation, semantic search, and analysis of listening behavior [10].

Similar to other multi-label tasks, music tagging suffers from the label imbalance problem, as most tags have only a few songs as positive samples. Unlike multi-label image classification, music tagging is notably more subjective because different individuals may perceive the same song differently. Additionally, it faces the multiple instance problem; for instance, a song tagged with *guitar* does not imply that the guitar is present in every segment of the song. The lack of constraints on the content of tags further exacerbates the issue, resulting in tagging datasets that are extremely messy and noisy. These factors collectively make achieving high precision in music tagging challenging.

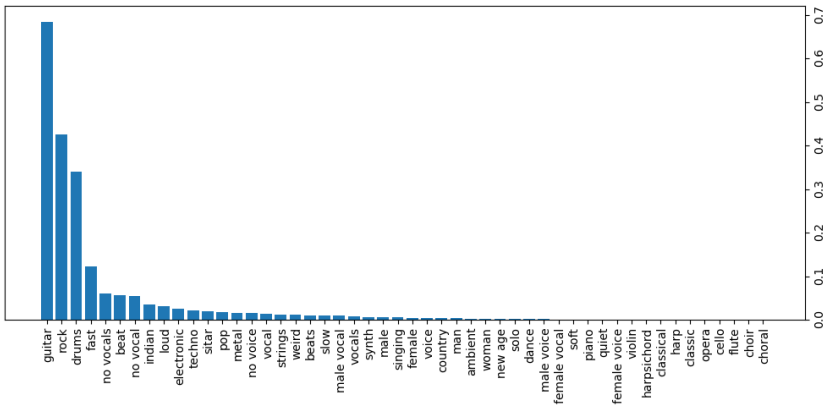

Figure 1: An example of music tagging: the output of our model for an instrumental segment of *Hey Jude* by the Beatles

A spectrogram visualizes the spectrum of frequencies of a signal as it varies with time, capturing essential information such as pitch and beats for a music signal. Features of music are reflected in the spacial features of its spectrogram. Intuitively, we can classify music by analyzing their spectrograms as images. Indeed, with the success of convolutional neural networks (CNNs) in image classification, many researchers have adopted CNNs on spectrograms for music tagging, and achieved state-of-the-art results in this task [9]. However, their precision is far from perfect and they have relatively slow convergence during training.

To address these challenges, we propose further transferring other methods from image classification to music tagging. Motivated by the success of the Query2Label (Q2L) [5] framework in image multi-label classification, we aim to improve the performance of CNN-based music tagging models by adopting a similar architecture. In particular, we modify a simple CNN with residual connections [3] (ResCNN) implemented in [9] and incorporate a Q2L transformer with learnable label embeddings as queries to predict the tags.

We train and evaluate our model and several state-of-the-art CNN-based models on MagnaTagATune [4] dataset with a consistent experimental setup. Experiments show that our model has superior performance than other models, particularly in terms of mean average precision. Compared to the original ResCNN, our model shows a faster convergence rate and acceptable training overhead.

The rest of this paper is organized as follows. Section 2 provides an overview of the related work. Section 3 describes the architecture of our model in detail. Section 4 presents the design and results of our experiments. Section 5 concludes the paper.

2 Related Work

2.1 CNN-based Music Tagging Models

Recent advances in deep learning have significantly impacted the field of automatic music tagging, with convolutional neural networks (CNNs) becoming the predominant architecture due to their success in other domains such as image and speech recognition. [9] presented and evaluated several state-of-the-art CNN-based music tagging models thoroughly.

One notable model is the Harmonic CNN [8], which utilizes a harmonically stacked trainable representation to preserve spectro-temporal locality in the convolution layers. This model has shown effectiveness in capturing the intricate harmonic structures in music, making it a strong baseline for comparison. Another influential model is Musicnn [6], which incorporates domain knowledge into its filter designs to capture both timbral characteristics and temporal patterns. This model uses vertical filters to capture timbral features and horizontal filters to capture temporal features, thereby leveraging the unique aspects of music signals for better tagging performance. Finally, a simple 7-layer CNN with residual connections has shown exceptional results when it is trained with short

chunks of audio. It is named "short-chunk CNN with residual connections" in [9], but we'll call it "ResCNN" for simplicity.

These three models have state-of-the-art performance on MagnaTagATune. In our work, we use them as baselines for evaluation. Besides, we also modify the architecture of ResCNN to be the backbone of our framework.

2.2 Query2Label Framework

Query2Label [5] is a simple and effective Transformer approach to multi-label image classification. It's a two-stage framework. In the first stage, it uses an image classification backbone to extract the spacial features of the input image. In the second stage, it leverages multiple Transformer decoders with the label embeddings as queries to predict the existence of the corresponding label, where the extracted features are used as the keys and values of the multi-head cross-attention layers. This method allows the model to focus on specific labels and related features during training, which can enhance both the precision and the interpretability of multi-label image classification tasks.

By integrating the Q2L framework into our CNN-based music tagging model, we aim to leverage its strengths in handling multi-label tasks. The incorporation of learnable label embeddings as queries allows the model to dynamically adjust its focus based on the tags being predicted, potentially leading to improved performance in music tagging.

3 Method

3.1 Architecture

The overall architecture of our model is shown in Figure 2. For a given audio segment, we first transform it into the corresponding spectrogram by short-time Fourier transform (STFT), then feed it into our backbone (modified ResCNN) to get extracted features as input of the Q2L Transformer. Label embeddings are sent to the Transformer as queries. Finally, the output of the Transformer (through sigmoid function) is the prediction for labels.

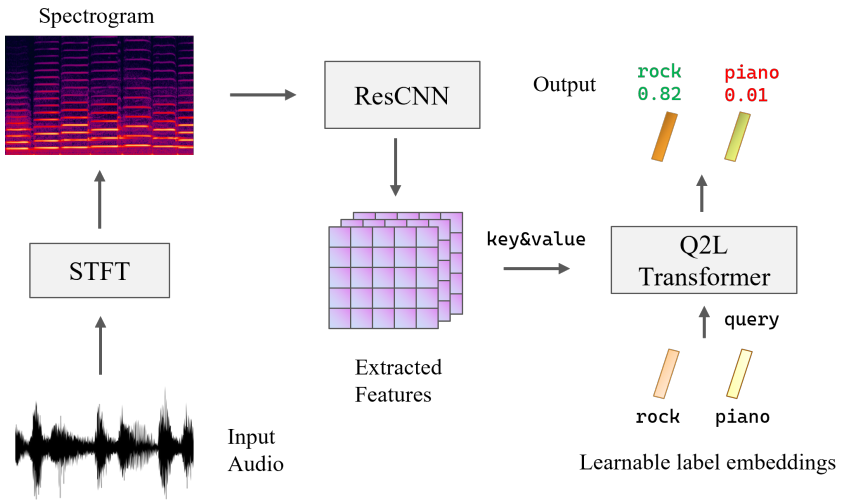

Figure 2: The architecture of our model

Input transform. We use 3.69s audio segment as input, then transform it into spectrogram through STFT. Specifically, we convert the linear frequency scale into the *mel scale*, which mimics our perception of pitch, resulting in Melspectrogram, the most popular input representation for music tagging [10]. We also apply logarithmic compression and normalization to make the final Melspectrogram suitable for feature extraction.

Backbone. The original ResCNN has 7 CNN layers, each contains two 3×3 filters with residual connections. It incorporates a fully connected layer, a linear projection layer and sigmoid function to generate the output. We decrease the number of CNN layers to 5 to reduce the computation cost and remain more positional feature for Q2L Transformer. We also modify the output layer to get an intermediate representation instead of a final prediction.

Q2L Transformer. We adopt an almost standard Transformer architecture [7]. A Transformer encoder is used to further extract audio features, but it is optional. There are two tranformer decoders, which accept learnable label embeddings as input and use the extracted features as keys and values of the multi-head cross-attention sublayers. At last we add a linear layer and a sigmoid function to produce the final prediction for each label.

3.2 Loss Function

We adopt the common mean binary cross entropy loss for multi-label classification. For number of tags N and the output tag prediction $p = [p_1, \ldots, p_N]^T$, we have the loss function

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \left(y_i \log p_i + (1 - y_i) \log(1 - p_i) \right)$$

where y_i indicates whether the input audio has the tag i.

4 Experiments

4.1 Dataset

MagnaTagATune (MATA) [4] is a widely utilized dataset for benchmarking automatic music tagging models. It provides multi-label annotations based on genre, mood, and instrumentation for 25,877 audio segments, each 30 seconds in length. The audio files are in MP3 format with a 32 Kbps bitrate and a 16 kHz sample rate. The dataset is originally divided into 16 folders, with a common practice being to use the first 12 folders for training, the 13th folder for validation, and the last three folders for testing. We follow the same data split and only use the top 50 most frequent tags for our experiment.

4.2 Evaluation Metrics

The area under receiver operating characteristic curve (ROC-AUC) is a common metric for binary classification, while the area under precision-recall curve (PR-AUC) may be a more informative metric for imbalanced datasets. Besides, the macro PR-AUC, acquired by averaging the PR-AUC scores across all labels, equals to mean average precision which we aim to improve. Therefore, we adopt both macro ROC-AUC and macro PR-AUC as our evaluation metrics.

4.3 Design

We train and test our model along with ResCNN, Harmonic CNN and Musicnn as baselines. For each model, we randomly initialize its parameters, and train for an appropriate number of epochs such that the performance on validation set is nearly optimum and not over-fitting. Then we use the best model on validation set for test. We repeat 4 times and average the test score to make results more reliable.

Besides, since our backbone is modified ResCNN, we additionally collect the validation PR-AUC scores and time for 15 epochs of training and validation to compare their convergence rate and training efficiency.

4.4 Results

Model	ROC-AUC	PR-AUC
Our Model	**0.9072**	**0.4491**
ResCNN	0.8960	0.4204
Harmonic CNN	0.8943	0.4171
Musicnn	0.8824	0.3890

Table 1: Test results for four models.

As shown in Table 1, our model outperforms all other models both in terms of ROC-AUC and PR-AUC. The significantly superior PR-AUC of our models provides a strong validation for the improvement on the precision of music tagging.

Model	Time per Epoch
Our Model	48.2s
ResCNN	42.9s

Table 2: Average training and validation time per epoch of two models.

Table 2 shows average training and validation time per epoch of our model and ResCNN. Our model only introduces around 12% of time overhead compared to ResCNN, which is quite acceptable. The validation PR-AUC curves of our model and ResCNN are shown in Figure 3. It's clear that our model has a faster convergence rate and higher precision than ResCNN.

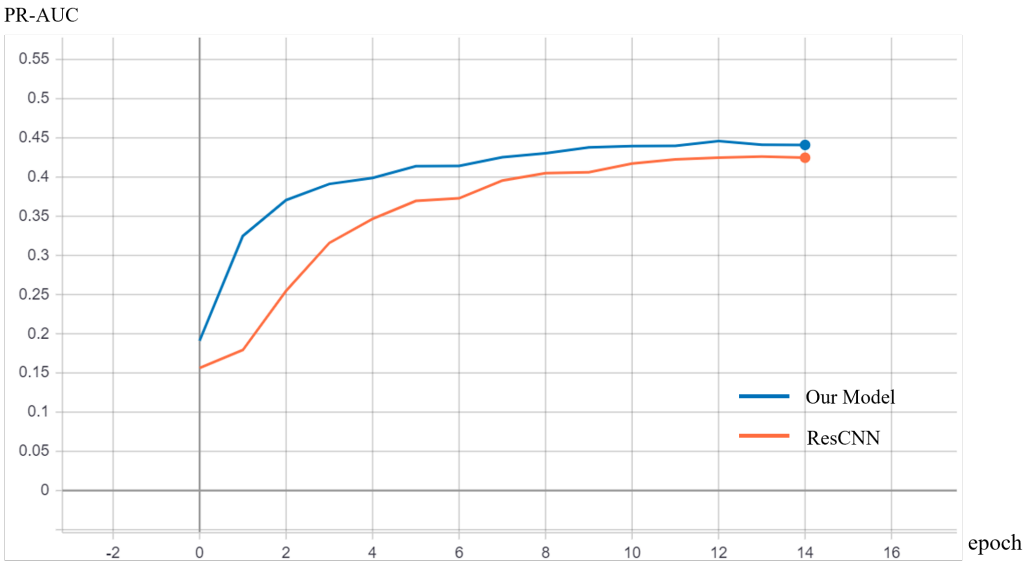

Figure 3: Validation PR-AUC curves of our model and ResCNN

5 Conclusion

In this paper, we proposed a novel approach to automatic music tagging by integrating the Query2Label (Q2L) framework with a modified ResCNN backbone based on spectrograms. By leveraging the strengths of the Q2L Transformer architecture, our model effectively improves the precision of CNN-based music tagging on MATA dataset. Our model also exhibits a faster convergence rate compared to the original ResCNN, with only a modest increase in computational overhead.

Due to limited time and resources, we only conducted experiments on MATA. Future work could involve extending this approach to larger and more diverse music datasets such as Million Song Dataset [1] and MTG-Jamendo Dataset [2]. Additionally, investigating the applicability of our model to other multi-label audio classification tasks could provide valuable insights and further validate the generalizability of our approach.

References

[1] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. The million song dataset. *In Proc. of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, 2011.

[2] Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. The mtg-jamendo dataset for automatic music tagging. *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML)*, 2019.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] Edith Law, Kris West, Michael I Mandel, Mert Bay, and J Stephen Downie. Evaluation of algorithms using games: The case of music tagging. In *In Proc. of International Society for Music Information Retrieval Conference (ISMIR)*, pages 387–392, 2009.

[5] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. Query2label: A simple transformer way to multi-label classification, 2021.

[6] Jordi Pons, Oriol Nieto, Matthew Prockup, Erik Schmidt, Andreas Ehmann, and Xavier Serra. End-to-end learning for music audio tagging at scale. *In Proc. of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, 2018.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[8] Minz Won, Sanghyuk Chun, , Oriol Nieto, and Xavier Serra. Data-driven harmonic filters for audio representation learning. *In Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[9] Minz Won, Andres Ferraro, Dmitry Bogdanov, and Xavier Serra. Evaluation of cnn-based automatic music tagging models. In *Proc. of 17th Sound and Music Computing*, 2020.

[10] Minz Won, Janne Spijkervet, and Keunwoo Choi. *Music Classification: Beyond Supervised Learning, Towards Real-world Applications*. https://music-classification.github.io/tutorial, 2021.