深層学習と転移学習に基づく音楽の埋め込み抽出に関する研究

Research on Music Embedding Extraction Based on Deep Learning and Transfer Learning

電気電子情報通信工学専攻　宗先哲

22N5100027I　Zong Xianzhe

Research Abstract

In the era of rapid technological advancement, the efficient search and management of ever-expanding volumes of information have become particularly crucial. Although traditional search engines rely on text-related search mechanisms, Content-Based Image Retrieval (CBIR) technology, which integrates text and image embeddings for search purposes, has emerged as a hot research topic.

However, the application of this method in the audio domain remains underdeveloped. This study proposes an innovative approach that leverages image processing techniques to extract embedding information from audio, to be utilized in future vector databases or graphic search engines, thereby enhancing search efficiency[1].

This study adopts advanced Convolutional Neural Network (CNN)[3] and Residual Network (ResNet)[4] architectures, this research successfully compressed audio information into a *1×128* feature vector. The model was initially trained on a large-scale sample dataset and then transferred to a new humming dataset for recognition and classification tasks in the vector space[2], ultimately achieving a highest accuracy rate of 53.45%. This feature vector was extracted at the end, effectively completing the process of extracting audio embedding information.

Contents

This thesis consists of 5 chapters. The contents and results of each chapter are summarized as follows…

Chapter 1: Introduction

Highlights the importance of using embedding technology for searching text, images, and audio. Subsequently, it discusses the current challenges faced in feature extraction. Following this, the research objectives of this thesis are elaborately outlined, and the methods chosen to overcome these challenges are described. Finally, the structure and content of each chapter are summarized.

Chapter 2: Previous Research and Related Studies

While there are no direct precedents for this study, closely related fields include music similarity comparison and the application of CLIP (Contrastive Language–Image Pre-training) models. Music similarity retrieval focuses on identifying similarities between music files, whereas CLIP models explore the relationship between text and images.

In the Related Studies part, the main models, methods, and their features used in this research are introduced. This includes the concept of pre-training in Transfer Learning, the advantages of local perception and weight sharing in Convolutional Neural Networks, and how Residual Networks prevent gradient vanishing problems caused by overly deep layers, thereby enhancing the model's representational power and performance. Additionally, the document introduces the Mel Spectrogram and explains its suitability for image recognition networks.

Chapter 3: Preliminary Preparation

This chapter primarily discusses the selection and processing of datasets, the handling of audio files, the construction of neural networks, and the overall experimental process. The article proposes two strategies for transfer training and methods for dataset processing, but in this section, we only list the most effective strategy (included in Chapter 4).

Firstly, the study selected 2 datasets: AudioSet[5], as the pre-training dataset, and Lyra-QBH[6], as the dataset for validating transfer learning. AudioSet is a dataset with a large volume of samples and many categories, from which we selected the top 10 categories with the highest number of samples. To prevent training bias due to imbalanced sample sizes, we set a maximum limit of 10,000 samples per category and referred to this dataset as the "*10-category Task Dataset*." The specific details are shown in Table 1. The Lyra-QBH dataset, is a song humming dataset. Due to the imbalance in the number of samples across categories, we selected the 10 categories with the most samples and more than 20 samples each to form a new dataset for validation in subsequent experiments. The specific details are shown in Table 2.

Table 1: 10-category task Dataset

Label	Sample Counts
Music	10,000
Speech	10,000
Vehicle	10,000
Animal	10,000
Drum	10,000
Singing	10,000
Engine	10,000
Water	8,994
Tools	8,107
Silence	7,662

Table 2: Lyra-QBH Dataset

Label	Sample counts
s024	22
s031	25
s032	24
s034	23
s036	33
s038	34
s039	20
s065	20
s086	21
s098	20

Secondly, for the training set of the Lyra-QBH dataset, we selected original songs downloaded from the Internet. Starting from an arbitrary point in the middle of the song, we continuously sampled 20 music segments, each 10 seconds long, within a span of 5 seconds at intervals of 250 ms, and repeated this process 100 times to obtain 2000 samples per category. Subsequently, all audio files were transformed into Mel Spectrograms, yielding a 64×1001 dimensional matrix in **npy** format. This matrix was then subjected to data augmentation, normalization, and feature expansion processing.

Thirdly, we constructed the network architecture as illustrated in Figure 1. Each Convolutional Block (ConvBlock) within the network executes the operation defined by Formula:

$$\boldsymbol{conv2d(a, b, kernel\_size = (c, d), stride = (e, f), padding = (g, h))}$$

Where $\boldsymbol{a}$ is the number of input channels, $\boldsymbol{b}$ is the number of output channels, $\boldsymbol{kernel\_size}$ is the size of the convolution kernel (frequency dimension $\boldsymbol{c}$ and time dimension $\boldsymbol{d}$), $\boldsymbol{stride}$ is the step size (frequency dimension $\boldsymbol{e}$ and time dimension $\boldsymbol{f}$), and $padding$ is used to fill the frequency dimension ($\boldsymbol{g}$) and time dimension ($\boldsymbol{h}$) if necessary. The network is designed to output a 1×128 feature vector, which is then utilized to perform classification tasks.

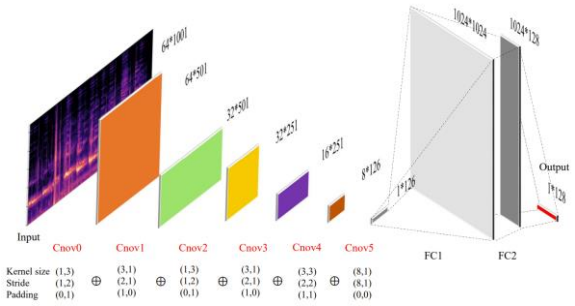

Figure 1: ResNet18 CNN Architecture

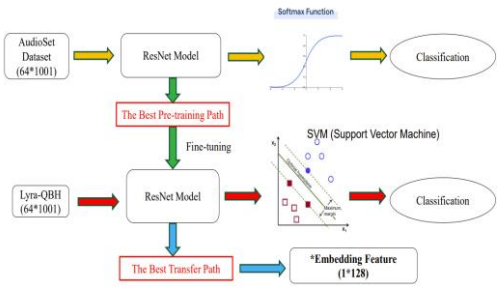

Figure 2: Experimental Process

Finally, the experimental procedure is introduced, with the process illustrated in Figure 2. Initially, we train our model on the 10-category Task Dataset to identify the optimal training path. Following this, we freeze the weights of each layer along this training path and subsequently transfer them to the training on the Lyra-QBH dataset through fine-tuning and by using SVM for regression. Ultimately, this approach yields the best path for extracting the embedding feature vector.

Chapter 4: Experimental and Result

The experiment evaluates the model using accuracy and confusion matrix metrics.

Initially, in the pre-training phase, the best generalization results were obtained when the batch size was set to **16**. After 512 iterations of training, the trend of the model's loss and accuracy is depicted in Figure 3. Among all the paths tested, the path with the best performance achieved a test accuracy of **91%**, with the average test accuracy being **85%**. The confusion matrix for the best-performing path on the test set is displayed in Figure 4. (In Figure 3,blue line represents test ,red represents training)

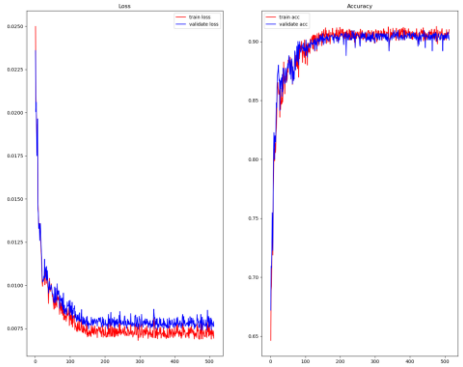

Figure 3: Pre-training Model of Loss and Accuracy Trend

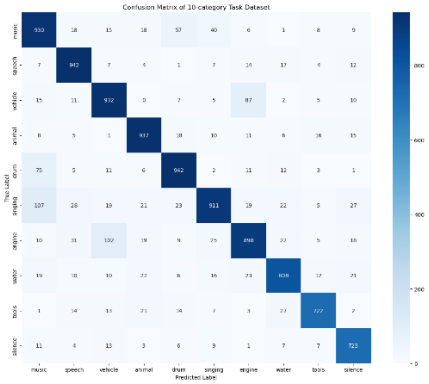

Figure 4: Confusion Matrix of the best Pre-training Path

During the model transfer process, the best generalization was achieved by only modifying the weights of the fully connected layer while freeze the weights of the other layers. After an additional 100 iterations of training, the model's loss and accuracy trends are illustrated in Figure 5. The path with the highest accuracy in the validation set achieved an accuracy of **53.45%,** with its confusion matrix displayed in Figure 6. (In Figure 6,blue line represents training ,red represents validation)

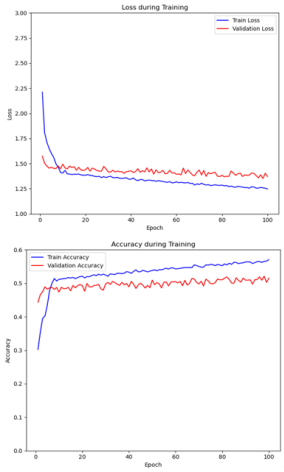

Figure 5: Transfer Model of Loss and Accuracy Trend

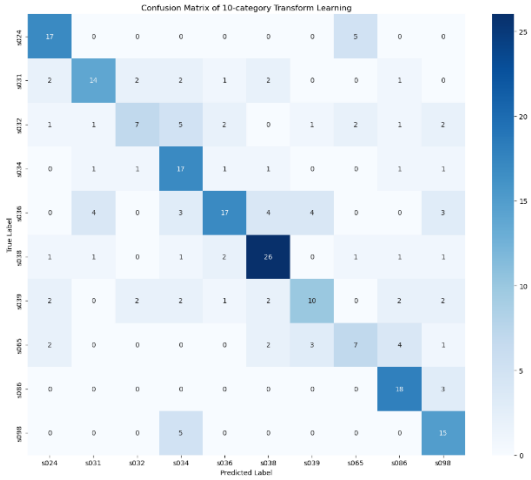

Figure 6: Confusion Matrix of the best Transfer Path

Chapter 5: Conclusion

We explored a novel method for mapping music feature vectors through the integration of deep learning and transfer learning. While this approach has shown potential, the maximum accuracy of the model only reached 53.45%, indicating that several underlying factors may be contributing to suboptimal precision. The following issues have been identified, along with challenges that future research needs to address. (1) Limitations of Dataset Size, (2) Complexity of Network Structure,
(3) Lack of Comparative Experiments with Other Models, (4) Insufficient Information Learned during Pre-training.

In the future, as model accuracy improves, it will become feasible to compare the similarity of embedding feature vectors and to explore the integration of audio embeddings with those of text and images. These efforts will significantly contribute to the advancement of the field of music information retrieval.

References

[1] Wold E, Blum T, Keislar D, et al. "Content-Based Classification, Search, and Retrieval of Audio." *IEEE Multimedia*, 2002,3(3):27-36.

[2] Pan, Sinno Jialin, and Qiang Yang. "A survey on transfer learning." *IEEE Transactions on knowledge and data engineering* 22.10 (2009): 1345-1359.

[3] LeCun, Yann, et al. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE* 86.11 (1998): 2278-2324.

[4] He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[5] Gemmeke, Jort F., et al. "Audio set: An ontology and human-labeled dataset for audio events." *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE*, 2017.

[6] Tencent Music Lyra Lab. "Lyra-Query By Humming Dataset(Lyra-QBH Dataset) ".
[Retrieved 2023-09]. https://lyracobar.y.qq.com/hum_dataset_eng.html.