DeepSound: Genre Recognition through Deep Learning

Thanh Dang **Chi Nguyen**
Department of Computer Science, Colgate University
Department of Computer Science, Connecticut College
thdang@colgate.edu, cnguyen2@conncoll.edu

Abstract

Music genre classification is a challenging task requiring effective extraction of discriminative features from audio signals to capture complex patterns and nuances in various music styles. Traditional techniques like Mel-Frequency Cepstral Coefficients (MFCCs) may not fully capture intricate music data characteristics. In recent years, deep learning techniques, particularly neural networks, have shown remarkable success in automatically learning meaningful representations from raw data across various domains, including image and speech recognition. This study investigates leveraging pre-trained neural network models, specifically VGGish, a pre-trained convolutional neural network (CNN) originally developed for audio classification tasks, for feature extraction in music genre classification. VGGish's performance in extracting discriminative features for music genre recognition is evaluated against a baseline approach using MFCCs. The objective is to assess VGGish's efficacy in capturing relevant audio characteristics and its potential for transfer learning in music-related tasks. The research contributes to ongoing efforts in developing robust feature extraction techniques for audio data analysis and highlights the significance of leveraging pre-trained models for knowledge transfer across domains.

1 Introduction

The ability to automatically extract meaningful features from audio data has become an increasingly important task in various domains, such as music information retrieval, speech recognition, and environmental sound classification (Ghoraani and Sankur, 2017). With the recent advancements in deep learning and neural network architectures, there has been a growing interest in leveraging these powerful models for feature extraction and representation learning from raw audio signals (Choi et al., 2017b).

One particular challenge in audio data processing is the effective extraction of discriminative features that can capture the relevant characteristics of the input signal (Choi et al., 2017a). Traditional feature extraction techniques, such as Mel-Frequency Cepstral Coefficients (MFCCs), have been widely used and proven effective in various audio analysis tasks (Ghoraani and Sankur, 2017). However, these hand-crafted, raw features may not fully capture the complex patterns and nuances present in audio data, particularly in domains with high variability, such as music genre classification.

In this study, we aim to investigate the effectiveness of a pre-trained neural network model, VGGish (Hershey et al., 2017a), in extracting meaningful features from audio data for music genre classification. Specifically, we compare the performance of VGGish with a baseline approach using MFCCs as input features. The objective is to testify the success of existing neural networks in deep feature extraction and assess the performance and effectiveness of VGGish in music feature extraction and feature summarization.

The significance of this research lies in the potential of leveraging pre-trained models for transfer learning in audio-related tasks. By studying the capabilities of pre-trained models to generalize knowledge about music genres, we can explore their applicability to other music classification and recognition tasks (Choi et al., 2017b). Furthermore, this work contributes to the ongoing efforts in developing efficient and robust feature extraction techniques for audio data, which has implications in various domains, such as multimedia retrieval, audio signal processing, and human-computer interaction.

2 Research Questions

Two specific key questions that we hope to answer in this study are:

1. How effective is the VGGish pre-trained convolutional neural network model in extracting discriminative features for music genre classification compared to traditional feature extraction techniques like Mel-Frequency Cepstral Coefficients (MFCCs)?

2. To what extent can the knowledge learned by the VGGish pre-trained model be transferred and adapted for music genre classification, and what are the potential benefits of leveraging transfer learning in this domain?

3 Previous Solutions

The task of music genre classification has been extensively studied in the field of music information retrieval, and various approaches have been proposed to tackle this challenging problem. The general pipeline for music classification using deep learning models typically involves the following steps:

1. **Audio Preprocessing**: The raw audio signal is preprocessed to prepare it for feature extraction. This may include resampling, converting to a suitable format, and segmenting the audio into smaller chunks or frames.

2. **Feature Extraction**: Discriminative features are extracted from the preprocessed audio data to capture relevant characteristics that can distinguish between different music genres. Several feature extraction methods have been explored in the literature, ranging from traditional hand-crafted features to learned representations using deep neural networks.

   (a) **Mel-Frequency Cepstral Coefficients (MFCCs)**: MFCCs are a widely used traditional feature extraction technique in audio signal processing. They are derived from the short-term power spectrum of the audio signal and are designed to mimic the human auditory system's response to sound (Ganchev et al., 2005). MFCCs have been extensively used as input features for music genre classification tasks due to their ability to capture relevant timbral and spectral characteristics of music signals (Tzanetakis and Cook, 2002).

   (b) **VGGish**: Developed by researchers at Google, VGGish is a pre-trained convolutional neural network (CNN) model designed for audio classification tasks (Hershey et al., 2017b). It is based on the VGGNet architecture, originally developed for image classification, but adapted for audio data by treating log-mel spectrograms as input "images." VGGish is trained on a large-scale audio dataset and can be used as a feature extractor by leveraging the activations from one of its intermediate layers, providing a learned representation of the input audio signal.

3. **Model Training**: The extracted features are fed into a deep learning model, such as a convolutional neural network (CNN) or a recurrent neural network (RNN), which is trained on a labeled dataset of music genres. The model learns to map the input features to the corresponding genre labels during the training process.

4. **Evaluation and Testing**: The trained model is evaluated on a separate test dataset to assess its performance in music genre classification. Common evaluation metrics include accuracy, precision, recall, and F1-score.

Both MFCCs and VGGish have demonstrated strengths in extracting meaningful features for music genre classification. MFCCs, as hand-crafted features, capture low-level spectral and timbral characteristics that have proven useful for various audio analysis tasks, including music genre recognition (Xu et al., 2003). On the other hand, VGGish, as a learned representation, has the potential to capture higher-level and more complex patterns in the audio data, leveraging the power of deep neural networks and their ability to learn hierarchical representations from raw data (Choi et al., 2017c).

The choice between MFCCs and VGGish (or other feature extraction methods) often depends on the specific task, dataset, and computational resources available. While MFCCs offer a computationally efficient and well-established approach, VGGish has shown promising results in audio classification tasks and may provide a more powerful and discriminative feature representation, particularly for complex audio signals like music (Gwardys and Gwardys, 2021).

4 Dataset

For the evaluation of our proposed music genre classification pipelines, we utilize 2 datasets.

GTZAN dataset is widely adopted in the field of machine listening research for music genre recognition (MGR) tasks (Tzanetakis and Cook, 2002). The dataset was collected in 2000-2001 from various sources, including personal CDs, radio broadcasts, and microphone recordings, to represent a diverse range of recording conditions (gtz). GTZAN Genre Dataset represents a total of 1000 audio tracks with a 30-second duration contained in the dataset. The dataset is divided into a total of 10 genres: blues, classical, country, disco, hiphop, jazz, metal, pop, reggae, and rock, each with 100 tracks. All the tracks are 22050Hz Mono 16-bit audio files in .wav format. In our study, we employ the GTZAN dataset as a in-distribution testbed to assess the performance of our proposed feature extraction and classification pipelines based on MFCCs and the VGGish pre-trained model.

We also handcraft our own out-of-distribution dataset of 31 audio files across 7 genres (country, hiphop, classical, blues, jazz, pop, rock) that are released from 2010s. All the tracks are 29- to 30-second long, 22050Hz Mono 16-bit audio files in .wav format. The out-of-distribution dataset will test the generalizability and learning capabilities of the models when tested on unseen, out-of-domain data.

5 Propsed Methods & Evaluation

In this work, we propose and evaluate two distinct music classification pipelines utilizing different feature extraction methods: Mel-Frequency Cepstral Coefficients (MFCCs) and VGGish. Our methodology and contributions can be summarized as follows:

- We introduce and compare two music classification pipelines based on the MFCCs and VGGish feature extraction methods, respectively.
- The GZTAN dataset is preprocessed using each feature extraction method, and two custom neural network architectures are employed for genre classification predictions.
- For the MFCCs preprocessed data, we return 13 coefficients computed by MFCC, each coefficient represents a different aspect of the spectral shape of the audio clip. The choice of n_mfcc=13 is backed by previous research in the speech recognition community, which reported that these coefficients were found to provide a good representation of the spectral shape of speech signals. We then train a deep neural network model comprising several convolutional and pooling layers, followed by dense layers for classification. This model serves as a baseline for comparison.
- For the VGGish preprocessed data, we leverage a 2D convolutional neural network (CNN) architecture trained on top of the VGGish feature extractor. VGGish converts the input audio features into a semantically meaningful, high-level 128-dimensional embedding, which can be fed as input to a downstream classification model. The downstream 2D CNN model can be shallower than usual since the VGGish embedding is more semantically compact and informative compared to raw audio features.
- We establish a pipeline for training, hyperparameter tuning, and testing, which both the MFCCs-based dense model and the VGGish-based 2D CNN model will follow. The MFCCs-based dense model serves as a baseline to evaluate the performance of the VGGish-based CNN.
- We evaluate the performance of both models using the following metrics: accuracy, precision, recall, and F1-score. These metrics provide a comprehensive assessment of the models' classification capabilities.

By introducing and comparing these two distinct feature extraction and classification pipelines, we aim to assess the effectiveness of the VGGish pretrained model in extracting discriminative features for music genre classification.

6 Results

The experimental results obtained from our study aim to address the two research questions posed:

1. The VGGish-based model achieved a high accuracy of 85%, outperforming the baseline MFCCs-based model, which attained an accuracy of 51%. This significant improvement of 34% demonstrates the effectiveness of the

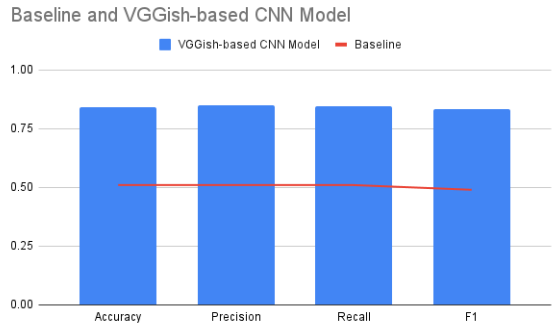

Figure 1: VGGish-based 2D CNN Outperforms Baseline for In-Distribution GZTAN Dataset.

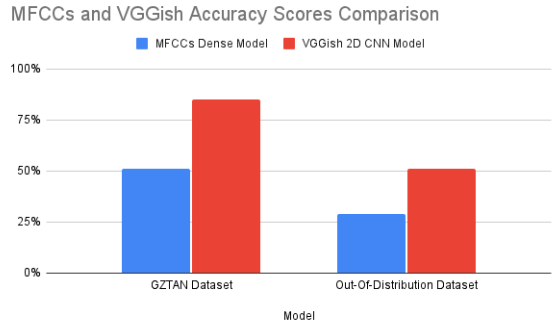

Figure 2: VGGish-based 2D CNN Still Outperforms Baseline Model for Out-of-Distribution Testing.

VGGish model in converting audio input features into a semantically meaningful, high-level 128-dimensional embedding suitable for the downstream classification task. The compact and informative nature of the VGGish embedding surpasses the raw audio features extracted by the MFCCs method, leading to superior performance in music genre classification.

2. When evaluated on unseen, new data, the generalizability of both the VGGish-based CNN model and the MFCCs-based Dense model was limited, with accuracies of 51% and 29%, respectively. This decrease in performance could be attributed to the small sample size of the out-of-distribution test set and potential differences between the new dataset and the GTZAN dataset used for training, such as variations in content, distribution, recording quality, or instrumentation.

   However, a notable finding is that the VGGish-based model still outperformed the MFCCs-based model by 22% on the new dataset. This result demonstrates the robustness and transferability of the knowledge learned by the VGGish pre-trained model, highlighting the potential benefits of leveraging transfer learning in the music genre classification domain.

The superior performance of the VGGish-based model on both the GTZAN dataset and the new dataset underscores the effectiveness of deep learning techniques and pre-trained models in extracting discriminative features from audio data. While further improvements may be necessary to enhance generalizability, the results indicate the promising potential of leveraging transfer learning and pre-trained models for music genre classification and related tasks in the field of music information retrieval.

7 Discussion

Findings underscore leveraging pre-trained neural network models like VGGish for feature extraction in music genre classification. Harnessing deep learning and transfer learning, the VGGish-based model demonstrated superior performance over traditional MFCCs-based approach, achieving 85% accuracy on GTZAN dataset.

Significant improvement highlights VGGish's ability to capture discriminative features and intricate music data characteristics more effectively than hand-crafted features like MFCCs, attributed to deep neural networks' hierarchical learning capabilities enabling automatic extraction of high-level, semantically meaningful representations from raw audio data (Choi et al., 2017c).

Results indicate potential for transferring knowledge learned by pre-trained models like VGGish to new music tasks. Despite limited generalizability on unseen data, VGGish-based model consistently outperformed MFCCs-based approach, suggesting robustness and transferability of learned representations. The results also align with broader success of deep learning and transfer learning across domains like image and speech recognition (LeCun et al., 2015; Hinton et al., 2012). Leveraging pre-trained models and knowledge transfer can achieve significant performance gains in challenging tasks like music genre classification.

Acknowledging limitations and potential factors contributing to reduced generalizability on unseen data, including differences in dataset characteristics, recording quality, instrumentation, or content distribution between training and testing datasets. Inherent complexities and nuances in music data

pose challenges for effective feature extraction and classification.

Further research needed to enhance generalizability and robustness of pre-trained models for music genre classification. Techniques like domain adaptation, data augmentation, and multi-task learning could mitigate challenges posed by dataset variations and improve transferability of learned representations.

Limitations

The manual selection of out-of-distribution data is not perfect and might be subjected to personal bias.

Acknowledgements

We would like to thank Professor Bálint Gyires-Toth for the supportive guidance and insightful conversations.

References

Gtzan genre collection. http://marsyas.info/downloads/datasets.html. Accessed: [Date accessed].

Keunwoo Choi, George Fazekas, Kyunghyun Cho, and Mark Sandler. 2017a. Deep audio representation learning for music genre recognition. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*.

Keunwoo Choi, George Fazekas, Mark Sandler, and Kyunghyun Cho. 2017b. Transfer learning for music classification and regression tasks. *arXiv preprint arXiv:1703.09179*.

Keunwoo Choi, George Fazekas, Mark Sandler, and Kyunghyun Cho. 2017c. Transfer learning for music classification and regression tasks. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, pages 141–149.

Todor Ganchev, Nikos Fakotakis, and George Kokkinakis. 2005. Comparative evaluation of various mfcc implementations on the speaker verification task. In *Proceedings of the 10th International Conference on Speech and Computer (SPECOM 2005)*.

Behnaz Ghoraani and Boulbaba Sankur. 2017. A survey on audio feature extraction and classification. *Journal of Signal Processing Systems*, 89(3):399–420.

Grzegorz Gwardys and Daniel Gwardys. 2021. Deep transfer learning for music genre recognition. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR)*, pages 647–654.

Shawn Hershey, Sourabh Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017a. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE.

Shawn Hershey, Sourabh Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017b. Cnn architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135. IEEE.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. 2012. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal processing magazine*, 29(6):82–97.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, 521(7553):436–444.

George Tzanetakis and Perry Cook. 2002. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302.

Changsheng Xu, Namunu C Maddage, Xi Shao, Fang Cao, and Qi Tian. 2003. Musical genre classification using support vector machines. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 429–432. IEEE.