Empirical Analysis of Music Genre Classification using CNNs

Aditya Vikram Singh (avsingh@umass.edu) Sahil Gupta (sahgupta@umass.edu)
University of Massachusetts Amherst

Abstract

The paper addresses the challenge of music genre classification using the GTZAN dataset and explores the effectiveness of deep learning techniques, specifically convolutional neural networks (CNNs). The objective is to develop a robust model for automatic music categorization, with applications in recommendation systems and content organization. The study compares the performance of a custom-designed CNN (SpectralCNN) trained from scratch on spectral features with pre-trained models (ResNet-18 and EfficientNetv2-S) originally learned on natural images. The evaluation includes metrics such as accuracy, precision, recall, and F1 score, along with confusion matrices and AUC-ROC curves. The results suggest that the choice of input feature influences model performance, with ResNet-18 excelling in Mel-Spectrogram and MFCC inputs, while SpectralCNN performs best with normal Spectrogram input. The findings imply that there may not be a significant difference in the learned representations of spectral and natural images, highlighting the need for further exploration and experimentation on larger datasets.

1. Introduction

As the digital age continues to revolutionize the way we consume and create music, the need for automated systems capable of classifying music into various genres has become increasingly important. Music genre classification is a fundamental task with diverse applications, including music recommendation, content organization, and artist discovery. This paper aims to address the challenge of classifying music from the GTZAN dataset into ten distinct genres and leverages the power of deep learning techniques. Our objective is to create a robust and accurate model that can provide valuable insights into the automatic categorization of music, allowing for improved music management and enriched user experiences. In this paper, we outline our approach to music genre classification by training a deep convolutional neural network from scratch on spectral features and comparing its performance with pre-trained models learned on natural images for the genre classification task. We discuss the observations and report the evaluation metrics to infer how well our model learns and answers our research problem.

2. Problem Statement

The problem we aim to address in this research is the classification of music into ten distinct genres using the GTZAN dataset. The GTZAN dataset consists of 1,000 audio tracks, each 30 seconds in duration, equally divided into ten different genres: blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock. Each audio file is a `.wav` file. Our goal is to train a deep learning model that can correctly classify these audio tracks into the correct genre categories with a high degree of accuracy. We expect the model to generalize well to unseen music samples and provide reliable predictions for real-world applications.

To evaluate the model's performance, we will employ standard classification metrics such as accuracy, precision, recall, and F1-score. We will also use techniques like cross-validation to ensure the model's robustness and reduce the risk of overfitting.

3. Data

We will use the GTZAN Dataset, which consists of 1000 audio files, each of which is 30 seconds long and belongs to one of 10 genres. It also consists of corresponding spectrogram images for each audio file. The dataset size is around 1 GB, therefore we will not require heavy computational resources. The dataset was selected due to the following reasons:

1. No noise in the data sets
2. Highly reliable: popular and credible dataset
3. High-quality audio files
4. Consistent WAV format

4. Literature Review

A comparison of audio signal preprocessing methods for deep neural networks on music tagging

The study conducted titled "A Comparison of Audio Signal Preprocessing Methods for Deep Neural Networks on

Music Tagging" [4] delves into the intricate task of music tagging, an essential component in the broader domain of music genre classification. The paper systematically investigates various audio signal preprocessing methods and their impact on the performance of deep neural networks (DNNs) in the context of music tagging. The study addresses the critical aspect of preprocessing in the context of deep learning applications for music analysis, shedding light on the pivotal role that the initial stages of signal processing play in subsequent classification tasks. The authors found that out of all the preprocessing techniques, only logarithm scaling of the input resulted in signicant improvement for the models.

Convolutional Recurrent Neural Networks for Music Classification

In this study [1], the authors explore more advanced neural network architectures for music classification, using Convolutional Recurrent Neural Networks (CRNNs) as a prominent strategy. CRNNs leverage Convolutional Neural Networks (CNNs) for localized feature extraction and Recurrent Neural Networks (RNNs) for temporal summarization of the extracted features. They did an analysis with three distinct CNN structures, historically employed in music tagging, with a focus on controlling the number of parameters and their associated performance and training time. The findings emphasize the robust performance of CRNNs in terms of parameter efficiency and training time, highlighting the efficacy of their hybrid structure in concurrently addressing feature extraction and summarization in the context of music classification. Notably, CRNNs exhibit a nuanced balance, outperforming CNNs in scenarios with equivalent parameters, while also revealing a trade-off between computation speed and memory utilization under varying parameter sizes.

The study also introduces the concept that CNNs inherently assume hierarchical features extractable by convolutional kernels, spanning from low-level attributes such as onset to higher-level patterns like percussive instrument intricacies. The authors propose a CRNN specifically tailored for music tagging and systematically regulate network size by varying parameters for controlled comparisons based on memory and computation considerations. The experimentation outcomes demonstrate comparable performance between 2D convolution with 2D kernels (k2c2) and CRNN. However, the study illuminates an inherent trade-off between computation speed and memory efficiency, revealing nuanced dynamics dependent on the scale of parameters employed. This nuanced understanding of the interplay between CRNNs and CNNs, elucidated in the study, contributes valuable insights to the ongoing discourse on optimal neural network architectures for music genre classification.

Forest Species Recognition using Deep Convolutional Neural Networks

The paper "Forest Species Recognition using Deep Convolutional Neural Networks" [3] explores the application of Convolutional Neural Networks (CNNs), to identify different types of trees in forests. They focus on two sets of data: one with close-up pictures of trees (macroscopic images) and another with extremely detailed images (microscopic). The challenge lies in handling these high-resolution images effectively to achieve accurate tree identification without overwhelming the system with too many complex settings.

Deep learning, a powerful technology, has gained popularity for its ability to improve accuracy in recognizing patterns. One of its key goals is to enable models to automatically learn various layers of representation from raw data, instead of manually designing features.

The authors investigate the use of CNNs for recognizing forest species and propose a method to adapt these networks to high-resolution images without changing the architecture used for lower-resolution ones. The authors highlight that the deep learning model they employed not only achieved state-of-the-art performance but also learned useful feature detectors, capable of identifying edges, color-based features, and gaps in the woods.

Bottom-up Broadcast Neural Network For Music Genre Classification

The paper titled "Bottom-up Broadcast Neural Network For Music Genre Classification" [5] addresses the challenges in music genre classification by proposing a novel convolutional neural network (CNN) architecture called Bottom-up Broadcast Neural Network (BBNN). Unlike traditional approaches that often lose critical low-level features during the abstraction of high-level semantic information, BBNN aims to simultaneously preserve both low-level and high-level features throughout the network. The architecture includes a unique Broadcast Module (BM) consisting of Inception blocks with dense connectivity, allowing effective extraction of information embedded in the time-frequency of audio signals at different scales simultaneously. The BBNN design minimizes the need for extensive data augmentation, making it more efficient in handling smaller datasets without compromising performance. The paper emphasizes the importance of considering various time-frequency scales for effective music genre recognition.

The proposed BBNN architecture takes into account the long contextual information from spectrograms, providing a more suitable basis for decision-making in music genre classification. The authors conducted experiments on benchmark datasets, including GTZAN, Ballroom, and Extended Ballroom, demonstrating the superior performance of their neural network compared to existing methods. The paper concludes by considering the Bottom-up Broadcast Neural Network a promising solution to the challenges of music genre classification.

5. Technical Approach

In [2], they achieved nearly identical classification accuracy between simple Short-Time Fourier Transforms and Mel Spectrograms, so we chose to use Mel Spectrograms as they were computationally tractable (due to dimensionality reduction) and consequently more informative per compute time unit for complex CNN architectures, given that we can store the pre-processed assets. [2] also mentions that log-scaling (essentially converting amplitude to a decibel scale) improves accuracy, thus, we chose to employ that in our initial model.

Thus, to address the problem of music genre classification, we propose a comprehensive technical approach that combines data preprocessing, model architecture, and training techniques.

5.1. Data Preprocessing

1. **Data Validation**: We first checked the dataset for inconsistencies and corrupt files. We noticed that out of the total 1000 audio instances in the dataset, 1 audio file was corrupt (in an unknown format, not decodable). Due to this reason, we removed that instance, and our working dataset only consists of the remaining 999 training instances.
2. **Feature Extraction**: We extracted some features from the audio files for use with our array of deep learning models. Our underlying plan is to build an ensemble of models based on these well-defined features:
   (a) Waveforms
   (b) Spectrogram
   (c) Mel Spectrogram
   (d) Mel Frequency Cepstral Coefficients (MFCCs)
3. **Feature Scaling/Normalization**: We performed conversion of amplitudes to decibels (as empirically reasoned above) and batch normalization was integrated to ensure the model will converge faster.

5.2. Model Development

1. **Model Architecture**: We architected a Convolutional Neural Network, titled SpectralCNN, that uses the extracted features as input, with 5 Convolutional layers followed by 3 Dense (Fully-Connected) layers. We also applied Batch normalization and Dropout layers interspersed between such that our model was amenable to faster convergence and robust to overfitting. We have a diagram of the model architecture tagged 1.
2. **Training**: We trained the model on batch sizes of 32 instances for 50 epochs using (state-of-the-art) Adam optimizer with an initial predefined learning rate of 1×10^{-3} and multi-class cross-entropy loss. We also used a validation set for fine-tuning hyperparameters and prevent overfitting.
3. **Transfer Learning**: We also applied transfer learning to two pre-trained models ResNet-18 and EfficientNetv2-s where their last fully connected layer was modified for the music genre classification task.

6. Evaluation Metrics and Experiments

Since the genre classification task entails multi-class classification, we chose the following evaluation metrics to analyze the various networks' (SpectralCNN, ResNet-18, EfficientNetv2-s) performance on the test set across the various input features (Mel-Spectrogram, Spectrogram, MFCC) used for training:

1. Accuracy
2. Precision
3. Recall
4. F1 Score

We also created confusion matrices for each configuration of the network and input feature. Using these confusion matrices and the evaluation metrics defined above, we were able to examine our hypothesis regarding the performance of a deep convolutional network trained from scratch on spectral images, as well as ascertain the confidence level for our assumption that natural images have a different learned representation from the spectral images. We conducted these experiments to gain a better understanding of these problem statements and provide a baseline benchmark for a deep convolutional network for spectral image classification.

7. Results

We have the constructed features for an audio sample from the `blues` genre.

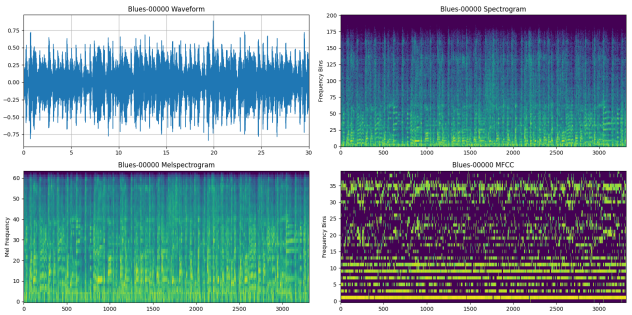

Figure 2. Various features extracted for a single audio instance

We also constructed a plot across classes for the 4 features we had extracted.

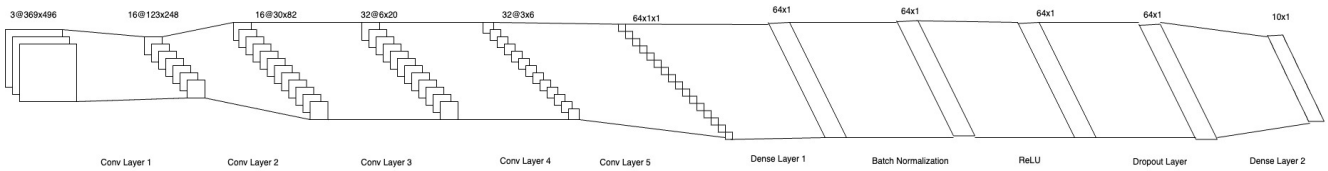

Figure 1. CNN Model Architecture for GTZAN

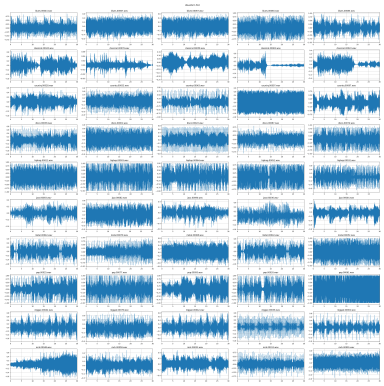

Figure 3. Waveplots for 5 random samples for each class

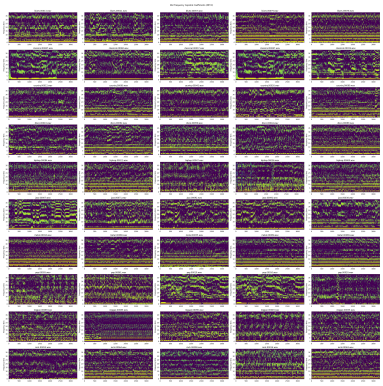

Figure 6. MFCCs for 5 random samples for each class

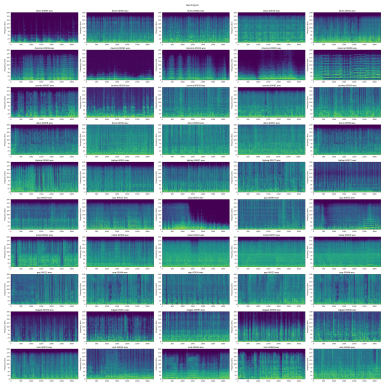

Figure 4. Spectrograms for 5 random samples for each class

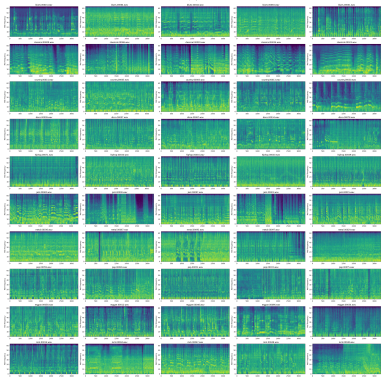

Figure 5. Mel Spectrograms for 5 random samples for each class

7.1. Evaluation Metrics

Model Type	Accuracy	Precision	Recall	F1 Score
SpectralCNN	0.4552	**0.5522**	0.4585	0.4379
ResNet-18	**0.5759**	0.5328	**0.5839**	**0.5472**
EfficientNetv2-S	0.4759	0.4677	0.4850	0.4618

Table 1. Comparative performance on Mel-Spectrogram Input

Model Type	Accuracy	Precision	Recall	F1 Score
SpectralCNN	**0.5069**	**0.5342**	**0.5063**	**0.4959**
ResNet-18	0.4793	0.5030	0.4857	0.4850
EfficientNetv2-S	0.4828	0.4611	0.4920	0.4639

Table 2. Comparative performance on Spectrogram Input

Model Type	Accuracy	Precision	Recall	F1 Score
SpectralCNN	0.4724	0.5242	0.4782	0.4646
ResNet-18	**0.5345**	**0.5394**	**0.5384**	**0.5332**
EfficientNetv2-S	0.4483	0.4390	0.4485	0.4315

Table 3. Comparative performance on MFCC Input

7.2. Confusion Matrices

7.2.1 Mel-Spectrogram Input

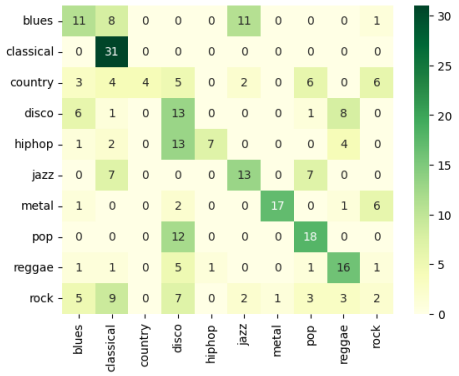

Figure 7. Confusion Matrix for Spectral CNN for Mel-Spectrogram Input

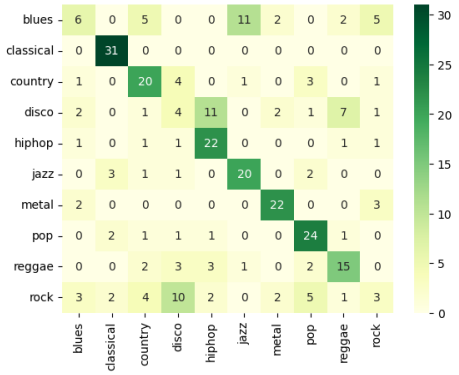

Figure 8. Confusion Matrix for ResNet-18 for Mel-Spectrogram Input

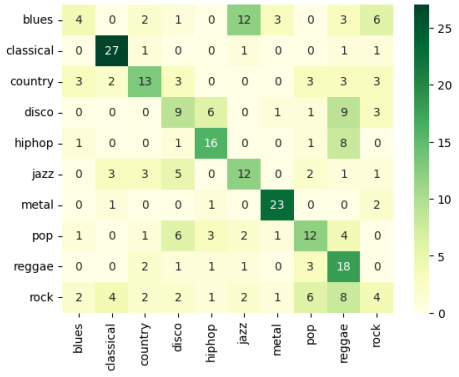

Figure 9. Confusion Matrix for EfficientNetv2-s for Mel-Spectrogram Input

7.2.2 Spectrogram Input

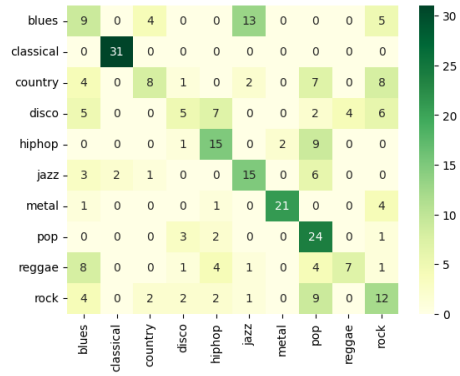

Figure 10. Confusion Matrix for Spectral CNN for Spectrogram Input

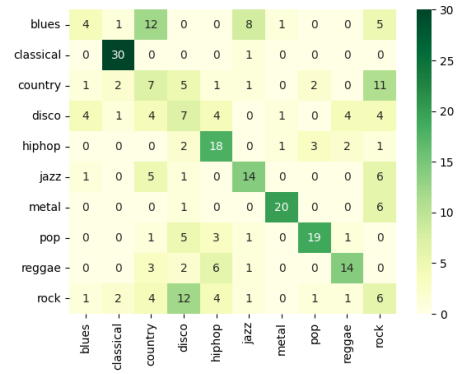

Figure 11. Confusion Matrix for ResNet-18 for Spectrogram Input

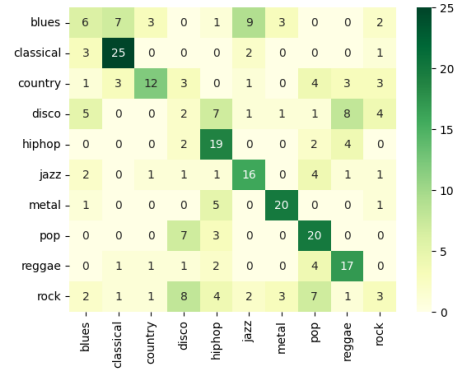

Figure 12. Confusion Matrix for EfficientNetv2-s for Spectrogram Input

7.2.3 MFCC Input

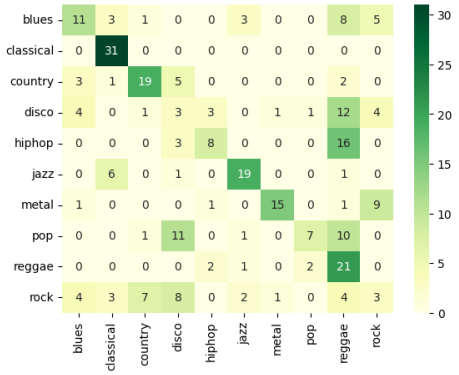

Figure 13. Confusion Matrix for Spectral CNN for MFCC Input

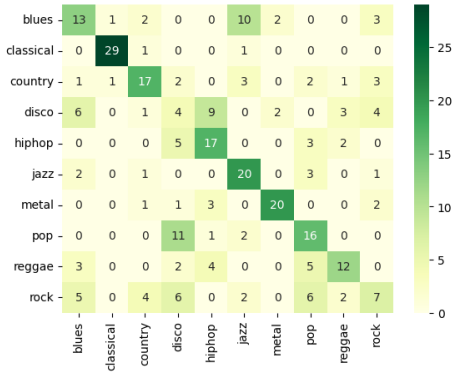

Figure 14. Confusion Matrix for ResNet-18 for MFCC Input

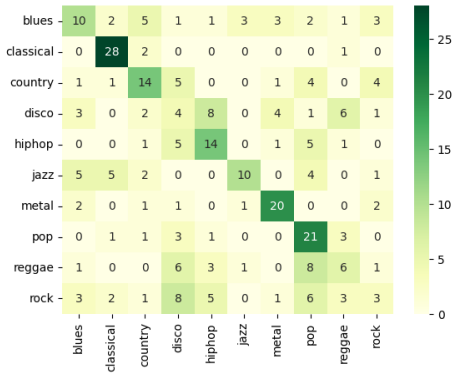

Figure 15. Confusion Matrix for EfficientNetv2-s for MFCC Input

8. Conclusion

Through our observations, we posit that there is not a significant difference in the feature representation between natural images (of objects in reality) and spectral images (generated by applying audio transformations to an audio clip) learned by deep convolutional neural networks. We selected the three well-defined spectral image features for audio clips: Mel-Spectrogram, Spectrogram, and MFCC (Mel Frequency Cepstral Coefficients), and developed a deep convolutional neural network model **SpectralCNN** that was specifically trained from scratch to classify the music genre associated with an audio clip through its spectral features (proponent for spectral image learned representation).

To examine our hypothesis about the difference between natural and spectral images, we then selected two pre-trained models (on ImageNet, therefore, on natural images) **ResNet-18** and **Efficientv2-s** and applied the concept of transfer learning to imbibe the new classification task of music genre classification with spectral features. Since these models were pre-trained on natural images, they served as proponents of natural image learned representation. Thus, we compared the performance of these three models on the GTZAN dataset and realized that **ResNet-18 has the best performance for Mel-Spectrogram or MFCC inputs**, but **SpectralCNN was the best model for normal Spectrogram input**.

Thus, as there was no clear best model for all input features, our research concluded with the idea that there is no significant difference in the learned representations of spectral images and natural images. We do understand the limitations of our experiments as we were bounded in terms of computing power and the scope of our analysis was restricted to a small audio dataset. We aim to conduct more extensive experimentation on a larger audio dataset and with more computing resources.

References

[1] Keunwoo Choi, György Fazekas, Mark Sandler, and Kyunghyun Cho. Convolutional recurrent neural networks for music classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2392–2396, 2017. 2

[2] Keunwoo Choi, György Fazekas, Kyunghyun Cho, and Mark Sandler. A comparison of audio signal preprocessing methods for deep neural networks on music tagging, 2021. 3

[3] Luiz Gustavo Hafemann. Forest species recognition using deep convolutional neural networks. 2014. 2

[4] M. Hasan, S. Ullah, M. J. Khan, and K. Khurshid. Comparative Analysis of Svm, ANN and Cnn for Classifying Vegetation Species Using Hyperspectral Thermal Infrared Data. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4213:1861–1868, 2019. 2

[5] Caifeng Liu, Lin Feng, Guochao Liu, Huibing Wang, and Shenglan Liu. Bottom-up broadcast neural network for music genre classification, 2019. 2