Comparing Music Genre Classification Methods

Tajuar Bhuiyan
University of Michigan
tajuarb@umich.edu

In-Young Chang
University of Michigan
inyoungc@umich.edu

Sunny Wang
University of Michigan
sunngwng@umich.edu

Zhiyu Zhang
University of Michigan
zhiyuzha@umich.edu

Abstract

In this paper, we propose a data augmentation method for music genre classification using Convolutional Neural Networks (CNNs). Our approach combines Mel-spectrograms with Short-Time Fourier Transform (STFT) spectrograms as joint data inputs, aiming to enhance classification accuracy. Leveraging DenseNet in both pre-trained and non-pre-trained configurations, our experiments reveal the efficacy of transfer learning and fine-tuning within this framework. Significant improvements in test accuracy, AU-ROC scores, and confusion matrix results are demonstrated, particularly with the pre-trained DenseNet121 model using transfer learning. The study concludes that this innovative dual-spectrogram augmentation offers a promising direction for music genre classification.

1. Introduction

Music genre classification, an emerging cornerstone for enhancing user experiences while navigating and discovering music contents, presents a complex challenge that is, to many, subjective. A burgeoning niche, it is gaining traction in machine learning, especially when compared to the more established domain of image classification. Despite the differences in signal forms, there has been a rapidly growing interest in applying Convolutional Neural Network (CNN) architectures to audio challenges [5].

This paper introduces a novel data augmentation methodology specifically tailored for music genre classification, leveraging Convolutional Neural Networks (CNNs) for empirical analysis. Addressing the challenge of limited and costly labeled audio data, this innovative approach is designed to mitigate overfitting and enhance classification accuracy. This exploration is not only pivotal in advancing the field of audio-based machine learning but also in setting new benchmarks for the efficient use of data resources.

2. Related Work

In the field of music genre classification, deep learning and audio signal analysis have evolved significantly. Despite being foundational in capturing the timbral aspects of music, the Mel Frequency Cepstral Coefficient's (MFCC) lossy representation of the audio signal limits its ability to retain detailed information crucial for fine-grained genre classification [9]. Recent advancements have brought forth diverse data preprocessing techniques, ranging from utilizing raw audio [4], extracting acoustic features [1], and employing visual representations [3], including wavelets and spectrograms.

Spectrogram data, particularly Short-Time Fourier Transform (STFT) and Mel-scale spectrograms, have shown superior performance in classification tasks [4]. The STFT spectrogram, a widely adopted method, maintains relevance by striking a balance between information retention and data dimensionality [11]. Additionally, the Mel-spectrogram, which transforms audio frequencies to the Mel scale, offers an effective means of preprocessing by capturing the perceptual nuances of human hearing and converting audio files into images [7].

In the realm of data augmentation for music genre classification, there have been mainly two well-documented methods. Sound transformation expands a dataset by altering a music track into various versions through processes like pitch-shifting, time-stretching, and filtering [8]. Each transformation creates a distinct version of the original track, thus enriching the dataset with varied audio properties. Sound segmentation, conversely, involves dividing a longer audio signal into shorter, discrete segments [8]. This approach increases the size of the dataset by creating multiple samples from a single track. Building upon these established techniques, we hope to introduce a novel method that combines two types of spectrogram data, namely, Mel-spectrograms and STFT spectrograms, as data augmentation.

3. Method

As shown in Figure 1, we adopt a similar methodology created by Bian *et al*. [2], who transformed audio clips into spectrograms and employed both sound transformation and sound segmentation for data augmentation. Specifically, we align with their use of DenseNet for sophisticated feature extraction, particularly in the context of audio analysis, and also agree upon their approach of utilizing spectrograms. Our methodology diverges and contributes to the field of music genre classification by introducing a novel preprocessing technique, which involves the augmentation of Mel-spectrograms with Short-Time Fourier Transform (STFT) spectrograms as combined data inputs. The central hypothesis of our research is that this dual-spectrogram approach can lead to superior classification outcomes, as opposed to relying solely on a single category of feature representation, such as the Mel-spectrogram.

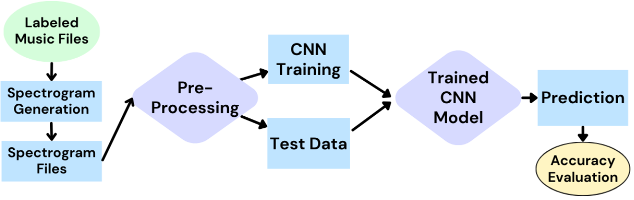


Figure 1: Our methodology examining the Dual-spectrogram Approach.

3.1. Data Preprocessing

Spectrograms, representing sequences of spectra varied along the time axis, are crucial for applying CNN to music genre classification. This method effectively transforms music audio tagging into an image classification task. We propose juxtaposing Mel-spectrograms with Short-Time Fourier Transform (STFT) spectrograms as the visual feature representations for each audio clip.

Our methodology, which amalgamates both spectrograms, is designed to harness the inherent strengths of both formats. This hybrid dual-spectrogram approach distinguishes itself from traditional data augmentation methods by providing multiple forms of feature representation, instead of simply expanding the dataset through variations and subdivisions of source materials that only constitute a single category of feature representation.

3.2. Model Architecture

Baseline Model. Our Convolutional Neural Network (CNN) for music genre classification consists of five convolutional layers with ReLU activation functions, addressing non-linearity and the vanishing gradient problem. The layers increase from 3 to 128 channels, each followed by Batch Normalization for output stabilization and Max pooling to reduce spatial dimensions and highlight features.

A Dropout layer with a 0.3 rate is introduced before the final classification stage to mitigate overfitting. The architecture concludes with a fully connected layer, receiving a 9856-feature vector from the final pooling layer, ensuring comprehensive feature analysis. The softmax activation function in this layer converts outputs into a probability distribution across music genres, showcasing the CNN's strength in feature extraction and classification.

DenseNet. Building upon the basic CNN structure, we transition to a more sophisticated model: DenseNet [6]. The core philosophy of DenseNet (Densely Connected Convolutional Network) lies in its unique connectivity pattern. This design is mathematically represented as follows in Eq. (1):

$$x = H([x_0, x_1, ..., x_{i-1}]), \tag{1}$$

where x_ℓ denotes the output feature-map of the ℓth layer, and H_ℓ represents the composite function of layer ℓ which may include operations such as batch normalization, ReLU, and convolution. The term $[x_0, x_1, \ldots, x_{\ell-1}]$ signifies the concatenation of the feature maps produced by all preceding layers up to layer $\ell - 1$.

To further examine our dual-spectrogram technique in more detailed contexts, we experiment with both pre-trained and non-pre-trained versions of DenseNet. Pre-trained models can accelerate convergence and improve performance, especially when the available dataset is limited. Specifically, we are interested in the model's performance with transfer learning, as opposed to fine-tuning.

Fine-tuning involves the strategic adaptation of a pre-existing, pre-trained DenseNet network to align closely with the specific nuances of our dataset and classification task. This method hinges on the premise that tailored adjustments to the network, attuned to the distinct characteristics of music genres, can significantly enhance its predictive accuracy and relevance to our specialized classification challenge.

On the other hand, transfer learning also employs the DenseNet model in its pre-trained state, capitalizing on its generic feature extraction capabilities developed through exposure to extensive and diverse datasets. This approach presupposes that the pre-learned features are sufficiently versatile to be applicable to our domain-specific task.

In addition, training from scratch is also evaluated to serve as a comparison for its trained counterparts. The progression from untrained to different trained versions of the same model allows us to incrementally distinguish and determine the most effective approach for the deployment of CNN architectures in music genre classification with our dual-spectrogram data augmentation technique.

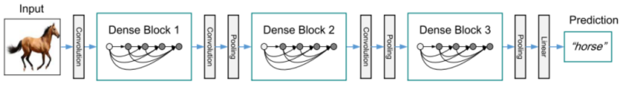


Figure 2: Illustration of a DenseNet architecture featuring three dense blocks, adapted from Huang *et al*. [6].

4. Experiments

In the constrained data environment of music genre classification, DenseNet121 is selected for its parameter efficiency and dense connectivity, which aid in mitigating overfitting while maximizing feature extraction. The use of DenseNet121 allows for a sophisticated yet not overly complex model that is capable of achieving high accuracy on our dual-spectrogram augmented dataset, proving to be a balanced choice for robust music genre classification.

4.1. Dataset

The dataset employed in our study is anchored by the renowned GTZAN dataset collected by Tzanetakis and Cook [10], often likened to the MNIST of sounds in music genre classification research. GTZAN is composed of a meticulously curated collection of 10 distinct music genres, including Blues, Classical, Country, Disco, Hiphop, Jazz, Metal, Pop, Reggae, and Rock. Each genre is represented by 100 audio tracks, with each track being exactly 30 seconds long. The tracks are in .au format with 22,050 Hz monophonic and 16-bit audio files.

Furthermore, each track is accompanied by a Mel-spectrogram of size 432×288. In addition, we add STFT spectrograms for each audio clip. These spectrograms are carefully crafted to match the dimensions of the Mel-spectrograms, indispensable for our dual-spectrogram data augmentation approach. The dataset is divided into training, validation, and test sets in an 80:10:10 ratio.

4.2. Adaptive Training Strategies

Early Stopping. This mechanism is employed to halt the training process once the validation loss plateaued. By monitoring the validation loss, training cases when no significant improvement was observed, preventing the model from learning noise and overfitting to the training data.

Model Checkpointing. To safeguard against potential loss of optimal model states, we implement model checkpointing. This process involves saving the model at different epochs, ensuring the ability to recover the most effective state of the model during the training process.

4.3. Hyperparameters and Model Configuration

In our study, we opt for the Adam optimizer, renowned for its adaptive learning rate handling and efficiency with sparse gradients. The learning rate is calibrated at 0.0005 to strike a balance between quick convergence and high accuracy. For the loss function, we choose Cross-Entropy Loss, a standard in multi-class classification problems due to its effectiveness in comparing the predicted probabilities with actual class labels. Additionally, we set a patience level of 20 for early stopping, halting training process if there is no improvement in the model's performance on the validation set for 20 consecutive epochs. This also helps in preventing overfitting and ensures that the model generalizes well to new data.

4.4. Evaluation Metrics

Area Under the Receiver Operating Characteristic Curve (AUROC). AUROC is a comprehensive metric used to evaluate a model's performance in distinguishing between different genres. This approach calculates the AUROC score for every possible pair of genres using a one-vs-one (OVO) method. After determining these scores, a macro averaging technique is applied where each genre is given equal weight in the final calculation. This method is particularly effective as it accounts for both the true positive and false positive rates, providing a balanced view of the model's ability to differentiate between genres, irrespective of their frequency in the dataset.

Confusion Matrix. For more detailed insights, the confusion matrix is employed. This tool helps us understand the model's performance on a per-genre basis, highlighting the predictions and plotting them against the true value classes. It is instrumental in identifying specific genres where the model performed exceptionally well or did not meet expectations.

4.5. Results

We began with the baseline model, setting a preliminary standard with a modest test accuracy of 0.44 and an AUROC score of 0.7962. As our research unfolded, we systematically integrated more sophisticated techniques, ultimately culminating in our most advanced model: a pre-trained DenseNet121 utilizing transfer learning coupled with our dual-spectrogram data augmentation. This final arrangement reached a pinnacle in performance, achieving a remarkable test accuracy of 0.75 and an AUROC score of 0.9563, which vividly showcases the substantial impact of our methodological innovations.

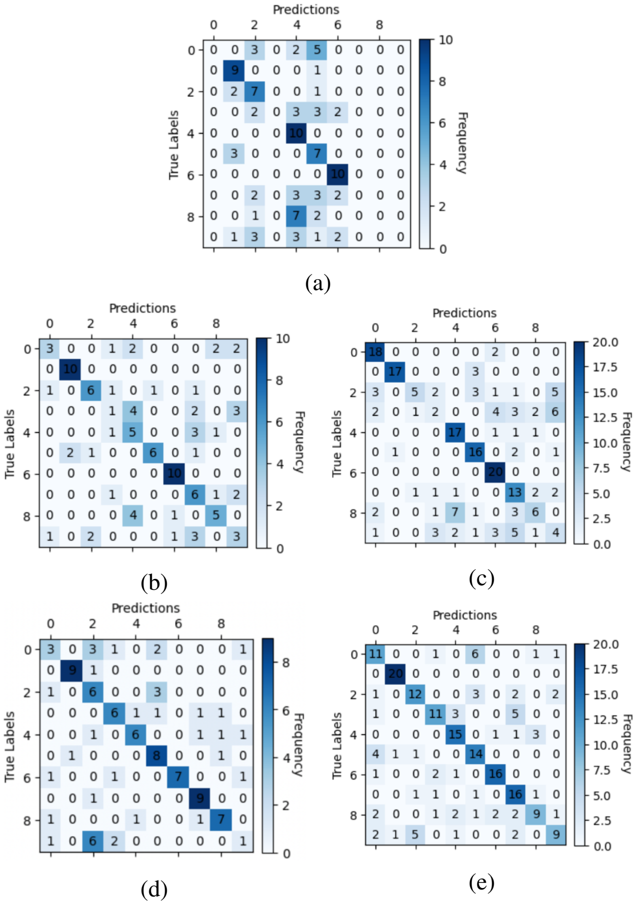
The confusion matrices displayed in Figure 3 provide an analysis of model efficacy across different configurations, with the corresponding labels for each matrix enumerated in Table 1. Furthermore, a detailed comparison of each model's performance, focusing on test accuracy, loss, and AUROC scores across different models and data augmentation strategies, is presented in Table 2.

Our analysis reveals that the application of our dual-spectrogram augmentation technique, even without the advantages of pre-training, significantly enhances the DenseNet model's capability to classify music genres. This is evidenced by a marked decline in classification errors, particularly noticeable within the blues, hip-hop, and jazz genres, as indicated by the confusion matrices.

Delving deeper into the realm of pre-trained DenseNet models, we dissect and evaluate the relative merits of fine-tuning versus transfer learning methodologies. The confusion matrix not only reflects an uptick in test accuracies and AUROC scores but also decisively demonstrates that transfer learning, especially when integrated with our dual-spectrogram augmentation, triumphs over fine-tuning. This is manifested by an increased concentration of correct predictions along the matrix diagonal and a corresponding decrease in off-diagonal elements, indicative of reduced misclassification instances.

	0	1	2	3	4	5	6	7	8	9
Genre	Blues	Classical	Country	Disco	Hip-Hop	Jazz	Metal	Pop	Reggae	Rock

Table 1: Labels for the Confusion Matrices in Figure 3.



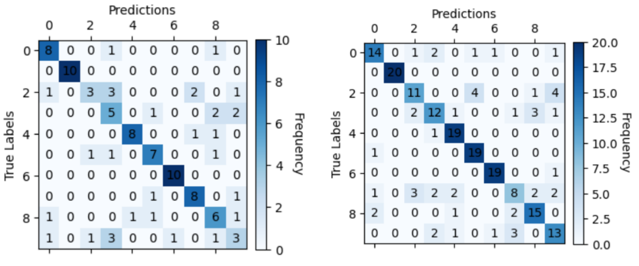
(a)

(b)

(c)

(d)

(e)



(f)

(g)

Figure 3: (a) Base Model. (b) DenseNet with No Dual-Spectrogram Augmentation. (c) DenseNet with Dual-Spectrogram Augmentation. (d) Pre-trained DenseNet with Fine-Tuning and No Dual-Spectrogram Augmentation. (e) Pre-trained DenseNet with Fine-Tuning and Dual-Spectrogram Augmentation. (f) Pre-trained DenseNet with Transfer Learning and No Dual-Spectrogram Augmentation. (g) Pre-trained DenseNet with Transfer Learning and Dual-Spectrogram Augmentation.

Model	**Accuracy**	**Loss**	**AUROC**
Baseline Model	44%	2.0131	0.7962
DenseNet without Pre-training (No Aug.)	55%	1.3496	0.8897
DenseNet without Pre-training (Dual-Aug.)	59%	1.5611	0.9058
DenseNet with Fine-Tuning (No Aug.)	62%	1.1298	0.9299
DenseNet with Fine-Tuning (Dual-Aug.)	66%	0.9508	0.9404
DenseNet with Transfer Learning (No Aug.)	68%	0.9796	0.9409
DenseNet with Transfer Learning (Dual-Aug.)	75%	0.7898	0.9563

Table 2: Model Performances on Music Genre Classification.

5. Conclusions

We conclude that the integration of Mel-spectrograms with STFT spectrograms as a dual-spectrogram data augmentation technique significantly enhances music genre classification using CNNs. This innovative approach, tested with extensive experiments on the GTZAN dataset, not only improves classification accuracy but also showcases the model's robustness in handling diverse musical features. Moreover, our results highlight the potential of transfer learning in conjunction with dual-spectrogram augmentation, which proves to be more effective than fine-tuning given the size of the dataset. This study provides a promising direction for future research in music genre classification and audio signal processing, emphasizing the potential in creative data augmentation strategies.

6. Acknowledgements

We are grateful to Dr. Andrew Owens, Sarah Jabbour, Dr. Zeynep Özcan, and Molly Jones for their invaluable guidance, corrections, and inspiration throughout this project.

References

[1] N. Auguin, S. Huang, and P. Fung, *Identification of live or studio versions of a song via supervised learning*, *Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 1–4, 2013.

[2] W. Bian, J. Wang, B. Zhuang, Y. Jian-Kui, S. Wang, and J. Xiao, *Audio-Based Music Classification with DenseNet and Data Augmentation*, in *Lecture Notes in Computer Science*, 2019, pp. 56–65. doi: `10.1007/978-3-030-29894-4_5`.

[3] Y. M. G. Costa, L. S. Oliveira, A. L. Koericb, and F. Gouyon, *Music genre recognition using spectrograms*, *18th International Conference on Systems, Signals and Image Processing*, pp. 1–4, 2011.

[4] S. Dielman and B. Schrauwen, *End to end deep learning for music audio*, *IEEE International Conference on Music Information Retrieval (ISMIR)*, 2011.

[5] S. Hershey et al., *CNN Architectures for Large-Scale Audio Classification*, arXiv:1609.09430 [cs, stat], Jan. 2017. Available: `https://arxiv.org/abs/1609.09430`.

[6] Gao Huang, Zhuang Liu and Laurens van der Maaten, *Densely Connected Convolutional Networks*, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.4700-4708. Honolulu, Hawaii, 2017.

[7] D. Joshi, J. Pareek, and P. Ambatkar, *Comparative study of MFCC and Mel-spectrogram for RAGA classification using CNN*, *Indian Journal of Science and Technology*, vol. 16, no. 11, pp. 816–822, Mar. 2023. doi: `10.17485/ijst/v16i11.1809`.

[8] R. Mignot and G. Peeters, *An analysis of the effect of data augmentation Methods: Experiments for a Musical Genre Classification task*, *Transactions of the International Society for Music Information Retrieval*, vol. 2, no. 1, pp. 97–110, Jan. 2019. doi: `10.5334/tismir.26`.

[9] O.M. Mubarak, E. Ambikai Rajah, and J. Epps, *Novel Features for Effective Speech and Music Discrimination*, *IEEE Engineering on Intelligent Systems*, pp. 342-346, 2006.

[10] B. L. Sturm, *The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use*, *VBN Forskningsportal (Aalborg Universitet)*, pp. 1–29, Jun. 2013. Available: `https://vbn.aau.dk/ws/files/185623960/SturmJNMR20131211.pdf`.

[11] Lonce Wyse, *Audio spectrogram representations for processing with Convolutional neural networks*, *Proceeding of the First International Workshop on Deep Learning for Music*, 2017.