Application of Deep Learning for Music Genre Image Classification Using MobileNetV2

Nelson Krisanto
Student ID: 23203786
nelson.krisanto@ucdconnect.ie

Abstract

This study explores using transfer learning with the MobileNetV2 architecture to classify images of spectrograms representing ten different music genres in the GTZAN dataset available on Kaggle. The dataset presents challenges such as class imbalance and high intra-class variability. To address these challenges, the researchers fine-tuned the pre-trained convolutional neural network and utilized advanced data preprocessing techniques. The modified MobileNetV2 model achieved a classification accuracy of 77.50%, with notable improvements in precision and recall. This result demonstrates the effectiveness of transfer learning for enhancing classification accuracy and the model's ability to generalize across diverse image datasets.

1 Introduction

Deep learning has shown great potential in the automated music genre classification domain. It effectively recognizes complicated patterns in audio and its derived data forms. The GTZAN dataset has become a comprehensive basis for exploring the application of image-based classification techniques in music genre recognition. It comprises images of audio spectrograms across ten music genres. This project uses the MobileNetV2 architecture, known for its efficiency in mobile vision applications, within a transfer learning framework to address the challenges posed by the dataset. These challenges include significant class imbalance and variability in spectrogram images. The study aims to optimize the balance between model complexity and computational resource usage by resizing images to 432x288 pixels for input to MobileNetV2 and employing techniques such as data augmentation and targeted model tuning. The experimental setup includes rigorous model evaluation metrics like accuracy, precision, recall, and AUC, providing insights into both the strengths and limitations of the approach and offering a robust foundation for future enhancements in similar classification tasks.

2 Related Work

The intersection of deep learning and audio analysis has been a fertile ground for research, especially in the field of music genre classification. In recent times, researchers have focused on the use of convolutional neural networks (CNNs) on transformed representations of audio data, such as spectrograms or Mel-Frequency Cepstral Coefficients (MFCCs), in addition to direct audio waveforms. This approach resembles the use of image recognition methods for audio data, as it recognizes the patterned nature of spectrograms as similar to visual inputs. Notable works have included variations of popular CNN architectures, such as VGG and ResNet, which have been successfully adapted for audio classification tasks. Although the use of MobileNetV2 in this domain is less common, our approach highlights its novelty. Due to its efficiency and effectiveness in handling mobile-based

Preprint. Under review.

vision tasks, the model is a promising candidate for handling spectrogram images, which share similar properties with traditional image data.

3 Experimental Setup

We conducted our experiments using the GTZAN dataset, which is widely known for evaluating music genre classification systems. The dataset includes 1000 audio tracks, with each track being 30 seconds long and equally distributed across ten musical genres. In our project, we converted audio tracks into spectrogram images in order to apply CNN-based image classification techniques.

3.1 Initial Model Configuration

We started with a pre-trained MobileNetV2 model on ImageNet, which has a lightweight architecture that's suitable for processing high-resolution images with low computational overhead. Initially, we set the input image size to 128x128 pixels.

3.2 Model Adjustments and Challenges

During our first experiments, we faced various challenges that required us to make adjustments to the model setup:

- **Input Resolution:** The initial resolution of 128x128 pixels was not enough to capture the detailed features required for accurate genre classification from spectrograms. This limitation was evident from suboptimal performance metrics in the early stages of training. After conducting several tests, we increased the resolution to 256x256, which was then followed by the image dimension of 432x288 pixels. This significantly improved the model's ability to distinguish finer details in the spectrograms, but it also increased computational demands.
- **Overfitting:** In the initial stages, our model showed overfitting, which means that there was a significant difference between the accuracy of the training and validation data. To mitigate this issue, we introduced more rigorous data augmentation techniques and added extra dropout layers. We experimented with different dropout rate parameters ranging from 0.2 to 0.7, and found that a dropout rate of 0.6 significantly improved the model's ability to generalize.
- **Optimizer Adjustments:** After experiencing unstable training dynamics with the standard Adam optimizer, switching to a legacy version of the optimizer stabilized the training process and improved convergence.

3.3 Hyperparameter Tuning and Evaluation

We conducted thorough hyperparameter tuning to enhance the performance:

- **Learning Rate Adjustments:** Fine-tuning the learning rate was crucial, with lower rates yielding better convergence, indicating the model's sensitivity to the step size in the complex spectrogram-based feature space.
- **Inclusion of Advanced Metrics:** We included precision, recall, and AUC in our evaluation metrics to provide deeper insights into the model's performance across the different classes, which is particularly important given the inherent imbalances in the dataset.

3.4 Reverting Changes

Experiments with additional convolutional layers and alternative activation functions did not consistently improve performance and were removed to simplify the model without sacrificing effectiveness.

3.5 Final Model Selection

After several times testing and validating, we selected a model setup that balanced computational efficiency and predictive performance, resulting in substantial improvements over the initial configuration.

4 Results

4.1 Model Performance

The final model was configured to process larger images of 432x288 pixels, which proved crucial for accurate music genre classification by effectively extracting necessary features.

The performance of our final model on the validation set exceeded previous configurations:

- **Loss:** The model achieved a low loss of 0.8798, indicating a minimal average error per prediction.
- **Accuracy:** It reached a high accuracy of 77.50%, demonstrating the model's effectiveness in classifying music genres correctly.
- **Precision:** The precision was recorded at 80.32%, reflecting the model's accuracy in predicting positive labels and its ability to minimize false positives.
- **Recall:** The recall achieved was 75.50%, showcasing the model's ability to identify all relevant instances across the dataset effectively.
- **AUC:** An AUC of 0.9534, demonstrating an excellent ability of the model to distinguish between the various music genres.

4.2 Visual Evaluation

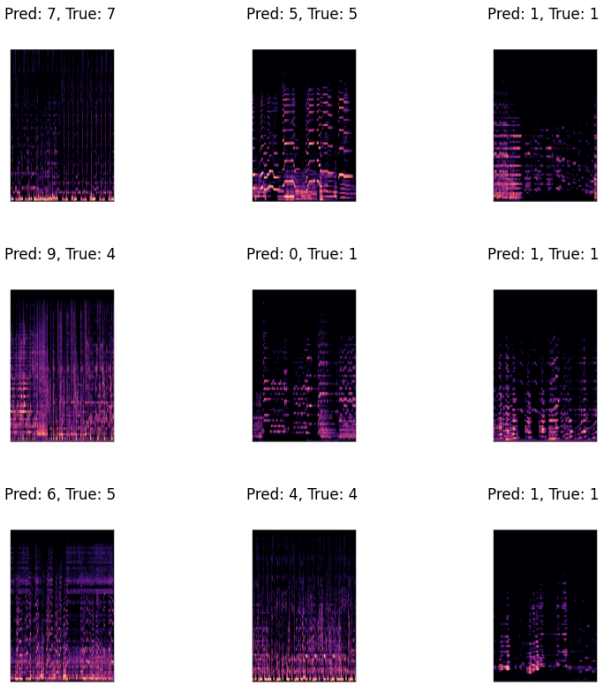

Figure 1: Sample predictions from the validation set showing the predicted and true labels. The model's predictions showcase its robustness in recognizing certain genres more clearly than others, with high consistency across samples.

4.3 Discussion

These results highlight the complex challenges involved in music genre classification, especially in terms of handling the diversity and subtlety in musical patterns. While the model demonstrated high performance metrics, the slight discrepancies in precision and recall across different genres underline the potential for further refinement. Future efforts could explore more sophisticated models or enhanced feature extraction techniques to capture the nuanced characteristics of musical genres better.

5 Conclusion & Future Work

This study showed that a transfer learning approach using MobileNetV2 is effective in classifying music genres from spectrogram images. The modified MobileNetV2 model achieved a high accuracy rate of 77.50%, with precision and recall scores indicating its strong ability in recognizing musical genres from visual data. **Future Work** includes several areas for improvement and exploration:

- **Data Augmentation:** With the use of more advanced data augmentation techniques, the model's capacity to generalize from the training data and handle overfitting can potentially be improved.
- **Hyperparameter Optimization:** Automated approaches such as Bayesian optimization could be used to fine-tune the model's hyperparameters more efficiently.
- **Model Ensembling:** Combining predictions from multiple models may improve the accuracy and reliability of music genre classification from spectrograms.
- **Feature Engineering:** Exploring different methods of feature extraction from audio files, such as using different types of spectrograms or including temporal features, could provide richer information for the classification task.
- **Deep Learning Architectures:** Investigating more complex architectures like EfficientNet or Transformer-based models could lead to improvements in classification performance.

Due to the intrinsic variability in musical expression, musical genre classification remains a challenging task. However, with the help of these enhancements, we can develop more nuanced models that can more effectively capture the complexities of musical genre classification.

References

[1] Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing, 10*(5), 293-302.

[2] Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2017). Transfer learning for music classification and regression tasks. *In Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*.

[3] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv preprint arXiv:1704.04861*.

[4] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. *In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.

[5] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. *http://www.deeplearningbook.org*

[6] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.