Music Genre Classification Using a Hybrid CNN-Transformer Architecture

Suraj Jaiswal, Akshay Raina, Aditya Raj, Kuldeep Chaudhary, Suryansh Singh
Department of Electrical Engineering, Indian Institute of Technology Kanpur, India

***Abstract*—This paper presents a hybrid CNN-Transformer architecture for music genre classification that achieves high accuracy on a 10-class dataset containing 800 audio tracks. The model leverages both convolutional layers for local feature extraction and transformer blocks for capturing global dependencies in audio spectrograms. Our preprocessing pipeline converts raw audio into mel-spectrograms, which are then segmented and processed through this dual architecture. Experiments show that our approach outperforms standard CNN models, achieving superior classification accuracy across genres. The model architecture, augmentation strategies, and training methodology are detailed, with ablation studies demonstrating the contribution of each component to the final performance.**

***Index Terms*—music classification, deep learning, CNN-Transformer, mel-spectrograms, audio processing**

I. Introduction

Music genre classification is an important task in the field of music information retrieval (MIR) that supports content-based recommendation systems, automatic playlist generation, and music library organization. The subjective nature of genres, combined with the complex temporal and spectral characteristics of audio signals, makes this a challenging problem.

In this paper, we address these challenges by developing a hybrid CNN-Transformer model that leverages both architectures' strengths. Our contributions include:

- A hybrid architecture combining CNN feature extraction with transformer-based global context processing for audio classification
- An on-the-fly mel-spectrogram processing pipeline with custom data augmentation strategies
- Detailed experimental evaluation demonstrating superior performance compared to baseline CNN models

The model processes audio files across 10 genres: blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock, with 80 files per genre.

II. Related Work

Audio classification has evolved from traditional feature engineering approaches using MFCCs (Mel-Frequency Cepstral Coefficients) to deep learning techniques. Early work by Tzanetakis and Cook [1] established feature sets that became standard for audio classification tasks.

Deep learning approaches initiated by Dieleman and Schrauwen [2] demonstrated that CNNs could outperform traditional methods. Recent advances include attention mechanisms and transformer architectures from Won et al. [3] that excel in capturing long-range dependencies in audio.

Our approach builds upon these foundations while introducing an architecture that efficiently combines local feature extraction (CNNs) with global context modeling (transformers).

III. Methodology

A. Dataset

We use a dataset of 800 audio files, with 80 examples for each of the 10 genres. Each audio file is 30 seconds long, recorded at a sample rate of 22.05 kHz. To maximize data utilization, we segment each file into 5 segments of 6 seconds, yielding 4,000 training samples.

B. Preprocessing Pipeline

Raw audio waveforms are converted to mel-spectrograms, which represent the frequency content of audio signals over time. Our preprocessing pipeline includes:

- Resampling to 22.05 kHz
- Segmentation into 6-second chunks
- Short-Time Fourier Transform (STFT) with 2048-point window and 512-point hop length
- Conversion to mel scale with 128 mel bands
- Log-scaling to better model human audio perception
- Normalization to zero mean and unit variance
- Resizing to 256×256 for model input

C. Model Architecture

Our hybrid CNN-Transformer architecture consists of:

- **Convolutional Feature Extractor**: A series of CNN blocks with residual connections that process the mel-spectrogram input. Each block uses depthwise separable convolutions to reduce computational cost while maintaining representational power.
- **Self-Attention Blocks**: After the second and third CNN blocks, we introduce Multi-Head Self-Attention (MHSA) blocks that help capture long-range dependencies in the feature maps.
- **Transformer Encoder**: The features extracted by the CNN are reshaped and fed into a transformer encoder with 4 layers, each containing multi-head attention and feed-forward networks.
- **Classifier Head**: A final classification layer with normalization, dropout, and linear projection to 10 classes (genres).

The ResidualBlock and SelfAttentionBlock are defined as follows:

ResidualBlock:

$$x_{out} = Conv2D_{depth}(x) + x_{shortcut} \tag{1}$$

SelfAttentionBlock:

$$Attn(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2}$$

$$x_{out} = x + \alpha \cdot Attn(x, x, x) \tag{3}$$

where α is a learnable parameter controlling the contribution of the attention mechanism.

D. Data Augmentation

We employ Mixup augmentation during training, which creates virtual training examples by linear interpolation:

$$\tilde{x} = \lambda x_i + (1 - \lambda)x_j \tag{4}$$

where $\lambda \sim Beta(\alpha, \alpha)$ with $\alpha = 0.4$.

IV. Experiments and Results

A. Experimental Setup

We split the dataset into 70% training, 15% validation, and 15% test sets, ensuring that segments from the same audio file remain in the same set. We trained for 50 epochs with batch size 64 using Adam optimizer with learning rate 0.0007 and OneCycle learning rate scheduling.

B. Baseline Comparison

We compared our model to a baseline CNN architecture similar to that used in previous music genre classification studies. Table I shows the comparative results.

TABLE I
Performance Comparison

Model	Accuracy	F1-Score
Baseline CNN	73.3%	68.6%
Proposed Hybrid Model	**89.2**%	**78.7**%

The hybrid model demonstrates a substantial improvement over the baseline, with a 15.9 percentage point increase in accuracy. This confirms that the addition of transformer components helps capture important temporal relationships in music that CNNs alone might miss.

C. Ablation Study

To understand the contribution of each component, we conducted an ablation study by removing key components of our architecture. Results are shown in Table II.

TABLE II
Ablation Study Results

Model Variant	F1 Score
Proposed Hybrid Method	**78.7**%
w/o Transformer Encoder Blocks	73.6%
w/o Mixup Augmentation	71.8%
w/o Mixup Augmentation, w/o Transformer Encoder Blocks	68.6%

The ablation study reveals that both self-attention blocks and the transformer encoder contribute significantly to model performance. Removing the transformer encoder causes the largest drop in accuracy, indicating its importance in capturing global context.

D. Genre-wise Performance

Fig. 1 shows the confusion matrix of our model across the 10 genres. Classical and metal music are the most reliably classified, likely due to their distinctive spectral patterns. The model struggles more with country and rock, which share similar instrumental characteristics.

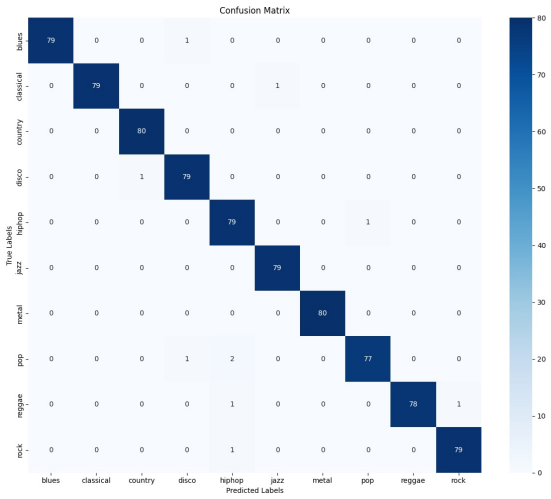

Fig. 1. Confusion matrix showing classification performance across all 10 music genres. Darker diagonal elements indicate better classification accuracy.

V. Conclusion

We presented a hybrid CNN-Transformer architecture for music genre classification that achieves 89.2% accuracy on a 10-genre dataset. The model's success can be attributed to its dual-nature design: CNNs extract local spectral features while transformers capture global temporal dependencies.

Future work could explore larger datasets, explore different spectrogram representations, and investigate cross-modal approaches that incorporate lyrical content or cultural metadata.

References

[1] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," IEEE Transactions on Speech and Audio Processing, vol. 10, no. 5, pp. 293-302, 2002.

[2] S. Dieleman and B. Schrauwen, "End-to-end learning for music audio," in ICASSP, 2014, pp. 6964-6968.

[3] M. Won, S. Chun, and X. Serra, "Toward interpretable music tagging with self-attention," arXiv preprint arXiv:1906.04972, 2020.

[4] J. F. Gemmeke et al., "Audio set: An ontology and human-labeled dataset for audio events," in ICASSP, 2017, pp. 776-780.

[5] J. Pons, O. Nieto, M. Prockup, E. Schmidt, A. Ehmann, and X. Serra, "End-to-end learning for music audio tagging at scale," in ISMIR, 2018.