Music Genre Classification

Omar Khaled *221027817*
Nada Ayman *221007645*

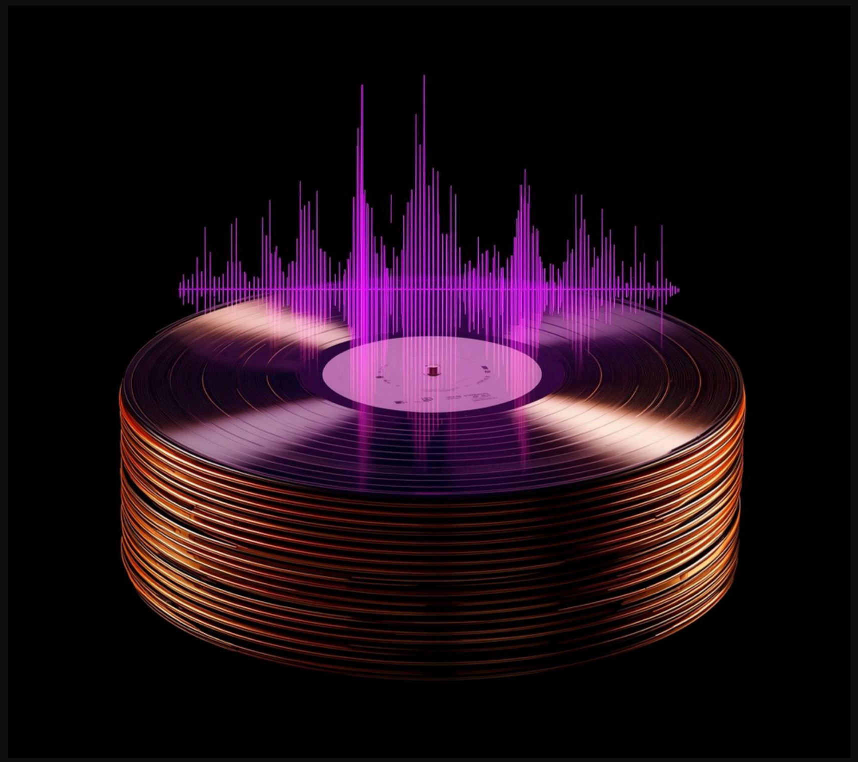
MUSIC ENTHUSIAST

Music genre classification

- Still mostly done manually
- Time consuming for large libraries
- Subjective
- Inconsistent when scaled across platforms or datasets

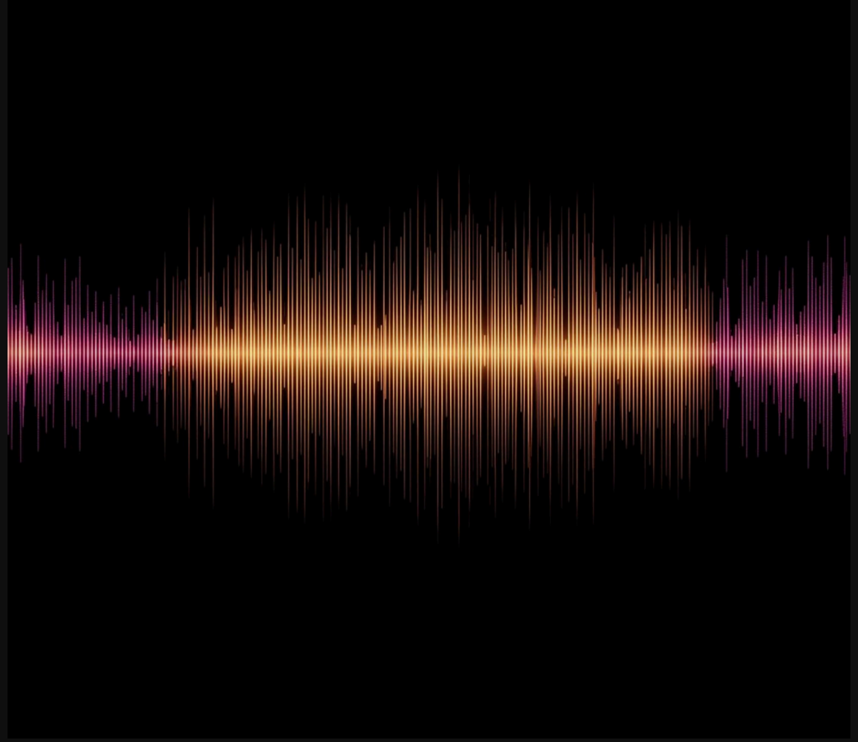
Background

- Music is a time varying signal
  - Raw waveforms are difficult for machine learning models to interpret
- Mel-spectrogram
  - Captures both time and frequency
- MFCC
  - A set of features that describes the shape of the frequency spectrum
- GradCam
- Integrated Gradients



Dataset

- GTZAN dataset:
  - 1000 audio files
  - 10 genres
  - 30 seconds clips
- Audio preprocessing:
  - Audio resampled to 22,050 Hz
  - Converted into Mel-Spectrograms
- Train/validation split: 80/20



Proposed Model

- CONV → LSTM → Attention → Classifier
- 3 convolution layers: extract local patterns in spectrogram
- 1 LSTM layer: capture temporal dependencies
- 4 Attention heads: Highlights important segments
- Classifier: maps model output to predictions

Training setup

- Trained from scratch on Apple MPS
- Batch size: 8
- Epochs: 50
- Early stopping applied
- Data Augmentation applied on training set

Results

- Quantative Results
  - Best Validation Accuracy: 44.5% (Epoch 47)
    - Unbalanced baseline: ~15%
    - After initial fixes: 32%
    - Final: 44.5% (39% relative improvement)
  - Balanced Training
    - Best epoch: Train 35.5%, Val 44.5%
    - Minimal overfitting (~5% gap
- Training Progression
  - Epochs 1-16: Struggling (~8-20%)
    - Regularization preventing learning
  - Epochs 17-30: Breakthrough (~24-33%)
    - LR reduced to 5e-3, model adapts
  - Epochs 31-47: Sustained improvement (~33-44.5%)
    - Progressive convergence, peak at epoch 47
  - Epochs 48-50: Minor overfitting (~41.5%)

Challenges and limitations

- GTZAN is fairly small for deep learning standards
- Limits generalization and makes overfitting easier
- Temporal modeling may be limited by segment length
- Some genres overlap heavily

Conclusion

- Key Achievements
  - Balanced CNN-LSTM-Attention Architecture
    - Real temporal sequences (16 time steps via LSTM)
    - Attention mechanism for dynamic feature weighting
    - Proper regularization preventing overfitting
  - Effective Temporal Data Augmentation
    - Time-stretch, pitch-shift, Gaussian noise
    - Spectrogram masking (time & frequency)
    - 5-10% improvement in generalization
  - Hyperparameter Optimization
    - Optimal balance between capacity and regularization
    - Learning rate scheduling improved convergence
    - Early stopping maximized training potential
- Model Performance
  - 44.5% accuracy on 10-class GTZAN dataset
    - 4.4x above random baseline (10%)
    - Strong custom architecture (no pre-training)
    - Competitive with published results