DeepSound

Genre Recognition Through Deep Learning

Thanh Dang - Chi Nguyen

Dataset

1. GTZAN Dataset (In-Distribution):
   - Collected from various sources in 2000-2001
   - Contains 1000 audio tracks, each 30 seconds long
   - Divided into 10 genres (blues, classical, country, disco, hiphop, jazz, metal, pop, reggae, rock)
2. Out-of-Distribution Dataset
   - Handcrafted dataset with 31 audio files across 7 genres (country, hiphop, classical, blues, jazz, pop, rock).
   - Tracks released in the 2010s.

   ⇒ We will be able to test models on unseen data to assess generalization capabilities.

Feature Extractions

Mel-Frequency Cepstral Coefficients (MFCCs)

- Compute 13 coefficients per audio clip to represent different aspects of spectral shape.
- We will use MFCCs for our baseline Dense model

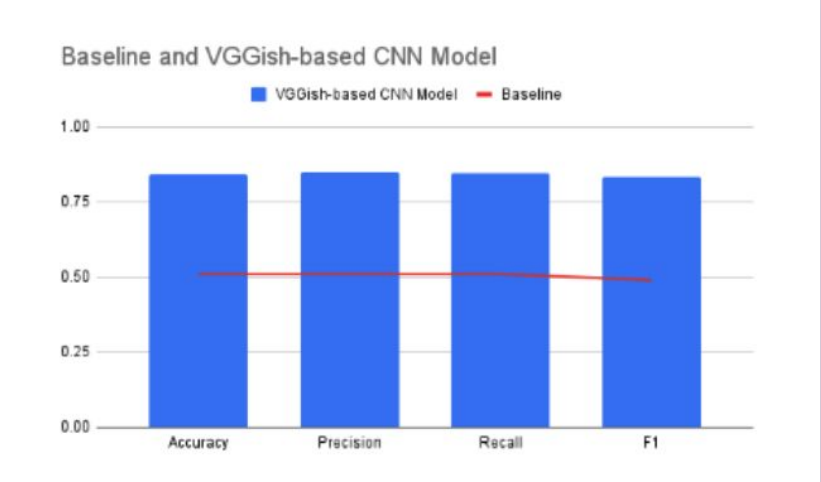
VGGish

- VGGish is a pre-trained CNN model designed for audio classification tasks
- Convert audio features into a 128-dimensional embedding
- We will leverage the VGGish for our CNN model

Results

VGGish Model Performance:

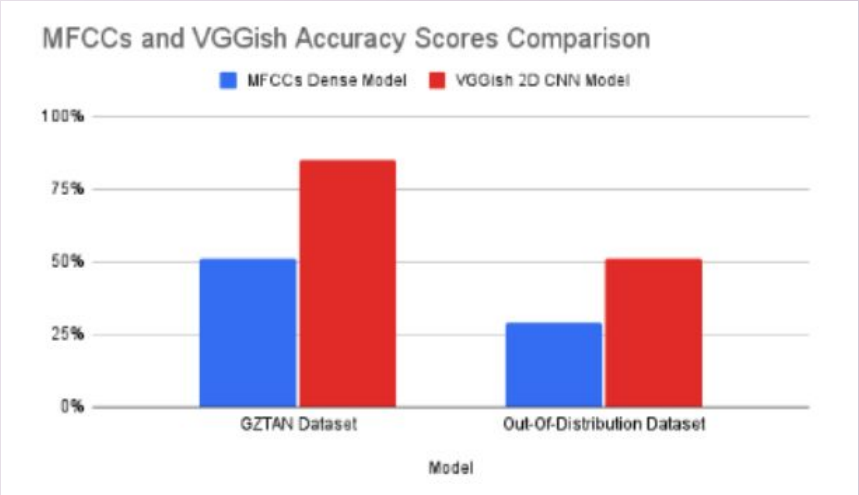
- Achieved an impressive accuracy of 85%
- Outperformed the baseline MFCCs-based model (51% accuracy)



Test on Unseen Data

- VGGish-based CNN model: 51% accuracy
- MFCCs-based Dense model: 29% accuracy

⇒ Performance drop due to differences between the new dataset and GTZAN (e.g., content, distribution, recording quality)



Thank you

for your attention !