Comparing the Performance of Traditional Deep Learning Models on the GTZAN Dataset

Tianhao Liu
School of Computer Science
University College Dublin
tianhao.liu@ucdconnect.ie

Abstract

In recent years, deep learning models have gained widespread adoption for music genre classification tasks, demonstrating superior performance in handling complex data. As a newcomer to the field of deep learning, I am intrigued by exploring the effectiveness of various traditional deep learning models on the GTZAN dataset, a renowned benchmark dataset for music genre classification tasks. The evaluated models encompass Multilayer Perceptron neural networks, convolutional neural networks (CNNs), and recurrent neural networks (RNNs) such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks. This study aims to deepen my understanding of the characteristics of these different traditional deep learning models and provide insights into the design of future deep learning architectures.

1 Introduction

The GTZAN dataset has been widely utilized as a standard benchmark for evaluating music genre classification tasks. This experiment aims to assess the performance of various deep learning models on the GTZAN dataset. The models range from simple feedforward neural networks to more complex architectures, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs) such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks. The subsequent sections of this paper will delve into related work in the field of music genre classification, followed by a detailed description of the methodology employed in our experiment, presentation of the results obtained, and conclusions drawn from the findings.

2 Related Work

In the field of music genre classification, various traditional deep learning models have been extensively explored. Convolutional neural networks (CNNs), initially successful in image classification, have also shown promise in audio processing (Hershey et al. [1]). Apart from CNNs, different variants of recurrent neural networks (RNNs) have demonstrated strong performance in audio classification tasks. For instance, Banuroopa et al. [2] achieved an impressive 98.8% accuracy using an LSTM model for classifying the UrbanSound8K dataset. Similarly, GRU (Gated Recurrent Unit) networks have exhibited notable efficacy in audio recognition tasks. According to Shubham Khandelwal et al. [3], GRUs surpassed LSTMs across all network depths in a large vocabulary continuous speech recognition task.

While much of these research has been conducted on large audio datasets, the performance of these models on smaller datasets remains unclear, especially on GTZAN database. This experiment aims to compare the effectiveness of these models on the GTZAN database, employing a simple Multilayer Perceptron (MLP) as a baseline for comparison.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

3 Experimental Setup

3.1 Dataset

The GTZAN dataset is a benchmark music genre classification dataset consisting of 1000 audio tracks, each lasting 30 seconds and sampled at 22050Hz. This dataset encompasses 10 distinct genres, with each genre represented by 100 tracks in .wav format. The genres covered include blues, classical, country, disco, hiphop, jazz, metal, pop, reggae, and rock.

3.2 Data pre-process

References

[1] Hershey, Shawn and Chaudhuri, Sourish and Ellis, Daniel P. W. and Gemmeke, Jort F. and Jansen, Aren and Moore, R. Channing and Plakal, Manoj and Platt, Devin and Saurous, Rif A. and Seybold, Bryan and Slaney, Malcolm and Weiss, Ron J. and Wilson, Kevin. (2017) CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 131-135.

[2] Banuroopa, K. and Shanmuga Priyaa, D. (2021) MFCC based hybrid fingerprinting method for audio classification through LSTM. *International Journal of Nonlinear Analysis and Applications*, 12(Special Issue), pp. 2125-2136.

[3] Shubham Khandelwal, Benjamin Lecouteux, Laurent Besacier. COMPARING GRU AND LSTM FOR AUTOMATIC SPEECH RECOGNITION. [Research Report] LIG. 2016. ⟨hal-01633254⟩