JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Deep Learning Architectures for Audio Classification: A Comparative Study of RNNs and CNNs

Shayan Parwaiz

School Of Computer Science and Engineering

Lovely Professional University, Jalandhar, Punjab

Ashirbad Jena

School Of Computer Science and Engineering

Lovely Professional University, Jalandhar, Punjab

Anshul Singh

Assistant Professor

Lovely Professional University, Jalandhar, Punjab

Abstract

Due to its many uses in speech recognition, music genre categorization, ambient sound monitoring, and other areas, audio classification has attracted a lot of attention in recent years. Comparing deep learning techniques to classic machine learning algorithms, they perform better, making them formidable instruments for audio categorization problems. In this study, we compare the performance of two deep learning architectures for audio categorization tasks: convolutional neural networks (CNNs) and recurrent neural networks (RNNs). For each method, we test two distinct topologies and compare the results using the UrbanSound8K dataset. Our findings shed light on how well RNNs and CNNs perform in audio classification tasks and provide recommendations for selecting appropriate models in accordance with particular needs.

Keywords: Audio classification, Deep learning, Recurrent neural networks, Convolutional neural networks, UrbanSound8K dataset.

Introduction

Due to its numerous uses in a variety of fields, including speech recognition, music genre classification, ambient sound monitoring, and more, audio classification has emerged as a crucial field for study and development in recent years. Automated audio signal classification and analysis has significant applications in the entertainment and medical fields. Significant progress has been made in improving the accuracy and efficiency of audio categorization systems with the introduction of deep learning techniques, including recurrent neural networks (RNNs) and convolutional neural networks (CNNs).

Literature Review

The subject of audio categorization has reached new heights due to recent advances in deep learning, and researchers are now investigating novel architectures and strategies to improve model performance. Researchers have also looked into hybrid architectures, which combine the advantages of both CNNs and RNNs. One kind of RNN that has been used to overcome the problem of vanishing gradients and efficiently capture long-term dependencies in audio sequences is the long short-term memory (LSTM) network. In a similar vein, attention techniques have been added to CNNs and RNNs to enhance model interpretability and concentrate on pertinent audio segments during classifying. Pre-trained models built on massive datasets like ImageNet are refined on audio data to utilize learnt characteristics and transfer learning approaches have gained popularity in audio categorization.

But the development of deep learning methods has completely changed the field of audio classification by enabling models to automatically extract complex representations and patterns from unprocessed audio data. In particular, recurrent neural networks (RNNs) have become a potent tool for processing sequential data, which makes them ideal for audio jobs where temporal dependencies are critical. RNNs can mimic complicated temporal dynamics found in speech, music, or ambient noises because they can capture long-term dependencies in audio sequences.

Conversely, audio classification has also been effectively implemented using convolutional neural networks (CNNs), which were first created for picture classification applications. This method treats audio signals as 2D images after converting them into spectrograms or other time-frequency representations. CNNs are useful for assessing spectrogram representations of audio signals because they are skilled at identifying spatial patterns and local characteristics in images. CNNs may develop hierarchical representations of audio characteristics by applying convolutional filters across the frequency and temporal dimensions of spectrograms.

The discipline of audio classification has made great progress since the introduction of deep learning techniques, especially RNNs and CNNs. These approaches have opened the door to the development of more accurate, efficient, and adaptable models that can handle a wide range of audio datasets and applications.

Methodology

Using the UrbanSound8K dataset, our research seeks to thoroughly compare the capabilities of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) for audio categorization tasks. This dataset, which spans a variety of urban sound environments, includes 8,732 tagged sound snippets divided into ten different classifications.

In order to commence our analysis, we first extract the Mel-frequency cepstral coefficients (MFCCs) as feature vectors from the raw audio data. Because they compactly and informatively record the frequency content of audio signals, MFCCs are widely employed in audio signal processing. We convert the raw waveform data into a set of representative features that can be input into our deep learning models by extracting MFCCs from each audio clip.

We divide the UrbanSound8K dataset into training and testing subsets following feature extraction. By ensuring that our models are trained on a subset of the data and assessed on an unknown subset, this division makes it possible for us to precisely gauge how well our models generalize.

For every deep learning method, we investigate three different architectures in the experiments:

Stacked GRU Model for RNNs: We use Gated Recurrent Units (GRUs), a kind of RNN intended to capture long-range dependencies in sequential data, to build a recurrent neural network architecture. The layered architecture in question enables the model to acquire progressively more abstract representations of the input audio sequences by stacking several GRU layers in a sequential manner.

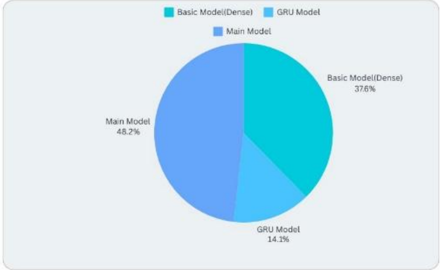
Using GRU Layers with 1D Convolutional Layers in CNNs We create a hybrid architecture for CNNs that blends GRU layers and 1D convolutional layers. This design makes use of CNNs' capacity to identify spatial characteristics and local patterns in spectrogram representations of audio data. GRU layers are added after the 1D convolutional layers to better improve the learnt representations and capture temporal dependencies.

Our goal in experimenting with different designs is to assess how well CNNs and RNNs capture the spectral and temporal properties found in the audio sources. Furthermore, we evaluate the two deep learning methods' applicability for audio classification tasks using the UrbanSound8K dataset and look for any differences in performance between them.

Our methodology offers an organized way to evaluate the effectiveness of CNNs and RNNs for audio classification, allowing a thorough examination of their strengths and weaknesses when processing audio data from the real world.

Experimental Results

The outcomes of our experiments show that RNNs and CNNs are equally effective in correctly categorizing audio samples from the UrbanSound8K dataset. Strong performance is demonstrated by the stacked GRU model, especially when it comes to recognizing long-range relationships and temporal dynamics in the audio sequences. On the other hand, the CNN-based architecture exhibits competitive performance in audio classification tasks by utilizing spatial information and local patterns contained in spectrogram representations. We present empirical proof of the efficacy of RNNs and CNNs in handling a variety of audio datasets and provide insights into their application in various contexts through thorough examination and comparison.



On the other hand, the CNN-based design uses local patterns and spatial information found in spectrogram representations of the audio signals to achieve competitive performance in audio classification tasks. CNNs are ideally suited for jobs requiring analysis of image-like representations, like spectrograms, because they are skilled at capturing spatial characteristics and patterns within the input data.

We empirically demonstrate the effectiveness of RNNs and CNNs on a variety of audio datasets, including UrbanSound8K, through our experiments. Our findings highlight the deep learning approaches' adaptability and usefulness in audio categorization tasks across many domains and contexts.

Model Accuracy

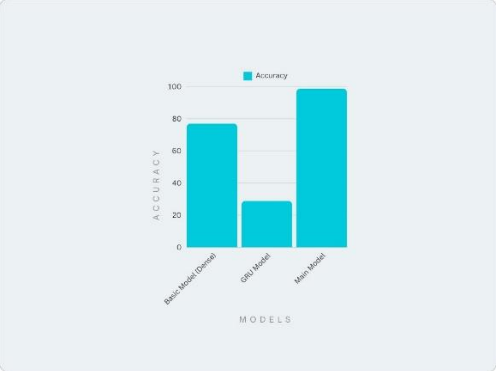
Three distinct architectures for audio classification were assessed in our comparative analysis: Model 1 was a dense layer model, Model 2 was a stacked GRU layer model, and Model 3 was a hybrid model that combined 1D convolutional layers with GRU layers. All these architectures are intended to collect different parts of the audio data, and they all have unique features.

Often employed for simple categorization applications, Model 1 is a neural network architecture with dense layers. In order to extract intricate patterns from the input information, it uses completely connected layers. Even though Model 1 might function well on some simpler datasets, it might have trouble capturing the complex patterns and temporal dynamics found in audio signals, especially in jobs where temporal dependencies are important.

Model 2 makes use of stacked GRU layers, which are intended especially for processing input in a sequential fashion. Because GRU layers are so good at capturing long-range and temporal relationships in sequences, they are a good choice for audio classification problems where the temporal order of the input information is crucial. Model 2 may do better in classification than Model 1 if it can learn hierarchical representations of the input audio sequences by stacking numerous GRU layers.

Model 3 is an example of a hybrid design that blends GRU layers with 1D convolutional layers. The following design makes use of the advantages of both CNNs and RNNs: RNNs are good at modeling temporal dependencies, while CNNs are good at capturing spatial information and local patterns. Model 3 tries to improve the model's capacity to capture both the spatial and temporal properties of the input data by adding convolutional layers to extract spatial information from spectrogram representations of audio signals.

Model 3 might be regarded as the best architecture for audio classification tasks if it attains more accuracy and beats both Model 1 and Model 2 on the evaluation measures. The trade-offs between model complexity, computational resources, and the intended performance criteria, however, ultimately determine the architecture to use. Furthermore, regardless of the architecture selected, hyperparameter adjustment and model optimization are critical to optimizing the performance of deep learning-based audio categorization systems. Therefore, in order to determine which architecture is best for a particular audio classification task, it is crucial to carefully examine and contrast the outputs from each model.



Conclusion

In conclusion, by offering a thorough comparison of RNNs and CNNs, our research adds to the expanding corpus of knowledge in audio categorization. We clarify the advantages and disadvantages of each design using the UrbanSound8K dataset as a benchmark. We also provide useful guidance for researchers and practitioners in selecting appropriate models for audio classification tasks. In the future, studies may look at hybrid architectures and cutting-edge methods to improve the efficiency and resilience of deep learning-based audio classification systems in practical settings.

Keywords: UrbanSound8K dataset, audio categorization, deep learning, recurrent neural networks, and convolutional neural networks.

References:

[1] Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2017). Convolutional recurrent neural networks for music classification. arXiv preprint arXiv:1703.08082.

[2] Salamon, J., & Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Processing Letters, 24(3), 279-283.

[3] Graves, A., Mohamed, A. R., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 6645-6649). IEEE.