Comparing CNN and LSTM Architectures for Environmental Sound Classification

Luka Kurešević

Abstract

Environmental sound classification is a key task in auditory scene analysis with applications in security, surveillance, and human-computer interaction. In this paper, we compare the performance of Convolutional Neural Networks (CNNs) based on the ResNet architecture (ResNet18, ResNet34, ResNet50, and ResNet101) with Recurrent Neural Networks (RNNs) using Long Short-Term Memory (LSTM) cells for sound classification on the UrbanSound8K dataset. The audio data was preprocessed into Mel spectrograms, providing a time-frequency representation as input to the models. Through experimentation, we assess model accuracy and loss across training and testing phases to evaluate efficacy of different approaches for environmental sound classification.

1 Introduction

Environmental sound classification (ESC) is a crucial component of computer audition systems that aim to interpret everyday sounds. It finds applications in areas such as smart homes, autonomous vehicles, and environmental monitoring, where identifying sounds like sirens, dog barks, and car horns is essential.

In this study, we explore the UrbanSound8K dataset, which contains 8732 labeled audio samples from 10 different sound classes, including gun shots, children playing, and street music. We aim to compare the efficacy of deep learning models, particularly convolutional and recurrent networks, in classifying these environmental sounds. Specifically, we evaluate ResNet-based CNNs (ResNet18, ResNet34, ResNet50, and ResNet101) and LSTMs to understand their relative performance.

2 Dataset Overview

The UrbanSound8K dataset is a widely used benchmark for environmental sound classification tasks. It contains 8732 audio files, each labeled as one of 10 different classes. The dataset is organized into 10 folds, allowing for cross-validation. All excerpts are taken from field recordings uploaded to www.freesound.org, hence the dataset is particularly challenging due to the high variability in sound recordings, both in terms of acoustic environments and recording devices.

2.1 Audio Preprocessing

For our experiments, each audio sample in the UrbanSound8K dataset was converted into a Mel spectrogram. The spectrogram provides a time-frequency representation of the sound, capturing both temporal dynamics and frequency content, which are to be used for discriminating between different sound classes. Key parameters include:

- **n_fft:** 1024 (size of fast Fourier transforms window)
- **hop_length:** 512 (distance between frames)
- **n_mels:** 128 (number of Mel bands)

3 Model Selection and Motivation

For this study, we selected both convolutional and recurrent neural network architectures for their distinct strengths in capturing different aspects of audio data.

3.1 ResNet Architectures

ResNet CNNs are known for their powerful feature extraction capabilities due to their deep architectures and residual connections, and have proven to perform well at image classification tasks. We considered four variants:

- **ResNet18:** A relatively shallow network with 18 layers, often used as a baseline due to its lower computational complexity.
- **ResNet34:** A slightly deeper version, still employing basic residual blocks but with increased depth for better feature extraction.
- **ResNet50:** Employs bottleneck blocks, increasing depth and feature representation without a proportional increase in computation.
- **ResNet101:** The deepest model in our comparison, with 101 layers and bottleneck blocks, aimed at extracting the most detailed features.

3.1.1 Preliminary experiments

During preliminary experiments, both ResNet50 and ResNet101 showed a tendency towards overfitting, with ResNet101 struggling to reach accuracy above 50% outside its training set. In the case of our relatively small dataset, this opposed the hypothesis that more complex networks extract more detailed features. For this reason, we dropped ResNet101 out of consideration for final training cycles, but stuck with ResNet50 in case that it generalizes better over a longer training period.

When comparing efficiency of ResNet18 and ResNet34 over a smaller number of training epochs, the simpler model converged faster once again, regardless of different hyperparameters. Hence, ResNet18 was chosen to be trained in the final experiment.

3.2 Long Short-Term Memory (LSTMs) Networks

LSTMs are designed to remember long-term dependencies, making them effective at modeling sequential data, which audio is. For this task, we configured the LSTM models with two different hidden layer sizes (64 and 128 units) and used two LSTM layers:

- **LSTM (64 units, 2 layers):** A smaller LSTM architecture to capture temporal dependencies with fewer parameters.
- **LSTM (128 units, 2 layers):** A larger LSTM configuration to better capture complex dependencies in sound sequences.

3.2.1 Preliminary experiments

Having noticed a tendency of CNNs to overfit training data across trial experiments, we configured the LSTMs with a dropout rate of 0.2 - this improved test accuracy over a smaller number of epochs and effectively prevented the model from overfitting training sets.

Introduction of more LSTM cells per model was considered, but ultimately disregarded as it failed to show improvements that would justify the increase in training time.

4 Experimental Setup

The models were trained using 10-fold cross-validation, over 30 epochs in each fold (the ResNet model experiments were cut short, due to evident overfitting). In this way, we attempt to account for the unevenness of the dataset, as well as comply with the established benchmarks.We monitored both training accuracy and loss, as well as test accuracy and loss at each epoch. The key hyperparameters used in the experiments were:

- **Batch size:** 128
- **Optimizer:** Adam optimizer with a learning rate of 0.001
- **Loss function:** Cross-entropy loss

Following are graphs depicting the worst and best performing folds for each model:

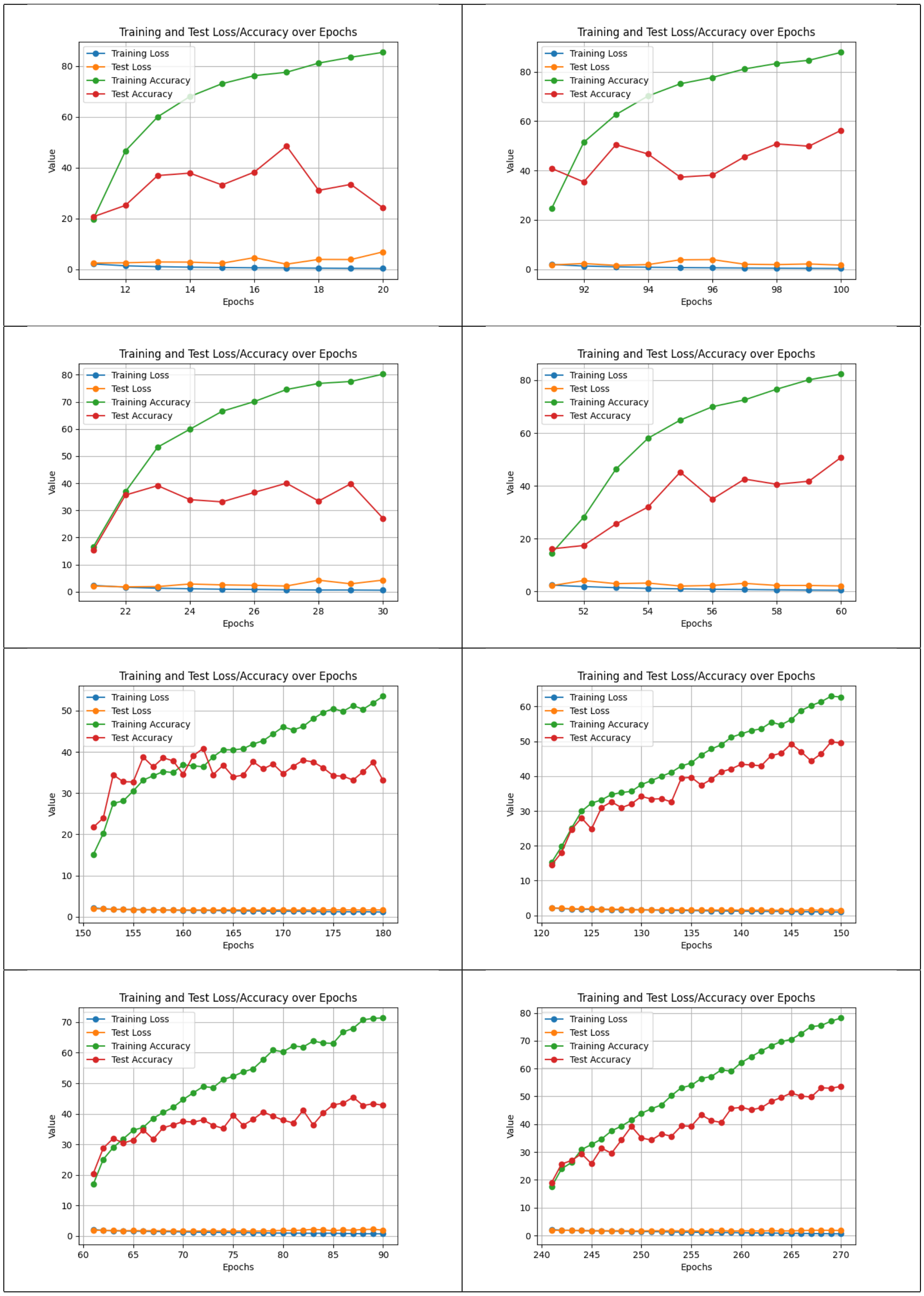

Figure 1: Table of 8 images arranged in 2 rows and 4 columns.

5 Results

Model	**Train Acc (%)**	**Train Loss**	**Test Acc (%)**	**Test Loss**
ResNet18	**85.99**	0.38	42.56	3.46
ResNet50	80.36	0.55	37.33	4.05
LSTM (64, 2 layers)	55.96	1.12	41.64	1.65
LSTM (128, 2 layers)	76.37	0.65	**47.76**8	1.88

Table 1: Training and Validation Accuracy and Loss for ResNet and LSTM Models

6 Discussion

Upon examining the graphs and tabular data, it is evident that Convolutional Neural Networks (CNNs) tend to overfit the data, whereas Long Short-Term Memory networks (LSTMs) demonstrate better generalization capabilities. This improved generalization in LSTMs can be partly attributed to their configuration, which included measures to mitigate overfitting, such as a dropout rate of 0.2.

When comparing ResNet18 and ResNet50, both models exhibited overfitting to the training data, resulting in highly variable accuracy across different epochs. For relatively small datasets, such as UrbanSound8K, shallower models like ResNet18 tend to perform better, as evidenced by their ability to achieve higher test accuracy scores. The performance of ResNet models could potentially be improved in future work by incorporating dropout layers, weight decay, and learning rate scheduling. The lower test accuracy and higher variance observed in the ResNet models may be attributed to their ability to effectively extract spatial features, which does not pair well with the size of our specific dataset.

The LSTM models managed to avoid significant overfitting and predicted testing data with higher precision, which can be partly credited to the use of dropout and partly to their configurations presenting inherently better balance between model complexity and dataset size, allowing for superior generalization, which the lower variance of prediction accuracy across neighbouring epochs demonstrates.

It is important to note that more reliable test accuracy would be extracted from evaluating on environmental sounds outside of the US8K dataset, and that somewhat lower accuracy when predicting labels of real world data is to be expected. Although containing complex data, the UrbanSound8K dataset is extremely unlikely large enough for reliable generalization.

7 Conclusion

In summary, while ResNet models excel at spatial feature extraction in spectrogram-based data, LSTMs demonstrate a superior ability to capture temporal dynamics. Taking steps to combat overfitting proved to significantly impact validation accuracy. For environmental sound classification tasks, a hybrid approach that integrates the strengths of both architectures could further improve performance.

References

[1] UrbanSound8K Dataset, Available: https://urbansounddataset.weebly.com/urbansound8k.html, Accessed: [September 9, 2024].

[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016. Available: https://www.deeplearningbook.org/, Accessed: [September 9, 2024].

[3] A. Khanna, ”Understanding LSTM for Sequence Classification: A Practical Guide with PyTorch,” Medium, 2023. Accessed: [September 9, 2024].

[4] E. B. Thomas, ”Understanding LSTM: An In-depth Look at its Architecture, Functioning, and Pros & Cons,” Medium, 2023. Accessed: [September 9, 2024].

[5] GeeksforGeeks, ”Residual Networks (ResNet) - Deep Learning,” Available: https://www.geeksforgeeks.org/residual-networks-resnet-deep-learning/, Accessed: [September 9, 2024].

[6] PyTorch Documentation, Available: https://pytorch.org/docs/stable/index.html, Accessed: [September 9, 2024].