Proceedings of the 2nd International Conference on Functional Materials and Civil Engineering
DOI: 10.54254/2755-2721/89/20241128

Neural Networks for Audio Classification: Multi-scale CNN-LSTM Approach to Animal Sound Recognition

Cong Xu

Hong Kong University, Hong Kong, China

u3630742@connect.hku.hk

Abstract. This research explores the application of neural networks, specifically CNN-LSTM models, for classifying sound signals from dogs, frogs, and cats, selected from the ESC-50 dataset. The sound data was preprocessed using Mel-frequency cepstral coefficients (MFCCs) and augmented through time stretching, pitch shifting, and noise addition to enhance model generalization in varied acoustic environments. We compared two deep learning models: a traditional CNN-LSTM and an improved version with multi-scale feature extraction, allowing for capturing both short-term and long-term sound patterns. Our findings show that the multi-scale CNN-LSTM architecture outperforms the traditional CNN-LSTM, achieving a test accuracy of 86.11% compared to 80.56%. These results highlight the effectiveness of multi-scale feature extraction for handling complex audio signals. This research offers valuable insights into bioacoustics and has broader applications in areas such as environmental sound monitoring, wildlife preservation, and animal behavior analysis.

Keywords: Audio classification, CNN-LSTM, Multi-scale convolution, Animal sound recognition, MFCC.

1. Introduction

This study investigates how neural networks can be used to classify sound signals, with a focus on distinguishing between sounds made by dogs, frogs, and cats. The analysis utilizes the ESC-50 dataset [1], which is a comprehensive collection of environmental sound recordings encompassing 50 different categories. For this research, we specifically selected three sound types—dog barking, frog croaking, and cat meowing—because their unique acoustic features make them ideal for assessing the effectiveness of classification models.

To improve the models' ability to generalize, we employed several data augmentation techniques such as time stretching, pitch shifting, and noise addition [2]. These methods introduce variability into the dataset, better reflecting real-world conditions and helping the models adapt to various acoustic environments and distortions. For preprocessing, we utilized Mel-frequency cepstral coefficients (MFCCs) [3], a common feature extraction approach in audio analysis that effectively captures key spectral features. Further preprocessing included standardization and normalization to enhance the models' performance and stability.

We conducted a comparison between two deep learning models: a traditional CNN+LSTM[4] model and an enhanced version that integrates multi-scale feature extraction within the CNN+LSTM framework. The multi-scale convolution technique enables the model to capture audio features across

DOI: 10.54254/2755-2721/89/20241128

various temporal resolutions, which enhances its ability to detect both short and long-duration sound patterns. Our results show that neural networks, especially those with multi-scale feature extraction, are highly effective in classifying animal sound patterns, such as those from dogs, frogs, and cats. These insights contribute to the broader fields of bioacoustics, environmental sound monitoring, wildlife conservation, and the study of animal behavior through audio analysis [5].

2. Model Performance and Analysis

2.1. Preprocessing

The preprocessing workflow for this audio classification task begins with loading sound data from specific folders, each representing a different class (dogs, frogs, and cats). The audio files are read using the librosa library [6], which is widely adopted for audio signal processing in machine learning research. After loading, the data is normalized to a range of [-1, 1] to ensure uniform amplitude scaling across different audio samples. To improve feature consistency, the signals are centered by shifting the maximum amplitude to the middle of the signal.

Next, Mel-frequency cepstral coefficients (MFCCs) are extracted, capturing essential frequency characteristics of the audio signals. MFCCs have been shown to be highly effective in tasks like speech recognition and music classification, making them a popular choice for sound-related tasks [7]. These MFCC features are normalized to ensure zero mean and unit variance across the dataset, stabilizing the training process and ensuring faster convergence.

To improve generalization and reduce overfitting, several data augmentation techniques were applied. These include adding random noise, shifting the signal in time, pitch shifting, and time stretching [8]. These augmentations were applied randomly to the training, validation, and test sets, creating a more robust dataset for model training. The dataset was divided into 30 files for training, 5 for validation, and 5 for testing. Finally, to ensure consistent input dimensions for the neural network, the extracted MFCC features were padded to a fixed length using the pad_sequences function from TensorFlow [9].

The preprocessed data, including the augmented training, validation, and test sets, was saved in .npz format for easy loading into the neural network models. This comprehensive preprocessing pipeline ensures that the data is clean, diverse, and structured for efficient model training and evaluation

Training data shape: (1470, 256, 13), Labels shape: (1470, 3)

Validation data shape: (36, 256, 13), Labels shape: (36, 3)

Test data shape: (36, 256, 13), Labels shape: (36, 3)

2.2. Results and Discussion

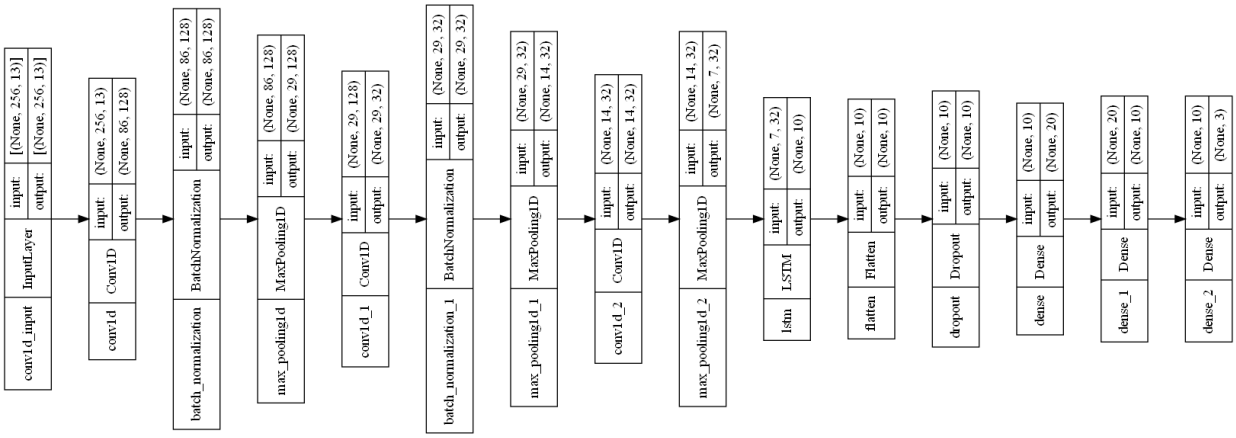

Figure 1. CNN-LSTM Architecture for Animal Sound Classification

DOI: 10.54254/2755-2721/89/20241128

The CNN-LSTM model combines the strengths of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks to extract both spatial and temporal features from the audio data. The convolutional layers in this model are responsible for capturing localized patterns, such as short-duration sounds like barking, croaking, or meowing, which are typical of animals. This makes CNNs effective in identifying such short bursts of audio signals. The subsequent LSTM layers, on the other hand, process these time-ordered features, helping to capture long-term dependencies and sequential characteristics of the sound data.

During the training process, this CNN-LSTM model demonstrated robust learning behavior. The accuracy results, as depicted in Figure 2, show that the model converges well, achieving a test accuracy of 80.56%. This is a solid performance given the complexity of animal sound classification tasks, which often involve subtle differences between sound classes. The relatively high test accuracy suggests that the model was effective at generalizing from the training data to unseen data.

However, the training and validation accuracy curves reveal some important trends. For example, the training accuracy increased steadily across epochs, but the validation accuracy plateaued early, suggesting the model may have reached the limits of its generalization ability with the given architecture. This plateau indicates that the model might be somewhat under-optimized in capturing long-term sequential dependencies, or it may have benefitted from additional regularization techniques, such as early stopping or further dropout [10].

In terms of loss, as shown in Figure 3, the training and validation loss decreased in tandem during the initial epochs, indicating smooth convergence. However, after a certain point, the validation loss stagnated, implying that the model's capacity to improve further on unseen data was limited. This could suggest that while the CNN-LSTM combination performs well for localized feature extraction, it may struggle to fully capture the intricate temporal relationships in the audio data, especially when it comes to long-duration sounds or subtle variations between classes.

Best model test accuracy: 0.8056

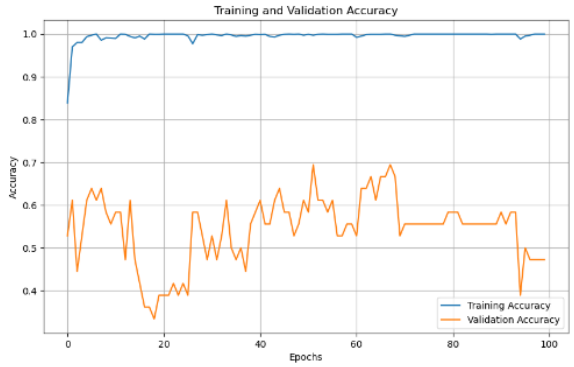

Figure 2. Training and Validation Accuracy Over Epochs for Audio Classification Model

Figure 3. Training and Validation Loss Over Epochs for Audio Classification Model

DOI: 10.54254/2755-2721/89/20241128

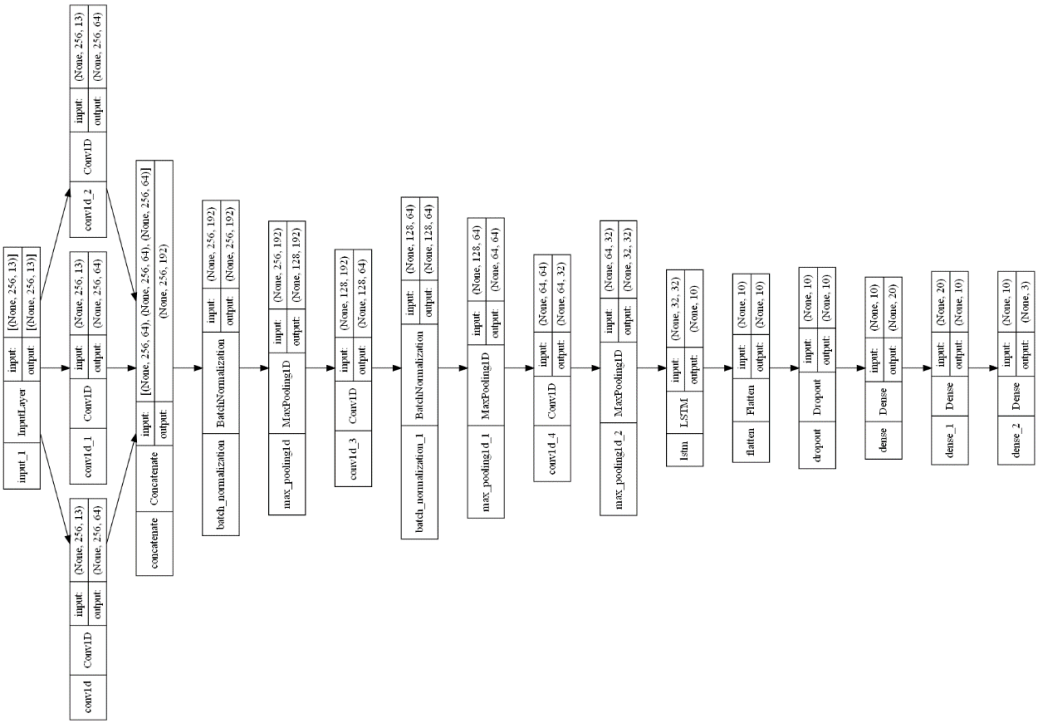

Figure 4. Architecture of the CNN-LSTM Model for Audio Classification

To address the limitations observed in the traditional CNN-LSTM model, the multi-scale CNN-LSTM architecture introduces convolutional layers with multiple kernel sizes to capture features at various temporal resolutions, shown in Figure 4. This design improvement enables the model to handle both short and long-term dependencies in the audio signals more effectively, as each kernel size targets different time-scales of the sound data. Smaller kernels focus on finer, short-term details (such as quick animal sounds), while larger kernels capture broader, long-term patterns that may span over larger timeframes in the audio sequence.

As reflected in the results, the multi-scale CNN-LSTM model achieved a higher test accuracy of 86.11%, significantly outperforming the traditional CNN-LSTM model. The introduction of multiple kernel sizes played a key role in improving feature extraction, allowing the model to be more versatile in learning from diverse sound signals. This performance boost can be attributed to the model's ability to simultaneously learn from both detailed and high-level temporal patterns, offering a more comprehensive representation of the sound data.

Figure 5 illustrates the training and validation loss over epochs for the multi-scale CNN-LSTM model. Here, the validation loss continues to decrease more steadily than in the standard CNN-LSTM model, indicating that the model generalizes better to unseen data and is less prone to overfitting. This is supported by the smaller gap between training and validation losses, suggesting that the multi-scale architecture promotes better generalization across different data sets. The regularization techniques applied, such as dropout, also help mitigate overfitting by reducing the likelihood that the model memorizes specific patterns rather than learning generalizable features.

In Figure 6, the training and validation accuracy curves for the multi-scale CNN-LSTM show a smoother, more sustained improvement throughout the epochs. Unlike the traditional CNN-LSTM model, where validation accuracy stagnated early on, the multi-scale model consistently improved over time, highlighting its enhanced capacity to adapt to varying sound characteristics. This suggests that the model's multi-scale feature extraction ability enables it to better capture the complexities inherent in audio data.

DOI: 10.54254/2755-2721/89/20241128

The multi-scale CNN-LSTM model's superior performance can be attributed to its ability to simultaneously analyze both short-term and long-term temporal features. The architecture's flexibility in capturing various temporal scales allows it to better differentiate between sound classes, leading to improved classification accuracy, particularly in more challenging cases where the audio patterns of different classes overlap or exhibit subtle differences.

Best model test accuracy: 0.8611

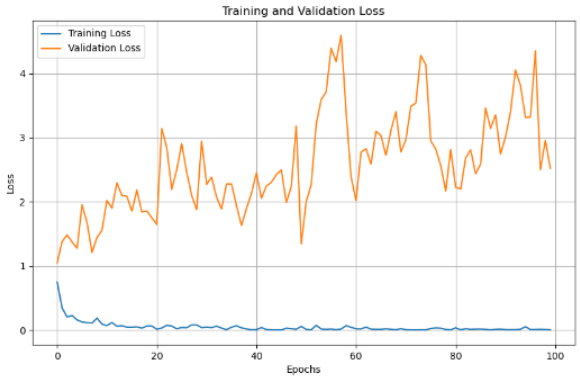

Figure 5. Training and Validation Loss Over Epochs for Multi-scale CNN-LSTM Model

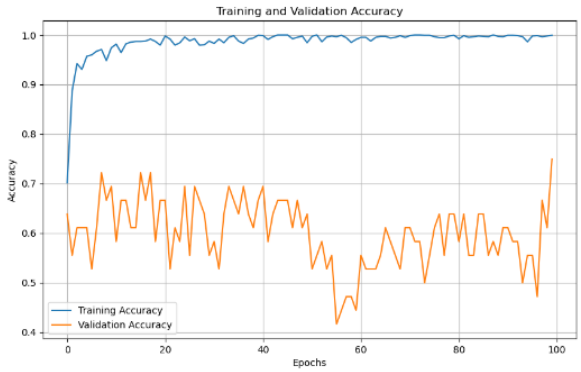

Figure 6. Training and Validation Accuracy Over Epochs for Multi-scale CNN-LSTM Model

3. Conclusion

The implementation of the multi-scale CNN-LSTM architecture significantly boosted classification accuracy. By integrating multi-scale feature extraction—where convolutional layers with different kernel sizes capture sound patterns across multiple temporal scales—the model's ability to detect and distinguish between various sound signals improved considerably.

In comparison to the original CNN-LSTM model, which achieved an accuracy of 80%, the multi-scale version raised the accuracy to 86%. This increase is due to the model's enhanced ability to learn both short-term and long-term temporal features, leading to a more detailed representation of the audio signals. The multi-scale convolution layers enabled the model to better adapt to different sound patterns, which improved its generalization on the test data. These results underscore the effectiveness of multi-scale feature extraction in processing complex audio data and demonstrate its potential to enhance performance in sound classification tasks.

References

[1] K. J. Piczak. ESC: Dataset for Environmental Sound Classification. Proceedings of the 23rd Annual ACM Conference on Multimedia, Brisbane, Australia, 2015.

DOI: 10.54254/2755-2721/89/20241128

[2] McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in python. In Proceedings of the 14th python in science conference (pp. 18-25).

[3] Gourisaria, M.K., Agrawal, R., Sahni, M. et al. Comparative analysis of audio classification with MFCC and STFT features using machine learning techniques. Discov Internet Things 4, 1 (2024). https://doi.org/10.1007/s43926-023-00049-y

[4] Swapna, G., Kp, S., & Vinayakumar, R. (2018). Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals. Procedia computer science, 132, 1253-1262.

[5] Stowell, D., & Plumbley, M. D. (2014). Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. PeerJ, 2, e488.

[6] Shanmuga Sundari, M., Priya, K. S. S., Haripriya, N., & Sree, V. N. (2023, March). Music genre classification using librosa implementation in convolutional neural network. In Proceedings of Fourth International Conference on Computer and Communication Technologies: IC3T 2022 (pp. 583-591). Singapore: Springer Nature Singapore.

[7] Prabakaran, D., & Sriuppili, S. (2021). Speech processing: MFCC based feature extraction techniques-an investigation. In Journal of Physics: Conference Series (Vol. 1717, No. 1, p. 012009). IOP Publishing.

[8] Ko, T., Peddinti, V., Povey, D., & Khudanpur, S. (2015, September). Audio augmentation for speech recognition. In Interspeech (Vol. 2015, p. 3586).

[9] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... & Zheng, X. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467.

[10] Dibbo, Sayanton V., et al. "Lcanets++: Robust audio classification using multi-layer neural networks with lateral competition." 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW). IEEE, 2024.