Urban Sound Classification using Convolutional Neural Networks

Stage Progress Report

Jinghuai Tang

SID 520378067

Github ID: JinghuaiTang https://github.com/JinghuaiTang/elec5305-project-520378067

12/10/2025

1 Introduction

In the initial stage of this project, the main idea was to use neural networks to recognize and extract special sound signals. However, this concept was too broad and theoretical, which made it difficult to identify a clear research objective or evaluate progress in a measurable way. The feedback received pointed out that the project lacked a specific target and needed to focus on a concrete application supported by a suitable dataset and baseline model. In response to this, the project has now been redefined with a clear and achievable goal: to classify different types of urban environmental sounds using a Convolutional Neural Network (CNN). This updated direction transforms the idea from a general exploration of sound recognition into a focused study with real data, clear metrics, and practical outcomes.

Urban sound classification represents a relevant and realistic application of deep learning in the field of acoustic signal processing. In modern cities, environmental sounds such as sirens, car horns, dog barks, drilling, and street music carry significant contextual information. Automatically identifying these sounds can benefit public safety systems, smart city monitoring, and environmental noise management. Therefore, this new project direction not only gives the work a well-defined scope but also connects it to meaningful real-world applications that can demonstrate the practical value of neural networks in sound analysis.

2 New Focus Points

The updated project now has a clear research objective, a fixed dataset, and a practical implementation plan. The goal is to develop a CNN-based classifier that can recognize multiple categories of urban sounds from short audio clips. The project adopts the *UrbanSound8K* dataset, which contains 8,732 labeled audio samples distributed across ten everyday urban sound classes such as sirens, car horns, drilling, and street music. These sounds were recorded in real environments, providing natural variations in background noise and loudness, making the dataset suitable for deep learning experiments. Unlike the previous plan, which proposed to analyze sound signals in general, this refined project establishes a specific and measurable target: to train a model that can distinguish among sound types with quantifiable accuracy and robustness. [1]

The motivation for this change is twofold. First, focusing on environmental sounds makes the project realistic within the available time frame and computational resources, while still maintaining academic and practical value. Second, this direction allows for reproducibility and fair comparison with existing studies, since the dataset and baseline CNN models are publicly available. By working on a well-defined dataset and objective, the project becomes technically feasible and academically meaningful, bridging theoretical understanding with hands-on experimentation in sound signal analysis.

3 Methodology Overview

The new methodology is based on a deep learning workflow that integrates data preparation, feature extraction, model design, training, and performance evaluation. The data preparation process involves downloading and organizing the UrbanSound8K dataset, converting each audio file into a time-frequency representation such as a spectrogram or Mel-Frequency Cepstral Coefficients (MFCCs). These visual-like representations capture both temporal and spectral characteristics of sounds, allowing neural networks to learn distinctive patterns between different classes. Additional preprocessing, such as normalization, silence trimming, and data augmentation through pitch shifting or time stretching, is applied to improve the model's robustness. [2]

The model design centers around a Convolutional Neural Network implemented in Python using the PyTorch framework. The CNN architecture is chosen because of its strong ability to detect local features in spectrograms, similar to how it processes spatial features in images. The network includes convolutional and pooling layers for hierarchical feature extraction, followed by fully connected layers for classification. During training, the cross-entropy loss function is used to measure the difference between predicted and actual labels, and optimization is performed using the Adam algorithm. Mini-batch training and validation monitoring are employed to avoid overfitting, while hyperparameters such as learning rate, batch size, and number of epochs are fine-tuned to achieve better generalization. The performance evaluation will rely on multiple metrics, including accuracy, precision, recall, and F1-score, along with confusion matrices that visualize class-level prediction results. This step-by-step methodology ensures that the project remains structured, reproducible, and technically achievable within the course timeline.

4 Current Progress

At the current stage, substantial progress has been made toward the implementation of the refined project plan. The UrbanSound8K dataset has been successfully downloaded, and an initial inspection of metadata confirmed the presence of ten balanced classes with diverse acoustic characteristics. Preprocessing scripts have been developed using the Librosa library to extract MFCCs and spectrograms from each sound clip, and normalization techniques have been tested to stabilize input data. A simple baseline CNN model has been built and compiled in PyTorch, forming the foundation for the classification experiments. Preliminary training was conducted on a smaller subset of the data to verify the model pipeline, and the initial results achieved around 60% classification accuracy after one epoch. Although these are only early results, they indicate that the system design and workflow are functioning correctly. [3]

To implement the new project direction and verify its feasibility, I wrote and uploaded a baseline code file capable of performing initial training and evaluation. This code completes a full workflow on the UrbanSound8K dataset, including audio loading, fixed-length trimming and padding, Log-Mel feature extraction and normalization, as well as a lightweight CNN model implemented in PyTorch for training, validation, and testing. It also produces a confusion matrix and a classification report, which allow quick inspection of class-level recognition accuracy and provide a basis for future improvement.

The baseline includes simple data augmentation techniques such as time shifting and additive noise and provides clear hyperparameter entries (e.g., learning rate, batch size, training epochs, number of Mel filters) so that the experiment can be scaled up to the full dataset. The structure allows for further optimization through regularization, enhanced augmentation strategies, or deeper model designs.

This code file has now been added to the GitHub repository as the foundation for future extensions and performance improvements.

5 Next Steps

The next phase of the project will involve completing the full training and evaluation process while improving model performance and robustness. The CNN model will be trained on all folds of the UrbanSound8K dataset to ensure comprehensive coverage of sound variations. Experiments will be conducted to test different architectures, including variations in convolutional depth and filter size, to observe how they affect accuracy and generalization. Once training is completed, results will be visualized using confusion matrices and accuracy curves, and feature activation maps will be analyzed to understand what the network learns from the sound representations. [4]

Furthermore, a comparison will be made between the CNN model and a traditional baseline approach that uses MFCC features with a Support Vector Machine (SVM) classifier. This comparison will help highlight the advantages of deep learning over conventional methods in terms of performance and adaptability. Finally, all findings will be compiled into a detailed report, supported by visual and quantitative results, to demonstrate that the project meets its intended goals. The completion of these steps will mark the transition from an exploratory idea to a fully realized and assessable research project. [5]

6 Reflection on Feedback

Reflecting on the feedback received from the earlier stage, it is clear that the most significant improvement lies in defining a focused and achievable objective. The previous proposal introduced an interesting but vague concept, while this new version specifies a clear dataset, methodology, and evaluation process. By narrowing the scope to urban sound classification, the project now aligns well with the course expectations and ensures tangible outcomes. The introduction of the Urban-Sound8K dataset, along with the CNN-based model design, provides both a practical foundation and measurable progress criteria.

This refinement has made the project much more coherent and purposeful. The process of revising the objective and redefining the scope has also been a valuable learning experience. It demonstrated that in engineering research, a successful project often depends more on clarity of purpose and feasibility than on theoretical ambition. Through this experience, I have gained a better understanding of how to translate a broad concept into a concrete technical plan, and I am confident that the project will now continue to develop along a productive and achievable path.

References

[1] J. Cao, M. Cao, J. Wang, C. Yin, D. Wang, and P.-P. Vidal, "Urban noise recognition with convolutional neural network," *Multimedia Tools and Applications*, vol. 78, no. 20, pp. 29 021–29 041, 2019.

[2] J. K. Das, A. Ghosh, A. K. Pal, S. Dutta, and A. Chakrabarty, "Urban sound classification using convolutional neural network and long short term memory based on multiple features," in *2020 fourth international conference on intelligent computing in data sciences (ICDS)*. IEEE, 2020, pp. 1–9.

[3] K. Jaiswal and D. K. Patel, "Sound classification using convolutional neural networks," in *2018 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)*. IEEE, 2018, pp. 81–84.

[4] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th international workshop on machine learning for signal processing (MLSP)*. IEEE, 2015, pp. 1–6.

[5] J. Sang, S. Park, and J. Lee, "Convolutional recurrent neural networks for urban sound classification using raw waveforms," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 2444–2448.