CSCI-SHU 360: Machine Learning

Final Competition Report - Fall 2025

Echoes of the City: Urban Sound Classification

Qisheng Zhou
ql2698@nyu.edu

Abstract

Environmental sound classification in noisy urban settings faces significant challenges when training from scratch without pre-trained models. We propose a robust pipeline leveraging trainable PCEN for adaptive feature extraction, FiLM for deep metadata fusion, and Sharpness-Aware Minimization (SAM) to ensure superior generalization. This approach successfully resolves acoustic ambiguities and prevents overfitting in data-constrained regimes, ultimately securing 1st place with 91.3% accuracy in the "Echoes of the City" competition.

1 Introduction

1.1 Task Description

The primary objective of this competition, "Echoes of the City," is to develop a robust machine learning system capable of classifying short urban audio clips into 10 distinct acoustic categories. These categories include *air_conditioner, car_horn, children_playing, dog_bark, drilling, engine_idling, gun_shot, jackhammer, siren,* and *street_music*. The input consists of raw audio waveforms (≤ 4 seconds), often characterized by significant background noise, overlapping sound events, and varying recording qualities. The challenge lies in extracting discriminative features from these complex acoustic scenes without relying on pre-trained weights or external data.

1.2 Dataset Overview

The dataset provided for this competition is derived from the standard UrbanSound8K dataset, a widely used benchmark in environmental sound classification.

- **Training Set:** The labeled training data corresponds to Folds 1 through 8 of the original UrbanSound8K organization. This structure allows for standardized cross-validation strategies during the development phase.
- **Test Set:** The unlabeled test data, used for the Kaggle leaderboard evaluation, comprises samples corresponding to the original Folds 9 and 10.
- **Metadata:** In addition to the audio files, metadata including subjective "salience" ratings (foreground vs. background) was provided, which I explicitly leveraged to guide model training.

1.3 Evaluation Metrics

To comprehensively assess model performance, the competition employs a weighted composite metric:

- **Overall Accuracy (80%):** Measures the global correct classification rate.
- **Macro-F1 Score (20%):** Ensures that the model performs well across all categories, penalizing bias towards majority classes in potentially imbalanced splits.

My final submission achieved a **91.3%** accuracy on the private leaderboard, securing the **1st place** ranking.

2 Method

To address the challenges of the UrbanSound8K dataset—specifically the strict prohibition on pre-trained models, class imbalance, and complex environmental noise—I designed a comprehensive pipeline. My approach focuses on robust feature extraction, metadata-driven architecture, and sharpness-aware optimization.

2.1 Data Pre-processing and Augmentation

Unlike standard approaches that rely on fixed Log-Mel spectrograms, I implemented a dynamic and learnable feature extraction pipeline.

- **Trainable PCEN (Per-Channel Energy Normalization):** Instead of static logarithmic compression, I utilized Per-Channel Energy Normalization (PCEN). I implemented a trainable PCEN layer (accelerated via TorchScript JIT) to adaptively compress the dynamic range of the audio. This allows the model to better handle varying gain levels and background noise typical in urban environments.
- **Multi-Resolution Spectral Analysis:** To balance the trade-off between time and frequency resolution, I employed a multi-resolution strategy in my advanced models. I computed spectrograms using two different window sizes simultaneously: a **Large FFT** ($N = 2048$) to capture fine-grained frequency details, and a **Small FFT** ($N = 1024$) to capture transient temporal events. These features were concatenated channel-wise before entering the backbone.
- **Hierarchical Augmentation Pipeline:** Given the constraint of training from scratch, heavy augmentation was crucial to prevent overfitting:
  - *Waveform Level:* I applied Gaussian Noise and Gain perturbations directly to the raw waveform to simulate recording quality variations.
  - *Spectrogram Level:* I utilized SpecAugment (Frequency and Time Masking) to force the network to learn robust features rather than relying on specific frequency bands.
  - *Batch Level (Mixup):* I employed Mixup, generating virtual training examples by linearly interpolating between two samples. Crucially, to handle the metadata consistent with Mixup, I implemented an **Embedding-level Mixup** strategy: instead of mixing raw labels, I interpolated the salience embedding vectors, ensuring mathematical consistency during the gradient update. </enditemize

2.2 Model Architectures and Modules

I established a robust baseline using ResNet-34 and progressively evolved the architecture to incorporate attention mechanisms and metadata fusion.

- **Deep Audio Stem & Backbone Variations:** I modified the standard ResNet stem, replacing the initial 7×7 convolution/max-pool layer with a **Deep Audio Stem** (three stacked 3×3 convolutions). This modification, combined with removing the initial MaxPool layer, preserved high-resolution time-frequency details essential for short audio clips.
- I experimented with three diverse backbones to ensure heterogeneity for the ensemble:
  - *ResNet-34:* Optimized for general feature extraction.
  - *WideResNet-50-2:* Increased channel width (up to 2048) to capture complex acoustic patterns.

    - *Res2Net-50:* Utilized multi-scale receptive fields within single residual blocks to handle both wide-band (impact sounds) and narrow-band (sirens) signals.
- **Attention and Sequence Modeling:**
    - *CBAM (Convolutional Block Attention Module):* Integrated into the residual blocks to refine features along both channel and spatial dimensions.
    - *BiGRU:* Inserted after the convolutional stages to model long-term temporal dependencies in the feature sequence.
    - *ASP (Attentive Statistics Pooling):* Replaced Global Average Pooling with ASP, calculating the weighted mean and standard deviation to capture the statistical distribution of non-stationary audio signals.
- **Metadata Fusion via FiLM:** To leverage the "salience" (foreground/background) metadata, I moved beyond simple concatenation. I implemented **FiLM (Feature-wise Linear Modulation)**. The salience metadata is mapped to an embedding, which then generates scale (γ) and shift (β) parameters to affine-transform the feature maps of the backbone. This allows the metadata to fundamentally modulate how the network processes acoustic features based on whether the sound is in the foreground or background.

2.3 Optimization and Refinement

To maximize the generalization capability of my models without external data, I employed advanced optimization and ensemble strategies.

- **Sharpness-Aware Minimization (SAM):** The core of my training strategy was the SAM optimizer. Instead of solely minimizing the training loss value, SAM seeks parameters that lie in a "flat minimum" neighborhood. By simultaneously minimizing the loss value and the loss sharpness (perturbing weights by ϵ in the direction of gradient ascent), SAM significantly improved the model's generalization on the unseen test set.
- **Training Strategy:** All models were trained using **8-Fold Cross-Validation**, ensuring every data point contributed to the learning process. Hyperparameters (learning rate, weight decay, Mixup alpha) were rigorously tuned using Optuna over 300+ trials.
- **Ensemble Methodology:** My final submission was an ensemble of the **top 12 best-performing models**, covering different architectures (ResNet, WideResNet, Res2Net) and fusion strategies (Concat vs. FiLM). I averaged the softmax probabilities of these models to produce the final prediction, achieving a robust accuracy of **91.3%**.

3 Experiment

My experiments were conducted using 8-fold cross-validation on the UrbanSound8K dataset. The final submission achieved a top-1 accuracy of **91.3%** on the test set. This section details the quantitative breakdown of my improvements, architectural refinements, and insights regarding validation pitfalls.

3.1 Step-wise Ablation Study: From Baseline to 87%

To rigorously evaluate the contribution of each component, I conducted a cumulative ablation study starting from a standard baseline (Log-Mel Spectrogram + SpecAugment).

- **Impact of PCEN (+3%):** Replacing standard Log-Mel spectrograms with Trainable PCEN yielded an immediate 3% performance gain. This confirmed that adaptive dynamic range compression is far more effective than fixed logarithmic scaling for handling the varying gain levels of urban recordings.
- **The Multi-Channel Misstep (-1%):** Early in the process, I attempted to use 3-channel inputs (RGB-style stacking) on the ResNet-34 baseline. Surprisingly, this resulted in a 1% performance drop. I hypothesize that for a model of this capacity trained from scratch, the single-channel input already contained sufficient information. Forcing a 3-channel representation at this stage likely introduced more noise than signal, complicating the optimization landscape. Consequently, the majority of my ensemble models utilized optimized single-channel inputs.
- **Regularization via Mixup (+5%):** Adding Mixup to the PCEN-based model provided a massive 5% boost, bringing the accuracy to 82%. This highlights that in the absence of pre-training, linear interpolation between samples is critical for learning robust decision boundaries.
- **Optimization via SAM (+5%):** The final leap came from the SAM optimizer. By seeking flat minima ($w + \epsilon$), SAM contributed another 5% improvement, pushing the single-model performance to 87.1%.

3.2 Architectural Refinements: Preserving Details

Standard computer vision architectures often discard fine-grained details that are critical for audio. I implemented two key modifications to the ResNet backbone:

- **Deep Audio Stem:** I replaced the standard 7×7 input convolution with a stack of three 3×3 convolutions. This allowed the model to extract more complex frequency textures from the raw input.
- **No-MaxPool:** I removed the initial MaxPooling layer. Since my input spectrograms (e.g., 128 Mel bands) are much smaller than standard images, aggressive pooling caused a loss of vital transient information. Preserving the spatial resolution at the early stages proved essential for distinguishing short-duration events like gunshots.

3.3 Insights from Failed Strategies

My process was also defined by identifying what *did not* work, which provided deep insights into the dataset characteristics.

- **The Sub-model Failure:** I attempted to train specialized binary classifiers for persistent confusion pairs. This approach failed for two distinct reasons:
  - *Label Noise (Drilling vs. Jackhammer):* The acoustic similarity coupled with subjective ground truth labels meant that even a specialized model could not disentangle the noise.
  - *Missing Information (Air Conditioner vs. Street Music):* The confusion here was not due to feature similarity but semantic ambiguity. Adding a sub-model failed because the input lacked context. This failure directly motivated my successful Salience/FiLM strategy, proving that adding the **correct metadata** was more effective than adding **more models**.
- **The Trap of Over-Optimized Hyperparameters:** In one experiment (Model 2), I used Optuna to rigorously tune hyperparameters, achieving a massive +8% gain in Cross-Validation (CV) score. However, this yielded only a 0.3% improvement on

the test set. This revealed a crucial insight: on this specific dataset, aggressive hyperparameter tuning led to **overfitting the validation folds** rather than learning robust features. This finding pivoted my focus from parameter tuning to structural improvements (e.g., SAM, FiLM) which offered genuine generalization gains.

4 Conclusion

In this competition, I presented a comprehensive audio classification framework that achieved **1st Place** with a test accuracy of **91.3%**. My approach successfully overcame the constraints of "no pre-training" and small-scale data through a synergy of robust feature engineering, sharpness-aware optimization, and deep metadata fusion.

My findings offer several critical insights for urban sound classification:

- **Metadata is as Vital as Audio:** The integration of salience metadata via FiLM (Feature-wise Linear Modulation) proved to be a decisive factor. It demonstrated that when acoustic features are ambiguous (e.g., overlapping foreground/background sounds), external context is essential. Deeply modulating features with metadata is far more effective than simple late fusion.
- **Generalization Geometry Matters:** In low-data regimes where pre-training is prohibited, the geometry of the loss landscape is paramount. The significant gains from SAM (Sharpness-Aware Minimization) indicate that seeking "flat minima" is the most effective defense against overfitting, outperforming complex architectural changes.
- **Resolution Trade-offs:** My experiments with Multi-Resolution PCEN and Deep Audio Stems highlight that urban sounds require a balance of high temporal resolution (for transient events) and fine frequency detail (for harmonic textures), which standard computer vision architectures often neglect.

For future work, I plan to investigate **Class-Specific Adaptive Weighting**. Since my experiments showed that baseline models already achieve high accuracy on "easy" categories, future training iterations could dynamically assign higher loss weights to difficult classes to further resolve persistent confusion pairs. Additionally, I aim to expand my feature set beyond PCEN by incorporating complementary acoustic descriptors such as MFCCs (for timbre) and Chroma features (for pitch profiles). Fusing these multi-dimensional inputs could provide a richer representation for resolving semantic ambiguities in complex urban soundscapes.

5 Kaggle Score Screenshot

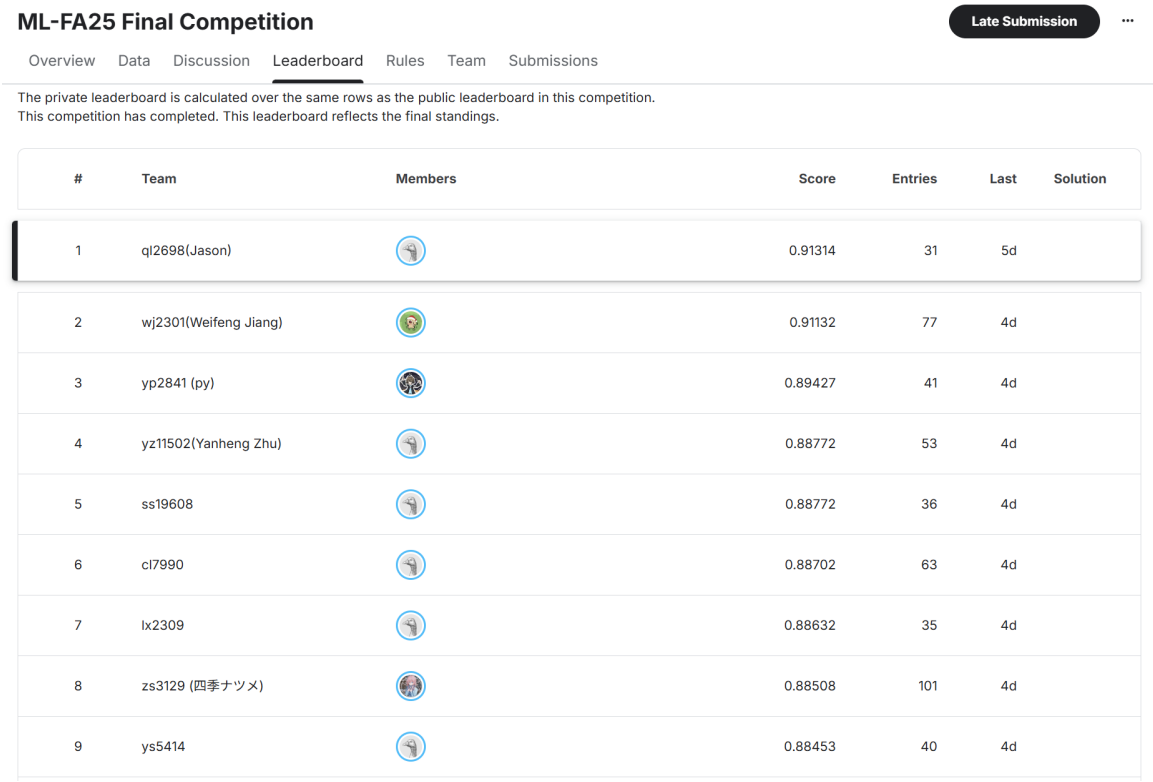

Figure 1: Final Leaderboard Standing: 1st Place with 0.91314 Score