Urban Sound Classification

A CNN Exclusive

Brijesh Patel, Victoria Chiu
{brijeshp, vchiu94}@stanford.edu

Stanford | ENGINEERING
Electrical Engineering

Project Overview

- Our project seeks to classify the ten most frequently observed urban sounds.
- Environmental sound classification is a growing field of research with applications to large-scale, content-based multimedia indexing and retrieval.
- Current challenges include a lack of a common taxonomy and a scarceness of real world, annotated data.

Dataset

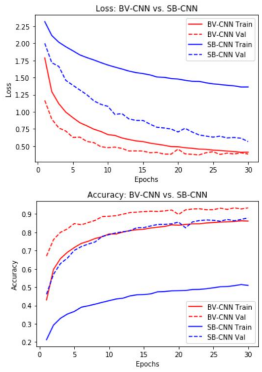
- UrbanSound8K[1] dataset: 10 categories, 8732 urban sound excerpts of up to 4s in duration taken from real field recordings.
- Each clip was split into overlapping windows of 633ms.
- From each frame, a Mel-Frequency spectrogram and its delta were extracted with 128 mel bands.
- Mostly even class distribution except for decreased representation of 2 classes: car horn and gun shot.

Summary of Results

Model	Training	Validation	Test
Logistic Reg. (MFC)			55%
SB-CNN (MFC-D)			79%
SB-CNN1 (MFC-D)	50.96%	87.83%	87%
SB-CNN2 (MFC-D)	84.34%	92.01%	92%
SB-CNN3 (MFC-D)	85.05%	91.78%	93%
BV-CNN (MFC-D)	86.13%	93.33%	94%

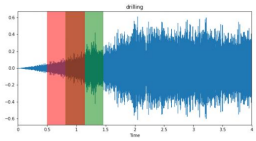
Notes for this section:
[1] MFC: Mel Frequency Cepstrum spectrogram, MFC-D: Mel Frequency Cepstrum spectrogram with corresponding deltas
[2] SB-CNN1 modifies window size, SB-CNN2 modifies dropout, SB-CNN3 modifies filter size, BV-CNN aggregates improvements

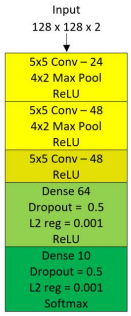
1. **The linear classifier is a simple model providing a baseline for performance**
2. **SB-CNN is a shallow network which seems to overuse dropout parameters**
3. **SB-CNN* has a smaller window size which captures features better**
4. **BV-CNN has fine tuned dropout, smaller filters, and a deeper network**
5. **Overlapping, smaller windows and better hyperparameters increase accuracy**



SB-CNN

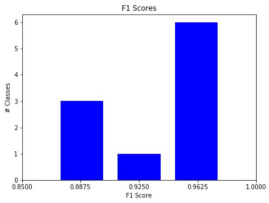
- Reimplementation of Salamon et. Al's custom CNN from 2016 IEEE Signal Processing Letters[2]
- 79% test accuracy
- MFC spectrogram and deltas as input
- Slight modifications: no data augmentation
- We used a 4.5 ms frame rate as opposed to a 23 ms frame rate for the MFC calculation.
- Performs markedly better with 87% test accuracy.
- Still large discrepancy between train and validation accuracies

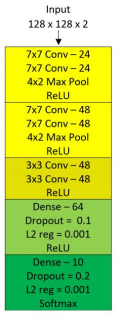




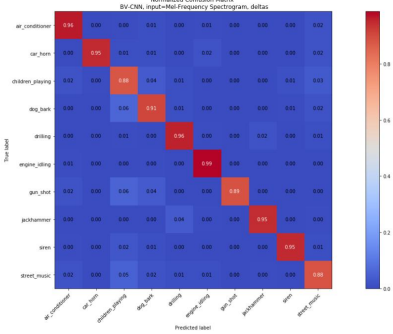
BV-CNN

- Hyperparameter tuning of dropout rate, filter size, and number of convolutional layers
- Slightly larger filter sizes and deeper network showed improvements
- Decreased overall dropout and fine tuned rate based on depth of the specifc dropout layers





Confusion Matrix (BV-CNN)



References

- [1] J. Salamon, et al. "A Dataset and Taxonomy for Urban Sound Research", 22nd ACM International Conference on Multimedia, Orlando USA, Nov. 2014.
- [2] Salamon, et al. "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," IEEE Signal Processing Letters 24.3 (2017):279-283. Web