WreckingNet: Neural approach for the classification of audio signals in construction sites

Neural Networks Project A.Y 2018/2019

Alessandro Maccagno 1653200
Andrea Mastropietro 1652886
Umberto Mazziotta 1647818

DIPARTIMENTO DI INGEGNERIA INFORMATICA AUTOMATICA E GESTIONALE ANTONIO RUBERTI

SAPIENZA
UNIVERSITÀ DI ROMA

Structure

- Introduction
- Dataset and Data Preprocessing
- Architecture
- Experiment and Results
- Conclusions and Future Work

Structure

- **Introduction**
- Dataset and Data Preprocessing
- Architecture
- Experiment and Results
- Conclusions and Future Work

Introduction

- Perform classification of vehicles and tools in construction sites using audio signals
- Neural approach
  - Combination of two CNNs (raw data and spectrogram)
- Working on real data
  - Provided by Professor Yongcheol Lee from Louisiana State University
- Based on the work of Li et al. that performed environmental sound classification with a neural approach
- High classification accuracy (97-98%)

Structure

- Introduction
- **Dataset and Data Processing**
- Architecture
- Experiment and Results
- Conclusions and Future Work

Dataset (1/2)

- Provided by Prof. Yongcheol Lee of LSU
- Live recording of construction event sounds
  - Tools and machines performing a certain action
- Real world data
  - Difficulties arise: noise and low quality recordings
- Large number of classes (~20)

Dataset (2/2)

- 5 classes selected:

Excavator Cat 320E

Excavator Hitachi 50U

Concrete Mixer

Backhoe JD50D Compact

Compactor Ingersoll Rand

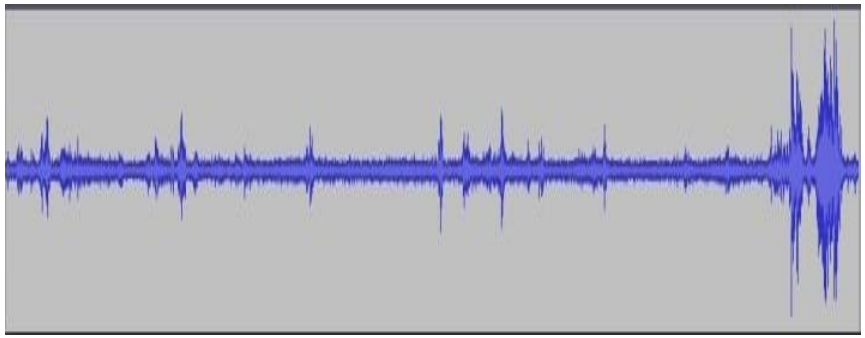
Data Preprocessing (1/2)

- Each track segmented in two partitions
  - 70% of the track for training, 30% for test
- Each partition split into smaller fragments of 30ms
  - 15ms overlap
- Training/Testing partition done before splitting
  - Done to avoid the model to be tested on samples on which it was trained due to the overlap
- Use pickle module to serialize data into files
  - No need to regenerate components from audio files
  - Noticeable reduction of the time required to obtain the dataset
  - Noticeable reduction of the size of the data

Data Preprocessing (2/2)

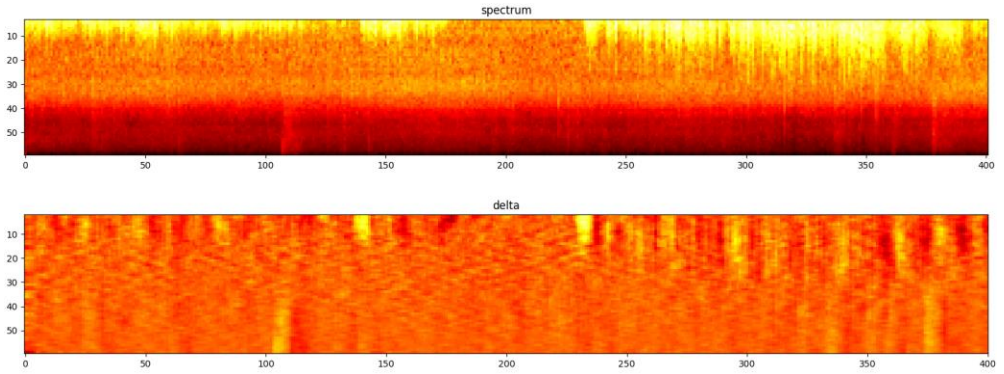
We use 2 types of information

- Raw data
  - Waveform of the audio
  - Resampling at 2*22050Hz
  - Value of the wave of each instant is in the interval [-1.0, 1.0]
  - 662 float values



Waveform

- Spectral data
  - Log-scale mel-spectrogram of the audio
  - 60 mel bands
  - First-order time derivative of the mel-spectrogram
  - 2 columns (time buckets)



Mel spectrogram and delta

Structure

- Introduction
- Dataset and Data Preprocessing
- **Architecture**
- Experiment
- Results
- Conclusions and Future Work

Architecture (1/4)

- Inspired by the works of **Li et al.** and **Piczak**

An Ensemble Stacked Convolutional Neural Network Model for Environmental Event Sound Recognition

Shaobo Li [1,2], Yong Yao [1,*], Jie Hu [3], Guokai Liu [3], Xuemei Yao [3] and Jianjun Hu [1,4,*]

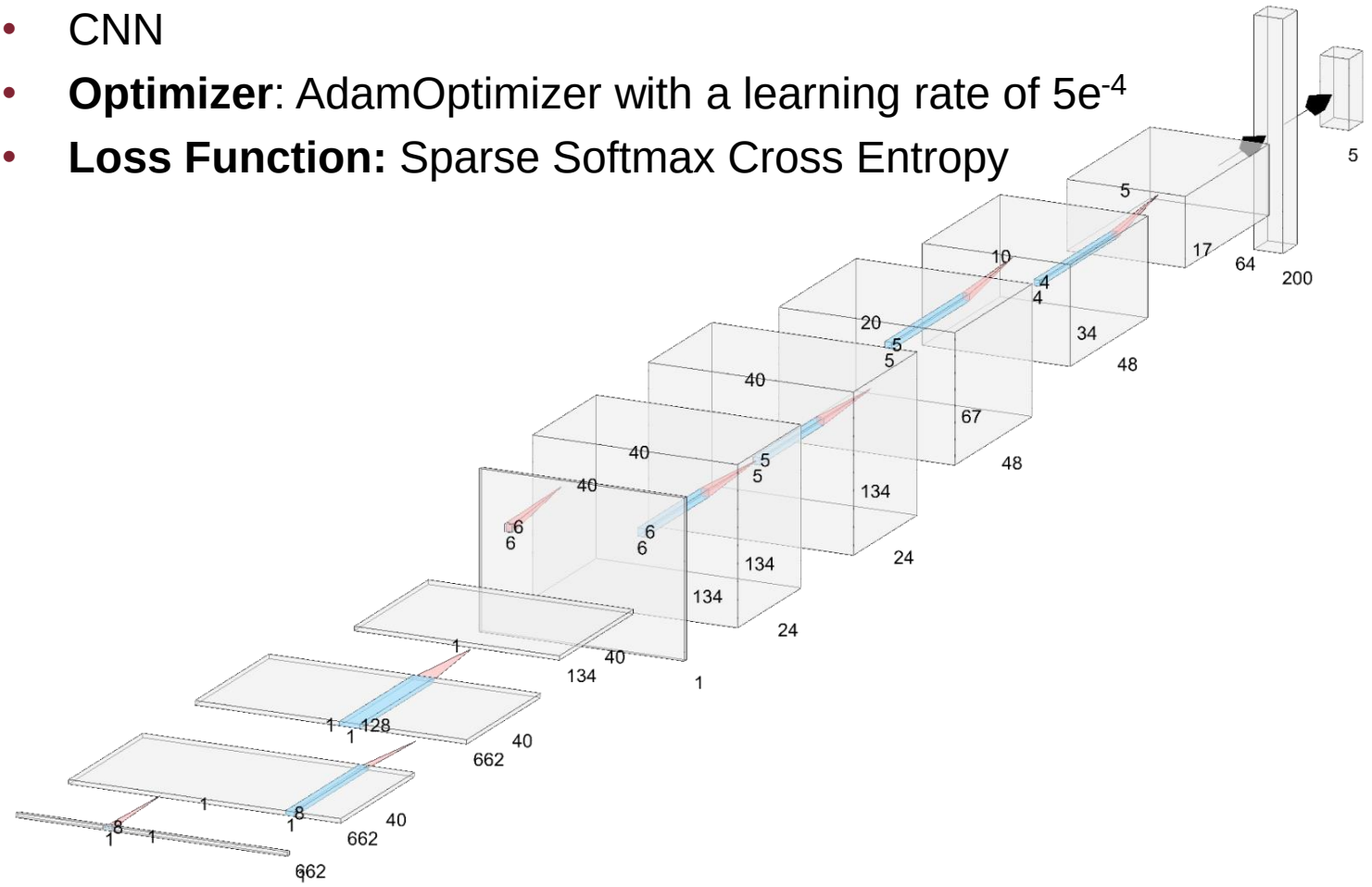
ENVIRONMENTAL SOUND CLASSIFICATION WITH CONVOLUTIONAL NEURAL NETWORKS

Karol J. Piczak

- Use two different convolutional neural networks
- Each neural network makes use of a different type of data
- Unify the results using Dempster-Shafer theory to improve results

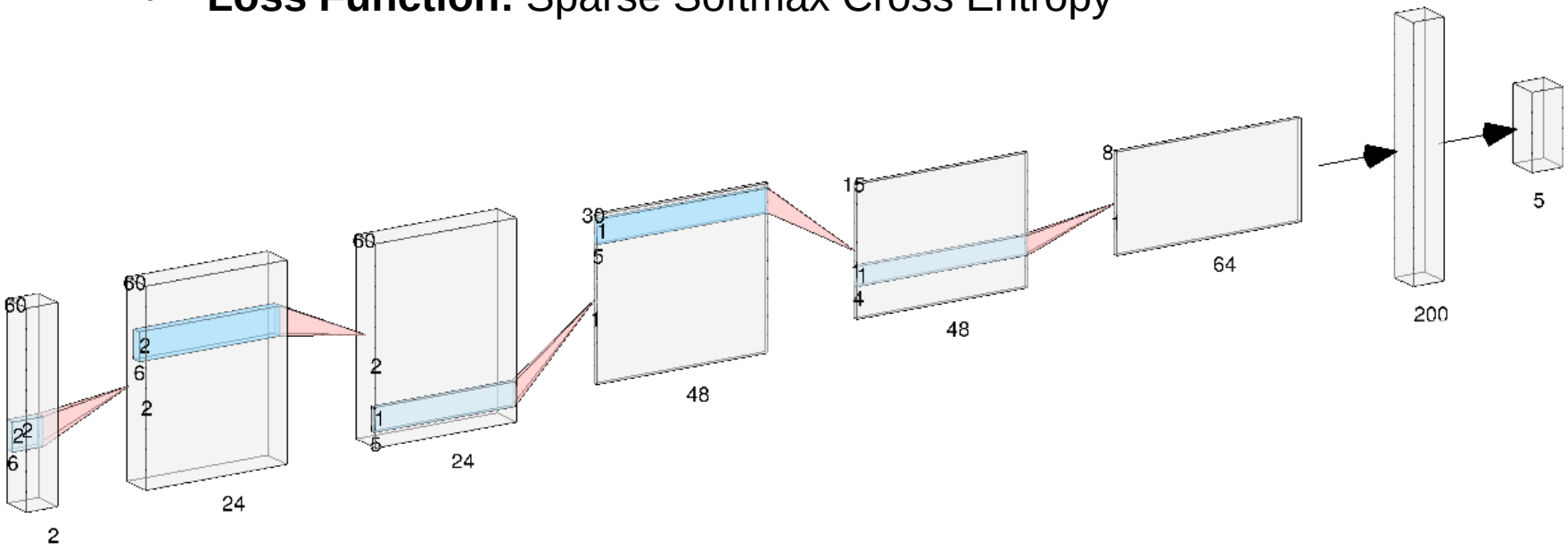
Architecture (2/4) - RawNet

- Uses the waveform of the audio fragment
- CNN
- **Optimizer**: AdamOptimizer with a learning rate of $5e^{-4}$
- **Loss Function:** Sparse Softmax Cross Entropy



Architecture (3/4) - SpectroNet

- Uses the spectrograms of the audio fragment
- CNN
- **Optimizer**: AdamOptimizer with a learning rate of $5e^{-4}$
- **Loss Function:** Sparse Softmax Cross Entropy



Architecture (4/4) – DSE Module

- Implements Dempster's rule of combination:

$$m(C) = \begin{cases} 0 & \text{if } C = \emptyset \\ \frac{1}{K} \sum_{C_1 \cap C_2 = C} m_1(C_1) m_2(C_2) & \text{otherwise} \end{cases}$$

$$K = \sum_{C_1 \cap C_2 = \emptyset} m(C_1) m(C_2)$$

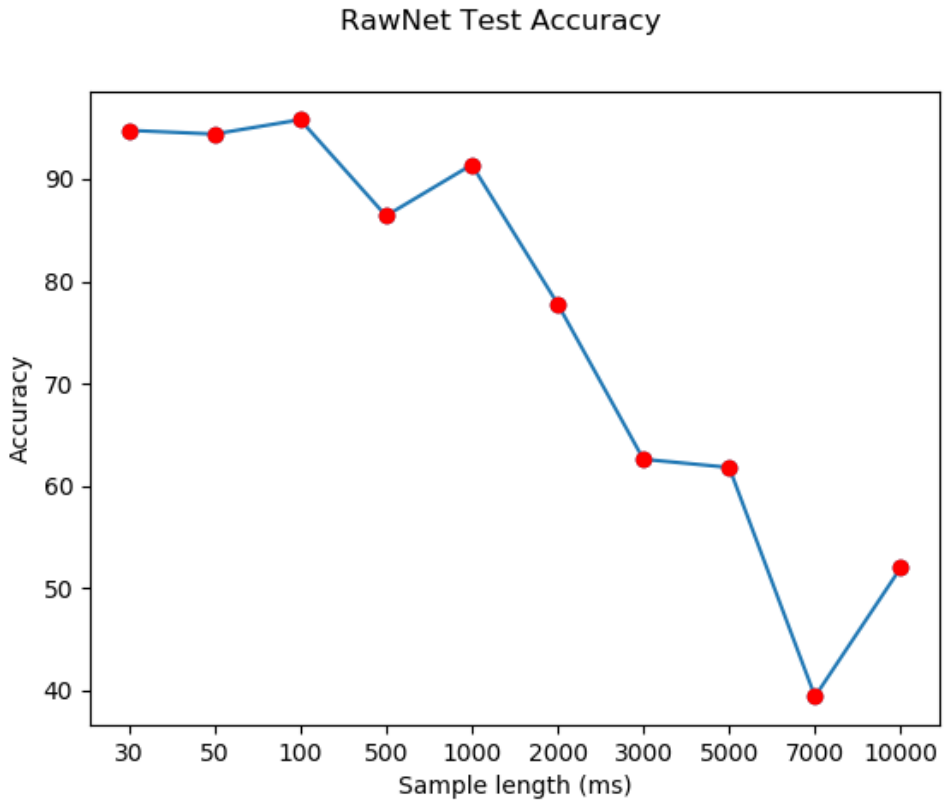
- Classes are mutually exclusive
- Each network outputs a class probability distribution via its softmax layer
- The probability that a segment belongs to a single class is equal to the normalized product of the two networks probabilities

Structure

- Introduction
- Dataset and Data Preprocessing
- Architecture
- **Experiment and Results**
- Conclusions and Future Work

Experiment (1/5) - Setup

- Choice of the most suitable sample length
- Several models built and tested
- Chosen 30ms samples
  - High accuracy
  - Higher number of subsamples



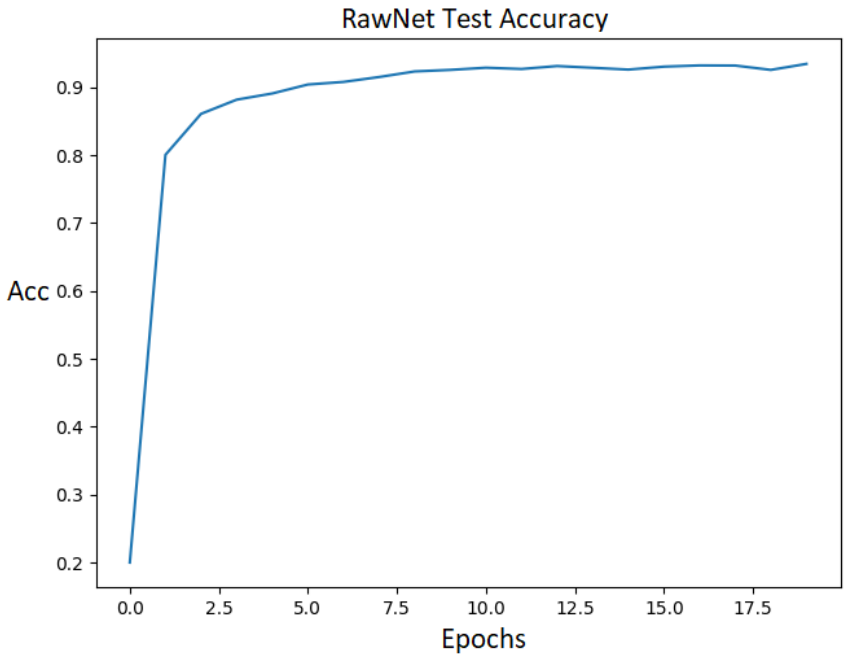
Experiment (2/5) - Classification

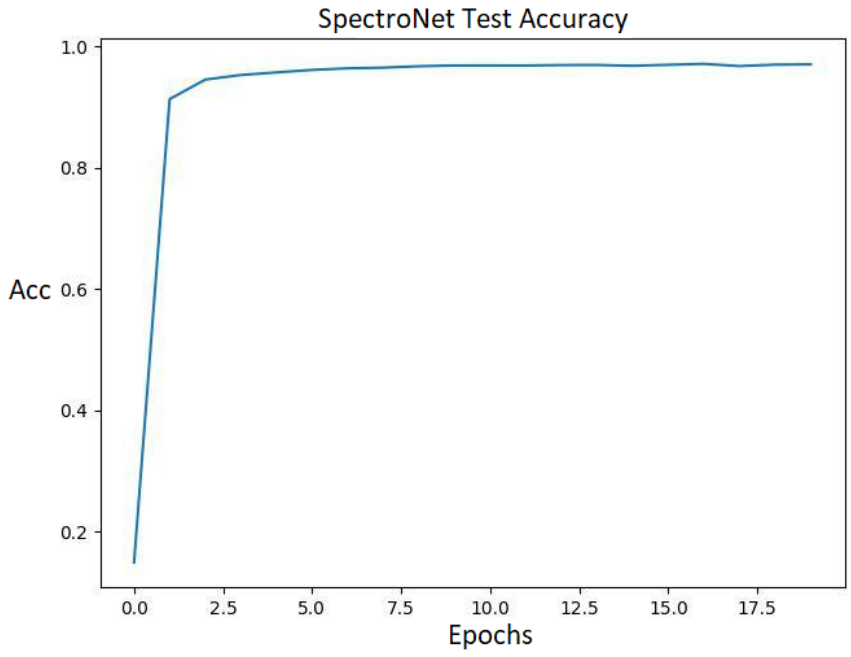
- Perfomed 5-Fold cross validation
  - 80% training
  - 20% test
  - Disjoint subset

Network	Accuracy	Precision	Recall	F1
RawNet	93.59	93.68	93.55	93.61
SpectroNet	97.08	97.34	**97.30**	97.32
WreckingNet (DSE)	**98.27**	**97.84**	97.10	**97.46**

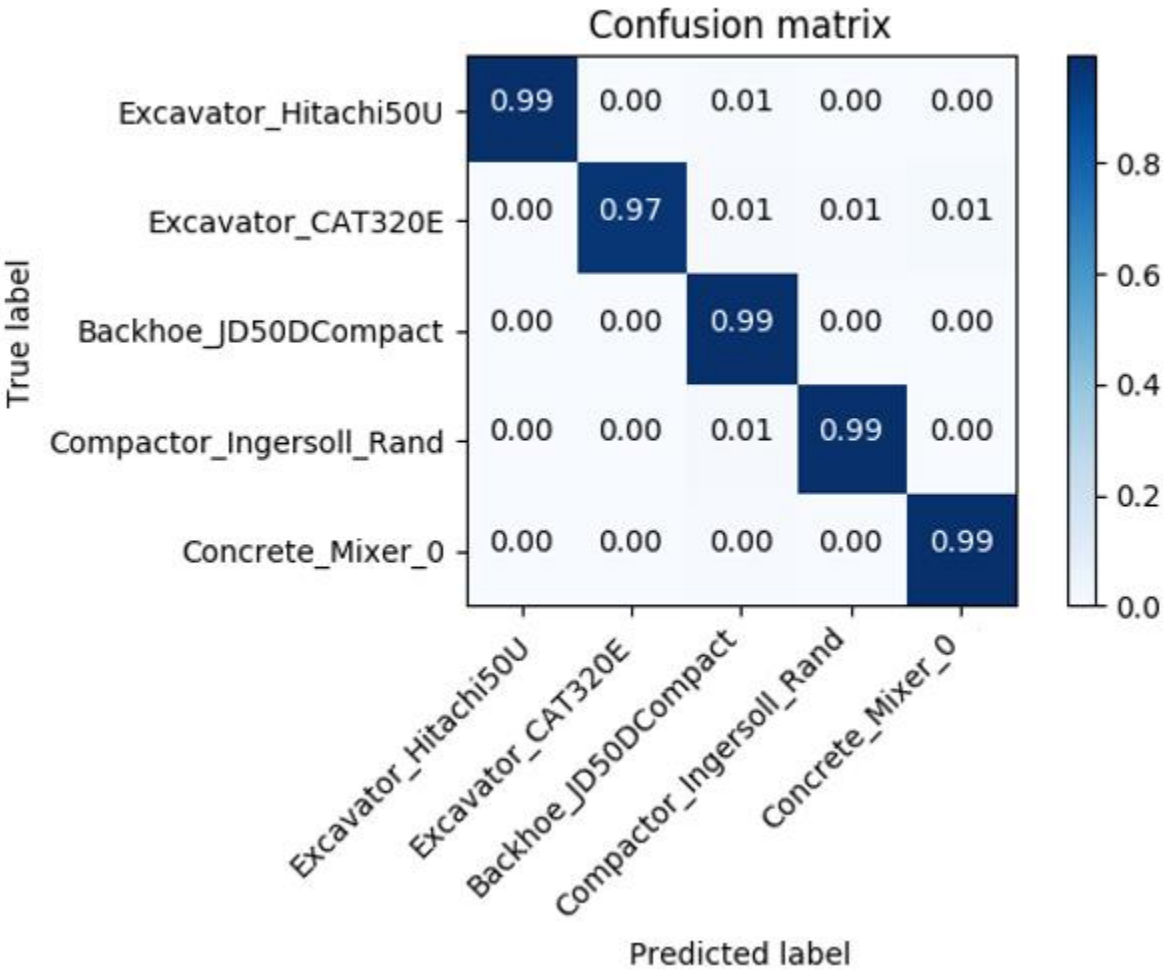
- Using DSE we obtain results higher than the highest of the two

Experiment (3/5) – Test Accuracy





Experiment (4/5) – Confusion Matrix



Experiment (5/5) - Prediction

- Audio track split into 30ms subsamples
- Each subsasmple is calssified by the network as belonging to one of the classes
- Overall classification determined by the majority of labels among all the fragment
- Performed on a sample of a concrete mixer recorded by us: correctly classified

Structure

- Introduction
- Dataset and Data Preprocessing
- Architecture
- Experiment and Results
- **Conclusions and Future Work**

Conclusions and Future Work

- Extension of a general environmental sound classification model to a more specific field
- High classification accuracy obtained on 5 classes

- Extension to a higher number of classes
- Activity monitoring
- Recognition of hazard during constrution activity

THANK YOU FOR YOUR ATTENTION