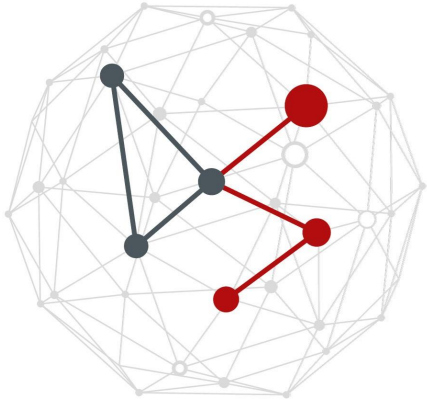
MIME

Master's degree ICT Internet Multimedia Engineering

PROJECT 5



DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

DIPARTIMENTO MATEMATICA

1222 · 2022 800 ANNI

UNIVERSITÀ DEGLI STUDI DI PADOVA

Project no. 5
“environmental sound classification”

Reference papers

[Piczak15] K.J. Piczak, ESC: Dataset for Environmental Sound Classification, in Proceedings of the 23rd ACM International Conference on Multimedia, Brisbane, Australia, 2015.

[Piczak15-1] K. J. Piczak, Environmental sound classification with convolutional neural networks, in Proceedings of the IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), Boston, MA, 2015.

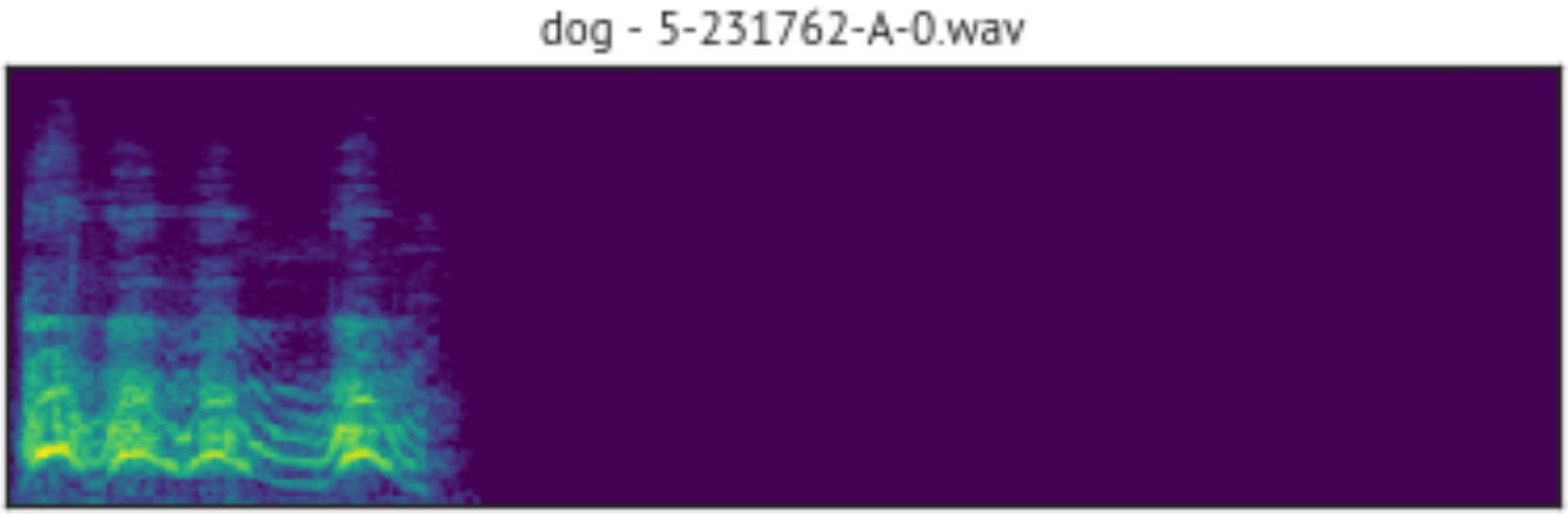
ECS-50 dataset (884 MB uncompressed)

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/YDEPUT

- Annotated collection of 2000 short clips comprising 50 classes of various common sound events

High level description of the dataset

- 5-second-long clips, 44.1 kHz, single channel
- Arranged into 5 uniformly sized cross-validation folds, ensuring that clips originating from the same initial source file are always contained in a single fold



High level description of the dataset

- 50 classes in the dataset

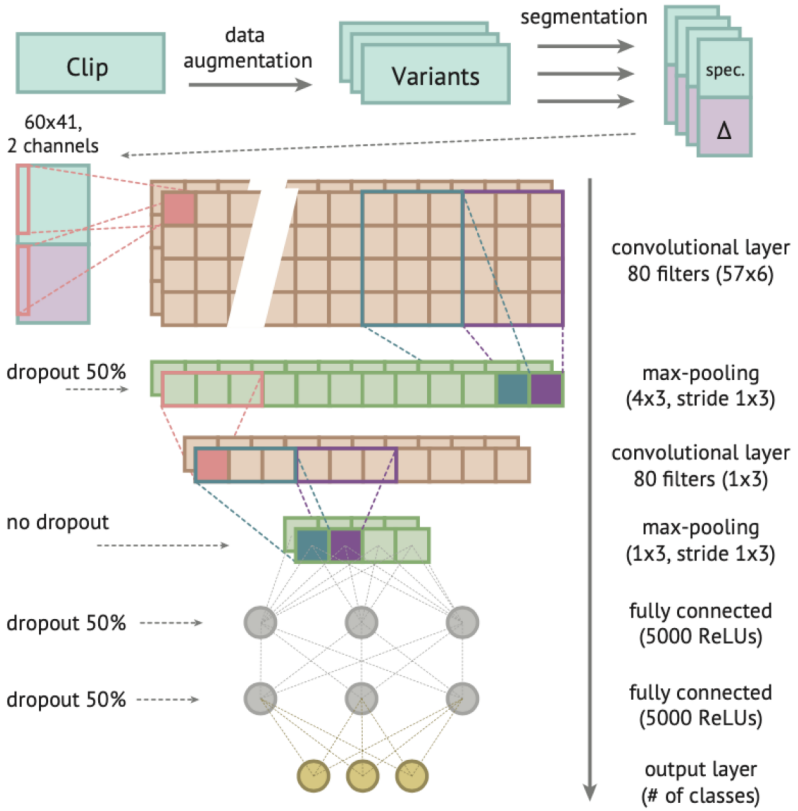
Animals	Natural soundscapes & water sounds	Human, non-speech sounds	Interior/domestic sounds	Exterior/urban noises
Dog	Rain	Crying baby	Door knock	Helicopter
Rooster	Sea waves	Sneezing	Mouse click	Chainsaw
Pig	Crackling fire	Clapping	Keyboard typing	Siren
Cow	Crickets	Breathing	Door, wood creaks	Car horn
Frog	Chirping birds	Coughing	Can opening	Engine
Cat	Water drops	Footsteps	Washing machine	Train
Hen	Wind	Laughing	Vacuum cleaner	Church bells
Insects (flying)	Pouring water	Brushing teeth	Clock alarm	Airplane
Sheep	Toilet flush	Snoring	Clock tick	Fireworks
Crow	Thunderstorm	Drinking, sipping	Glass breaking	Hand saw

High level description of the dataset

- ESC-10: selection of 10 classes from the bigger dataset
  - The differences between classes are much more pronounced, with limited ambiguity
  - Classes: *sneezing, dog barking, clock ticking, crying baby, crowing rooster, rain, sea waves, fire crackling, helicopter, chainsaw*
- meta/esc50.csv data description, the “esc10” column indicates if a given file belongs to the *ESC-10* subset
- meta/esc50-human.xlsx contains the human classification accuracy

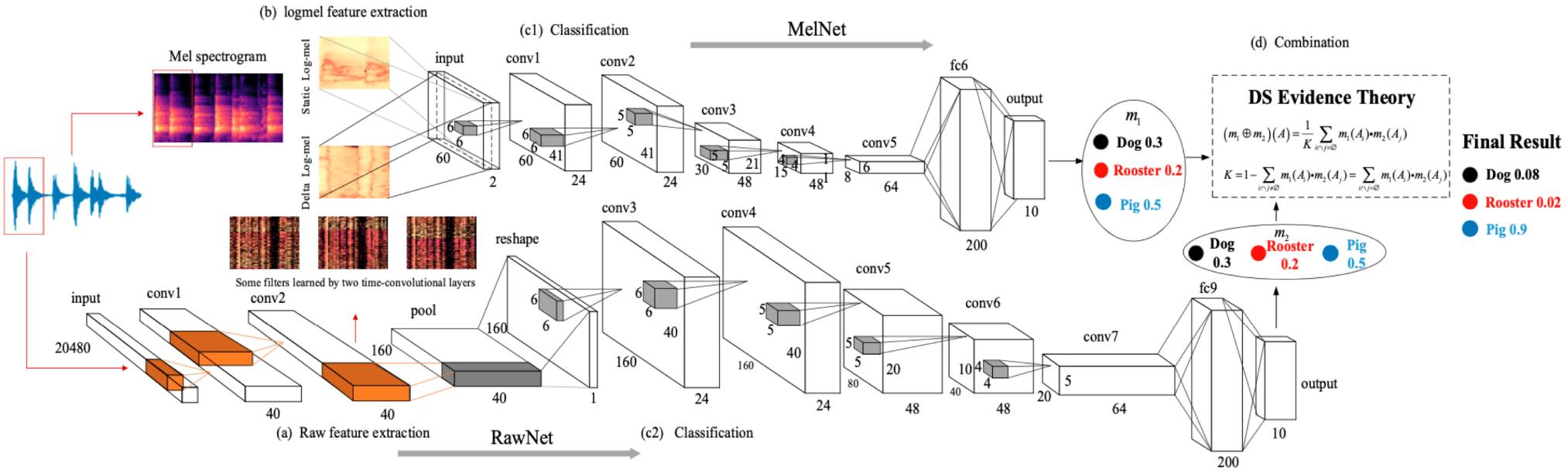
Approach in [Piczak15-1]

- Data augmentation: apply random time delays to the original recordings
- Feature extraction: log-scaled mel-spectrograms with 60 mel-bands
  - resampled to 22,050 Hz
  - windows size 1024
  - hop length 512
- Learning architecture: CNN



Other reference [Li18]

- Combines mel-spectrogram features and raw audio waveform



[Li18] S. Li, Y. Yao, J. Hu, G. Liu, X. Yao and J. Hu, An Ensemble Stacked Convolutional Neural Network Model for Environmental Event Sound Recognition, Applied Science, vol. 8, no. 1152, July 2018.

Useful links

- Dataset GitHub repository
  https://github.com/karolpiczak/ESC-50
- Some useful functions
  https://nbviewer.jupyter.org/github/karoldvl/paper-2015-esc-dataset/blob/master/Notebook/ESC-Dataset-for-Environmental-Sound-Classification.ipynb

Project proposal

- Classification tasks
  - on the entire ESC-50 dataset
  - on the restricted ESC-10 dataset
  - on each of the 5 groups of sounds:
    - animals
    - natural soundscapes & water sounds
    - human, non-speech sounds
    - interior/domestic sounds
    - exterior/urban noises
- Features: try with different approaches: mel-spectrogram, other manual-extracted features, raw data, combinations
- Architecture
  - different possibilities: CNN, RNN, …