MONITORING OF DOMESTIC ACTIVITIES BASED ON MULTI-CHANNEL ACOUSTICS A TIME-CHANNEL 2D-CONVOLUTIONAL APPROACH

Technical Report

Marco Tiraboschi

Università degli Studi di Milano,
Department of Computer Science, Laboratorio di Informatica Musicale
via Celoria 18, Milan, MI 20133, Italy
marco.tiraboschi@studenti.unimi.it

ABSTRACT

This approach is meant to be an extension of the DCASE 2018 task 5 baseline system for domestic activity recognition exploiting multi-channel audio: the Convolutional Neural Network model has been slightly restructured for this purpose by using two-dimensional convolutions along the dimensions of time and channel.

1. PREPROCESSING

The CNN input layer is the matrix of the short-time log mel-band energies for each one of the audio channels: 500 windows of 40ms with 50% overlap are used for the STFT computation and 40 log mel-band energies are extracted for each time window.

The result is a three-dimensional $4 \times 500 \times 40$ matrix: 4 audio channels, 500 time windows and 40 log mel-band energies.

2. SYSTEM

The network architecture is based on the baseline system convolutional network [1] and extended by operating the convolutions along the two dimensions of time and channel.

The rationale behind this choice is that the network will learn patterns that are equivariant along the dimensions of time and channel, but not along the frequency axis, as similar patterns in different frequency bands don't necessarily belong to similar acoustic events.

- Input data: $4 \times 500 \times 40$
- Architecture:
  - 2D Convolutional layer (filters: 64, kernel size: 3×5, stride: 1, axes: channel, time) + Batch Normalization + ReLU activation
  - 1D Max Pooling (pool size: 5, stride: 5, axis: time) + Dropout (rate: 20%)
  - 2D Convolutional layer (filters: 64, kernel size: 2×3, stride: 1, axes: channel, time) + Batch Normalization + ReLU activation
  - 1D Global Max Pooling (axis: time) + Dropout (rate: 20%)
  - Dense layer (neurons: 128) + ReLU activation + Dropout (rate: 20%)
  - Softmax output layer (classes: 9)

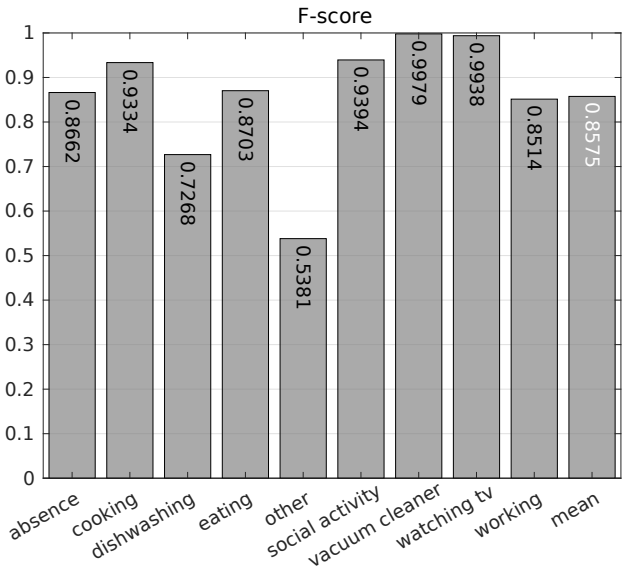

Figure 1: Class-wise and macro-averaged F-scores resulted from cross-validation.

Because of time and computational resources constraints, the model has been trained for 80 epochs only. In the future, it would be interesting to train the model for 420 epochs more to match the baseline training time for a fairer measure of its performance.

- Learning:
  - Optimizer: Adam (learning rate: 0.0001)
  - Epochs: 80
  - Batch size: 512

The proposed system has been implemented in Matlab, using the Neural Network Toolbox [2] .

3. CROSS-VALIDATION

As output of the training, four convolutional networks have been obtained, each by training the model on a different fold of the cross-validation setup defined by the challenge coordinators. Here are

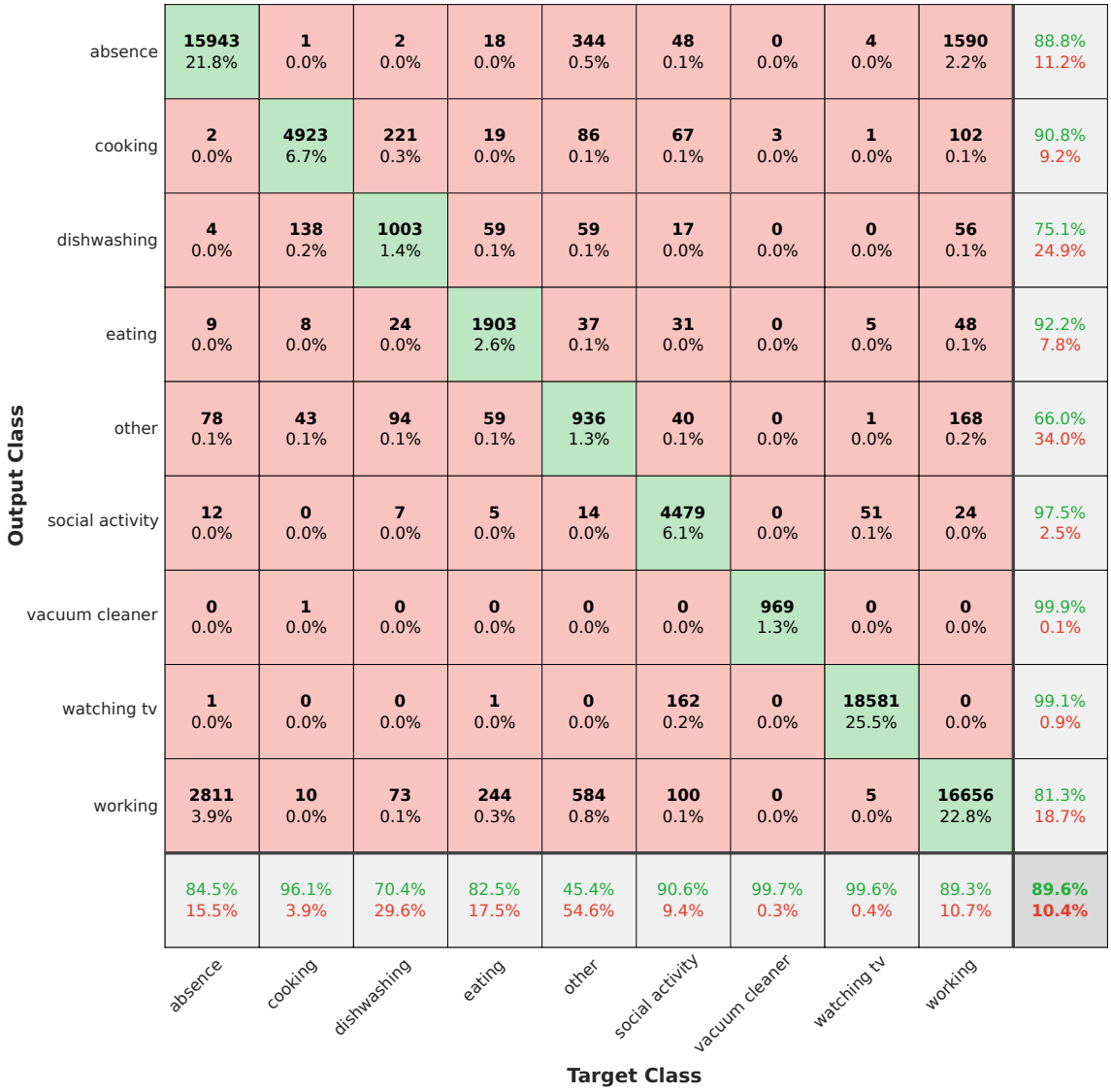

Figure 2: Confusion matrix of the cross-validation results.

presented the class-wise and mean F-scores (figure 1) and the confusion matrix (figure 2) resulted from cross-validation.

4. EVALUATION OUTPUT

For each file in the evaluation dataset, the output class is the class that maximizes the sum of the prediction scores from the four convolutional networks.

5. CONCLUSIONS

The results obtained with this approach have shown that exploiting inter-channel acoustic features can significantly improve the performance of a system of acoustic event recognition.

In effect, the proposed system exceeds the baseline by 1.25% macro-averaged F-score (85.75% against 84.50%) in 16% of the training time (80 epochs against 500).

6. REFERENCES

[1] G. Dekkers, L. Vuegen, T. van Waterschoot, B. Vanrumste, and P. Karsmakers, "DCASE 2018 Challenge - Task 5: Monitoring of domestic activities based on multi-channel acoustics," KU Leuven, Tech. Rep., July 2018

[2] Matlab Neural Network Toolbox
`https://www.mathworks.com/help/nnet`