GCNN FOR CLASSIFICATION OF DOMESTIC ACTIVITIES

Technical Report

Lionel Delphin-Poulat, Cyril Plapous, Rozenn Nicol

Orange Labs, Lannion, France, {lionel.delphinpoulat, cyril.plapous, rozenn.nicol}@orange.com

ABSTRACT

A model of classifier for processing multi-channel audio segments into nine classes categorizing daily activities (Task 5 of Challenge DCASE 2018) is presented. Its framework is based on Gated Convolutional Neural Network (GCNN). Four models are proposed with different learning strategies. They achieve a macro-averaged F1-score between 85.95 and 88.72%.

Index Terms— DCASE2018, Gated Convolutional Neural Network, microphone array

1. INTRODUCTION

This report describes a model of classifier for the task 5 of the Challenge DCASE 2018 [1, 2]. This task focuses on the audio monitoring of domestic activities. The model aims at classifying multichannel audio segments into one of the nine predefined classes, which inventory daily activities observed in home environments (i.e. ”other”, ”social activity”, ”eating”, ”working”, ”absence”, ”vacuum cleaner”, ”dishwashing”, ”watching tv”, ”cooking”). A novelty brought by this task is the properties of the audio recording which is based on a set of microphone arrays [3]. Each array is referred to as a ”node” and is composed of a linear array of four microphones (i.e. four channels) spaced by 5 cm. Thus one issue raised by this task is to handle multichannel audio in the model, for instance by including array processing. The whole database includes thirteen nodes distributed in an appartment. In the DCASE challenge, only seven nodes (located in the living room and kitchen area) are considered. The development set is restricted to four of these latter and contains 72984 audio segments of 10s duration. The nine classes are unequally represented among this selection. Each segment consists of four channels resulting from the four microphones. The sampling frequency is 16 kHz.

The report is organized as follows. Section 2 presents the structure of the proposed models, from the feature extraction to the classification. The learning strategy is also detailed. Then performances are illustrated in Section 3, before concluding.

2. PROPOSED MODELS

2.1. Feature

Before extrating features, the audio waveforms are prepocessed to suppress the DC component. Then one log MEL spectrogram is extracted per channel (see Fig. 1). In this extraction, phase information is discarded.

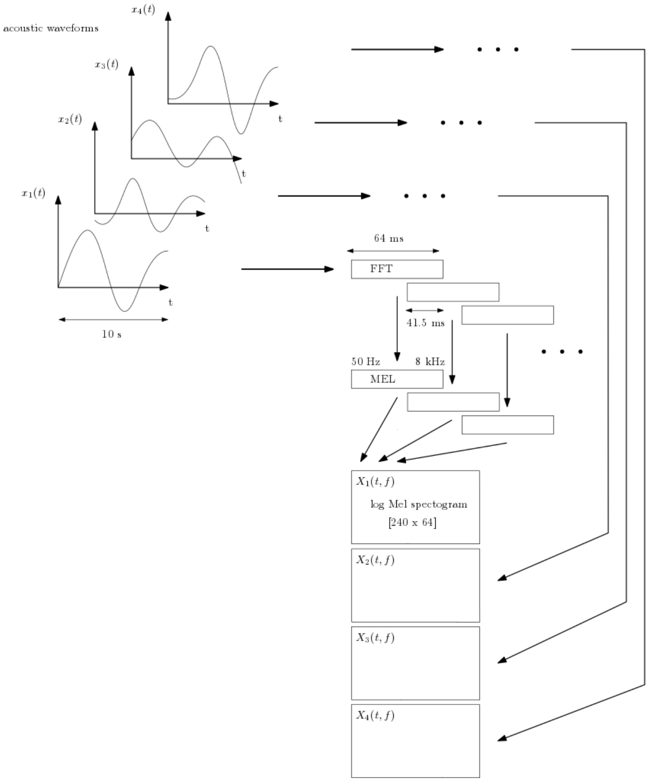

Figure 1: Extraction of features from the raw waveforms.

2.2. Model structure

Our model is based on Convolutional Neural Networks (CNNs) in combination with Gated Linear Units (GLUs), leading to Gated Convolutional Neural Network (GCNN) [4, 5]. Three elementary GCNN blocks are concatened. Each GCNN block is composed of the following operations:

- convolution,
- activation by GLUs,
- maxpooling,
- drop-out.

The first GCNN block (see Fig. 2 and 3) is applied to the log Mel spectrogram. It is followed by two similar blocks, which are respectively applied to the output (i.e. d=256 or d=128 images) of the first (see Fig. 4) or second block. The last step is performed by a Multi Layer Perceptron (MLP) (see Fig. 5) which outputs the predicted

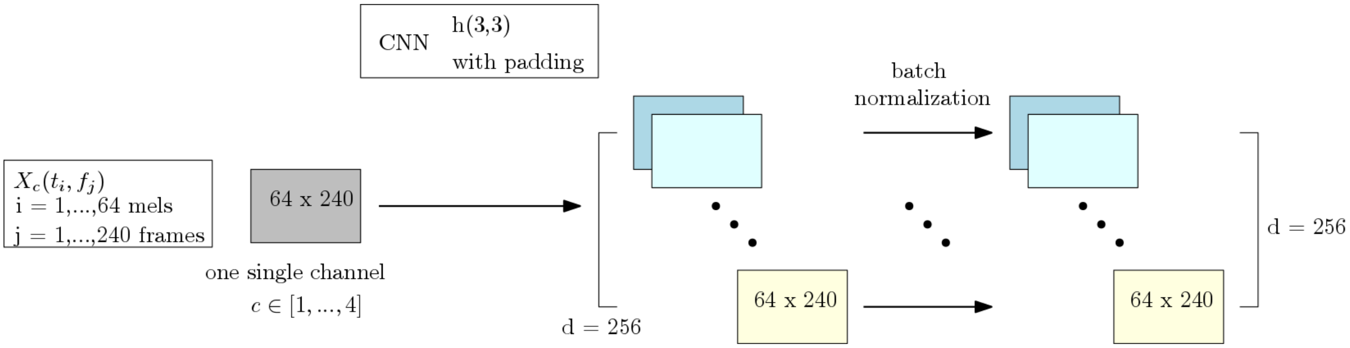

Figure 2: Scheme of the first Gated Convolutional Neural Network Block: first sub-block consisting of the convolutional unit.

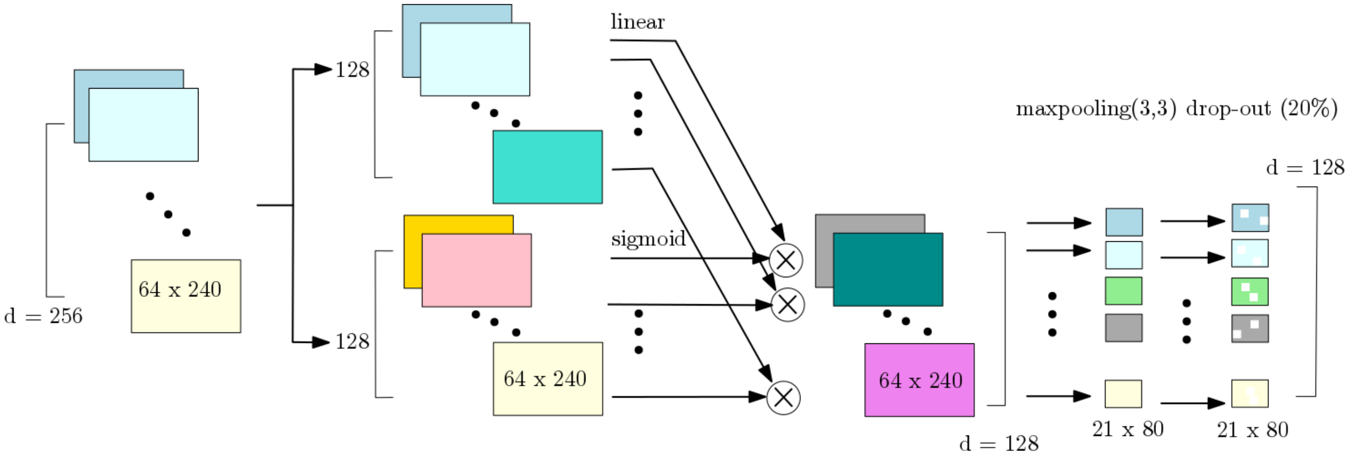

Figure 3: Scheme of the first Gated Convolutional Neural Network Block: second sub-block consisting of the Gated Linear Units, maxpooling and drop-out operations.

probability of the nine classes for each channel. The probability for an audio segment is computed as the average of the values obtained for the four channels.

2.3. Model training

The training phase consists in optmizing the categorical cross-entropy between the predicted probability and the ground truth of the audio segments. Optimization is peformed by the ADAM algorithm [6] with a learning rate of 10^{-4}. The learning process is organized by batch of 64 audio segments. In the data set, the number of segments per class highly depends on the class (i.e. ”other”: 2060, ”social activity”: 4944 segments, ”eating”: 2308 segments, ”working”: 18644 segments, ”absence”: 18860 segments, ”vacuum cleaner”: 972 segments, ”dishwashing”: 1424 segments, ”watching tv”: 18648 segments, ”cooking”: 5124 segments), which biases learning. To avoid this, each batch is balanced : each class is represented by an equal number of segments. For the overall learning, 200 epochs are considered. At each epoch, the model which achieves the highest macro-averaged F1-score for the validation set is kept. The final model is selected as the one which maximizes the F1-score at the end of the last epoch.

Three learning strategies were used:

- The model (model 1) is trained and validated exclusively on the training set of the Fold 1 provided by the development set. Training is performed with 90% of the Fold and the remaining 10% are used for the validation.
- The model (model 2) is trained on the whole training set of Fold 1 and is validated on the test set of Fold 1.
- As for the first strategy, the model (model 3) is trained and validated on the training set of the Fold 1 (training: 90%, validation: 10%), except that the data are augmented by a noisy version of Fold 1, in which a gaussian noise (mean = 0, variance = 2 dB) is added to the log MEL Spectrogram. The objective is to improve the robustness of the model, particularly regarding the three nodes that will be present in the evaluation set but are not included in the development set.

3. EXPERIMENTS AND RESULTS

The learning strategies described above define three models which were evaluated on the test subset of the development set (restricted to Fold 1). In addition a forth model is obtained as the average prediction of the first three ones. Their performances are given in Tab. 1. It should be noticed that in the case of models 2 and 4, there is a possible bias due to the fact that the test set was used as the validation set. Nevertheless it is observed that scores of models 1 and 2 are very close.

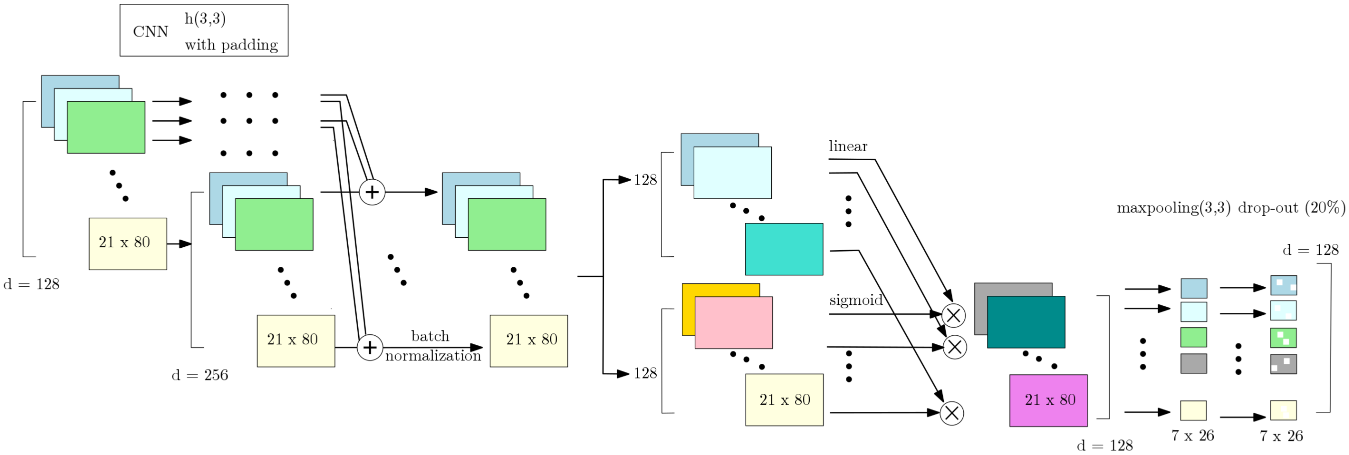

Figure 4: Scheme of the second Gated Convolutional Neural Network Block

Models	1	2	3	4
Other	52.15	51.56	43.81	53.58
Social activity	97.59	97.27	96.45	97.60
Eating	88.16	89.01	85.66	88.91
Working	86.82	86.92	82.70	86.35
Absence	89.46	88.68	86.45	88.66
Vacuum cleaner	100.00	100.00	100.00	100.00
Dishwashing	84.87	86.85	81.93	85.79
Watching tv	99.69	99.70	99.38	99.72
Cooking	97.94	97.50	97.16	97.88
Overall	88.52	88.61	85.95	88.72

Table 1: Macro-averaged F1-score of the proposed models for each class.

It is observed that some classes (i.e. ”social activity”, ”vacuum cleaner”, ”watching tv”, ”cooking”) are very well recognized, whereas the class ”other” presents low scores whatever the model. Indeed this latter contains probably too heterogeneous material to extract reliable features. Our best model (model 4) achieves an overall score of 88.72%.

4. CONCLUSION

Classification of multi-channel audio segments into classes of daily activities was investigated. A model based on Gated Convolutional Neural Network (GCNN) was developed and submitted to the Challenge DCASE 2018 (Task 5). Three learning strategies were compared. Finally the model resulting from the average prediction of the models associated to these three strategies achieves the best score (i.e. macro-averaged F1-score of 88.72%).

5. REFERENCES

[1] http://dcase.community/challenge2018/.

[2] 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.

[3] G. Dekkers, S. Lauwereins, B. Thoen, M. W. Adhana, H. Brouckxon, T. van Waterschoot, B. Vanrumste, M. Verhelst, and P. Karsmakers, “The SINS database for detection of daily activities in a home environment using an acoustic sensor network,” in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, Munich, Germany, November 2017, pp. 32–36.

[4] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, “Language modeling with gated convolutional networks,” *CoRR*, vol. abs/1612.08083, 2016. [Online]. Available: http://arxiv.org/abs/1612.08083

[5] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, “Large-scale weakly supervised audio classification using gated convolutional neural network,” *CoRR*, vol. abs/1710.00343, 2017. [Online]. Available: http://arxiv.org/abs/1710.00343

[6] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

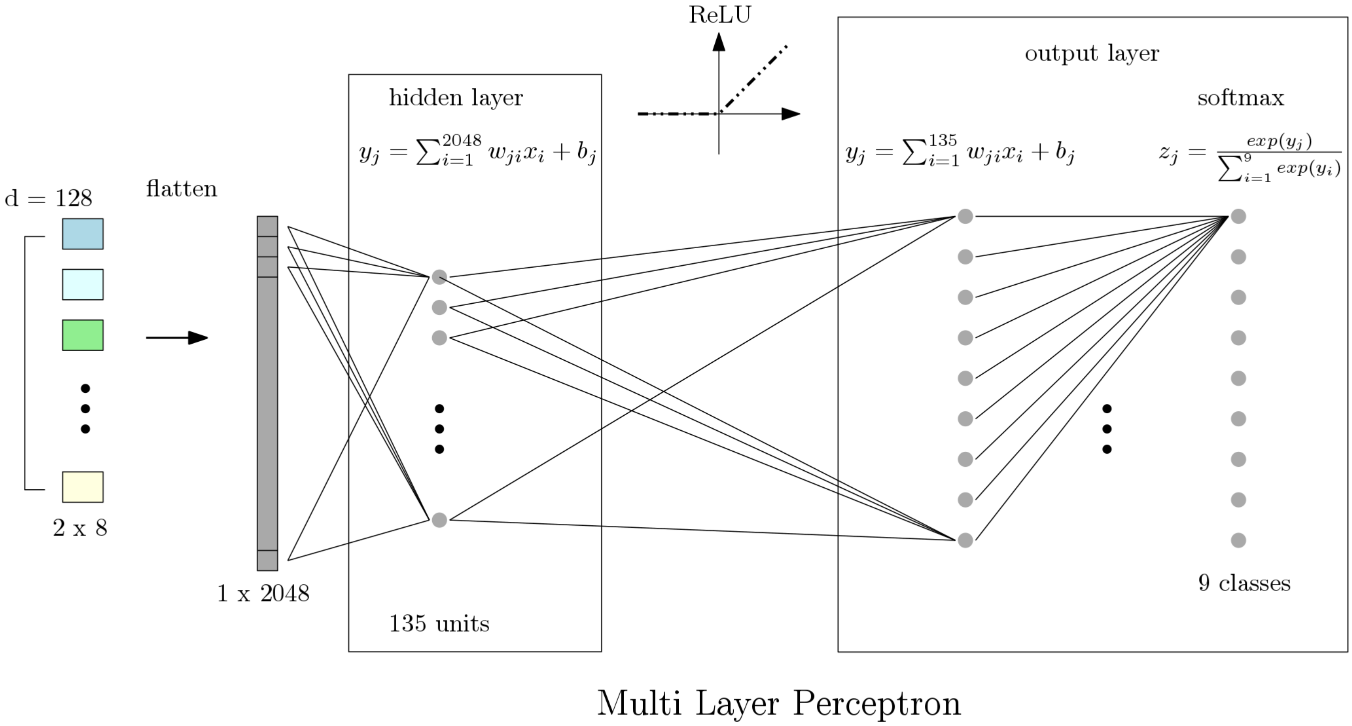
ReLU
hidden layer
$y_j = \sum_{i=1}^{2048} w_{ji}x_i + b_j$
output layer
softmax
$y_j = \sum_{i=1}^{135} w_{ji}x_i + b_j$
$z_j = \frac{exp(y_j)}{\sum_{i=1}^{9} exp(y_i)}$
d = 128
flatten
2 x 8
1 x 2048
135 units
9 classes
Multi Layer Perceptron

Figure 5: Scheme of the Multi Layer Perceptron.