MEAN TEACHER WITH DATA AUGMENTATION FOR DCASE 2019 TASK 4

Technical Report

Lionel Delphin-Poulat

Orange Labs Lannion
22307 Lannion, France
lionel.delphinpoulat@orange.com

Cyril Plapous

Orange Labs Rennes
35512 Cesson-Sevigne, France
cyril.plapous@orange.com

ABSTRACT

In this paper, we present our neural network for the DCASE 2019 challenge's Task 4 (Sound event detection in domestic environments) [1]. The goal of the task is to evaluate systems for the detection of sound events using real data either weakly labeled or unlabeled and simulated data that is strongly labeled. We propose a mean-teacher model with convolutional neural network (CNN) and recurrent neural network (RNN) together with data augmentation and a median window tuned for each class based on prior knowledge.

***Index Terms*—** DCASE 2019, RCNN, Mean Teacher, Data augmentation, Median window

1. INTRODUCTION

In this paper, we propose a sound event detector based on the provided baseline [2]. This baseline relies on a mean-teacher model [3] that is based on convolutional neural network (CNN) [4] and recurrent neural network (RNN) [5]. From this baseline, we increase the complexity of the architecture and use data augmentation, both in time and frequency domains. We also tune the median window filters for each class depending on prior knowledge, computed from each class training data.

2. DATASET

The dataset of DCASE 2019 is composed as follows: Labeled training set, Unlabeled in domain training set and Synthetic set with strong annotations.
The Labeled training set contains 1578 clips, the Unlabeled in domain training set contains 14412 clips and finally, the Synthetic strongly labeled set contains 2045 generated clips.
The audio clips are sampled at 44,100 Hz with a maximum duration of 10 seconds. Each audio clip contains at least one sound corresponding to one of the 10 possible classes.

2.1. Some statistics about the dataset

Table 1 gathers the mean and median duration of each sound class. This information is useful as it will be used as prior knowledge to post process the detection obtained by the neural network.

Table 1: Sound duration for each class.

Class	Occur-rences	Mean in s	Medi-an in s
Alarm/bell/ringing	755	1.07	0.38
Blender	540	2.58	1.62
Cat	547	1.07	0.88
Dishes	814	0.58	0.37
Dog	516	0.98	0.48
Electric shaver/toothbrush	230	4.52	4.07
Frying	137	5.17	5.13
Running water	157	3.91	3.60
Speech	2132	1.16	0.89
Vacuum cleaner	204	5.29	5.30

We define three sound categories:

- Impulsive sound: "Alarm/bell/ringing", "Dishes" and "Dog" for which the median duration is less than 0.5s.
- Intermediate sound: "Blender", "Cat" and "Speech" for which the median duration is around 1s.
- Background sound: "Electric shaver/toothbrush", "Frying", "Running water" and "Vacuum cleaner" for which the median duration is greater than 3s.

2.2. Audio preprocessing

First, we resample the audio clips at 22,050 Hz (after conversion to single-channel, when necessary) and then we extract the log mel-spectrogram from the audio clips. The size of the analysis window is 2048, the hop length is 365 and the number of mels is chosen to be 128. We also noticed that some audio files contain only numerical zeros, and decided to remove these files. We also remove the DC component as some files contain strong DC level which is useless. Finally, we normalize the mel-spectrograms for each mel-bin by the global mean and the standard deviation of the value for this bin. The mean and standard deviation are computed on the training set.

3. PROPOSED SOLUTIONS

The solution we propose is based on the provided baseline, *i.e.* a mean-teacher model, itself based on the solution proposed by the winner of DCASE 2018 Task 4 challenge [6]. This model relies on two same RCNN networks (CNN + RNN). We trained two models, the difference between them relies in the data augmentation part, as detailed in section 3.3.

3.1. Modifications to the baseline

We propose to modify the baseline as follows:

- The maximum noise level used in the mean-teacher approach is set to 15 dB for the additive Gaussian noise.
- The architecture of the mean-teacher model is set to be more complex than the baseline's:
    - The RNN part is composed by two layers of RNN cells, each layer contains 128 cells.
    - The CNN part is composed by 7 layers. For each layer the kernel size is [3, 3], padding and stride are both [1, 1]. The number of filters and pooling size for each layer are respectively [16, 32, 64, 128, 128, 128, 128] and [[2, 2], [2, 2], [1, 2], [1, 2], [1, 2], [1, 2], [1, 2]]. This configuration leaves us with only one value per frame at the RNN input.
    - The dropout is set to 50% and the activation is a GLU.
    - The training is set for 200 or 300 epochs with a rampup [7] of 50 epochs. The target learning rate is 0.001 and the optimizer is Adam (as in the baseline).
    - The batch size is set to 24.

The three databases (i.e. the labeled, unlabeled and synthetic data) were used for training in the same proportion in each batch as proposed in the baseline (1/4 for the labeled data, 1/2 for the unlabeled data and 1/4 for the synthetic data).

3.2. Median window length

In [6], the influence of the median window is underlined. This window is used to post process the frame by frame outputs. For each class, the network outputs a probability of detection of the considered class. A detection indicator is computed from the probability: it is set to 1 if the probability is greater than 0.5 and to 0 otherwise. For each class, this detection indicator is smoothed by a median filter along the time axis. This is done to avoid spurious detections, *e.g.* the vacuum cleaner has a rather continuous sound and it is highly improbable that it appears on one frame, disappears on the next one and reappears on the following one.

When analyzing the duration of the different sound classes on the synthetic database, we noticed that they greatly vary from one class to another as shown in section 2.1. We propose to adapt the length of the window size of the median filters to each of the groups defined in that section. The signal window hop corresponds to 16.6ms and, due to the pooling along the time axis, an output frame corresponds to 66.4ms of the original signal sampled at 22,050 Hz.

In practice, we use the following setup:

- Median window length of 41 (*i.e.* 2.7s) for background sounds "Electric shaver/toothbrush", "Frying", "Running water" and "Vacuum cleaner".
- Median window length of 13 (*i.e.* 0.9s) for intermediate sounds "Blender", "Cat" and "Speech".
- Median window length of 5 (*i.e.* 0.3s) for impulsive sounds "Alarm/bell/ringing", "Dishes" and "Dog".

3.3. Data augmentation

In addition to these modifications, we propose to introduce data augmentation by shifting the normalized mel-spectrogam along both the time and frequency axes. This kind of data augmentation along with the modifications from sections 3.1 and 3.2 constitutes Model 1. We also trained Model 2 which is the same as Model 1 with an additional form of data augmentation, *i.e.* adding noise to the mel-spectrogram values.

The time shift is set to a maximum of 270 frames (forward and backward with a Normal distribution with zero mean and a standard deviation of 90 frames). For time-shifting, the spectrogram is wrapped along the time axis, *e.g.* for a positive time-shift, the last frames of the spectrogram become the first frames of the shifted spectrogram. Moreover, for samples where strong labels are available, the strong labels are of course shifted accordingly.

We also propose to add a frequency shift in the mel domain. In that case, the shift is applied after feature extraction on the mel spectrogram. The maximum shift is set to a maximum of 8 bands (up and down in frequency) with a Normal distribution with a zero mean and standard deviation of 8/3.

The noise addition (for Model 2 only) is performed relatively to the normalized mel-spectrogram. A global signal level is computed for the whole normalized mel-spectrogram and a signal-to-noise ratio is set randomly according to a uniform law within a given range. A global noise level can then be computed and a random Gaussian noise with zero mean and a standard deviation matching the global noise level is then added.

4. RESULTS

In DCASE 2019 task 4, the event-based F1-score (macro-average) is used to evaluate the performance. In Table 2, the results that we obtained for our proposed models are given for validation 2019 database. The results are also available for each class, for information. Model 1 was trained on 200 epochs and Model 2 was trained on 300 epochs. In both cases, the master (smoothed) model was evaluated.

Table 2: F1-score (event-based) for proposed Models 1 and 2.

Class	F1-score Model 1 (%)	F1-score Model 2 (%)
Alarm/bell/ringing	46.4	47.2
Blender	39.6	43.9
Cat	41.6	43.6
Dishes	34.6	31.7
Dog	28.9	32.0
Electric shaver/toothbrush	57.4	60.7
Frying	42.2	45.1
Running water	35.8	35.7
Speech	41.2	44.2
Vacuum cleaner	52.8	52.0
Global	**42.1**	**43.6**

The performance reaches 42.1% (F1-score, event-based) for Model 1 and 43.6% for Model 2, which is significantly higher than the baseline (23.7%).

5. CONCLUSION

In this paper, we proposed a sound event detector based on a mean-teacher model (CRNN) and inspired by the provided baseline. The performance of the baseline (23.7%) could be significantly increased thanks to data augmentation (time, frequency and added noise), a new architecture and a class dependent median filter. Finally our proposed models respectively reach 42.1% of F1-score (event-based) for Model 1 and 43.6% for Model 2.

6. REFERENCES

[1] http://dcase.community/challenge2019/task-sound-event-detection-in-domestic-environments

[2] http://dcase.community/challenge2019/task-sound-event-detection-in-domestic-environments#baseline

[3] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results" in arXiv: 1703.01780, 2017.

[4] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in Machine Learning for Signal Processing (MLSP), 2015 IEEE 25th International Workshop on. IEEE, 2015.

[5] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. IEEE, 2016.

[6] L. JiaKai, "Mean teacher convolution system for DCASE 2018 Task 4," Technical Report, DCASE2018 Challenge, 2018.

[7] S. Laine and T. Aila, "Temporal Ensembling for Semi-Supervised Learning" in arXiv: 1610.02242, 2017.