HODGEPODGE: SOUND EVENT DETECTION BASED ON ENSEMBLE OF SEMI-SUPERVISED LEARNING METHODS

Technical Report

Ziqiang Shi

Fujitsu Research and Development Center, Beijing, China
shiziqiang@cn.fujitsu.com

ABSTRACT

In this technical report, we present the techniques and models applied to our submission for DCASE 2019 task 4: Sound event detection in domestic environments. We aim to focus primarily on how to apply semi-supervise learning methods efficiently to deal with large amount of unlabeled in-domain data. Three semi-supervised learning principles have been used in our system, including: 1) Consistency regularization applies data augmentation; 2) MixUp regularizer requiring that the prediction for a interpolation of two inputs is close to the interpolation of the prediction for each individual input; 3) MixUp regularization applies to interpolation between data augmentations. We also tried an ensemble of various models, trained by using different semi-supervised learning principles.

Index Terms— DCASE 2019, sound event detection, semi-supervised learning, convolutional recurrent neural networks

1. INTRODUCTION

The Detection and Classification of Acoustic Scenes and Events (D-CASE) is a series of challenges aimed at developing sound classification and detection systems [1, 2, 3]. Task 4 is sound event detection in domestic environments, the aim is to predict the presence or absence and the onset and offset times of sound events. Task 4 provides weakly-labelled data, unlabelled data and simulated strongly-labelled data for training. For the detailed information about the dataset and the challenge, please refer [3].

2. PROPOSED METHODS

Herein, we present a framework of our submission for task 4 of DCASE 2019.

2.1. Feature extraction

The dataset for task 4 is composed of 10 sec audio clips recorded in domestic environment or synthesized to simulate a domestic environment. The datasets are from AudioSet [4], FSD [5] and SIN-S dataset [6]. No preprocessing step was applied in the presented frameworks. The acoustic features for the 44.1kHz original data used in this system consist of 128-dimensional log mel-band energy extracted in Hanning windows of size 2048 with 431 points overlap. Thus the maximum number of frames is 1024.

2.2. Model architecture

For the detection of acoustic events, we apply a convolutional recurrent neural network (CRNN), which is used as the baseline system for task 4 of DCASE 2019. The log mel-band energy is then fed to the CRNN, which has seven 2-D convolutional layers and then two layers of bi-directional gated recurrent units (BiGRU) followed by a dense layer with sigmoid activation to compute posterior probabilities of the different sounds classes. Pooling along the time axis is used in training with the segment-level and frame-level labels. There are two outputs in this network. The output from BiGRU followed by dense layers with sigmoid activation is considered as sound event detection result. This output can be used to predict event activity probabilities. The other output is the weighted average of the element-wise multiplication of the attention, considering as audio tagging result. Final loss of the network is the weighted sum of these two outputs.

2.3. Semi-supervised learning

Three semi-supervised learning methods were used in our framework, the first one is the 'Mean Teacher' [7] used in the baseline. 'Mean Teacher' applies data augmentation to semi-supervised learning by leveraging the idea that a student model and a teacher model, which is the exponential moving average of student parameters, should output the same class distributions for the same unlabeled example even after it has been augmented. The goal of 'Mean Teacher' is to minimize

$$L = L_S + w(t)L_{US} \tag{1}$$

where L_S is the usual cross-entropy supervised learning loss over labeled samples, and L_{US} is the consistency regularization term over unlabeled data.

The second is the Interpolation Consistency Training (ICT) [8]. ICT learns a student network in a semi-supervised manner. To this end, ICT uses a 'Mean Teacher' $f_{\theta'}$. During training, the student parameters θ are updated to encourage consistent predictions

$$f_\theta(\text{Mix}_\lambda(u_j, u_k)) \approx \text{Mix}_\lambda(f_{\theta'}(u_j), f_{\theta'}(u_k)), \tag{2}$$

and correct predictions for labeled examples, where

$$\text{Mix}_\lambda(a, b) = \lambda a + (1 - \lambda)b \tag{3}$$

is called the interpolation or MixUp [9]. In our system, we perform interpolation of labeled sample pair and their corresponding labels on both the supervised loss on labeled examples and the consistency loss on unsupervised examples.

The third one draws on some of the ideas in MixMatch [10], but not exactly the same. MixMatch introduces a single loss that unifies entropy minimization, consistency regularization, and generic regularization approaches to semi-supervised learning. MixMatch can only be used for one-hot labels, not suitable for task 4, where there may be several events in a single audio clip. So we didn't use MixMatch in its original form, just borrowed some ideas from MixMatch, including generating $K(>1)$ different augmentations for unlabeled examples, then doing MixUp between these augmentations, and encourages the consistency of teacher and student on these MixUps.

2.4. Model ensemble and submission

For this challenge, We submitted 4 prediction results with different model ensemble, under the team name 'BossLee'.

- Shi_BossLee_task4_1.output.csv: Ensemble model is conducted by averaging the outputs of different models with different maximum consistency coefficients in 'Mean Teacher' principle. The f-score on validation data was 0.367.
- Shi_BossLee_task4_2.output.csv: Ensemble model is conducted by averaging the outputs of different models with different maximum consistency coefficients in ICT principle. The f-score on validation data was 0.425.
- Shi_BossLee_task4_3.output.csv: Ensemble model is conducted by averaging the outputs of different models with different maximum consistency coefficients in MixMatch principle. The f-score on validation data was 0.389.
- Shi_BossLee_task4_4.output.csv: Ensemble model is conducted by averaging the outputs of the models in Submission 1, 2, and 3. The f-score on validation data was 0.417.

3. REFERENCES

[1] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the dcase 2016 challenge," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 2, pp. 379–393, 2018.

[2] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "Dcase 2017 challenge setup: Tasks, datasets and baseline system," in *DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events*, 2017.

[3] http://dcase.community/challenge2019/.

[4] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[5] E. Fonseca, J. Pons Puig, X. Favory, F. Font Corbera, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra, "Freesound datasets: a platform for the creation of open audio datasets," in *Hu X, Cunningham SJ, Turnbull D, Duan Z, editors. Proceedings of the 18th ISMIR Conference; 2017 oct 23-27; Suzhou, China.[Canada]: International Society for Music Information Retrieval; 2017. p. 486-93.* International Society for Music Information Retrieval (ISMIR), 2017.

[6] 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 (D-CASE2017), Munich, Germany*, 2017, pp. 32–36.

[7] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in neural information processing systems*, 2017, pp. 1195–1204.

[8] V. Verma, A. Lamb, J. Kannala, Y. Bengio, and D. Lopez-Paz, "Interpolation consistency training for semi-supervised learning," *arXiv preprint arXiv:1903.03825*, 2019.

[9] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[10] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *arXiv preprint arXiv:1905.02249*, 2019.