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

Ziqiang Shi, Liu Liu, Huibin Lin, Rujie Liu

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

Anyan Shi

ShuangFeng First
Beijing, China

ABSTRACT

In this paper, we present a method called HODGEPODGE[1] for large-scale detection of sound events using weakly labeled, synthetic, and unlabeled data in the Detection and Classification of Acoustic Scenes and Events (DCASE) 2019 challenge Task 4: Sound event detection in domestic environments. To perform this task, we adopted the convolutional recurrent neural networks (CRNN) as our backbone network. In order to deal with the small amount of tagged data and the large amounts of unlabeled in-domain data, we aim to focus primarily on how to apply semi-supervise learning methods efficiently to make full use of limited 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 predictions 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, which are trained by different semi-supervised learning principles. Our approach significantly improved the performance of the baseline, achieving a event-based f-measure of 42.0% compared to 25.8% of the baseline on the official evaluation dataset. Our submissions ranked third among 18 teams in the task 4.

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

1. INTRODUCTION

Sound carries a lot of information about our everyday environment and the physical events that take place there. We can easily perceive the sound scenes we are in (busy streets, offices, etc.) and identify individual sound events (cars, footsteps, etc.). The automatic detection of these sound events has many applications in real life. For example, it's very useful for intelligent devices, robots, etc., in the environment awareness. Also a sound event detection system can help to construct a complete monitoring system when the radar or video system may not work in some cases.

To contribute to the sound event detection task, the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge has been organized for four years since 2013 [1, 2, 3]. DCASE is a series of challenges aimed at developing sound classification and detection systems [1, 2, 3]. This year, the DCASE 2019 challenge comprises five tasks: acoustic scene classification, audio tagging with noisy labels and minimal supervision, sound event localization and detection, sound event detection in domestic environments, and urban sound tagging [3]. Among them, this paper describes a method for the task 4 of the DCASE 2019 challenge, large-scale detection of sound events in domestic environments using real data either weakly labeled or unlabeled, or synthetic data that is strongly labeled (with time stamps). The aim is to predict the presence or absence and the onset and offset times of sound events in domestic environments. This task is the follow-up to DCASE 2018 task 4, which aims at exploring the possibility to exploit a large amount of unbalanced and unlabeled training data together with a small weakly annotated training set to improve system performance. The difference is that there is an additional training set with strongly annotated synthetic data is provided in this year's task 4. Thus it can be seen that we are faced with three difficult problems: 1) there is no real strongly labeled and only too few weakly labeled data, 2) the synthetic data is obviously different from the real one, and how is the effect of synthetic data on the detection results? and 3) there is too much unlabeled data. Although this task is difficult , there have been a variety of methods proposed to solve this problem [4, 5, 6]. Furthermore, a baseline system that performs the task is provided in the DCASE 2019 challenge [7, 5].

Based on these previous studies, we propose to apply a convolutional recurrent neural network (CRNN), which is used as the backbone network in the baseline system for task 4 of DCASE 2019 [3]. In order to make full use of small amount of weakly labeled and synthetic data, the principles in interpolation consistency training (ICT) [8] and MixMatch [9] has been adopted in the 'Mean Teacher' [7, 5] framework. To avoid overfitting, consistency regularization on the provided unlabeled data is incorporated.

The rest of this paper is organized as follows: Section 2 introduces details of our proposed HODGEPODGE. The experiment settings and results are displayed and discussed in Section 3. We conclude this paper in Section 4.

2. PROPOSED METHOD

Herein, we present the method of our submissions for task 4 of DCASE 2019. In the following sections, we will describe the details of our approach, including feature extraction, network structure, how to use ICT and MixMatch in the context of 'Mean Teacher', and how to use unlabeled data.

---

[1]HODGEPODGE has two layers of meanings. The first layer is the variety of training data involved in the method, including weakly labeled, synthetic, and unlabeled data. The second layer refers to several semi-supervised principles involved in our method.

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. 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. In order to prevent the system from overfitting on the small amount of development data, we added random white noise (before log operation) to the melspectrogram in each mini-batch during training. The input to the network is fixed to be 10-second audio clip. If the input audio is less than 10 seconds, it is padded to 10 seconds; otherwise it is truncated to 10 seconds.

2.2. Neural network architecture

Figure 1 presents the CRNN network architecture employed in our HODGEPODGE. The audio signal is first converted to [128×1024] log-melspectrogram to form the input to the network. The first half of the network consists of the seven convolutional layers, where we use gated linear units (GLUs) instead of commonly rectified linear units (RELUs) or leaky ReLUs as nonlinear activations.
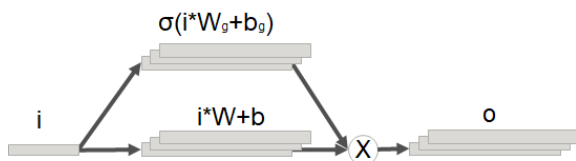

Figure 2: Architecture of a GLU.

Figure 2 shows the structure of a GLU :

$$o = (i * W + b) \otimes \sigma(i * W_g + b_g),$$

where i and o are the input and output, W, b, W_g, and b_g are learned parameters, σ is the sigmoid function and $\otimes$ is the element-wise product between vectors or matrices. Similar to LSTMs, GLUs play the role of controlling the information passed on in the hierarchy. This special gating mechanism allows us to effectively capture long-range context dependencies by deepening layers without encountering the problem of vanishing gradient.

For the seven gated convolutional layers, the kernel sizes are 3, the paddings are 1, the strides are 1, and the number of filters are [16, 32, 64, 128, 128, 128, 128] respectively, and the poolings are [(2, 2), (2, 2), (1, 2), (1, 2), (1, 2), (1, 2), (1, 2)] respectively. Pooling along the time axis is used in training with the clip-level and frame-level labels.

The gated convolutional blocks are followed by two bidirectional gated recurrent units (GRU) layers containing 64 units in the forward and backward path, their output is concatenated and passed to the attention and classification layer which are described below.

As depicted in Figure 1, the output of the bidirectional GRU layers is fed into both a frame-level classification block and an attention block respectively. The frame-level classification block uses a sigmoid activation function to predict the probability of each occurring class at each frame. Thus bidirectional GRUs

followed by a dense layer with sigmoid activation to compute posterior probabilities of the different sounds classes. In that case there are two outputs in this CRNN. The output from bidirectional GRUs 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. Thus the final prediction for the weak label of each class is determined by the weighted average of the element-wise multiplication of the attention and classification block output of each class c.

2.3. Semi-supervised learning

Inspired by the DCASE 2018 task 4 winner solution [5] and the baseline system [10], in which it uses the 'Mean Teacher' model [7], we also used 'Mean Teacher' as the main framework of our system.. 'Mean Teacher' is a combination of two models: the student model and the teacher model. At each training step, the student model is trained on synthetic and weakly labeled data with binary cross entropy classification cost. While the teacher model is an exponential moving average of the student models. The student model is the final model and the teacher model is designed to help the student model by a consistency mean-squared error cost for frame-level and clip-level predictions of unlabeled audio clips. That means good student should output the same class distributions as the teacher for the same unlabeled example even after it has been perturbed by Gaussian noise augmentation. The goal of 'Mean Teacher' is to minimize:

$$L = L_w + L_s + w(t)L_{cw} + w(t)L_{cs}$$

where L_w and L_s are the usual cross-entropy classification loss on weakly labeled data with only weak labels and synthetic data with only strong labels respectively, L_{cw} and L_{cs} are the teacher-student consistence regularization loss on unlabeled data with predicted weak and strong labels respectively, and $w(t)$ is the balance of classification loss and the consistency loss. Generally the $w(t)$ changes over time to make the consistency loss initially accounts for a very small proportion, and then the ratio slowly becomes higher. Since in the beginning, neither the student model nor the teacher model were accurate on predictions, and the consistency loss did not make much sense. $w(t)$ has a maximum upper bound, that is, the proportion of consistent loss does not tend to be extremely large. With different maximum upper bound of consistence weight $w(t)$, the trained model has different performances. In the next section, we ensemble the models trained under different maximum consistence weights to achieve better results.

HODGEPODGE did not change the overall framework of the baseline. It only attempts to combine several of the latest semi-supervised learning methods under this framework.

The first attempt is the interpolation consistency training (ICT) principle [8]. ICT teaches the 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

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

and correct predictions for labeled examples, where
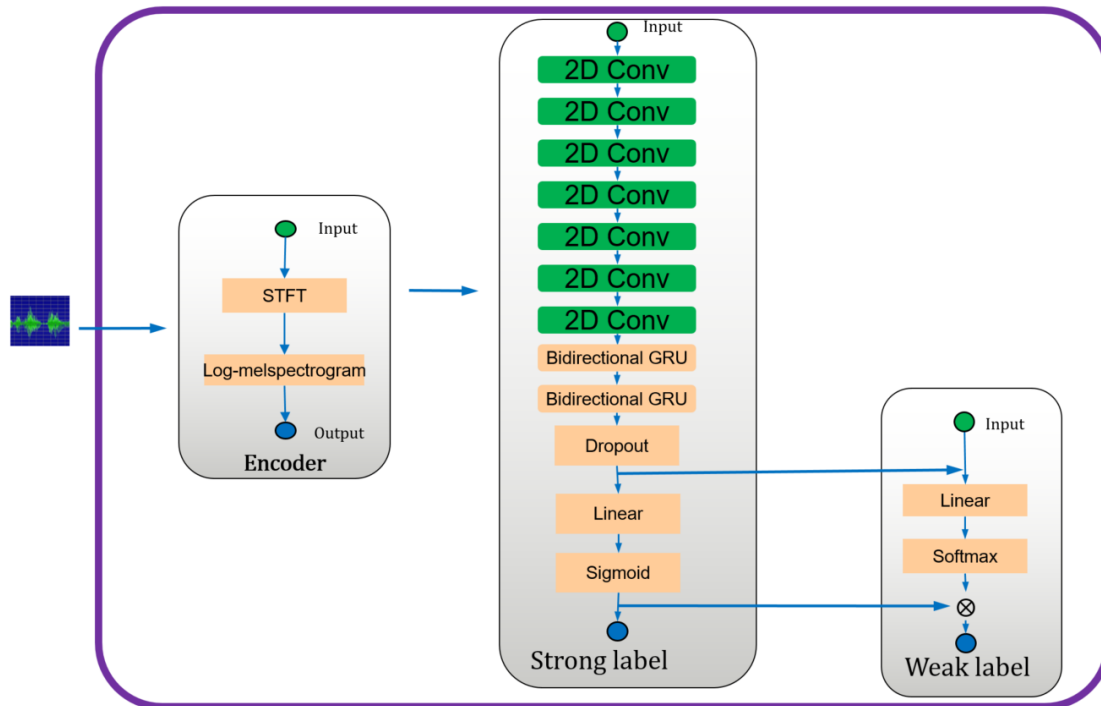
$$\text{Mix}_\lambda(a, b) = \lambda a + (1 - \lambda)b$$

Figure 1: Architecture of the CRNN in HODGEPODGE.

is called the interpolation or MixUp [11] of two log-melspectrograms u_j and u_k, and on each batch we sample a random λ from Beta(α; α) (e.g. $\alpha = 1.0$ in all our settings). In sum, the population version of our ICT term can be written a. In our system, we perform interpolation of sample pair and their corresponding labels (or pseudo labels predicted by the CRNNs) in both the supervised loss on labeled examples and the consistency loss on unsupervised examples. In each batch, the weakly labeled data, synthetic data, and unlabeled data are shuffled separately to form a new batch. There are 24 audio log-melspectrograms in each batch, of which 6 are weakly labeled, 6 are synthetic audio data, and the remaining 12 are unlabeled. Then we use the ICT principle to generate new augmented data and labels with the corresponding clips in the original and new batches. It should be noted that λ is different for each batch. Thus the loss

$$L_{ict} = L_{w,ict} + L_{s,ict} + w(t)L_{cw,ict} + w(t)L_{cs,ict}$$

where $L_{w,ict}$ and $L_{s,ict}$ are the classification loss on weakly labeled data with only weak labels and synthetic data with only strong labels using ICT respectively, $L_{cw,ict}$ and $L_{cs,ict}$ are the teacher-student consistence regularization loss on ICT applied on unlabeled data with predicted weak and strong labels respectively.

The second try draws on some of the ideas in MixMatch [9], but not exactly the same. MixMatch introduces a single loss that unifies entropy minimization, consistency regularization, and generic regularization approaches to semi-supervised learning. Unfortunately MixMatch can only be used for one-hot labels, not suitable for task 4, where may be several events in a single audio clip. So we didn't use MixMatch in its original form. In each batch, $K(> 1)$ different augmentations are generated, then the original

MixMatch does mixup on all data, regardless of whether the data is weakly labeled, synthetic or unlabeled. In our experiments, we found that the effect is not good, so we fine-tuned the MixMatch to do MixUp only between the augmentations of the same data type. The loss function is similar to the loss in the ICT case.

2.4. Model ensemble and submission

To further improve the performance of the system, we use some ensemble methods to fuse different models. The main differences of the single models have two dimensions, one is the difference of the semi-supervised learning method, and the other is the difference of the maximum value of the consistency loss weight. For this challenge, we submitted 4 prediction results with different model ensemble:

- HODGEPODGE 1: Ensemble model is conducted by averaging the outputs of 9 different models with different maximum consistency coefficients in 'Mean Teacher' principle. The F-score on validation data was 0.367. (Corresponding to Shi_FRDC_task4_1 in official submissions)

- HODGEPODGE 2: Ensemble model is conducted by averaging the outputs of 9 different models with different maximum consistency coefficients in ICT principle. The F-score on validation data was 0.425. (Corresponding to Shi_FRDC_task4_2 in official submissions)

- HODGEPODGE 3: Ensemble model is conducted by averaging the outputs of 6 different models with different maximum consistency coefficients in MixMatch principle. The F-score on validation data was 0.389. (Corresponding to

Shi_FRDC_task4_3 in official submissions)

- HODGEPODGE 4: Ensemble model is conducted by averaging the outputs of all the 24 models in Submission 1, 2, and 3. The F-score on validation data was 0.417. (Corresponding to Shi_FRDC_task4_4 in official submissions)

3. EXPERIMENTS AND RESULTS

3.1. Dataset

Sound event detection in domestic environments [11] is a task to detect the onset and offset time steps of sound events in domestic environments. The datasets are from AudioSet [12], FSD [13] and SINS dataset [14]. The aim of this task is to investigate whether real but weakly annotated data or synthetic data is sufficient for designing sound event detection systems. There are a total of 1578 real audio clips with weak labels, 2045 synthetic audio clips with strong labels, and 14412 unlabeled in domain audio clips in the development set, while the evaluation set contains 1168 audio clips. Audio recordings are 10 seconds in duration and consist of polyphonic sound events from 10 sound classes.

3.2. Evaluation Metric

The evaluation metric for this task is based on the event-based F-score [15]. The predicted events are compared to a list of reference events by comparing the onset and offset of the predicted event to the overlapping reference event. If the onset of the predicted event is within 200 ms collar of the onset of the reference event and its offset is within 200 ms or 20% of the event length collar around the reference offset, then the predicted event is considered to be correctly detected, referred to as true positive. If a reference event has no matching predicted event, then it is considered a false negative. If the predicted event does not match any of the reference events, it is considered a false positive. In addition, if the system partially predicts an event without accurately detecting its onset and offset, it will be penalized twice as a false positive and a false negative. The following equation shows the calculation of the F-score for each class.

$$F_c = \frac{2TP_c}{2TP_c + FP_c + FN_c},$$

where F_c, TP_c, FP_c, FN_c are the F-score, true positives, false positives, false negatives of the class c respectively. The final evaluation metric is the average of the F-score for all the classes.

3.3. Results

First we did some experiments to determine the best size of the median window. The median window is used in the post-processing of posterior probabilities to results in the final events with onset and offset. Table 1 shows the performance of HODGEPODGE systems on validation data set under different median window size. Coincidentally, all methods achieve the best performance when the window size is 9.

Table 2 shows the final macro-averaged event-based evaluation results on the test set compared to the baseline system. In fact, HODGEPODGE 1 is the ensemble of baselines, the only difference is that we use a deeper network, as well as higher sampling rate and larger features. It can be seen that both ICT and MixMatch principles can improve performance, especially ICT, which performs best in all HODGEPODGE systems.

Table 1: The performance of HODGEPODGE systems on validation data set under different median window size.

Median window size	5	7	9	11	13
HODGEPODGE 1	35.7%	36.4%	36.7%	36.5%	36.1%
HODGEPODGE 2	41.4%	42.1%	42.5%	42.2%	42.1%
HODGEPODGE 3	38.1%	38.7%	38.9%	38.3%	37.9%
HODGEPODGE 4	40.8%	41.5%	41.7%	41.3%	40.9%

Table 2: The performance of our approach compared to the baseline system.

Method	Evaluation	Validation
HODGEPODGE 1	37.0%	36.7%
HODGEPODGE 2	42.0%	42.5%
HODGEPODGE 3	40.9%	38.9%
HODGEPODGE 4	41.5%	41.7%
Baseline	25.8%	23.7%

4. CONCLUSIONS

In this paper, we proposed a method called HODGEPODGE for sound event detection using only weakly labeled, synthetic and unlabeled data. Our approach is based on CRNNs, whereby we introduce several latest semi-supervised learning methods, such as interpolation consistence training and MixMatch into the 'Mean Teacher' framework to leverage the information in audio data that are not accurately labeled. The final F-score of our system on the evaluation set is 42.0%, which is significantly higher than the score of the baseline system which is 25.8%.

5. ACKNOWLEDGEMENT

Many thanks to the anonymous reviewers for their comments, which make our description better.

6. 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] R. Serizel, N. Turpault, H. Eghbal-Zadeh, and A. P. Shah, "Large-scale weakly labeled semi-supervised sound event detection in domestic environments," *arXiv preprint arXiv:1807.10501*, 2018.

[5] L. JiaKai, "Mean teacher convolution system for dcase 2018 task 4," DCASE2018 Challenge, Tech. Rep., September 2018.

[6] Q. Kong, T. Iqbal, Y. Xu, W. Wang, and M. D. Plumbley, "Dcase 2018 challenge baseline with convolutional neural networks," *arXiv preprint arXiv:1808.00773*, 2018.

[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] 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.

[10] N. Turpault, R. Serizel, A. P. Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," 2019.

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

[12] 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.

[13] 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.

[14] 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*, 2017, pp. 32–36.

[15] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.