

Hodge and Podge: Hybrid Supervised Sound Event Detection with Multi-Hot MixMatch and Composition Consistence Training

Ziqiang Shi*, Liu Liu, Huibin Lin, and Rujie Liu

Fujitsu Research and Development Center, Beijing, China

February 17, 2020

Abstract

In this paper, we propose a method called Hodge and Podge for sound event detection. We demonstrate Hodge and Podge on the dataset of Detection and Classification of Acoustic Scenes and Events (DCASE) 2019 Challenge Task 4. This task aims to predict the presence or absence and the onset and offset times of sound events in home environments. Sound event detection is challenging due to the lack of large scale real strongly labeled data. Recently deep semi-supervised learning (SSL) has proven to be effective in modeling with weakly labeled and unlabeled data. This work explores how to extend deep SSL to result in a new, state-of-the-art sound event detection method called Hodge and Podge. With convolutional recurrent neural networks (CRNN) as the backbone network, first, a multi-scale squeeze-excitation mechanism is introduced and added to generate a pyramid squeeze-excitation CRNN. The pyramid squeeze-excitation layer can pay attention to the issue that different sound events have different durations, and to adaptively recalibrate channel-wise spectrogram responses. Further, in order to remedy the lack of real strongly labeled data problem, we propose multi-hot MixMatch and composition consistency training with temporal-frequency augmentation. Our experiments with the public DCASE2019 challenge task 4 validation data resulted in an event-based F-score of 43.4%, and is about absolutely 1.6% better than state-of-the-art methods in the challenge. While the F-score of the official baseline is 25.8%.

1 Introduction

Recently, sound event detection (SED) has become more and more popular in the field of acoustic signal processing, since it can be widely used in everyday life. By leveraging large corpora of strongly annotated sound data, where the onset and offset times of sound events have been annotated, that allow state-of-the-art models such as deep neural networks (DNN) can learn the characteristics of sound events to tackle the problem of SED. However, acquiring such large amounts of strongly labeled data is practically infeasible, since annotating the onset and offset times of sound events take more time than annotating audio clips for classification.

Obviously, if there is no real strongly labeled data, then the data can use is only weakly labeled data, maybe synthesized data, and unlabeled data. In SED, weak labels refer to ground truth labels that where only the presence of the sound events is labeled, with no temporal information on the onset nor offset of audio events. In addition to the weakly labeled data, we can also synthesize sound data to simulate the target application environment. Manually verified foreground events are embedded in the background texture, which is similar to the target environment. Then the distribution of sound events per class, the number of sound events per clip and the sound event class co-occurrence can be designed to be similar to the real recordings. At the same time, a large amount of unlabeled audio data can be recorded in the target environment. These unlabeled in-domain data may help us to augment supervised training methods, for example by semi-supervised learning method [17, 2]. This is the situation that we are concerned in this

*Corresponding author: shiziqiang@cn.fujitsu.com; shiziqiang7@gmail.com

paper, and also is precisely the challenge proposed by the Detection and Classification of Acoustic Scenes and Events (DCASE) 2019 task 4 [14].

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 provided in this year’s task 4. Furthermore, a baseline system that performs the task is provided in the DCASE 2019 challenge [13, 9], and there have been a variety of methods proposed to solve this problem [11, 9, 8].

This paper is based on the ‘Mean Teacher’ approach [13, 9], which has achieved the best result in DCASE 2018 task 4 [11], and also based on HODGEPODGE [12], which has ranked third in DCASE 2019 task 4. We propose to introduce a multi-scale squeeze-excitation mechanism into the standard convolutional recurrent neural network (CRNN) as the backbone network for both student and teacher in ‘Mean Teacher’ framework. In order to make full use of a small amount of weakly labeled and synthetic data, different temporal-frequency augmentation methods are utilized to supplement supervised training. In parallel, we introduce composition consistency training and multi-hot MixMatch training to encourage the model to produce the same mixed output distribution when its inputs are perturbed and interpolated, to generalize well and avoid overfitting on the labeled data, and to output confident predictions on unlabeled data.

2 Hybrid Supervised Sound Event Detection

Compared to the strongly or weakly supervised SED task, where strongly labeled onset and offset annotations or clip-level labels for the training set are given, the hybrid supervised SED task contains clip-level labels for weakly labeled data, strong labels for synthetic data, and unlabeled in-domain data.

In order to deal with this ill-posed problem, [9] introduces the ‘Mean Teacher’ approach [13] to do a similar task. The difference in situation is that there is no out-of-domain unlabeled audio here, but instead, synthetic strongly labeled data is easily obtained and provided. ‘Mean Teacher’ employed and adapted in Hodge and Podge is a combination of two models: the student model and the teacher model as shown in Fig. 1. At each training step, the student model is trained on synthetic and weakly labeled data with binary cross-entropy (CE) 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 on frame-level and clip-level predictions of unlabeled audio clips. That means a good student should output the same class distributions as the teacher for the same unlabeled example even after it has been perturbed by augmentation which will be introduced in Section 3.1.

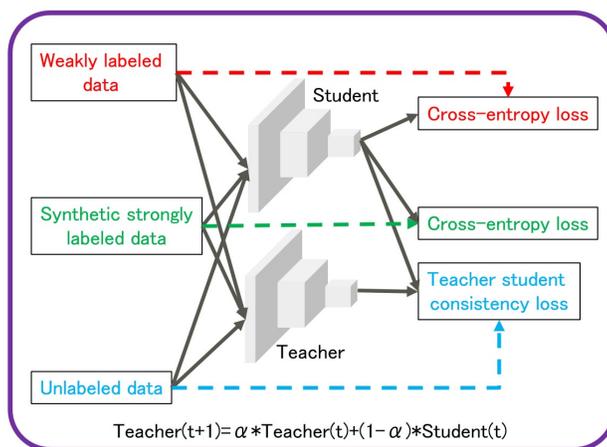


Figure 1: ‘Mean Teacher’ framework employed and adapted in Hodge and Podge.

3 Hodge and Podge

Herein, in the following sections, we will describe the details of Hodge and Podge, including feature extraction, pyramid squeeze-excitation CRNN, multi-hot MixMatch training, and composition consistency training.

3.1 Feature Extraction and Augmentation

Our proposal Hodge and Podge are developed and evaluated on the dataset provided by the DCASE 2019 challenge task 4 [14]. The dataset for task 4 is composed of 10-sec audio clips recorded in a 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. The input to the network is fixed to be a 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.

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 Mel spectrogram in each mini-batch during training. We also propose to introduce data augmentation by temporal-frequency shift. The temporal shift augmentation is a random shift of the signal by rolling the signal along the time axis. The frequency shift augmentation is a random roll in the range $\pm 5\%$ around the frequency axis in the Mel domain. A wrap-around both temporal-frequency shifts to preserve all information. Here $\pm 5\%$ wrap-around in frequency does not affect the sound much physically or perceptually, but can generate a lot of augmented data. One thing to note is that the frame-level labels of strongly annotated synthetic data also have to be shifted accordingly over the temporal shift.

3.2 Pyramid Squeeze-Excitation CRNN

A CRNN is used to map the Mel spectrogram to the frame-level and clip-level posterior probabilities of sound event presence. In order to improve the quality of representations by explicitly modeling the interdependencies between the channels of convolution layers in CRNN, the squeeze-excitation mechanism [7] is introduced to CRNN for SED. The structure of 2D Conv with squeeze-excitation mechanism, which is called ‘SE 2D Conv’ building block, is depicted in Fig. 2. The squeeze-excitation can learn to use global information to selectively emphasize informative features and suppress less useful ones. The feature maps are passed through a squeeze operation (Global Pooling) and an excitation operation (a simple gating mechanism with a Sigmoid activation and ReLU function using two fully connected Linear layers) to produce a collection of per-channel modulation weights, which are applied to the original feature maps to generate the output of the ‘SE 2D Conv’ which can be fed directly into subsequent layers of the network.

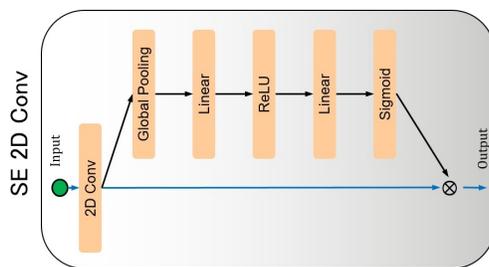


Figure 2: SD Conv with squeeze-excitation mechanism (SE 2D Conv).

With the squeeze-excitation mechanism and GLUs, Fig 3 presents the network architecture employed in our Hodge and Podge. In order to distinguish it from the traditional CRNN, here we call this backbone pyramid squeeze-excitation CRNN (PSE-CRNN). The audio signal is first converted to $[128 \times 1024]$ log-Mel spectrogram to form the input to the network. Here in the Fig 3, the horizontal arrow from the ‘encoder’ block to the ‘strong label’ block actually means the output of the ‘encoder’ block is the input to the ‘strong

label’ block. The first half of the ‘strong label’ block network consists of one pyramid plain convolutional layer with three parallel 2D convolutional modules of different kernel sizes 3, 5, and 7 (considering different sound events have different spans in time-frequency domain), and seven squeeze-excitation gated convolutional layers, for which the kernel sizes are 3, the paddings are 1, the strides are 1, and the numbers 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 SE convolutional blocks are followed by two bidirectional gated recurrent units (GRU) layers containing 64 units in the forward and backward path, their outputs are concatenated and passed to the attention and classification layer which are described below.

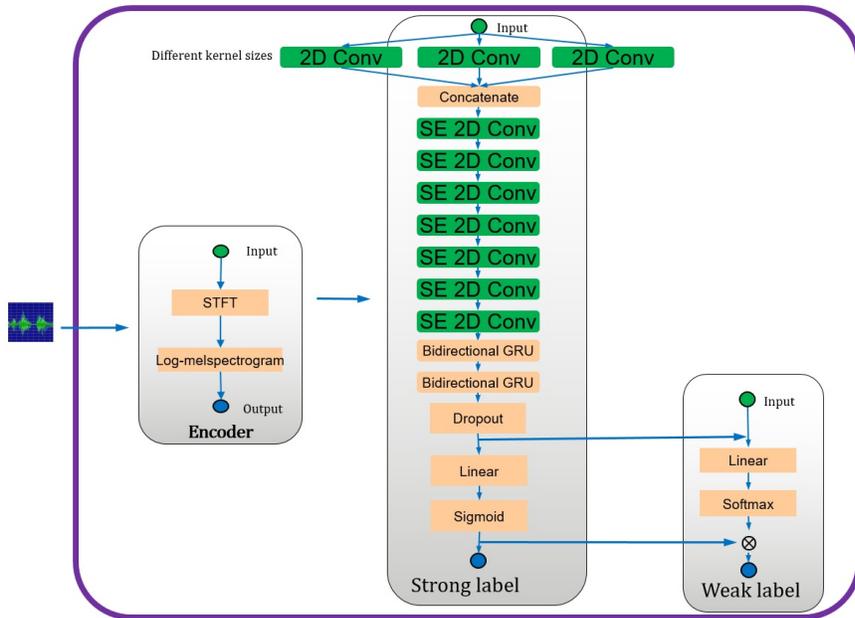


Figure 3: Architecture of the PSE-CRNN in Hodge and Podge.

As depicted in Fig. 3, the output of the bidirectional GRU layers is fed into both a frame-level classification block and an attention block (the ‘weak label’ block in the figure) respectively. Thus there are two outputs in this PSE-CRNN. The output from bidirectional GRUs followed by dense layers with sigmoid activation is considered as the sound event detection result. This output can be used to predict event activity probabilities. The other output from the ‘weak label’ block 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.

3.3 Composition Consistency Training

Under the framework of hybrid supervised sound event detection described in Section 2 and shown in Fig. 1, we propose to introduce composition consistency training (CCT) to make full use of unlabeled clips. The intuition of CCT is that by weighted combining the original data and the augmented data, the prediction results of the model will still be the same weighted combination, whether it is on labeled data or unlabeled data. During training, there are 24 audio log-Mel spectrograms in each batch, of which 6 are weakly labeled, 6 are synthetic, and the remaining 12 are unlabeled. The model 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

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

is called the MixUp [16] of two log-Mel spectrograms u_j and u_k , and in each batch we sample a random λ from $\text{Beta}(\alpha; \alpha)$ (e.g. $\alpha = 1.0$ in all our settings). It should be noted that λ is different for each batch. In CCT, we perform MixUp of sample pair and their corresponding labels (or pseudo labels predicted by the PSE-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. Then we use the CCT principle to generate new augmented data and labels with the corresponding clips in the original and new batches.

Thus in CCT, for weakly labeled data, clip level classification CE loss on MixUp of original data batch (D) and shuffled augmented data batch aug_D is computed as

$$L_{w,\text{CCT}} = \text{CE}(f_{\theta}^w(\text{Mix}_{\lambda}(\text{D}, \text{aug_D})), \text{Mix}_{\lambda}(f_{\theta}^w(\text{D}), f_{\theta}^w(\text{aug_D}))),$$

where θ is the parameter of PSE-CRNN model and f^w is the clip level posterior probability output. For synthetic strongly labeled data, frame level classification CE loss on MixUp data of original batch and shuffled aug_D is computed as

$$L_{s,\text{CCT}} = \text{CE}(f_{\theta}^s(\text{Mix}_{\lambda}(\text{D}, \text{aug_D})), \text{Mix}_{\lambda}(f_{\theta}^s(\text{D}), f_{\theta}^s(\text{aug_D}))).$$

where f^s is the frame level posterior probability output. For unlabeled data, clip level and frame level posterior probability consistency loss based between student model and teacher model on MixUp data of D and shuffled aug_D is calculated by mean square error

$$L_{cw,\text{CCT}} = \|f_{\theta}^w(\text{Mix}_{\lambda}(\text{D}, \text{aug_D})) - \text{Mix}_{\lambda}(f_{\theta'}^w(\text{D}), f_{\theta'}^w(\text{aug_D}))\|$$

and

$$L_{cs,\text{CCT}} = \|f_{\theta}^s(\text{Mix}_{\lambda}(\text{D}, \text{aug_D})) - \text{Mix}_{\lambda}(f_{\theta'}^s(\text{D}), f_{\theta'}^s(\text{aug_D}))\|$$

respectively, here θ' is the parameter of teacher model.

The total loss

$$L_{\text{CCT}} = L_{w,\text{CCT}} + L_{s,\text{CCT}} + w(t)(L_{cw,\text{CCT}} + L_{cs,\text{CCT}})$$

where generally the $w(t)$ changes overtime to make the consistency loss initially accounts for a very small proportion, and then the ratio slowly becomes higher. $w(t)$ has a maximum upper bound, that is, the proportion of consistency loss does not tend to be extremely large. With different maximum upper bound of consistency weight $w(t)$, the trained model has different performances.

3.4 Multi-Hot MixMatch Training

We also try to introduce the latest semi-supervised learning principles in MixMatch [1]. 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 classification, not suitable for our situation, where there may be several events in a single audio clip. Thus we introduce multi-hot MixMatch (M^3), our proposed semi-supervised learning method to simplify MixMatch for sound event detection.

Original MixMatch uses a Sharpen operation to let the average predictions across all augmentations approaching a Dirac (one-hot) distribution. Our multi-hot MixMatch discarded this average and Sharpen operation. During training, 2 different augmentations, for example, denoted as aug_A and aug_B , are generated for each sample in the batch, and multi-hot MixMatch to do MixUp only between the augmentations of the same data type. That means we perform interpolation of sample pairs and their corresponding labels (or pseudo labels predicted by the PSE-CRNNs) in both the supervised loss on labeled examples and the consistency loss on unsupervised examples. In each training batch, the original weakly

labeled data, synthetic strongly labeled data, and unlabeled data in `aug_A` and `aug_B` are shuffled separately to form a new batch. For original weakly labeled data, clip level classification CE loss on MixUp data from original data batch (D) and shuffled `aug_A` is computed as

$$L_{w,M^3} = \text{CE}(f_{\theta}^w(\text{Mix}_{\lambda}(D, \text{aug_A})), \text{Mix}_{\lambda}(f_{\theta}^w(D), f_{\theta}^w(\text{aug_A}))).$$

For synthetic strongly labeled data, frame level classification CE loss on MixUp data from original batch and shuffled `aug_B` is computed as

$$L_{s,M^3} = \text{CE}(f_{\theta}^s(\text{Mix}_{\lambda}(D, \text{aug_B})), \text{Mix}_{\lambda}(f_{\theta}^s(D), f_{\theta}^s(\text{aug_B}))).$$

For unlabeled data, clip level and frame level posterior probability consistency loss based between student model and teacher model on MixUp data from shuffled `aug_A` and shuffled `aug_B` is calculated by mean square error

$$L_{cw,M^3} = \|f_{\theta}^w(\text{Mix}_{\lambda}(\text{aug_A}, \text{aug_B})) - \text{Mix}_{\lambda}(f_{\theta'}^w(\text{aug_A}), f_{\theta'}^w(\text{aug_B}))\|$$

and

$$L_{cs,M^3} = \|f_{\theta}^s(\text{Mix}_{\lambda}(\text{aug_A}, \text{aug_B})) - \text{Mix}_{\lambda}(f_{\theta'}^s(\text{aug_A}), f_{\theta'}^s(\text{aug_B}))\|$$

respectively, here θ' is the parameter of teacher model.

The total loss

$$L_{M^3} = L_{w,M^3} + L_{s,M^3} + w(t)(L_{cw,M^3} + L_{cs,M^3}),$$

where generally the $w(t)$ changes, which is consistent with the functions and settings of CCT in Section 3.3.

3.5 Model Ensemble

To further improve the performance of the system, we use ensemble methods to fuse different models. The main differences between the single models are the maximum values of the consistency loss weight. An ensemble model is constructed by averaging the outputs of several different models with different maximum consistency coefficients.

4 Experiments and Results

4.1 Dataset

We evaluated our Hodge and Podge on the dataset of DCASE 2019 challenge task 4. The datasets are from AudioSet [6], FSD [5] and SINS dataset [4]. 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. For further information about that dataset, such as the number of items per class, distribution over classes, properties of sound events, and specific sound characteristics, please refer [14].

The evaluation metric for this task is based on the event-based F-score [10]. 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. The onset of the right predicted event should be 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.

Median window size	7	9	11	13
HODGEPODGE	41.1%	41.7%	41.7 %	41.4%
Hodge	43.0%	43.4%	43.2%	42.8%
Podge	42.7%	42.4%	42.2%	42.0%
DCASE baseline	25.8%			

Table 1: The performance of Hodge and Podge on validation data set under different median window size.

4.2 Systems

Four systems are evaluated and compared across the above conditions:

- **HODGEPODGE**: The solution provided by [12], which ranked third in the task 4 [3]. An ensemble model is constructed by averaging the outputs of different models with different maximum consistency coefficients in the interpolation consistency training principle [15].
- **Hodge**: The method proposed in Section 3 using PSE-CRNN with multi-hot MixMatch training principle 3.4.
- **Podge**: The method proposed in Section 3 using PSE-CRNN with composition consistency training principle 3.3.
- **Baseline**: The official solution provided by [14] is based on a ‘Mean Teacher’-type algorithm [9] with plain CRNN, which is as described in Section 2.

4.3 Results

The systems are evaluated with macro-averaged event-based F-score [10]. Practically the predicted frame-level strong label result should be continuous, thus we get a smooth prediction result through a 1-dimensional median filter in the time dimension. The median window size in Table 1 indicates the size, which gives the shape that is taken from the input original predictions, at every element position, to define the input to the filter function. Since the official test set is not public, we use the official validation set as the test set, and divide the training set into 9:1, as our own training set and validation set. In this experiment, HODGEPODGE [12] is chosen as the baseline. Table 1 shows the final macro-averaged event-based evaluation results on the official validation set compared to the baseline system. In the DCASE2019 task 4 challenge, on the official test set, HODGEPODGE achieved 42.0% performance, while the first place was 42.7% [3]. Compare them in parallel, HODGEPODGE showed 0.7% gap with the state-of-the-art on the official test set, while it can be seen from the Table 1 our Hodge achieved an average performance improvement of 1.6% compared to HODGEPODGE on the official validation set. This indicates that our Hodge has achieved state-of-the-art performance. It can be seen from the Table 1, the attempts made in Hodge and Podge do improve performance. Hodge and Podge are better than HODGEPODGE in all different median window sizes, and the average absolute improvement in F-score was about 1.6% and 0.8% respectively.

5 Conclusions

In this paper, we proposed a method called Hodge and Podge for sound event detection using only weakly labeled, synthetic and unlabeled data. Our approach is based on CRNN, whereby we introduce pyramid squeeze-excitation mechanism, composition consistency training, and multi-hot MixMatch consistency training with temporal-frequency augmentation to leverage the information in audio data that are not accurately labeled. The final F-score of our system on the official validation set is 43.4%, which is significantly higher than the score of the baseline system which is 25.8%.

References

- [1] Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.: Mixmatch: A holistic approach to semi-supervised learning. arXiv preprint arXiv:1905.02249 (2019)
- [2] Chapelle, O., Scholkopf, B., Zien, A.: Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. IEEE Transactions on Neural Networks 20(3), 542–542 (2009)
- [3] DCASE, R.: Detection and classification of acoustic scenes and events (dcase 2019) challenge task 4 results (2019)
- [4] Dekkers, G., Lauwereins, S., Thoen, B., Adhana, M.W., Brouckxon, H., van Waterschoot, T., Vanrumste, B., Verhelst, M., Karsmakers, P.: 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. pp. 32–36 (2017)
- [5] Fonseca, E., Pons Puig, J., Favory, X., Font Corbera, F., Bogdanov, D., Ferraro, A., Oramas, S., Porter, A., Serra, X.: 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] Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 776–780. IEEE (2017)
- [7] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
- [8] Kong, Q., Iqbal, T., Xu, Y., Wang, W., Plumbley, M.D.: Dcase 2018 challenge baseline with convolutional neural networks. arXiv preprint arXiv:1808.00773 (2018)
- [9] Lu, J.: Mean teacher convolution system for dcase 2018 task 4. Tech. rep., DCASE2018 Challenge (September 2018)
- [10] Mesáros, A., Heittola, T., Virtanen, T.: Metrics for polyphonic sound event detection. Applied Sciences 6(6), 162 (2016)
- [11] Serizel, R., Turpault, N., Eghbal-Zadeh, H., Shah, A.P.: Large-scale weakly labeled semi-supervised sound event detection in domestic environments. In: Workshop on Detection and Classification of Acoustic Scenes and Events (2018)
- [12] Shi, Z., Liu, L., Lin, H., Liu, R., Shi, A.: Hodgepodge: Sound event detection based on ensemble of semi-supervised learning methods. arXiv preprint arXiv:1907.07398 (2019)
- [13] Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in neural information processing systems. pp. 1195–1204 (2017)
- [14] Turpault, N., Serizel, R., Shah, A.P., Salamon, J.: Sound event detection in domestic environments with weakly labeled data and soundscape synthesis. In: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019) (2019)
- [15] Verma, V., Lamb, A., Kannala, J., Bengio, Y., Lopez-Paz, D.: Interpolation consistency training for semi-supervised learning. arXiv preprint arXiv:1903.03825 (2019)
- [16] Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
- [17] Zhu, X.J.: Semi-supervised learning literature survey. Tech. rep., University of Wisconsin-Madison Department of Computer Sciences (2005)