NCSP'22
RISP International Workshop on Nonlinear circuits, Communications and Signal Processing 2022
February 28 - March 1

Semi-Supervised Learning Using Weakly Labeled Data Generated by GAN in Sound Event Detection

Kazuya Ouma[1], Takeshi Yamada[1] and Shoji Makino[1,2]

[1]Graduate School of Science and Technology, University of Tsukuba, Ibaraki 305–8573, Japan
[2]Graduate School of Information, Production and Systems, Waseda University, Fukuoka 808–0135, Japan
E-mail: s2020573@s.tsukuba.ac.jp, takeshi@cs.tsukuba.ac.jp, s.makino@waseda.jp

Abstract

The performance of sound event detection (SED) has been considerably improved by applying deep learning-based methods. However, they still have one drawback that a large amount of data with strong labels consisting of occurring event classes, beginning and end times are needed. To solve the problem, some methods using data with weak labels consisting of only occurring event classes, which require less preparation, have been proposed. Furthermore, the use of unlabeled data has been attracting attention as a means of further reducing the cost of preparing training data. In this paper, we apply semi-supervised learning using a generative adversarial network (GAN) to SED and extend it to generate weakly labeled data. Unlike simple data augmentation with GAN, generated data are used not only for the training of SED but also for the training of GAN to obtain the effect of multi-task learning. To evaluate the effectiveness of the proposed method, we conducted a SED experiment using the dataset provided by Detection and Classification of Acoustic Scenes and Events (DCASE) 2020 Task 4. The proposed method improved the event-based F1 score by 3.76% and the segment-based F1 score by 6.59% compared with the baseline method without GAN.

1. Introduction

Sound event detection (SED) is important for applications such as automated vehicles and security systems. As shown in Fig. 1, SED detects the beginning and end of each sound event and identifies its class. It is expected to be used to monitor the situation outside automated vehicles and to detect anomalous sounds in security systems.

As highlighted in Detection and Classification of Acoustic Scenes and Events (DCASE) [1], an international competition for environmental sound recognition, deep learning using a neural network (NN), especially a convolutional neural network (CNN), is often used for SED. Training for SED requires a large amount of data with strong labels consisting of occurring event classes and beginning and end times, but strong labeling requires the analysis of acoustic data by listening and visual confirmation of frequency spectrograms.

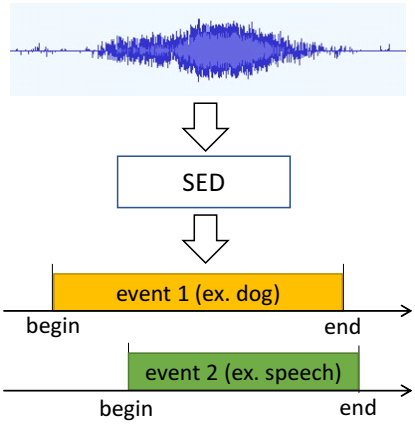


Figure 1: Overview of SED

Therefore, the preparation of thousands of data required for training has a huge human and time cost.

As a conventional solution, methods using data with weak labels consisting of only occurring event classes, which require less preparation, have been proposed to improve the detection accuracy [2, 3]. Recently, the use of unlabeled data has been attracting attention as a means of further reducing the cost of preparing training data. For example, a method that applies Mean-Teacher [4] to SED, i.e., semi-supervised learning has been proposed [5].

In the field of image recognition, a semi-supervised learning method based on a generative adversarial network (GAN) [6] that uses unlabeled image data has been proposed [7]. The original GAN consists of a generator that generates realistic data and a discriminator that determines whether the input data is realistic or not (binary real/fake classification). On the other hand, in semi-supervised learning using GAN [7], the image classification model is trained to solve the image recognition task (main task) using a small amount of labeled data and to perform the binary real/fake classification task (subtask) with a large amount of unlabeled data. This corresponds to multi-task learning [8], in which a single network is trained to solve multiple similar tasks. It is known that multi-task learning improves the accuracy and generalization

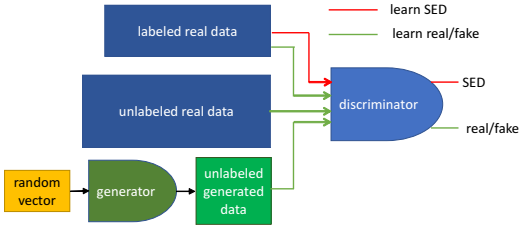


Figure 2: Apllication of semi-supervised learning using GAN to SED

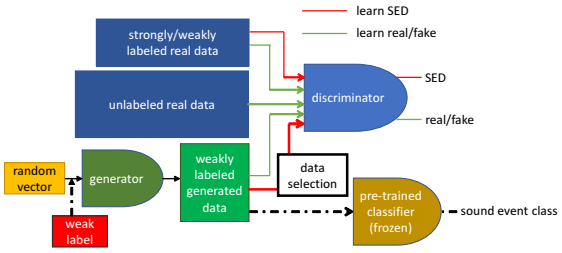


Figure 3: Training of discriminator using weakly labeled generated data

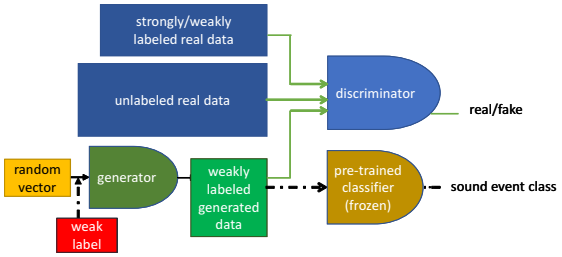


Figure 4: Training of generator to generate weakly labeled data

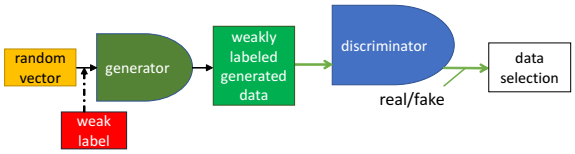


Figure 5: Selection of generated data by binary real/fake classification

performance of the main task even if there are not enough training data.

In this paper, we apply semi-supervised learning using GAN [7] to SED and extend it to generate weakly labeled data. Unlike simple data augmentation with GAN, generated data are used not only for the training of SED but also for the training of GAN to obtain the effect of multi-task learning. We verify the effectiveness of our proposed method through a comparison experiment.

2. Semi-supervised learning with GAN

In this paper, we apply semi-supervised learning using GAN [7] to SED. Fig. 2 shows its overview. The generator works in the same manner as in the original GAN, i.e., it generates data from an input random vector, and is trained so that the generated data are considered realistic by the discriminator. On the other hand, the discriminator is trained for the SED task (main task) using a small amount of strongly/weakly labeled real data and is also trained for the binary real/fake classification task using strongly/weakly labeled real data, unlabeled real data, and unlabeled generated data. The ability to extract features common to both binary real/fake classification and SED is improved; thus, it is expected to improve the detection accuracy and generalization performance of SED.

3. Proposed method

In the proposed method, as described above, the discriminator deals with SED as the main task and the binary real/fake classification task as the subtask. The difference from the method in Section 2. is that the weakly labeled generated data are directly used for training for the SED task, as well as for the binary real/fake classification task.

As shown in Fig. 3, the discriminator and the generator perform the training alternately. The discriminator uses the generated data not only for the subtask of binary real/fake classification but also for the main task of SED. It means that the discriminator is trained using the weakly labeled data generated in each training step, in addition to the strongly/weakly labeled real data and unlabeled real data. By updating the parameter with weakly labeled data generated in each step during training, we expect more accurate optimization for SED. Note that since the learning speed of the discriminator was faster than that of the generator in an experiment described below, we decided not to train for SED in the early stage of learning.

On the other hand, the generator is trained to generate data of a specified event class using the Auxiliary Classifier GAN method [9] as shown in Fig. 4. The generator maximizes the cross-entropy loss of the discriminator to generate more realistic data and minimizes the cross-entropy loss of the classifier to generate data where an event of a specified class occurs. Here, the classifier is pre-trained for the acoustic event class classifier, and its network parameters are frozen.

To ensure that the data generated by the generator improves the accuracy of SED, we select reliable generated data to be used for training for SED, as shown in Fig. 5. We perform batchwise binary real/fake classification on the generated data and use only the data that are considered realistic for SED training. As a result, we may augment the data using more realistic data.

The difference between the proposed method and the

method that performs simple data augmentation using GAN [10] is that the proposed method continues the training of GAN during the training for SED. It means that the proposed method enables to obtain the effect of multi-task learning [8].

4. Evaluation of the effectiveness of proposed method

4.1 Experiment overview

In this experiment, we used the dataset provided by DCASE 2020 Task 4 [11]. DCASE 2020 Task 4 involves the detection of acoustic events in a home environment. In this task, we are required to detect 10 classes of events, such as human voices and dog barks. The training data are all 10-sec acoustic data: 2584 strongly labeled real data generated using the impulse response of real measurements, 1578 weakly labeled real data from real recordings, and 14412 unlabeled real data from real recordings. The amount of unlabeled data is much larger than the amount of strongly and weakly labeled data, so deciding to utilize the unlabeled data is one of the challenges in this task.

The evaluation metric for this task is the event-based F1 score, and the score of the baseline method for this task is 34.80% [11]. The higher the F1 score, the higher the detection accuracy. Here, event-based refers to whether or not the detection was performed correctly, including the start and end times for each event. The segment-based F1 score is also used. This is a second-by-second evaluation of whether the occurring event has been detected correctly. The event-based score is a more stringent evaluation criterion than the segment-based score.

In this experiment, we compared the following four learning methods:

- **Baseline SED**
  It trains for SED using only strongly and weakly labeled real data; GAN is not used.
- **Proposed method A**
  It generates unlabeled data and uses them only for training for binary real/fake classification to improve the detection accuracy by multi-task learning. It corresponds to the method in Section 2..
- **Proposed method B**
  It generates weakly labeled data and uses them for training for SED as well as for binary real/fake classification to further improve the detection accuracy. It corresponds to the method in Section 3..
- **Simple data augmentation using GAN [10]**
  It uses weakly labeled data generated by the pre-trained generator for training for SED.

Table 1 summarizes the data used by each method for SED and binary real/fake classification. When training with weakly labeled data in each method, we used a method in which combinations of different multiple instance strategies and different poolings are considered [3].

Table 1: Training data used in each method

		Real			Generated	
		Strong	Weak	Unlabeled	Weak	Unlabeled
Baseline SED	SED	○	○	-	-	-
	Real/Fake	-	-	-	-	-
Proposed method A	SED	○	○	-	-	-
	Real/Fake	○	○	○	-	○
Proposed method B	SED	○	○	-	○	-
	Real/Fake	○	○	○	○	-
Simple data augmentation using GAN [10]	SED	○	○	-	○	-
	Real/Fake	-	-	-	-	-

Table 2: F1 scores in each method

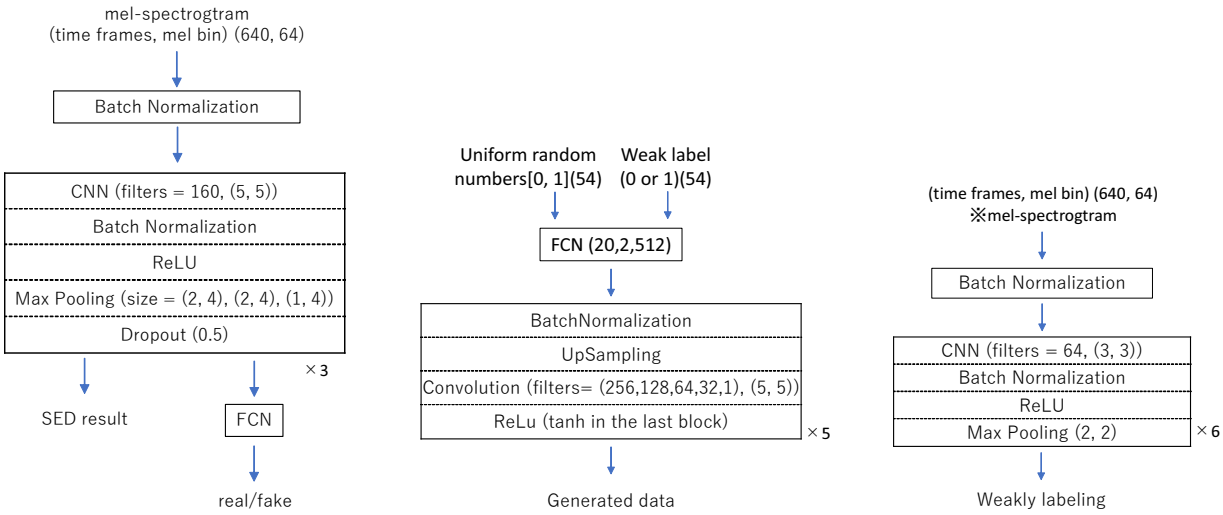
	event-based	segment-based
Baseline SED	36.70%	56.63%
Proposed method A	39.31%	61.84%
Proposed method B	40.46%	63.22%
Simple data augmentation using GAN [10]	37.82%	59.31%

Fig. 6 shows the network configurations for the discriminator, generator, and classifier in this experiment. These network configurations are common for all methods. The discriminator is based on the literature [3] and consists of three layers of CNN blocks. The generator is based on the GAN-based acoustic data generation method [12], which first generates a feature map from a random vector using a fully connected network, then expands the feature map using CNN and upsampling blocks, and finally outputs fake acoustic data. In accordance with the method used in DCASE 2020 Task 4, which uses weak label classification [13], the classifier consists of six layers of CNNs.

4.2 Experimental results and discussion

Table 2 shows the experimental results. From the table, we can see that the proposed method A improved the F1-score compared with the baseline SED method by utilizing unlabeled data. The proposed method B further improved the event-based F1 score by 1.15% and the segment-based F1 score by 1.38% compared with the proposed method A. The proposed method B also improved the F1-score compared with the method with simple data augmentation using GAN. These results confirmed that by integration of multi-task learning and data augmentation using GAN, further improvement of the detection accuracy was achieved.

Furthermore, to investigate the effect of the selection of the generated data in the proposed method B, we conducted an additional experiment using the proposed method B without the selection of the generated data. The result showed that the data selection processing improved the event-based F1 score by 0.55% and segment-based F1 score by 0.70%, respectively.



(a) discriminator (b) generator (c) classifier

Figure 6: Network configurations for the discriminator, generator, and classifier

5. Conclusion

In this paper, we applied semi-supervised learning using GAN to SED and extended it to generate weakly labeled data. Unlike simple data augmentation with GAN, generated data are used not only for the training of SED but also for training of GAN to obtain the effect of multi-task learning. To evaluate the effectiveness of the proposed method, we conducted a SED experiment using the dataset provided by DCASE 2020 Task 4. The results confirmed that the proposed method improved the event-based F1 score by 3.76% and the segment-based F1 score by 6.59% compared with the baseline SED method without GAN.

Acknowledgment

This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Nos. 20K11880 and 19H04131.

References

[1] http://dcase.community/

[2] T. Matsuyoshi, T. Komatsu, R. Kondo, T. Yamada, S. Makino, "Weakly labeled Learning using BLSTM-CTC for sound event detection," Proc. APSIPA ASC, pp. 1918–1923, Nov. 2018.

[3] Y. Huang *et al.*, "Multi-branch learning for weakly-labeled sound event detection," Proc. ICASSP, pp. 641–645, 2020.

[4] A. Tarvainen *et al.*, "Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results," in arXiv:1703.01780, 2017.

[5] J. Lu, "Mean Teacher Convolution System for DCASE 2018 Task 4," DCASE 2018 Task 4 Technical Report, 2018.

[6] I. J. Goodfellow *et al.*, "Generative adversarial nets," Proc. NIPS, vol. 2, pp. 2672–2680, 2014.

[7] T. Salimans *et al.*, "Improved techniques for training GANs," Proc. NIPS, pp. 2234–2242, 2016.

[8] R. Caruana, "Multitask learning: a knowledge-based source of inductive bias," Proc. ML, pp. 41–48, 1993.

[9] A. Odena *et al.*, "Conditional image synthesis with auxiliary classifier GANs," Proc. ML, vol. 70, pp. 2642–2651, 2017.

[10] X. Xia *et al.*, "Auxiliary classifier generative adversarial network with soft labels in imbalanced acoustic event detection," IEEE Trans. Multimedia, vol. 21, no. 6, pp. 1359–1371, 2019.

[11] http://dcase.community/challenge2020/task-sound-event-detection-and-separation-in-domestic-environments

[12] C. Donahue *et al.*, "Adversarial audio synthesis," Proc. ICLR, 2019.

[13] Y. Huang *et al.*, "Guided multi-branch learning systems for DCASE 2020 Task 4," DCASE 2020 Task 4 Technical Report, 2020.