Multitask Learning of Acoustic Scenes and Events Using Dynamic Weight Adaptation Based on Multi-focal Loss

Kayo Nada*, Keisuke Imoto*, Reina Iwamae, and Takao Tsuchiya
Doshisha University, Japan

***Abstract*—Acoustic scene classification (ASC) and sound event detection (SED) are principal tasks in environmental sound analysis. On the basis of the idea that acoustic scenes and sound events are closely relevant to each other, some groups previously proposed joint analysis of acoustic scenes and sound events utilizing multitask learning (MTL)-based neural network models. The MTL-based model shares information on acoustic scenes and sound events in mutual estimation. However, in the conventional methods, ASC and SED performances depend strongly on the learning weights of each ASC and SED task, and finding the appropriate balance between the learning weights of ASC and SED tasks is difficult. To address this problem, we therefore propose a dynamic weight adaptation method for multitask learning of ASC and SED based on multi–focal loss in this paper. Experimental results obtained using parts of the TUT Acoustic Scenes 2016/2017 and TUT Sound Events 2016/2017 show that the proposed method improves the scene classification and event detection performance by 3.52 and 3.27 percentage points in micro-Fscore compared with the conventional MTL-based method, respectively. Moreover, the experimental results also indicate that adapting the learning weights dynamically in accordance with the progress of model training improves the ASC and SED performances.**

I. Introduction

Environmental sound analysis has attracted increasing research interest in recent years, and it has significant potential in the development of various applications such as machine condition monitoring, automatic surveillance, media retrieval, and biomonitoring systems [1], [2], [3], [4]. Acoustic scene classification (ASC) and sound event detection (SED) are the primary tasks in environmental sound analysis. ASC is a task that predicts a predefined acoustic scene label in an audio recording, where the acoustic scene indicates the surroundings in which the audio is recorded, such as "office," "residential area," "train," and "indoor." SED involves detecting sound event labels and their time boundaries in the audio recording, where a sound event represents a sound class, such as "keyboard typing," "car," "cutlery," and "people talking."

For ASC and SED, neural-network-based methods, such as the convolutional neural network (CNN), convolutional recurrent neural network (CRNN), and Transformer, have been widely applied in recent works. For example, Valenti et al. have proposed a method for ASC based on CNN [5]. Liping et al. [6], Tanabe et al. [7], and Raveh et al. [8] have respectively proposed Xception, VGG, and ResNet-based scene classification methods, which have been widely used in image recognition. Hershey et al. have proposed an event detection method using CNN [9]. Çakır et al. have proposed a SED method utilizing CRNN, which can capture temporal information of sound events [10]. More recently, Kong et al. [11] and Miyazaki et al. [12] have proposed Transformer-based and Conformer-based methods for SED, respectively.

Most conventional methods of environmental sound analysis address acoustic scene and sound event analysis separately; meanwhile, acoustic scenes and sound events are related to each other. For example, in the acoustic scene "home," the sound events "dishes" and "glass jingling" are likely to occur, whereas the sound events "car" and "bird singing" occur infrequently. Therefore, when we recognize the sound events "dishes" and "glass jingling," information on the acoustic scene "home" helps identify these sound events, and vice versa. On the basis of this fact, Mesaros et al. [13] and Heittola et al. [14] have proposed methods for SED taking information on acoustic scenes into account in an unsupervised manner. Imoto and Shimauchi [15] and Imoto and Ono[16] have proposed scene classification methods considering sound events using Bayesian generative models. Bear et al. [17] and Tonami et al. [18], [19] have proposed methods for the joint analysis of acoustic scenes and sound events utilizing MTL-based neural network models of ASC and SED. These methods train the MTL model of ASC and SED using a linear combination of ASC and SED losses with constant weights. However, the conventional works reported that scene classification and event detection performances depend on the constant weights of ASC and SED losses, and discovering appropriate weights of ASC and SED losses is difficult. Moreover, it may be preferred to change the learning weights dynamically in accordance with the progress of model training. In this work, we thus propose a dynamic weight adaptation method for multitask learning of ASC and SED based on multi–focal loss.

The remainder of this paper is structured as follows. In section 2, we introduce the conventional methods for ASC, SED, and joint analysis of ASC and SED using multitask learning. In section 3, we propose a dynamic adaptive method of loss weighting factors in multitask learning. In section 4, we discuss an experiment carried out to evaluate the performance of the proposed method. Finally, we conclude this work in section 5.

*These authors contributed equally to this work.

TABLE I
SOUND EVENTS OCCURRING IN EACH ACOUSTIC SCENE IN TUT ACOUSTIC SCENES 2016, 2017, TUT SOUND EVENTS 2016, AND 2017 [24], [25]

	(object) banging	(object) impact	(object) rustling	(object) snapping	(object) squeaking	bird singing	brakes squeaking	breathing	car	children	cupboard	cutlery	dishes	drawer	fan	glass jingling	keyboard typing	large vehicle	mouse clicking	mouse wheeling	people talking	people walking	washing dishes	water tap running	wind blowing
City center	-	-	-	-	-	-	○	-	○	○	-	-	-	-	-	-	-	○	-	-	○	○	-	-	-
Home	-	○	○	○	-	-	-	-	-	-	○	○	○	○	-	○	-	-	-	-	-	○	○	○	-
Office	-	○	○	-	○	-	-	○	-	-	-	-	-	-	○	-	○	-	○	○	○	○	-	-	-
Residential area	○	-	-	-	-	○	-	-	○	○	-	-	-	-	-	-	-	○	-	-	○	○	-	-	○

II. CONVENTIONAL METHODS

A. Conventional Methods for ASC and SED

For conventional methods of ASC and SED, many neural-network-based methods based on CNN [5], [9], CRNN [10], and Transformer [11], have been proposed. In this section, we overview the conventional scene classification and event detection methods based on neural networks. In many conventional methods for ASC and SED, the time–frequency representation of the observed acoustic signal $X \in \mathcal{R}^{D \times T}$, such as the time series of mel frequency cepstrum coefficients (MFCCs) or the log mel-band spectrogram, is used as the acoustic feature. Here, D and T are the number of frequency bins and the number of time frames of the input acoustic feature, respectively. This time–frequency representation is fed to the ASC or SED network. In ASC, which estimates the predefined acoustic scene label with which a sound clip is most associated, the model parameters are optimized utilizing the output of the network and the following cross-entropy (CE) loss function $\mathcal{L}_{\text{scene}}$:

$$\mathcal{L}_{\text{scene}}(\boldsymbol{\theta}) = -\sum_{n=1}^{N} \Big\{ z_n \log(y_n) \Big\}, \tag{1}$$

where N, y_n, and z_n are the number of acoustic scene classes, the output of the network, and the target scene label, respectively. The target label is 1 if the sound clip is most associated with acoustic scene n and 0 otherwise.

On the other hand, in SED, which detects the sound event labels and their start and end times, the model parameters are optimized using the output of the network and the following binary cross-entropy (BCE) loss function $\mathcal{L}_{\text{event}}$:

$$\begin{aligned}\mathcal{L}_{\text{event}}(\boldsymbol{\theta}) &= -\sum_{t=1}^{T} \Big\{ \mathbf{z}_t \log(\mathbf{y}_t) + (1-\mathbf{z}_t) \log(1-\mathbf{y}_t) \Big\} \\ &= -\sum_{t,m=1}^{T,M} \Big\{ z_{t,m} \log(y_{t,m}) + (1-z_{t,m}) \log(1-y_{t,m}) \Big\},\end{aligned} \tag{2}$$

where T, M, $y_{t,m}$, and $z_{t,m}$ are the number of time frames, the number of sound event classes, the prediction of sound event m in time frame n, and the target label, respectively.

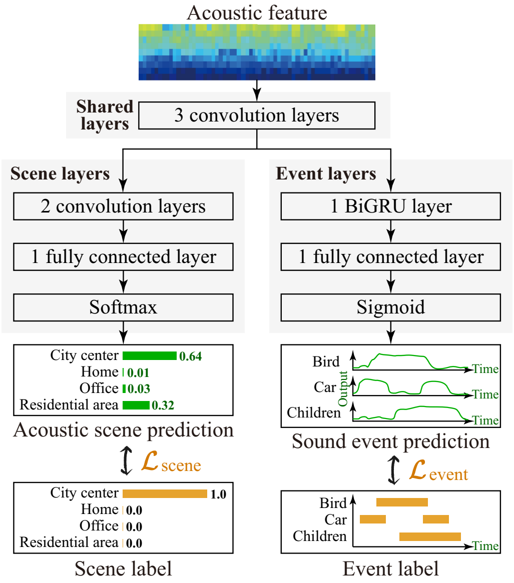

Fig. 1. Network structure of conventional MTL-based method [18]

B. Joint Analysis of Acoustic Scenes and Sound Events Using Multitask Learning

In most conventional methods, ASC and SED are studied separately. However, as shown in Table I, many acoustic scenes and sound events are closely related to each other; thus, information on acoustic scenes will help in detecting sound events, and vice versa. On the basis of this idea, environmental sound analysis based on multitask learning of ASC and SED has been proposed [17], [18], [19], [20]. In these methods, parts of the ASC and SED networks share the holding of information on acoustic scenes and sound events in common, as shown in Fig. 1.

In the conventional methods, the part of the network is shared to hold information of acoustic scenes and sound events in the shared layers. The CNN and bidirectional gated recurrent unit (BiGRU) layers are applied for scene classification network and event detection network, respectively. To train the multitask model of ASC and SED, the conventional method [18] utilizes the following loss function:

$$\mathcal{L}(\boldsymbol{\theta}) = \alpha \mathcal{L}_{\text{scene}}(\boldsymbol{\theta}) + \beta \mathcal{L}_{\text{event}}(\boldsymbol{\theta}), \tag{3}$$

where α and β are the constant weights of scene and event losses. In this work, $\beta = 1.0$ can be set without loss of generality.

III. Proposed Method

A. Motivation

The results of using conventional methods revealed that ASC and SED based on the multitask learning framework indeed improve the performances of classifying acoustic scenes and detecting sound events compared with the single task-based methods [18], [19]. However, ASC and SED performances depend on the weights α and β, and finding the appropriate balance between ASC and SED tasks is not easy. Moreover, in the conventional methods, the learning weights α and β are constant throughout model training. However, it may be preferable to change the learning weights dynamically in accordance with the progress of model training. To address this limitation of conventional methods, we thus propose a dynamic weight adaptation method for multitask learning of ASC and SED.

B. Dynamic Weight Adaptation of Multitask Learning Based on Multi–focal Loss

In the proposed method, we introduce focal loss [21], [22], [23] to dynamically adapt the weights of multitask learning of ASC and SED. Focal loss was originally proposed to dynamically adjust the training weight of the model in accordance with the difficulty/ease of training as follows:

$$\mathcal{L}(\boldsymbol{\theta}) = -\sum_{n=1}^{N} \Big\{ (1-y_n)^{\eta} z_n \log(y_n) \Big\}, \tag{4}$$

where η is the constant focusing parameter. Focal loss down-weights the training weight depending on the prediction error. To adopt this idea into the dynamic weight adaptation of multitask learning, we replace $\mathcal{L}_{\mathrm{scene}}(\boldsymbol{\theta})$ and $\mathcal{L}_{\mathrm{event}}(\boldsymbol{\theta})$ in Eq. (3) with the following multiple focal loss functions, respectively:

$$\mathcal{L}_{\mathrm{scene}}(\boldsymbol{\theta}) = -\sum_{n=1}^{N} \Big\{ (1-y_n)^{\eta} z_n \log(y_n) \Big\}, \tag{5}$$

$$\begin{aligned}\mathcal{L}_{\mathrm{event}}(\boldsymbol{\theta}) = -\sum_{t,m=1}^{T,M} \Big\{ & (1-y_{t,m})^{\gamma} z_{t,m} \log(y_{t,m}) \\ & + y_{t,m}^{\zeta} (1-z_{t,m}) \log(1-y_{t,m}) \Big\}, \end{aligned} \tag{6}$$

where γ and ζ are the constant focusing parameters. The proposed multi–focal loss function dynamically down-weights the training weights of ASC and SED and can determine the appropriate balance of the training weight between ASC and SED tasks automatically.

TABLE II
Detailed structure of MTL network of ASC and SED

Shared network	
Log-mel energy 500 frames × 64 mel bin	
3×3 kernel size / 128 ch. Batch norm., Leaky ReLU 1×8 Max pooling	
$\left(\begin{array}{c}3\times3 \text{ kernel size / 128 ch.}\\ \text{Batch norm., Leaky ReLU}\\ 1\times2 \text{ Max pooling}\end{array}\right) \times 2$	
Scene layers	**Event layers**
3×3 kernel size / 256 ch. Batch norm., Leaky ReLU 25×1 Max pooling	BiGRU w/ 32 units
3×3 kernel size / 256 ch. Batch norm., Leaky ReLU Global max pooling	FC w/ 32 units, Leaky ReLU
FC w/ 32 units, Leaky ReLU	FC w/ 25 units, sigmoid
FC w/ 4 units, Softmax	

TABLE III
Experimental conditions

Acoustic feature	Log-mel energy (64 dim.)
Frame length / shift	40 ms / 20 ms
Length of sound clip	10 s
Optimizer	RAdam [27]
Detection threshold of sound events	0.5

IV. Evaluation Experiments

A. Experimental Conditions

We evaluated the performance of the proposed dynamic weight adaptation of multitask learning. For the evaluation experiments, we built a dataset composed of parts of the TUT Acoustic Scenes 2016 and 2017 and TUT Sound Events 2016 and 2017 [24], [25]. From the TUT datasets, we selected sound clips including four acoustic scenes, "city center," "home," "office," and "residential area," for a total of 266 min of sounds (development set, 192 min; evaluation set, 74 min). These sound clips include the 25 types of sound events listed in Table I. The details of the dataset are found in [26].

As an acoustic feature, we applied the 64-dimensional log mel-band energy, which was calculated every 40 ms with a 20 ms hop size. The acoustic feature was fed into the MTL network proposed in [18]. The constant focusing parameters are set to $\gamma = 1.0$, $\zeta = 0.0625$, and $\eta = 0.5$ referring to [23]. For each method, we conducted the evaluation 10 times with random initial values of model parameters. The detailed network structure and other experimental conditions are listed in Tables II and III.

B. Experimental Results

1) Overall Performances of ASC and SED: Table IV shows the average experimental results for the conventional methods

TABLE IV
PERFORMANCES OF SCENE CLASSIFICATION AND EVENT DETECTION

Method	Scene classification		Event detection	
	Micro-Fscore	Macro-Fscore	Micro-Fscore	Macro-Fscore
CNN (ASC)	49.68%	45.67%	-	-
CNN-BiGRU (SED)	-	-	40.10%	7.39%
Conventional MTL ($\alpha = 0.001, \beta = 1.0$)	52.55%	43.41%	41.02%	6.83%
Proposed method ($\alpha = 0.001, \beta = 1.0, \gamma = 1.0, \zeta = 0.0625, \eta = 0.5$)	**55.07%**	**47.85%**	**44.29%**	**8.86%**

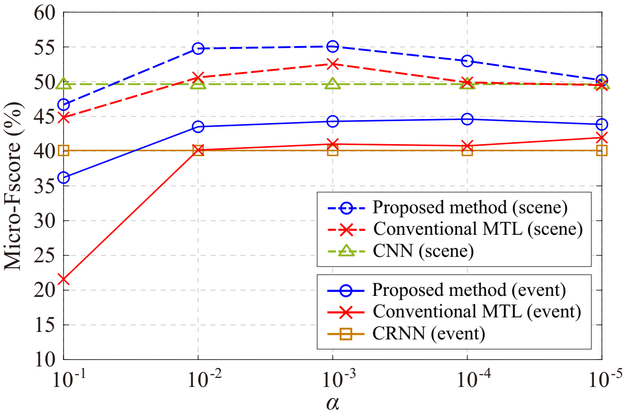

Fig. 2. Scene classification and event detection performance with various learning wights α

and the proposed method. For CNN and CNN-BiGRU, we applied the same network structures with shared + scene layers and shared + event layers, respectively. The results show that the proposed dynamic weight adaptation method achieves a reasonable performance in both ASC and SED tasks compared with the conventional MTL-based method. When $\alpha = 0.001$, $\beta = 1.0$, $\gamma = 1.0$, $\zeta = 0.0625$, and $\eta = 1.0$, micro-Fscores for scene classification and event detection with the proposed method are improved by 3.52 and 3.27 percentage points compared with those of the conventional MTL-based method.

To investigate how the proposed method determines the appropriate weights of ASC and SED, we evaluate the ASC and SED performances with various α values. Fig. 2 shows the scene classification and event detection performance with various learning weights α. The result indicates that the proposed method achieves a reasonable performance regardless of the weight α; thus, the proposed method enables us to apply the MTL-based method without considering the learning weights. Moreover, the result also indicates that adapting the learning weights dynamically in accordance with the progress of model training improves the ASC and SED performances.

2) Detailed Detection Results for Each Sound Event: To investigate the event detection performance in detail, we list the F-scores for selected sound events in Table V. The results show that the proposed method improves Fscores for many sound events. For instance, the proposed method detects the sound events "bird singing," "car," and "washing dishes" more accurately than the conventional MTL-based method. From Table I, these sound events are closely related to particular scenes, e.g., sound event "bird singing" only occurs in the acoustic scene "residential area." This indicates that when detecting sound events, the proposed method can take information on acoustic scenes into account more effectively than the conventional method. Meanwhile, the event detection performance for "people walking" is not improved. This is because the sound event "people walking" occurs in all acoustic scenes; thus, information on acoustic scenes may not help this sound event detection.

TABLE V
AVERAGE FSCORES FOR SELECTED SOUND EVENTS

Method	bird singing	car	people walking	washing dishes	water tap running
CNN-BiGRU	17.79%	43.85%	0.00%	0.41%	**43.23%**
Conventional MTL	23.72%	45.10%	0.00%	0.24%	6.55%
Proposed method	**32.73%**	**45.45%**	0.00%	**1.28%**	36.41%

V. CONCLUSIONS

In this paper, we proposed the dynamic weight adaptation method for multitask learning of ASC and SED. In the proposed method, we applied focal loss objective functions to each ASC and SED loss to dynamically adapt the learning weights of ASC and SED losses. Moreover, the proposed multi–focal loss can adapt the learning weights dynamically in accordance with the progress of model training. The experimental results obtained using parts of the TUT Acoustic Scenes 2016 and 2017 and TUT Sound Events 2016 and 2017 datasets indicate that the proposed multi–focal loss-based method outperforms the conventional MTL method by 3.52 and 3.27 percentage points of the micro-Fscore in scene classification and event detection tasks, respectively. Moreover, the experimental results also indicate that adapting the learning weights dynamically in accordance with the progress of model training improves the ASC and SED performances.

ACKNOWLEDGEMENT

This work was supported by JSPS KAKENHI Grant Number JP19K20304 and JP20H00613, and the Harris Science Research Institute.

References

[1] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda, N. Harada, Description and discussion on dcase2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring, Proc. Detection and Classification of Acoustic Scenes and Events (DCASE) (2020) 81–85.

[2] C. Chan, E. W. M. Yu, An abnormal sound detection and classification system for surveillance applications, Proc. European Signal Processing Conference (EUSIPCO) (2010) 1851–1855.

[3] E. Fonseca, M. Plakal, F. Font, D. P. W. Ellis, X. Favory, J. Jordi, X. Serra, General-purpose tagging of freesound audio with AudioSet labels: Task description, dataset, and baseline, Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE) (2018) 69–73.

[4] V. Morfi, R. F. Lachlan, D. Stowell, Deep perceptual embeddings for unlabelled animal sound, IEEE/ACM Trans. Audio Speech Lang. Process. 150 (1) (2020) 2–11.

[5] M. Valenti, S. Squartini, A. Diment, G. Parascandolo, T. Virtanen, A convolutional neural network approach for acoustic scene classification, Proc. International Joint Conference on Neural Networks (IJCNN) (2017) 1547–1554.

[6] Y. Liping, C. Xinxing, T. Lianjie, Acoustic scene classification using multi-scale features, Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE) (2018) 29–33.

[7] R. Tanabe, T. Endo, Y. Nikaido, T. Ichige, P. Nguyen, Y. Kawaguchi, K. Hamada, Multichannel acoustic scene classification by blind dereverberation, blind source separation, data augmentation, and model ensembling, Tech. Rep. DCASE Challenge 2018 Task5 (2018) 1–4.

[8] A. Raveh, A. Amar, Multi-channel audio classification with neural network using scattering transform, Tech. Rep. DCASE Challenge 2018 Task5 (2018) 1–4.

[9] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, K. Wilson, CNN architectures for large-scale audio classification, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2017) 131–135.

[10] E. Çakır, G. Parascandolo, T. Heittola, H. Huttunen, T. Virtanen, Convolutional recurrent neural networks for polyphonic sound event detection, IEEE/ACM Trans. Audio Speech Lang. Process. 25 (6) (2017) 1291–1303.

[11] Q. Kong, Y. Xu, W. Wang, M. D. Plumbley, Sound event detection of weakly labelled data with CNN-Transformer and automatic threshold optimization, IEEE/ACM Trans. Audio Speech Lang. Process. 28 (2020) 2450–2460.

[12] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, K. Takeda, Convolution-augmented transformer for semi-supervised sound event detection, Tech. Rep. DCASE Challenge 2020 Task4 (2020) 1–4.

[13] A. Mesaros, T. Heittola, A. Klapuri, Latent semantic analysis in sound event detection, Proc. European Signal Processing Conference (EUSIPCO) (2011) 1307–1311.

[14] T. Heittola, A. Mesaros, A. Eronen, T. Virtanen, Context-dependent sound event detection, EURASIP Journal on Audio, Speech, and Music Processing 2013 (1).

[15] K. Imoto, S. Shimauchi, Acoustic scene analysis based on hierarchical generative model of acoustic event sequence, IEICE Trans. Inf. Syst. E99-D (10) (2016) 2539–2549.

[16] K. Imoto, N. Ono, Acoustic topic model for scene analysis with intermittently missing observations, IEEE/ACM Trans. Audio Speech Lang. Process. 27 (2) (2019) 367–382.

[17] H. L. Bear, I. Nolasco, E. Benetos, Towards joint sound scene and polyphonic sound event recognition, INTERSPEECH (2019) 4594–4598.

[18] N. Tonami, K. Imoto, M. Niitsuma, R. Yamanishi, Y. Yamashita, Joint analysis of acoustic events and scenes based on multitask learning, Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) (2019) 333–337.

[19] N. Tonami, K. Imoto, R. Yamanishi, Y. Yamashita, Joint analysis of sound events and acoustic scenes using multitask learning, IEICE Trans. Inf. Syst. E104-D (02) (2021) 294–301.

[20] K. Imoto, N. Tonami, Y. Koizumi, M. Yasuda, R. Yamanishi, Y. Yamashita, Sound event detection by multitask learning of sound events and scenes with soft scene labels, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2020) 621–625.

[21] T. Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, Proc. IEEE International Conference on Computer Vision (ICCV) (2017) 2980–2988.

[22] K. Noh, J. H. Chang, Joint optimization of deep neural network-based dereverberation and beamforming for sound event detection in multichannel environments, Sensors 20 (7) (2020) 1–13.

[23] K. Imoto, S. Mishima, Y. Arai, R. Kondo, Impact of sound duration and inactive frames on sound event detection performance, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2021) 875–879.

[24] A. Mesaros, T. Heittola, T. Virtanen, TUT database for acoustic scene classification and sound event detection, Proc. European Signal Processing Conference (EUSIPCO) (2016) 1128–1132.

[25] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, B. Raj, T. Virtanen, DCASE 2017 challenge setup: Tasks, datasets and baseline system, Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE) (2017) 85–92.

[26] https://www.ksuke.net/dataset

[27] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, J. Han, On the variance of the adaptive learning rate and beyond, Proc. International Conference on Learning Representations (ICLR) (2020) 1–13.