AUDIO TAG REPRESENTATION GUIDED DUAL ATTENTION NETWORK FOR ACOUSTIC SCENE CLASSIFICATION

Ju-ho Kim[*], *Jee-weon Jung*[*] *Hye-jin Shim, and Ha-jin Yu*[†],

University of Seoul, School of Computer Science, Seoul, South Korea

ABSTRACT

Sound events are crucial to discern a specific acoustic scene, which establishes a close relationship between audio tagging and acoustic scene classification (ASC). In this study, we explore the role and application of sound events based on the ASC task and propose the use of the last hidden layer's output of an audio tagging system (*tag representation*), rather than the output itself (*tag vector*), in ASC. We hypothesize that the tag representation contains sound event information that can improve the classification accuracy of acoustic scenes. The dual attention mechanism is investigated to adequately emphasize the frequency-time and channel dimensions of the feature map of an ASC system using tag representation. Experiments are conducted using the Detection and Classification of Acoustic Scenes and Events 2020 task1-a dataset. The proposed system demonstrates an overall classification accuracy of 69.3%, compared to 65.3% of the baseline.

***Index Terms*—** tag representation, dual attention, acoustic scene classification

1. INTRODUCTION

Acoustic scene classification (ASC) classifies an input recording into one of the predefined scenes, and has been receiving increasing interest. The IEEE challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) provides a platform to facilitate research in related tasks with annual public datasets [1–4]. With recent advances in deep learning, many studies have adopted deep neural networks (DNNs) to compose an ASC system [5–7]. Various aspects have been studied, including DNN architecture, data augmentation, frameworks (end-to-end or explicit back-end classifier), and similarities between different scenes [8–12].

Among these studies, a few recent studies have focused on exploiting different tasks, such as audio tagging and sound event detection (SED), related to the ASC task [13–15] (detailed in Section 2). An audio tagging system predicts the posterior probability of predefined sound events in the input audio recording, whereas a SED system deduces the onset and offset of sound events in addition to the presence confirmation. Imoto *et al.* showed that the joint training of SED and ASC tasks via a multi-task learning [16] framework can increase the performance of the SED system [13,14]. Jung *et al.* introduced a novel framework [15] that utilizes outputs of an audio tagging system by either concatenating or applying multi-head attention [17].

In this study, we extend the work of [15] that uses the output of an audio tagging system (referred as *tag vector*) to derive an attention map for the ASC task, in several aspects. First, we analyze that tag vectors may pose an out-of-distribution problem [18]. That is, diverse undefined sound events that exist in an audio recording will cause the tag vectors to convey inaccurate information. Thus, we propose the use of the output of the last hidden layer (referred as *tag representation*) instead and validate its effectiveness.

We also propose the application of a dual attention mechanism to the feature map, one attention to the frequency-time dimension and the other to the channel dimension, inspired by [19, 20]. We hypothesize that the sound events, which constitute an important information regarding the classification of particular acoustic scenes scattered in the frequency-time and channel dimensions of a feature map, can be emphasized. Additionally, we change the input feature from raw waveform to Mel-spectrogram, modifying some details. Combining several proposals, the final proposed framework of this study trains an ASC system, in which a dual attention derived from a tag representation is applied. The proposed system demonstrates an overall classification accuracy of 69.3% in the DCASE2020 task1-a fold1 configuration test set.

2. EVENT DETECTION FOR ASC

Guastavino reported that humans perceive an acoustic scene utilizing the existence of sound events [21]. Thus, audio tagging and SED tasks are closely related to the ASC task. The audio tagging task predicts the existence of predefined sound events from an input recording. An audio tagging system outputs a *tag vector* with an equal dimensionality to the number of predefined sound events; the value of each dimension is between 0 and 1, denoting the predicted posterior probability of the existence of each event. The SED task predicts the existence of sound events and determines the onset and offset of each event. Both tasks are studied with different applications and are utilized to further improve the ASC system [13, 15].

In this study, we assume that audio tagging may be more adequate to help improve the ASC system because the ASC system does not require the onset and offset of sound events. Thus, we choose an audio tagging system to improve the ASC system and extend Jung *et al.*'s work [15], which proposed the use of tag vectors by either directly concatenating it with the representation vector of the ASC task for classification or producing an attention map for the channel domain.

3. BASELINES

3.1. Vanilla ASC system

The vanilla ASC system (i.e., *baseline*) is a variant of the squeeze-excitation (SE)-ResNet [22], which performs the ASC task in an

[*] Equal contribution

[†] Corresponding author

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Planning(2020R1A2C1007081).

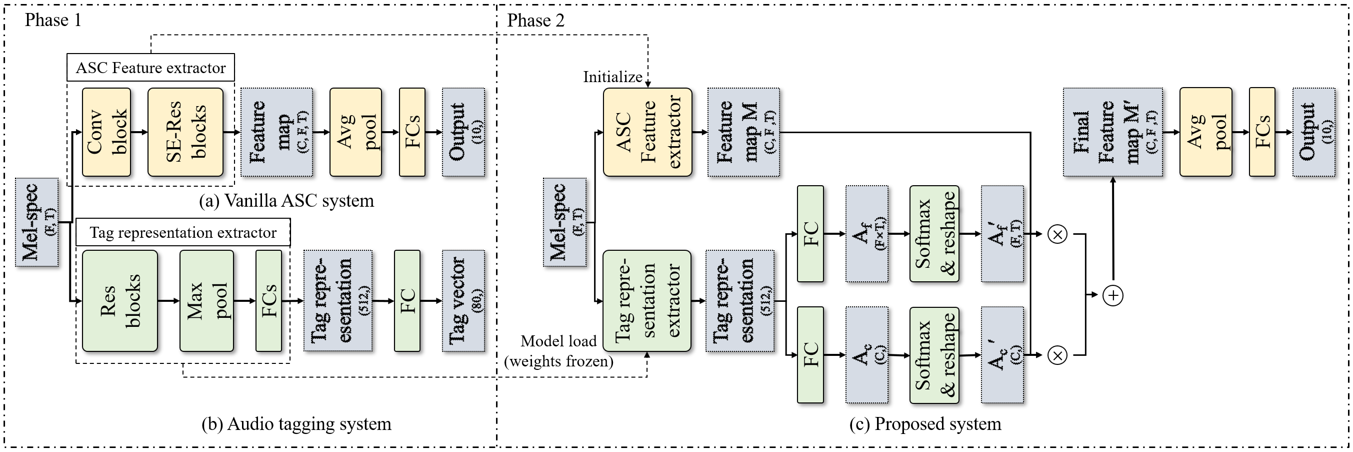

Figure 1: **Phase 1**: a vanilla ASC system (a) and an audio tagging system (b) are independently pre-trained. **Phase 2**: the proposed ASC framework using tag-representation-based dual attention (c) is initialized using (a) and (b), and then trained. The weight parameters of the tag representation extractor are frozen. The final feature map is derived by applying the dual attention to the feature map. (⊕: element-wise addition, ⊗: element-wise multiplication)

end-to-end manner. Figure 1-(a) shows the structure of the vanilla ASC system. First, when a Mel-spectrogram derived from an audio recording is provided as the input, the feature map is extracted using a convolution block and few SE-Res blocks. The convolution block comprises a convolution layer (Conv), batch normalization layer (BN) [23], and rectified linear unit (ReLU) layer. The SE-Res block consists of a sequence of layers, Conv-BN-ReLU-Conv-BN-SE, with a residual connection [24]. A global average pooling layer aggregates the feature map into a recording-level feature. The recording-level representation is fed through two fully-connected layers and then classified into one of the ten predefined scenes. Further details regarding the architecture of the vanilla ASC system are presented in our DCASE2020 challenge technical report [25].

3.2. Multi-label audio tagging system

An audio tagging system predicts the existence of predefined sound events. Using the knowledge that sound events are proactively identified when a human performs an ASC task [15, 21], this study utilizes an audio tagging system to improve the performance of the ASC system. We use the multi-label audio tagging system proposed by Akiyama *et al.* [26], the winning system of the DCASE2019 challenge task2. This system detects the existence of 80 sound events in an input recording. Figure 1-(b) depicts the structure of the audio tagging system. The system is trained in a semi-supervised manner using a set of training data consisting of a small amount of human-labeled data and a large amount of noisy labeled data. We use the Mel-spectrogram-based system among two systems proposed in the paper for consistency of input feature throughout this study.

4. PROPOSED FRAMEWORK

4.1. Tag representation

A tag vector, which represents predicted probabilities regarding the existence of predefined sound events, was used in a previous study [15]. Eighty predefined sound events are used in an audio tagging system. However, the actual number of sound events that can occur in an input recording for an ASC system outnumbers the predefined sound events. In an extreme scenario, a given audio recording may not include any predefined sound events. In such a case, the tag vector may mislead information, which corrupts the system, similar to an out-of-distribution problem [18].

To alleviate this issue, we use a tag representation instead of a tag vector, based on the assumption that a tag representation will involve rather abstract sound event information. It is inspired by [27], which uses the last hidden layer's output as the representation vector for a different task after training the DNN for a similar task. Table 1 shows the number of misclassified scene pairs among the baseline and the ASC systems utilizing the tag vector and tag representation. When using tag vectors, misclassification increases in two pairs ("shopping_mall-airport" and "public_square-street_pedestrian") among the top-7 most misclassified pairs of scenes. Contrarily, when using tag representation, misclassification decreases in all seven pairs. Furthermore, except "metro-tram", the tag representation consistently demonstrates a lower number of misclassified recordings.

4.2. Dual attention

Attention mechanism [17] was initially proposed for the machine translation task, and various attention mechanisms have been adopted across different tasks. The authors of [19] and [20] proposed the application of an attention mechanism to the positional and channel dimensions of the feature map simultaneously. They reported that long-range contextual information can be explored and the two-dimensional information can be refined. By leveraging this knowledge, we take advantage of the dual attention technique to improve ASC performance. We apply a dual attention method to the frequency-time and channel dimensions of the frame-level features before it is aggregated into a recording-level feature by an average pooling layer. Consequently, it is expected that scene information scattered in the two-dimension of the feature map will be emphasized. The attention map is generated directly using a single fully-connected layer, and not by a dot product between vectors derived from the convolution operation, as discussed in [19].

4.3. Tag representation guided dual attention network

Combining tag representation and dual attention mechanism, described in previous subsections, the overall proposed system is

Table 1: Comparison results of three systems for the top-7 confusing scene pairs. "Baseline" refers to the vanilla ASC system used in this paper; "Tag-vec" and "Tag-rep" refer to the systems using the tag vector and tag representation for ASC, respectively.

Scene pair	Baseline	Tag-vec	Tag-rep
Metro - Tram	114	**97**	105
Shop_mall - Airport	107	116	**104**
Shop_mall - Metro_st	84	56	**45**
Shop_mall - Street_ped	83	**69**	**69**
Public_sq - Street_ped	74	77	**71**
Public_sq - Park	74	65	**58**
Airport - Street_ped	66	63	**46**

Table 2: Comparison of the official baseline systems of the DCASE challenge and the two baselines used in this study.

System	Acc (%)
DCASE2019 task1-a baseline [3]	46.5
DCASE2020 task1-a baseline [4]	54.1
Ours-vanilla ASC	65.3
Ours-tag vector ASC	66.7

Table 3: Ablation and comparison experiments regarding the effects of frequency-time and channel attention and the methods used to derive attention maps using a single head ("self": self-attention, "tag": attention map derived using tag representation, "-": not applied).

System	Frequency-time	Channel	Acc (%)
#1	self	-	66.5
#2	-	self	65.9
#3	tag	-	65.8
#4	-	tag	67.6
#5	self	self	65.7
#6	self	tag	67.1
#7	tag	self	66.3
#8	tag	tag	**67.9**

shown in Figure 1-(c). Before training the proposed system, a vanilla ASC system and an audio tagging system are pre-trained to initialize the feature extractor and the tag representation extractor, respectively. Note that, for the tagging system, the output layer is removed and the weights are frozen. When an audio recording is entered as the input, an ASC feature map and tag representation are extracted in parallel.

The extracted tag representation is used to derive attention maps that are used to apply a dual attention mechanism. The assumption is that dispersed sound event information related to the characteristics of the scenes can be emphasized. Let M be a feature map, $M \in \mathbb{R}^{C\times F\times T}$ where C, F, and T refer to the number of feature map channels, frequency bins, and length of the sequence in the time dimension, respectively. Given a tag representation $T_{rep} \in \mathbb{R}^{512}$, we first feed it into two fully-connected layers to generate attention maps, A_f and A_c.

$$\begin{aligned} &A_f = T_{rep} \cdot W_{A_f},\ A_f \in \mathbb{R}^N, \\ &A_c = T_{rep} \cdot W_{A_c},\ A_c \in \mathbb{R}^C \end{aligned} \tag{1}$$

where $\cdot$ refers to the matrix multiplication; and $N = F \times T$. W_{A_f} and W_{A_c} are weight matrices of the fully-connected layers, $W_{A_f} \in \mathbb{R}^{512\times N}$, and $W_{A_f} \in \mathbb{R}^{512\times C}$. Subsequently, A_f and A_c are reshaped to $A_f \in \mathbb{R}^{h\times(N/h)}$ and $A_c \in \mathbb{R}^{k\times(C/k)}$, respectively, where h and k denote the number of heads in each case. The frequency-time attention map, A'_f, and channel attention map, A'_c, are denoted as:

$$\begin{aligned} &A'_f = [A'_{f1}, A'_{f2}, \cdots A'_{fh}],\ A'_{fh} \in \mathbb{R}^{N/h}, \\ &A'_{fh} = [A'_{f1h}, A'_{f2h}, \cdots, A'_{fih}], A'_{fih} \in \mathbb{R}^1, \\ &A'_c = [A'_{c1}, A'_{c2}, \cdots A'_{ck}],\ A'_{ck} \in \mathbb{R}^{C/k}, \\ &A'_{ck} = [A'_{c1k}, A'_{c2k}, \cdots, A'_{cik}], A'_{cik} \in \mathbb{R}^1, \end{aligned} \tag{2}$$

where A'_{fh} and A'_{ck} correspond to the softmax-applied attention map for a single head and i refers to the index for each dimension. Softmax is applied to each element of A'_{fh} and A'_{ck}, denoted as:

$$A'_{f_{ih}} = \frac{exp(A_{f_{ih}})}{\sum_{j=1}^{N/h} exp(A_{f_{jh}})}, \quad A'_{c_{ik}} = \frac{exp(A_{c_{ik}})}{\sum_{j=1}^{C/k} exp(A_{c_{jk}})} \tag{3}$$

where j refers to the index for counting elements in each head. Subsequently, A'_f and A'_c are reshaped back to $A'_f \in \mathbb{R}^{F\times T}$ and $A'_c \in \mathbb{R}^C$, respectively.

The frequency-time attention map and the channel attention map perform element-wise multiplication with M where unused dimensions are broadcasted. Individually calculated vectors are then added to comprise a final feature map $M' \in \mathbb{R}^{C\times W\times H}$, formally denoted as:

$$M' = M \otimes A'_f \oplus M \otimes A'_c, \tag{4}$$

where $\oplus$ and $\otimes$ denote element-wise addition and multiplication, respectively. The final feature map performs the ASC task through an average pooling layer and two fully-connected layers.

5. EXPERIMENTS AND RESULTS

5.1. Dataset

We use the DCASE2020 task1-a dataset for all experiments. The dataset contains single-channel audio recordings from 12 cities in 10 different acoustic scenes using 4 different devices and 11 augmented devices. Each recording has a duration of 10 s with 44.1kHz sampling rate and 24-bit resolution. The development set contains data from 10 cities and 9 devices. We use the official train/test split of the DCASE 2020 challenge which assigns 13,965 train recordings ($\approx$ 39 hours) and 2,970 test recordings. The evaluation dataset, used to submit our systems for the DCASE2020 Challenge, includes all 12 cities and 15 devices. Except in Table 6, all performances are reported using the official fold1 test set.

5.2. Experimental configurations

All ASC systems described in this paper are conducted under the same configurations. Mel-spectrograms are extracted using 128 Mel-filterbanks. The number of fast Fourier transform bins is 2,048, and the window length and shift size are 40 ms and 20 ms, respectively. The batch size and the number of epochs are set to 24 and 800, respectively. The optimizer is SGD; the learning rate is set to 0.001; we use the cosine learning rate scheduler. We only use categorical cross-entropy as the objective function. Mix-up [28] is applied by the data argumentation technique, and alpha is set to 0.1. Audio tagging system is identical to [26]. Detailed hyper-parameter

Table 5: Comparison of device and class-wise classification accuracies of the baseline and the proposed system on fold1 test set (baseline/proposed, %). ***Bold*** describes higher accuracy in each device or class.

Device	A	B	C	S1	S2	S3	S4	S5	S6	
Acc	70.9/**74.6**	65.6/**69.6**	69.6/**70.2**	64.4/**70.5**	61.7/**65.6**	65.2/**70.6**	64.8/**68.4**	66.2/**69.1**	59.6/**65.4**	
Class	Airport	Bus	Metro	Metro_st	Park	Public_sq	Shop_mall	Street_ped	Street_traf	Tram
Acc	61.4/**62.6**	**78.8**/74.7	68.2/**73.3**	58.0/**68.9**	74.6/**75.2**	63.3/**71.9**	53.3/**54.9**	49.3/**57.6**	82.0/**85.9**	64.3/**68.3**

Table 4: Comparison experiments on the number of heads when applying dual attention.

		# Frequency-time heads	
		3	9
# Channel heads	2	**69.3**	67.8
	4	69.2	68.2
	8	68.7	69.0
	16	68.7	68.1
	32	68.2	68.2

settings and configuration are further addressed in our technical paper [25].

5.3. Result analysis

Table 2 compares different baselines with the baselines of this study. The top two rows describe the performance of the official DCASE baselines which were trained using DCASE2020 task 1-a dataset. The third row refers to our vanilla ASC baseline. The bottom row shows the performance of our implementation of [15], which uses a tag vector for the ASC task. Results demonstrate that both our implemented systems outperform the community's baseline by over 10%. The tag-vector-based ASC system has a higher classification accuracy than that of the vanilla ASC system, which is consistent with the results in [15].

Table 3 addresses the effect of dual attention and the mechanism to derive an attention map with single head. A comparison of systems shows that using both attentions on the frequency-time and channel domains is more effective than applying an attention in one domain. For deriving attention maps, using tag representations persistently demonstrates a higher performance than self-derived methods, with an exception when applying only frequency-time domain attention. Herein, system #4 has the identical configuration as the last row of Table 2, but uses tag representation instead of tag vector. By comparing the results of the two systems, we confirm that using tag representation is more effective than using tag vectors for an ASC task. Applying both attentions using tag representation demonstrates the highest performance, with a classification accuracy of 67.9%.

Table 4 describes the result of a comparison study that shows the effect of the number of heads when deriving attention maps. The number of frequency-time attention head designates a range of frequencies to which attention is applied. In the case of three heads of frequency-time attention, attention is applied by dividing the feature map into three parts of the frequency bin. We choose the best performing system from Table 3 with an accuracy of 67.9% and change the number of heads. Overall, using fewer heads leads to a higher accuracy, with a few exceptions. The best result can be obtained using three heads for frequency-time attention and two heads for channel attention.

Table 5 addresses the classification accuracies of the vanilla ASC system and the best-performing proposed system for each scene and device. Across all nine devices including six augmented devices, the proposed system demonstrates higher classification accuracies. In terms of each acoustic scene, the proposed system outperformed the baseline in all scenes but Bus, in which the accuracy decreased from 78.8% to 74.7%.

Table 6: Results of our submitted systems for the DCASE2020 challenge task1-a.

System	# Param	Acc (%)
DCASE2020 baseline [4]	5M	51.4
Ours-tag_rep	0.6M	71.0
Ours-tag_rep+LCNN	1.6M	71.7

Finally, Table 6 shows our system's submission results for the DCASE2020 challenge. The performance of the submitted systems is the result of the score-sum ensemble in which systems were trained by constructing 4-fold cross validation. Support vector machine classifiers using radial basis function and sigmoid kernel are used for a score-level ensemble. Compared to the DCASE2020 baseline, the proposed system reported a 22.8% relative improvement in accuracy with one-eighth model complexity. The final row shows the result of the score-sum ensemble with another ASC system using an architecture referred to as LCNN [29], and the performance increased to 71.7%.

6. CONCLUSION

In this paper, we focused on the role of sound events included in an audio recording to improve the performance of an ASC system. A framework that uses a pre-trained audio tagging system was extended. We analyzed that tag representation yields more accurate attention maps compared to a conventional method that uses a tag vector. Leveraging the knowledge that a dual attention method can emphasize crucial information scattered in feature maps, we proposed a method of tag representation for guided dual attention. The proposed system demonstrated the superiority of the performance through several comparative experiments. Compared to baseline accuracy of 65.3%, the final proposed system shows an improved accuracy of ASC 69.3%. As our future work, we plan to study the relationship with other acoustic signal processing tasks such as SED using a multi-task learning method.

7. REFERENCES

[1] T. Virtanen, A. Mesaros, T. Heittola, A. Diment, E. Vincent, E. Benetos, and B. M. Elizalde, *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017).* Tampere University of Technology. Laboratory of Signal Processing, 2017.

[2] M. D. Plumbley, C. Kroos, J. P. Bello, G. Richard, D. P.

Ellis, and A. Mesaros, *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*. Tampere University of Technology. Laboratory of Signal Processing, 2018.

[3] M. Mandel, J. Salamon, and D. P. W. Ellis, *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*. NY, USA: New York University, October 2019.

[4] T. Heittola, A. Mesaros, and T. Virtanen, "Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, 2020, submitted. [Online]. Available: https://arxiv.org/abs/2005.14623

[5] K. Koutini, H. Eghbal-Zadeh, M. Dorfer, and G. Widmer, "The receptive field as a regularizer in deep convolutional neural networks for acoustic scene classification," in *2019 27th European signal processing conference (EUSIPCO)*. IEEE, 2019, pp. 1–5.

[6] J.-W. Jung, H.-S. Heo, I. Yang, S.-H. Yoon, H.-J. Shim, and H.-J. Yu, "Dnn-based audio scene classification for dcase 2017: dual input features, balancing cost, and stochastic data duplication," *DCASE2017 Workshop*, vol. 4, no. 5, 2017.

[7] T. Nguyen, F. Pernkopf, and M. Kosmider, "Acoustic scene classification for mismatched recording devices using heated-up softmax and spectrum correction," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 126–130.

[8] J. Huang, H. Lu, P. Lopez Meyer, H. Cordourier, and J. Del Hoyo Ontiveros, "Acoustic scene classification using deep learning-based ensemble averaging," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, 2019, pp. 94–98.

[9] S. Mun, S. Park, D. K. Han, and H. Ko, "Generative adversarial network based acoustic scene training set augmentation and selection using SVM hyper-plane," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, pp. 93–102.

[10] A. Vafeiadis, D. Kalatzis, K. Votis, D. Giakoumis, D. Tzovaras, L. Chen, and R. Hamzaoui, "Acoustic scene classification: From a hybrid classifier to deep learning," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, 2017, pp. 123–127.

[11] H.-S. Heo, J.-w. Jung, H.-j. Shim, and H.-J. Yu, "Acoustic scene classification using teacher-student learning with soft-labels," *Proc. Interspeech 2019*, pp. 614–618, 2019.

[12] J.-w. Jung, H. Heo, H.-j. Shim, and H.-J. Yu, "Distilling the knowledge of specialist deep neural networks in acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York University, NY, USA, October 2019, pp. 114–118.

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

[14] N. Tonami, K. Imoto, M. Niitsuma, R. Yamanishi, and Y. Yamashita, "Joint analysis of acoustic events and scenes based on multitask learning," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 338–342.

[15] J.-w. Jung, H.-j. Shim, J.-h. Kim, S.-b. Kim, and H.-J. Yu, "Acoustic scene classification using audio tagging (to appear)," *Proc. Interspeech 2020*, 2020.

[16] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[18] K. Lee, H. Lee, K. Lee, and J. Shin, "Training confidence-calibrated classifiers for detecting out-of-distribution samples," *arXiv preprint arXiv:1711.09325*, 2017.

[19] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[20] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[21] C. Guastavino, "Categorization of environmental sounds." *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, vol. 61, no. 1, p. 54, 2007.

[22] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[23] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[25] H.-j. Shim, J.-h. Kim, J.-w. Jung, and H.-j. Yu, "Audio tagging and deep architectures for acoustic scene classification: Uos submission for the DCASE 2020 challenge," DCASE2020 Challenge, Tech. Rep., 2020.

[26] O. Akiyama and J. Sato, "Dcase 2019 task 2: Multitask learning, semi-supervised learning and model ensemble with noisy data for audio tagging," *DCASE2019 Workshop*, 2019.

[27] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4052–4056.

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

[29] H.-j. Shim, J.-w. Jung, J.-h. Kim, and H.-j. Yu, "Capturing scattered discriminative information using a deep architecture in acoustic scene classification," *arXiv preprint arXiv:2007.04631*, 2020.