ATTENTIVE MAX FEATURE MAP AND JOINT TRAINING FOR ACOUSTIC SCENE CLASSIFICATION

Hye-jin Shim[1], *Jee-weon Jung*[2], *Ju-ho Kim*[1], *and Ha-Jin Yu*[1*]

[1]School of Computer Science, University of Seoul, [2]Naver Corporation

ABSTRACT

Various attention mechanisms are being widely applied to acoustic scene classification. However, we empirically found that the attention mechanism can excessively discard potentially valuable information, despite improving performance. We propose the attentive max feature map that combines two effective techniques, attention and a max feature map, to further elaborate the attention mechanism and mitigate the above-mentioned phenomenon. We also explore various joint training methods, including multi-task learning, that allocate additional abstract labels for each audio recording. Our proposed system demonstrates competitive performance with much larger state-of-the-art systems for single systems on Subtask A of the DCASE 2020 challenge by applying the two proposed techniques using relatively fewer parameters. Furthermore, adopting the proposed attentive max feature map, our team placed fourth in the recent DCASE 2021 challenge.

Index Terms— acoustic scene classification, attention, max feature map, joint training

1. INTRODUCTION

Acoustic scene classification (ASC) is the task of recognizing scenes based on environmental sounds. The detection and classification of acoustic scenes and events (DCASE) community hosted several challenges [1–3]. The DCASE 2020 challenge included two subtasks addressing different properties for ASC: 1) Subtask A requires the generalization to unknown devices, and 2) Subtask B demands a low-complexity solution in terms of model size (i.e., the number of parameters). Subtask A aims to identify a given audio clip recorded using multiple devices into one of the ten predefined acoustic scenes including the pedestrian street, airport, and tram. Subtask B aims to classify a given audio clip into one of the three relatively abstract level classes: outdoor, indoor, and transportation. The classification task for subtask A and subtask B are referred to as 10-class and 3-class classifications.

Recent studies in ASC can be primarily divided into two strands, data preprocessing and modeling, and most top-ranking systems focused on data preprocessing in the DCASE 2020 challenge. In data preprocessing, most recent systems have exploited delta and delta-deltas [4–8], diverse data augmentation techniques (e.g., mixup [9], and SpecAugment [10]), sub-frequency analysis [4, 11], feature investigation, and temporal division [12]. In modeling, ResNet [13] and convolutional neural networks (CNNs) are the most widely used architectures for ASC tasks. Techniques such as the adoption of the knowledge distillation framework [14, 15] and receptive field regularization [8] have been demonstrated as effective. Utilizing the representation of other tasks and training a general-purpose network have also been investigated [16–19]. Among various modeling approaches, the attention mechanism is one of the most effective techniques [20, 21].

In this paper, we propose two modeling techniques independent of data preprocessing: an attentive max feature map (AMFM) and joint training of the concrete classes of Subtask A and abstract classes of Subtask B. First, by visualizing the effect of attention on the feature map (see Figure 2), applying attention excessively discards potentially important information, despite improving the performance. Analyzing the visualization, we assume that preserving more information might further boost the discriminative power of a feature if adequately addressed. This assumption is in line with recent studies that have adopted sigmoid-based attention to alleviate discarding too much information [22]. We argue that the max feature map (MFM), a technique that adopts a competitive scheme via elementwise max operation [23], can be used for this purpose because MFM showed its effectiveness in [16, 24, 25]. Specifically, we design a new technique combining attention and MFM, which we refer to as the AMFM, in which attention emphasizes the most informative region and the MFM prevents excessive information loss using a max operation. The proposed AMFM compares the feature maps before and after the attention mechanism to mitigate excessive information loss.

Second, we propose a joint training scheme using the 10-class label from Subtask A and 3-class label from Subtask B to improve the performance of Subtask A. We explore four methods for training the two tasks. The underlying assumption is that additional labels can improve the supervision in the training process. The proposed approach is similar to the work by Hu et al. [5, 26], which employed the prediction of the classifiers of Subtask B to improve performance on Subtask A. The two classifiers were trained separately by Hu et al., whereas the proposed approach extends the framework by training a single model with the joint training scheme. To the best of our knowledge, this approach is the first to simultaneously train a single ASC system using labels for both Subtasks A and B. The performance of the proposed system is comparable to that of the state-of-the-art approaches without complex data preprocessing techniques. Besides, our team could take fourth place in the DCASE 2021 challenge adopting the proposed AMFM.

2. PROPOSED METHODS

2.1. Attentive max feature map

The MFM replaces the non-linear activation function, typically the rectified linear unit (ReLU), with a competitive scheme [23]. It was originally proposed for situations where noisy labels are dominant.

*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)

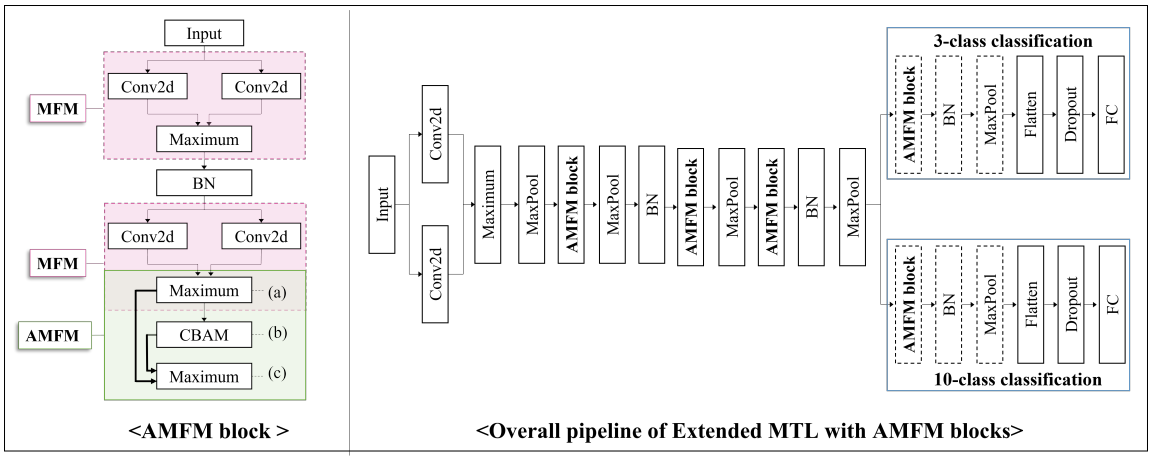


Fig. 1. Proposed attentive max feature map (AMFM) block structure (left) and overall pipeline of the extended multi-task learning (MTL) structure (right). The AMFM block on the left comprises both the max feature map (MFM) and AMFM, illustrated in pink and green boxes, respectively. The illustrated overall pipeline adopts the AMFM block and demonstrates the best results among the explored joint training methods.

Wu et al. [23] argued that using threshold-based non-linearity, such as ReLU, might not effectively generalize on unknown data distributions because an activation function separates noisy and informative signals using a threshold (or bias) learned from training data. Using threshold-based non-linearity might cause information loss, especially for the first few convolutional layers [23]. To overcome this issue, MFM adopts an elementwise max operation in place of ReLU non-linearity. The max operation selects relatively important features learned by different filters. Existing studies [16,24,25] have demonstrated that MFM is effective in ASC.

The implementation of the MFM operation can be described as follows. Let a be an output feature map of a convolutional layer, $a \in \mathbb{R}^{K \times T \times F}$, where K, T, and F refer to the number of output channels, time-domain frames, and frequency bins, respectively. In addition, a is first split into two equal-sized feature maps, a_1 and a_2, where a_1, $a_2 \in \mathbb{R}^{\frac{K}{2} \times T \times F}$. The MFM applied feature map is obtained by applying $\max(a_1, a_2)$ elementwise. On the left in Figure 1, the pink box describes the conventional MFM operation.

The recent literature on the ASC task has demonstrated the effectiveness of the attention mechanism [7, 20, 21]. The attention mechanism highlights important information and enriches the representation. Among various attention mechanisms, the convolutional block attention module (CBAM) [22] considers both channel and spatial attention, and has the advantage of seamless implementation regardless of the architecture. Our previous work confirmed that MFM and CBAM can be employed simultaneously [24]. However, through an analysis using the visualization of intermediate feature maps, we empirically found that information is excessively discarded, emphasizing only a small fraction (see Figure 2 (b)).

We hypothesize that, although the attention mechanism is an effective technique, preserving relatively more information might further improve the discriminative power of a feature. Thus, we argue that combining the two existing techniques (the attention mechanism and MFM) might leverage the effectiveness of the attention mechanism while restricting excessive information deletion. The proposed technique based on this inspiration is the AMFM, which competitively applies the attention mechanism. It compares two feature maps before and after the attention mechanism and outputs their elementwise maximum values as Figure 2 shows.

The proposed AMFM performs $\max(b, CBAM(b))$ elementwise. Here, b is $\max(a_1, a_2)$. We illustrate an AMFM block that involves both MFM (pink box) and AMFM (green box) on the left in Figure 1. Conv2d, Maximum, and BN refer to the 2D convolutional layer, maximum operation, and batch normalization, respectively. In the AMFM block, (a), (b), and (c) in Figure 1 are consistent with those in Figure 2. Figure 2 indicates that AMFM alleviates exaggerated attention and selects salient representation as intended. Moreover, AMFM can be applied to architectures where various attention mechanisms are used, although we combined MFM with the CBAM attention mechanism.

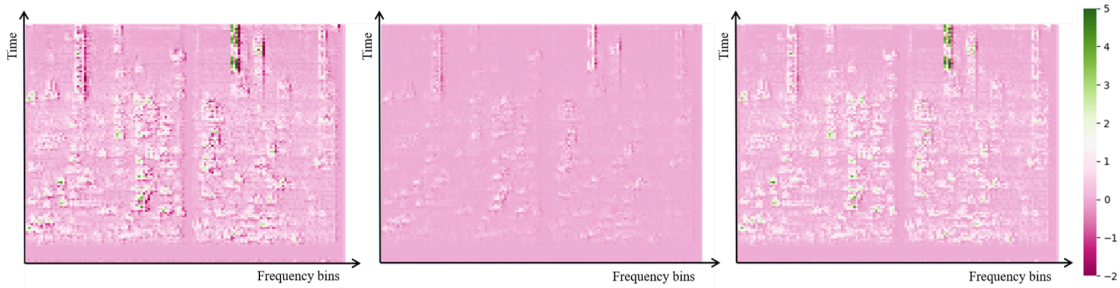
2.2. Joint training

When the tasks are highly related, joint training of related tasks provides better supervision as reported by [27, 28]. Thus, in this particular case of the ASC task, we assumed that adopting multi-task learning (MTL) using the two labels (3-class and 10-class) would be helpful. These two labels of the two subtasks only differ in the degree of abstraction, where a hierarchy exists (each class in the 3-class definition is further divided into three or four classes in the 10-class definition). Therefore, we propose to jointly train the model using both labels.

The term "joint training" in this paper includes pre-training, original MTL, and variants of MTL. The study by Hu et al. [5] is one of the most similar studies on ASC that considers the relationship between the two subtasks. The difference between our work and [5] is that [5] employed joint prediction where the final prediction is performed using the score fusion of the two separately trained classifiers. In contrast, our work trains an integrated model.

We explore four different methods for joint training of two subtasks where the first two methods directly apply the existing methods without modification, and the other two methods alter the MTL framework. First, we adopt the pre-training method. The system is first trained for the 3-class classification and then fine-tuned with the 10-class classification. Second, we apply the original MTL [29] for the two subtasks. In this case, the network is designed to learn the shared representation, and the last hidden layer is directly connected with the output layers of each task.

Third, we exploit the extended MTL architecture, which has additional layers allocated for each training task, after the last hidden



(a) Original feature map (b) Attention applied feature map (c) Output feature map of AMFM

Fig. 2. Each feature map indicates (a) before and (b) after the attention mechanism, and (c) the output of the attentive max feature map (AMFM). These were extracted from the first AMFM block at positions (a)-(c) on the left in Figure 1. This demonstrates that the attention mechanism can cause excessive information loss. After applying the AMFM (c) with a comparison of the two feature maps of (a) and (b), the features were found to enrich the representation preventing information loss (Best viewed in color.).

Table 1. Performance comparison according to the application of the attention mechanism and the structure of the convolutional neural network (CNN), max feature map (MFM), and attentive max feature map (AMFM).

System	Attention	Acc (%)
CNN w/ ReLU	×	70.2
	✓	68.3
CNN w/ LeakyReLU	×	69.6
	✓	68.2
MFM	×	69.4
	✓	70.3 ± 0.13
AMFM	✓	**70.7 ± 0.08**

layer shared by the two tasks. Finally, we investigate the sequential order of training for the two tasks, considering the hierarchical relationship between the two tasks. Unlike the original MTL, the two classifiers are not connected to the same layer. Instead, the proposed sequential MTL has a hierarchical design, where the output layer for the 3-class classification is placed after an intermediate hidden layer. This architecture design is in line with that proposed by [30]. Comparing the order of the 3-class and 10-class classifications, training the 3-class classification first, followed by the 10-class classification is more effective. This finding is similar to the majority of deep learning structures that deal with abstract representations in the hidden layers close to the input and specific/sophisticated representations in the hidden layer close to the output layer [31].

For further improvement, we additionally explore the joint prediction proposed in [5] for the extended and sequential MTL framework. Joint prediction exploits the score fusion of 3-class and 10-class classifications, where the 3-class classification result is used as prior knowledge. In terms of adjusting the weight ratio between the two tasks, we explored both intuitive and methodological approaches: a grid search and GradNorm [32], respectively.

3. EXPERIMENTAL SETTINGS AND RESULTS

3.1. Experimental settings

All experiments in this paper used the DCASE 2020 Task 1 Subtask A (1-A) dataset with the corresponding 3-class labels from Subtask B. We used 13,965 audio clips and 2,970 audio clips for training and evaluating the model, respectively, following the official protocol. The DCASE 2020 Task 1-A dataset consists of various audio clips collected from three real devices (A, B, and C) and six simulated devices (s1 to s6). Only devices A-C and s1 to s3 were used in the training set, and s4 to s6 were unavailable in the training phase. Each audio clip has a duration of 10s with a 44.1kHz sampling rate and 24-bit resolution. We used 256-dimensional Mel-spectrograms as the base feature. A short-time Fourier transform with 2,048 FFT points was applied with a 40 ms window size and 20 ms hop length. Mixup [9] and SpecAugment [10] were exploited for data augmentation. The initial learning rate was set to 0.001 and scheduled with a warm restart of the stochastic gradient descent (SGD). The SGD optimizer with a momentum of 0.9 was used. The batch size and number of epochs were set to 24 and 800, respectively. The architecture details are similar to those described in [24]. In the case of additional blocks for the MTL, the parameters of the AMFM block are identical to those for the other blocks. The last hidden layer for each task has 100 nodes followed by the output layer for each subtask. All models were implemented in PyTorch, a deep learning library in Python. Other data preprocessing techniques, such as the application of logarithms, deltas, delta-deltas, and sub-band frequency separation, were not used in this work, which leaves room for further improvements. The code is publicly available at `https://github.com/shimhz/AMFM`.

3.2. Result analysis

3.2.1. Attentive max feature map

Table 1 delivers the results of comparing the effectiveness of applying the CBAM in various deep neural network structures: CNN, MFM, and AMFM. In the experiments, CBAM decreased performance when applied to a CNN with the ReLU or leaky ReLU activations but improved performance when applied using MFM. Using the proposed AMFM, the best performance of 70.8% accuracy was achieved. By comparatively analyzing (a), (b), and (c) in Figure 2 jointly with their corresponding accuracies (70.2%, 68.3%, and 70.8%, respectively), we argue that AMFM emphasizes discriminative information while avoiding excessive information removal.

3.2.2. Joint training

Table 2 presents the comparison of various joint training strategies. The joint training experiments were conducted using the AMFM structure. Both pre-training and conventional MTL did not further improve the performance. In the experiments, unlike [5], joint pre-

Table 2. Comparison of various joint training strategies.

System	Joint prediction	# Params	Acc (%)
w/o joing training	×	1.5M	70.8
Pre-traing	×	1.5M	69.2
Conventional MTL	×	1.5M	69.7
Extended MTL	×	0.6M	**71.3**
	✓	0.6M	70.0
Sequential MTL	×	0.7M	71.0
	✓	0.7M	69.1
Separated Classifier [5]	✓	1.5M	69.4

Table 3. Experimental results adjusting the weight ratio of 3-class to 10-class classifications in joint training.

System	Ratio	Acc (%)
Proposed method	1 : 1	70.3
	1 : 2	69.6
	1 : 3	70.3
	1 : 4	70.7
	1 : 5	**71.3**
GradNorm [32]	-	70.1

diction resulted in worse performance for both extended and sequential MTL frameworks.

The best result was achieved using the extended MTL framework, which had additional layers for each training task after the last hidden layer shared by the two tasks. The sequential MTL also demonstrated a slight performance improvement. Through these results (original, extended, and sequential MTL), we conclude that the original MTL could not learn each task sufficiently. Instead, hidden layers solely assigned to solve a specific task are required. Based on the results, although the two tasks are similar, each output layer demands at least a few layers solely dedicated to each task. This interpretation is inspired by [31], where the authors reported performance degradation caused by excessive interference between two related tasks.

3.2.3. Additional experiments

Table 3 lists the ablation results of adjusting the weight ratio between the two subtasks for joint training. We explored both manual and methodological approaches to determine the optimal weight ratio. The best result was achieved when the ratio of the abstract to specific labels was 1:5.

Table 4 describes the accuracy of 3-class classification. The MTL is effective when related tasks are jointly trained, although it is difficult to determine whether those tasks are explicitly related. Therefore, we investigated 3-class classification performance to demonstrate that the two tasks (10-class and 3-class) are beneficial for each other.

The accuracy of the DCASE 2020 baseline of Subtask B was 88%, and the average accuracy of the submitted system was 87.3% [3]. When we trained the model that only performs the 3-class classification task using the same proposed AMFM structure without MTL, the accuracy was 91.4%. Compared with the aforementioned results, we could achieve a 92.2% accuracy with the extended MTL scheme while demonstrating 71.3% accuracy for the main task.

Table 4. 3-class classification accuracy. The proposed joint training also increases 3-class classification performance.

System	DCASE Baseline	Submitted systems' average	Ours, w/o MTL	Ours, w/ MTL
Acc (%)	88.0	87.3	91.4	**92.2**

Table 5. Comparison with recent state-of-the-art systems using the performance of individual systems without a score-level ensemble. The third to seventh rows list the top five best-performing systems on the DCASE2020 Task 1-A challenge.

System	Acc (%)	# Params
Proposed Method	**71.3**	**0.6M**
DCASE2020 Baseline [3]	54.1	5M
Suh et al. [4]	73.7	13M
Hu et al. [5]	76.9	-
Gao et al. [6]	71.8	4M
Liu et al. [7]	72.1	3M
Koutini et al. [8]	71.8	225M

3.2.4. Comparison with recent state-of-the-art systems

Table 5 presents a comparison with current state-of-the-art systems without applying ensemble techniques. The proposed system demonstrates comparable performance with state-of-the-art systems without complex data preprocessing techniques. In addition, the proposed method was performed with low-complexity architecture. Although it is outside the study scope, we expect that applying more data preprocessing methods might lead to further improvement.

3.2.5. Application to the DCASE 2021 challenge

To further demonstrate the effectiveness of the proposed AMFM, we introduce the recent result of the DCASE 2021 challenge (further details are in [33]). The DCASE 2021 ASC task requires lightweight systems with less than 128KB (≈65k parameters in 16-bit resolution). Our team ranked fourth place, with two submitted systems in which one system was ResNet-based and the other was AMFM-based. The two systems showed an accuracy of 70.1% and 70.0% respectively.

4. CONCLUSIONS

In this paper, we proposed a novel technique named AMFM that combines the merits of both the attention and MFM techniques, enabling the emphasis of discriminative information while avoiding excessive information deletion. We also explored four joint training methods, including extended and sequential MTL frameworks. Both proposed methods showed their effectiveness for the ASC task. Furthermore, the weight ratio and relevance of the two tasks in the MTL framework were experimentally investigated. The proposed model requires relatively few parameters compared to other state-of-the-art systems while performing competitively. Various data preprocessing techniques adopted in the recent literature leave room for further improvement of the proposed system because both proposals in this paper focus on building a better model architecture.

5. REFERENCES

[1] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *DCASE 2018 Workshop*.

[2] A. Mesaros, T. Heittola, and T. Virtanen, "Acoustic scene classification in dcase 2019 challenge: Closed and open set classification and data mismatch setups," in *DCASE 2019 Workshop*.

[3] T. Heittola, A. Mesaros, and T. Virtanen, "Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions," *DCASE 2020 Workshop*.

[4] S.-W. Suh, S.-Y. Park, Y.-H. Jeong, and T.-J. Lee, "Designing acoustic scene classification models with cnn variants," Tech. Rep., DCASE2020 Challenge.

[5] H. Hu, C.-H. H. Yang, X. Xia, X. Bai, X. Tang, Y. Wang, S. Niu, L. Chai, J. Li, H. Zhu, et al., "Device-robust acoustic scene classification based on two-stage categorization and data augmentation," Tech. Rep., DCASE2020 Challenge.

[6] W. Gao, M. McDonnell, and S. UniSA, "Acoustic scene classification using deep residual networks with focal loss and mild domain adaptation," Tech. Rep., DCASE2020 Challenge.

[7] J. Liu, "Acoustic scene classification with residual networks and attention mechanism," Tech. Rep., DCASE2020 Challenge.

[8] K. Koutini, F. Henkel, H. Eghbal-zadeh, and G. Widmer, "Cp-jku submissions to dcase'20: Low-complexity cross-device acoustic scene classification with rf-regularized cnns," Tech. Rep., DCASE2020 Challenge.

[9] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. ICLR*, 2018.

[10] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *in Proc. Interspeech*, pp. 2613–2617, 2019.

[11] S. S. R. Phaye, E. Benetos, and Y. Wang, "Subspectralnet–using sub-spectrogram based convolutional neural networks for acoustic scene classification," in *Proc. ICASSP*. IEEE, 2019, pp. 825–829.

[12] S.-K. Mun, S.-W. Shon, W.-I. Kim, D. K. Han, and H.-S. Ko, "Deep neural network based learning and transferring mid-level audio features for acoustic scene classification," in *Proc. ICASSP*. IEEE, 2017, pp. 796–800.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

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

[15] J.-W. Jung, H.-S. Heo, H.-J. Shim, and H.-J. Yu, "Knowledge distillation in acoustic scene classification," *IEEE Access*, vol. 8, pp. 166870–166879, 2020.

[16] 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," Tech. Rep., DCASE2020 Challenge.

[17] J.-W. Jung, H.-J. Shim, J.-H. Kim, S.-B. Kim, and H.-J. Yu, "Acoustic scene classification using audio tagging," *in Proc. Interspeech*, 2020.

[18] Q. Kong, T. Iqbal, Y. Xu, W. Wang, and M. D. Plumbley, "Dcase 2018 challenge surrey cross-task convolutional neural network baseline," *DCASE 2018 Workshop*.

[19] J.-W. Jung, H.-J. Shim, J.-H. Kim, and H.-J. Yu, "Dcasenet: An integrated pretrained deep neural network for detecting and classifying acoustic scenes and events," in *Proc. ICASSP*. IEEE, 2021, pp. 621–625.

[20] Z. Ren, Q. Kong, J. Han, M. D. Plumbley, and B. W. Schuller, "Attention-based atrous convolutional neural networks: Visualisation and understanding perspectives of acoustic scenes," in *Proc. ICASSP*, 2019, pp. 56–60.

[21] H. Phan, O. Y. Chén, L. Pham, P. Koch, M. De Vos, I. McLoughlin, and A. Mertins, "Spatio-temporal attention pooling for audio scene classification," *in Proc. Interspeech*, 2019.

[22] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. ECCV*, 2018, pp. 3–19.

[23] X. Wu, R. He, Z. Sun, and T. Tan, "A light cnn for deep face representation with noisy labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, 2018.

[24] H.-J. Shim, J.-W. Jung, J.-H. Kim, and H.-J. Yu, "Capturing discriminative information using a deep architecture in acoustic scene classification," *Applied Sciences*, vol. 11, no. 18, 2021.

[25] Y.-R. Lee, S.-Y. Lim, and I.-Y. Kwak, "Cnn-based acoustic scene classification system," *Electronics*, vol. 10, no. 4, pp. 371, 2021.

[26] H. Hu, C.-H. H. Yang, X. Xia, X. Bai, X. Tang, Y. Wang, S. Niu, L. Chai, J. Li, H. Zhu, et al., "A two-stage approach to device-robust acoustic scene classification," in *Proc. ICASSP*. IEEE, 2021, pp. 845–849.

[27] A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese, "Taskonomy: Disentangling task transfer learning," in *Proc. CVPR*, 2018, pp. 3712–3722.

[28] K. Dwivedi and G. Roig, "Representation similarity analysis for efficient task taxonomy & transfer learning," in *Proc. CVPR*, 2019, pp. 12387–12396.

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

[30] A. Jiménez, B. Elizalde, and B. Raj, "Sound event classification using ontology-based neural networks," in *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2018.

[31] K.-K. Maninis, I. Radosavovic, and I. Kokkinos, "Attentive single-tasking of multiple tasks," in *Proc. CVPR*, 2019, pp. 1851–1860.

[32] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich, "Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks," in *Proc. ICML*, 2018, pp. 794–803.

[33] H.-S. Heo, J.-W. Jung, H.-J. Shim, and B.-J. Lee, "Clova submission for the DCASE 2021 challenge: Acoustic scene classification using light architectures and device augmentation," Tech. Rep., DCASE2021 Challenge.