INCREMENTAL LEARNING OF ACOUSTIC SCENES AND SOUND EVENTS

Manjunath Mulimani, Annamaria Mesaros

Computing Sciences, Tampere University, Tampere, Finland
{manjunath.mulimani, annamaria.mesaros}@tuni.fi

ABSTRACT

In this paper, we propose a method for incremental learning of two distinct tasks over time: acoustic scene classification (ASC) and audio tagging (AT). We use a simple convolutional neural network (CNN) model as an incremental learner to solve the tasks. Generally, incremental learning methods catastrophically forget the previous task when sequentially trained on a new task. To alleviate this problem, we propose independent learning and knowledge distillation (KD) between the timesteps in learning. Experiments are performed on TUT 2016/2017 dataset, containing 4 acoustic scene classes and 25 sound event classes. The proposed incremental learner first solves the ASC task with an accuracy of 94.0%. Next, it learns to solve the AT task with an F1 score of 54.4%. At the same time, its performance on the previous ASC task decreases only by 5.1 percentage points due to the additional learning of the AT task.

Index Terms— Incremental learning, independent learning, knowledge distillation, acoustic scene classification, audio tagging

1. INTRODUCTION

The natural learning system of humans incrementally learns new concepts over time without forgetting the previously learned ones. This process of learning is known as continuous, incremental, or lifelong learning. In contrast, deep learning-based systems have the ability to learn a task efficiently, but fine-tuning the same system with a new task tends to override the previously acquired knowledge. This leads to a phenomenon of deteriorating performance on previously learned tasks known as catastrophic forgetting. Developing a robust system that should not degrade its performance significantly on previous tasks as new tasks are added is currently a pursued problem in many domains.

Most of the studies reported in the literature on incremental learning operate on images, e.g. object detection [1, 2], image classification [3, 4], and semantic segmentation [5, 6]. A few works report on incremental learning of audio such as environmental sound classification (ESC) [7, 8], audio captioning [9], and fake audio detection [10]. However, these methods are restricted to solving an initial base task followed by N incremental tasks of the same problem (e.g. ESC). In addition, these methods were designed to have the same number of classes in the incremental tasks, an assumption that does not hold in practical scenarios. Furthermore, most of these methods use a small portion of data from the previous task during the training of the system on current task data and employ complex postprocessing methods to alleviate catastrophic forgetting [11].

In this work, we propose an incremental learning method to solve distinct tasks over time: acoustic scene classification (ASC)
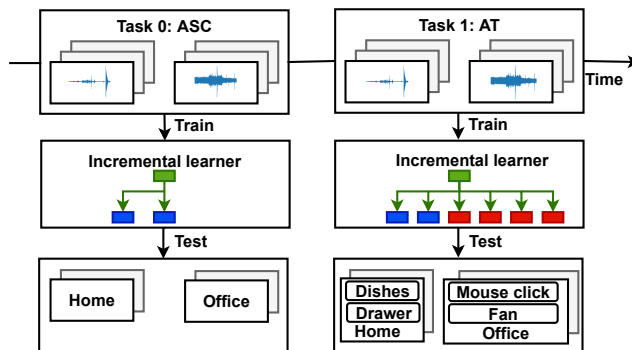
Figure 1: Incremental learning of distinct tasks: acoustic scene classification (ASC) and audio tagging (AT). Our incremental learner learns acoustic scene classes initially (blue units) and sound event classes incrementally (red units). After the learning process of each task, a learner is evaluated on classes of all tasks learned so far.

and audio tagging (AT) (see Fig. 1), to simulate the scenario of new information becoming available at a later time for the same audio material. In this case, the incremental learner first learns the acoustic scenes and then learns more detailed characterization of the acoustic content of the scene, i.e. the sound events active in the given acoustic scene. When learning the sound event classes, the same audio material is used, but the acoustic scene labels are no longer available in the learning process.

The proposed work is based on the class incremental approach [12] and uses a single classifier to learn both acoustic scenes and sound events rather than using a separate classifier for each task, as done in [13]. Using the same classifier makes the problem more challenging, because acoustic scene and sound event classes compete in a single classifier. In comparison [13] employs information on the task identity to get output from a specific classifier at the prediction step. Inspired by different incremental learning methods for image classification [3, 4, 12–14], we propose a simple CNN-based incremental learner to solve distinct tasks over time.

The main contributions of this work are summarized below.
- We design an independent learning (IndL) mechanism that allows a classifier to learn different tasks effectively.
- We combine IndL with Kullback-Leibler (KL) divergence-based distillation loss to learn new sound events while preserving the knowledge of acoustic scene classes.
- We conduct experiments on both two-step ASC-AT and three-step ASC-ASC-AT incremental setups to analyse the behavior of the proposed incremental learner over time.

The rest of the paper is organized as follows. Section 2 introduces the proposed framework for incremental learning of acoustic scenes and sound events tasks. Section 3 presents the different ex-

periments that compare incremental learning with multi-task learning and separate models for the same tasks. Finally, conclusions and future work are given in Section 4.

2. INCREMENTAL LEARNING

In our incremental learning setting, a set of ASC and AT tasks $\{\tau_0, \tau_1, \ldots, \tau_t\}$ is presented to a learner sequentially in incremental time steps t. The isolated task $\tau_t = \{(\mathbf{x}_i^{\tau_t}, \mathbf{y}_i^{\tau_t}) | 1 \leq i \leq m\}$ presented at time step t is composed of input features $\mathbf{x}_i^{\tau_t}$ and corresponding one-hot (for ASC) or multi-hot (for AT) ground truth label vectors $\mathbf{y}_i^{\tau_t} \in \{0,1\}^{C_t}$. C_t denotes the number of classes in tasks up to and including task τ_t. The two distinct ASC and AT tasks use the same audio clips but none of the tasks share the class labels. Specifically, the learner does not have access to labels of the τ_0 while learning the τ_1. Typically, this class imbalance makes the learner's predictions biased to focus on the classes in the current task and catastrophically forget the classes of the previous task.

In this work, our goal is to build a learner $\mathcal{P}^{\tau_t}$, which can solve all the tasks learned so far. A learner $\mathcal{P}^{\tau_t}$ is a deep network that includes a feature extractor $\mathcal{F}_\theta^{\tau_t}$ (parameterized by weights θ) and a fully connected layer $\mathcal{F}_\phi^{\tau_t}$ (parameterized by weights ϕ) for classification. Output logits of the network on a given input $\mathbf{x}$ are obtained by $\mathbf{o}(\mathbf{x}) = \mathcal{F}_\phi^{\tau_t}(\mathcal{F}_\theta^{\tau_t}(\mathbf{x}))$.

Generally, incremental learners learn to solve one initial base task followed by similar incremental tasks, for example like the setup in [7] where the initial task was learning 30 sound classes, then learning sets of 10 sound classes in incremental time steps for the same classification task. However, in this work, we solve distinct incremental tasks: single-label ASC and multi-label AT. We experiment with two scenarios. In the first scenario, the learner solves ASC in an initial step (τ_0 = ASC) and AT in the next time step (τ_1 = AT) as depicted in Fig. 1 (hereafter referred to as ASC-AT task). In the second scenario, the learner solves ASC in an initial step (τ_0 = ASC), and ASC and AT in subsequent incremental time steps (τ_1 = ASC and τ_2 = AT, hereafter referred to as ASC-ASC-AT).

2.1. Incremental ASC-AT learning

In the initial time step $t = 0$, a learner $\mathcal{P}^{\tau_0}$ learns $\mathcal{F}_\theta^{\tau_0}$ to classify the acoustic scene classes of a task τ_0. $\mathcal{P}^{\tau_0}$ is trained using cross-entropy loss $\mathcal{L}^{CE}$ computed using softmax σ over logits $\mathbf{o}$ as:

$$\mathcal{L}^{CE} = -\sum_{k=1}^{C} \mathbf{y}_k^{\tau_0} \cdot \log(\sigma(\mathbf{o}_k)), \qquad (1)$$

In the incremental time step $t = 1$, a new learner $\mathcal{P}^{\tau_1}$ is initialized by previous learner $\mathcal{P}^{\tau_0}$. The classifier $\mathcal{F}_\phi^{\tau_1}$ of $\mathcal{P}^{\tau_1}$ is extended to learn sound event classes of a task τ_1 by adding new output units. The output logits of $\mathcal{P}^{\tau_1}$ comprise $\mathbf{o} = \{\mathbf{o}^{old}, \mathbf{o}^{new}\}$. $\mathbf{o}^{old}$ and $\mathbf{o}^{new}$ denote the logits of acoustic scene classes and sound event classes respectively. We propose an independent learning process (IndL) through which the entire $\mathcal{P}^{\tau_1}$ is trained using separate losses for $\mathbf{o}^{old}$ and $\mathbf{o}^{new}$ logits (see Fig. 2).

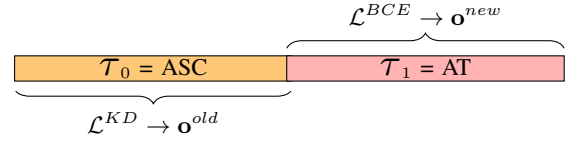A binary cross-entropy loss $\mathcal{L}^{BCE}$ is computed using sigmoid

Figure 2: Incremental ASC-AT learning; the different losses are calculated on logits of the previous ASC task and logits of the new AT task separately at the incremental time step $t = 1$.

σ over logits of the novel acoustic event classes $\mathbf{o}^{new}$ only:

$$\mathcal{L}^{BCE} = -\sum_{k=C_{t-1}+1}^{C_t} \mathbf{y}_k^{\tau_1} \cdot \log(\sigma(\mathbf{o}_k^{new}))$$
$$+ (1 - \mathbf{y}_k^{\tau_1}) \cdot \log(1 - \sigma(\mathbf{o}_k^{new})), \quad (2)$$

where C_{t-1} denotes the number of classes up to and excluding task τ_1. This independent learning of weights ϕ of $\mathcal{F}_\phi^{\tau_1}$ of the novel sound event classes from previous acoustic scene classes is meant to reduce the catastrophic forgetting. To be more specific, IndL allows $\mathcal{P}^{\tau_1}$ to learn from Eq. (2) using only the $\mathbf{o}^{new}$ logits of the sound event classes, while $\mathbf{o}^{old}$ logits of the acoustic scene classes are handled separately. This independence of the two tasks is also evident in the different loss functions per task ($\mathcal{L}^{CE}$ vs $\mathcal{L}^{BCE}$).

A distillation loss $\mathcal{L}^{KD}$ is computed using Kullback-Leibler divergence ($\mathcal{D}_{KL}$) between $\mathbf{o}^{old}$ logits of current $\mathcal{P}^{\tau_1}$ learner and output logits of previous frozen $\hat{\mathcal{P}}^{\tau_0}$ learner:

$$\mathcal{L}^{KD} = \mathcal{D}_{KL}(\hat{\mathbf{v}} || \mathbf{v}), \qquad (3)$$

where $\mathbf{v} = \sigma(\frac{\mathbf{o}^{old}}{T})$ denotes the $\mathbf{o}^{old}$ logits of $\mathcal{P}^{\tau_1}$ and $\hat{\mathbf{v}} = \sigma(\frac{\hat{\mathcal{P}}^{\tau_0}(\mathbf{x})}{T})$ denotes the logits of $\hat{\mathcal{P}}^{\tau_0}$. The σ is the softmax and T is the temperature hyperparameter to smooth the $\mathcal{L}^{KD}$. The $\mathcal{L}^{KD}$ acts as a forgetting constraint that penalizes the change concerning the output of the previous learner. Specifically, the learner $\mathcal{P}^{\tau_1}$ preserves the knowledge about the previous ASC task using $\mathcal{L}^{KD}$ and continues to learn the new AT task using $\mathcal{L}^{BCE}$. Therefore, $\mathcal{P}^{\tau_1}$ is trained using combined loss as:

$$\mathcal{L} = \mathcal{L}^{BCE} + \lambda \mathcal{L}^{KD}, \qquad (4)$$

where λ denotes the weight of the loss which we adaptively set to $\Omega \sqrt{C_t - C_{t-1}/C_t}$ as per the recommendation of [14]. Ω is a constant. $C_t - C_{t-1}$ denotes the number of new sound event classes.

We use two more techniques reported in the literature for incremental learning of images. One is, that the learning rate (LR) is reduced in incremental time steps, as done in [3]. This was shown to improve the transfer of knowledge from the old to the new learner and mitigate the adverse effect of imbalanced data in incremental time steps. Another is the use of cosine normalization in the classifier $\mathcal{F}_\phi^{\tau_t}$ [14]. It was observed that the magnitudes of the weight and bias of the previous and current classes in $\mathcal{F}_\phi^{\tau_t}$ are significantly different. Cosine normalization restricts the values of input distributions to $[-1, 1]$ and eliminates the bias that arises due to the magnitude difference.

2.2. Incremental ASC-ASC-AT learning

In this case, the learner $\mathcal{P}^{\tau_0}$ learns to solve an additional ASC task using $\mathcal{L}^{CE}$ at the initial time step $t = 0$. In the incremental time
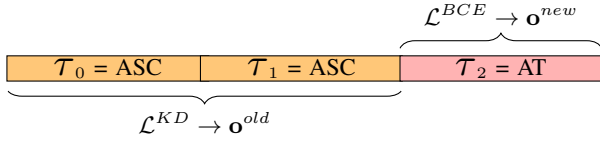
Figure 3: Incremental ASC-ASC-AT learning; the different losses are calculated on old ASC and new AT logits separately at an incremental time step $t = 2$.

step $t = 1$, $\mathcal{P}^{\tau_1}$ continues to learn a new ASC task in the absence of the old ASC task's data. Specifically, the two ASC tasks do not share audio clips nor scene classes. In the incremental time step $t = 1$, $\mathcal{L}^{CE}$ is computed independently using softmax σ over logits of the new acoustic scene classes $\mathbf{o}^{new}$ only, as per Eq. (1) and $\mathcal{L}^{KD}$ is used to hold the knowledge of old acoustic scenes, as per Eq. (3). $\mathcal{P}^{\tau_1}$ is trained using combined loss as:

$$\mathcal{L} = \mathcal{L}^{CE} + \lambda \mathcal{L}^{KD} \tag{5}$$

In the incremental time step $t = 2$, $\mathcal{P}^{\tau_2}$ learns the AT task (using $\mathcal{L}^{BCE}$) by preserving the knowledge of all old acoustic scenes (using $\mathcal{L}^{KD}$) as per the description given in 2.1 and illustrated in Fig. 3. For the AT task, the learner uses the same audio clips as the ASC task at $t = 1$.

3. EVALUATION AND RESULTS

3.1. Datasets

For ASC-AT, we use acoustic scenes and corresponding sound events from TUT 2016/2017 dataset [15, 16]. The dataset contains 192 minutes of audio recordings. Task 0 is composed of four acoustic scenes: home, residential area, city center, and office. Task 1 is composed of 25 sound events: bird singing, brakes squeaking, breathing, etc. Complete details about the data can be found in [15].

For ASC-ASC-AT, we use TUT Acoustic Scenes 2017 [17] and TUT 2016/2017. Task 0 is composed of 11 acoustic scenes: beach, bus, cafe/restaurant, car, forest path, grocery store, library, metro station, park, train, and tram. Tasks 1 and 2 are the ASC and AT from the previous experiment. The learner is trained and tested on official development and evaluation splits of the datasets in each step.

3.2. Implementation details

Input features in each time step are 40-dimensional log mel-band energies obtained from each audio segment in 40 ms frames with 50% overlap. The network architecture of the feature extractor $\mathcal{F}_\theta^{\mathcal{T}_t}$ includes three convolutional blocks, each consisting of two 3×3 convolutional layers, with batch-normalization and ReLU nonlinearity applied to each convolutional layer. 2×2 average pooling is applied to each convolutional block, and 20% dropout is applied after each average pooling to avoid overfitting. The number of feature maps of convolutional blocks is set to $\{16, 32, 64\}$. The flattened output of the last convolutional block is considered as the input to the cosine normalized fully-connected layer $\mathcal{F}_\phi^{\mathcal{T}_t}$, whose number of output units is equal to the number of classes in each time step.

The learner's network is trained using the SGD optimizer [18] with a momentum of 0.9 and a mini-batch size of 100 for 120 epochs. The initial learning rates for task 0 and incremental task(s)

Method	KD	$t = 0$ Task 0: ASC (Acc)	$t = 1$ Task 1: AT (F1)	$t = 1$ Task 0: ASC (Acc)
ASC	-	94.0	-	-
AT	-	-	53.0	-
Joint ASC-AT	-	72.0	50.4	-
Incremental ASC-AT	✗	94.0	54.4	84.1 (9.9↓)
Incremental ASC-AT	✓	94.0	54.4	88.9 (5.1↓)

Table 1: Incremental ASC (4 classes) and AT (25 classes) with and without KD, compared to individual and joint learning of tasks with a similar architecture. The value within () denotes the forgetting amount; ↓ indicates that lower is better.

are set to 0.1 and 0.01 respectively. CosineAnnealingLR [18] scheduler is used to update the optimizer in every epoch. Other hyperparameters: T and Ω are empirically set to 2 and 5 respectively.

3.3. Baseline systems and evaluation metrics

The performance of the proposed incremental ASC-AT system is compared with the individual ASC, AT, and joint ASC-AT baseline systems. The same network architecture as the incremental ASC-AT system is used in all the baseline systems for fair comparison. Individual ASC and AT systems solve only the ASC or AT task, respectively. A joint ASC-AT system is a multi-task system that is trained for ASC and AT tasks at the same time using cross-entropy loss and binary cross-entropy loss respectively, as proposed in [19]. The performance on ASC and AT tasks is evaluated using accuracy and F1 score (using a threshold of 0.5), respectively.

3.4. ASC-AT results

The experimental results provided in Table 1 compare the performance of the proposed incremental ASC-AT system with baseline systems. The learning of ASC and AT tasks at the same time as joint ASC-AT results in an overall performance lower than ASC and AT learned separately in different systems. Particularly, the accuracy of the ASC side in joint ASC-AT is significantly worse. This is also true with existing ASC-sound event detection (SED) multi-task models [16, 19].

On the other hand, the proposed incremental ASC-AT method can solve ASC and AT tasks with an accuracy of 84.1% (for the system without KD) and an F1 score of 54.4% respectively. The performance on the ASC task at the incremental time step is reduced by 9.9 percentage point (p.p.) as compared to the initial time step. Surprisingly, we see a small increase in the F1 score of the AT task as compared to the individual AT system. We hypothesize that this is due to the incremental learner already pre-trained using very relevant acoustic content (but different classes) in the initial time step, which may generate richer feature representations for the AT task. We observe that using KD is more efficient for preserving the knowledge of the previous task. The incremental ASC-AT system with KD outperforms the system without KD, having an average accuracy of 88.9% on the previous ASC task with only 5.1 p.p. reduction in performance.

Ablation setting			$t=0$	$t=1$			$t=2$		
			Task 0:	Task 1: ASC (Acc)			Task 2:	Task 0 and 1: ASC (Acc)	
IndL	KD	LRs	ASC (Acc)	Overall	Task-wise		AT (F1)	Overall	Task-wise
✓	✓	{0.1, 0.1}	65.3	40.0	Task 0: 34.6 (30.7↓)	Task 1: 54.9	53.0	38.8	Task 0: 32.9 (32.4↓)	Task 1: 54.7 (0.2↓)
✓	✓	{0.01, 0.01}	**68.1**	50.1	Task 0: 54.3 (13.8↓)	Task 1: 38.7	53.0	49.1	Task 0: 53.3 (14.8↓)	Task 1: 37.3 (1.4↓)
✓	✓	{0.001, 0.001}	49.1	37.2	Task 0: 23.4 (25.7↓)	Task 1: 77.0	43.0	34.0	Task 0: 17.7 (31.4↓)	Task 1: 75.2 (1.8↓)
✓	✓	{0.01, 0.001}	**68.1**	49.3	Task 0: 57.7 (10.4↓)	Task 1: 25.9	46.0	48.3	Task 0: 56.9 (11.2↓)	Task 1: 24.1 (1.8↓)
✓	✓	{0.1, 0.01}	65.3	**53.8**	Task 0: **54.6** (10.7↓)	Task 1: **53.1**	53.0	**52.2**	Task 0: **53.3** (12↓)	Task 1: **49.1** (4↓)
✗	✗	{0.1, 0.01}	65.3	26.4	Task 0: 0.1 (65.2↓)	Task 1: 85.0	53.0	23.2	Task 0: 0.1 (65.2↓)	Task 1: 80.4 (4.6↓)

Table 2: Incremental ASC-ASC-AT with and without independent learning (IndL) and knowledge distillation (KD) using different LR combinations in initial and incremental steps with Task 0: 11 classes, Task 1: 4 classes, Task 2: 25 classes. For $t=1$ and $t=2$ the overall ASC accuracy represents performance over all 15 scene classes, and separate task-wise accuracy is provided (over the 11 classes of Task 0 and over the 4 classes of Task 1).
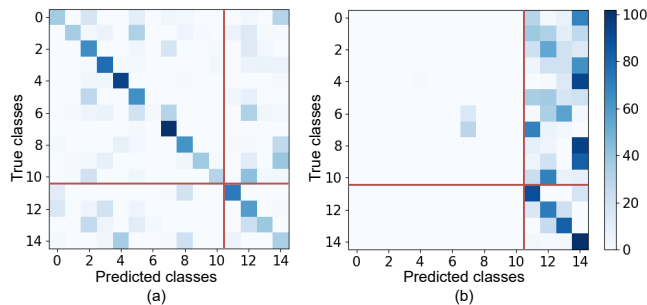

Figure 4: Confusion matrices w/ (a) and w/o (b) KD and IndL for ASC-ASC incremental learning (Task 0 to Task 1 of Table 2). Red lines separate the regions of new and old classes.

3.5. ASC-ASC-AT results

The experimental results for the ASC-ASC-AT scenario are presented in Table 2 and demonstrate the effectiveness of independent learning of acoustic scenes (as $\mathbf{o}^{old}$, $\mathbf{o}^{new}$) from tasks 0 and 1. Note that ASC-ASC tasks use different audio recordings, in contrast to ASC-AT using the same audio recordings with different labels.

As an ablation study, we experiment with a few combinations of LR, using the same and reduced LR in incremental tasks. Results show that the use of the same LR (0.1, 0.01, or 0.001) in initial and incremental time steps makes the learner either fail to learn new acoustic scene classes effectively (showing high-stability) or significantly forget the old acoustic scene classes (showing high-plasticity). This situation is also known as the stability-plasticity dilemma of a learner between new and old knowledge. Using a LR of 0.1 or 0.01 in all the steps does not affect the performance of AT task; LR 0.001 seems not suitable to solve either task, while 0.01 seems to be best for the initial ASC. Based on the results in Table 2, the combination using a LR of 0.1 and 0.01 provide a best balance between stability and plasticity, with a similar performance on the two ASC tasks at both $t=1$ and $t=2$.

Looking at learning with or without IndL, we observe a large difference in forgetting between tasks. A learner without IndL

learns from Eq. (1) using all logits $\mathbf{o}$ (combination of $\mathbf{o}^{old}$, $\mathbf{o}^{new}$) and class labels $\mathbf{y}^{\tau_1}$. Because when learning the incremental ASC task 1, the learner does not have access to the data $(\mathbf{x}^{\tau_0}, \mathbf{y}^{\tau_0})$ of previous ASC task 0, the values of the task 0 targets in $\mathbf{y}^{\tau_1}$ are zero. This makes the learner forget the old acoustic scene classes because it sees no examples of them. Hence, the amount of forgetting reached 65.2 p.p. and accuracy dropped to almost zero on the acoustic scene classes of task 0 at time step 1 (last row in Table 2). In contrast, IndL of ASC task 1 with KD achieves an accuracy of 54.6% on the acoustic scene classes of task 0, with only a 10.7 p.p. forgetting. The two cases are illustrated in Fig. 4: without KD and IndL the network mostly predicts new classes (Fig. 4b), while using KD and IndL rebalances the output (Fig. 4a).

It is worth noting that the F1 score of the AT task at $t=2$ is unaffected and it remains at 53.0%, the same as the individual AT system. This is because the AT task is always learned independently of acoustic scenes due to a different loss function, and takes advantage of a previous model trained on highly relevant acoustic material, even though the class information differs. This suggests that an independent learning mechanism is a suitable approach for learning all the tasks.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we presented incremental ASC-AT and ASC-ASC-AT systems to solve distinct tasks over time. Results show that the performance of the ASC-AT system is close to the individual ASC and AT systems and outperforms joint ASC-AT learning with a similar size architecture. Independent learning of previous and current tasks with knowledge distillation significantly decreased the problem of catastrophic forgetting. In the presented setup, the AT task is always independent of the ASC task; hence, learning these tasks would not much disturb the performances of one another irrespective of their order (whether ASC-AT or AT-ASC). Future work includes more detailed ablation studies of the different choices used to reduce forgetting and to improve overall performance, such as the order, size and type of the incremental tasks, and the use of cosine normalization.

5. REFERENCES

[1] J. Kj, J. Rajasegaran, S. Khan, F. S. Khan, and V. N. Balasubramanian, "Incremental object detection via meta-learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 9209–9216, 2021.

[2] J.-M. Perez-Rua, X. Zhu, T. M. Hospedales, and T. Xiang, "Incremental few-shot object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 846–13 855.

[3] S. Mittal, S. Galesso, and T. Brox, "Essentials for class incremental learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3513–3522.

[4] H. Ahn, J. Kwak, S. Lim, H. Bang, H. Kim, and T. Moon, "SS-IL: Separated softmax for incremental learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 844–853.

[5] L. Yu, X. Liu, and J. Van de Weijer, "Self-training for class-incremental semantic segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–12, 2022.

[6] F. Cermelli, D. Fontanel, A. Tavera, M. Ciccone, and B. Caputo, "Incremental learning in semantic segmentation from image labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4371–4381.

[7] Y. Wang, N. J. Bryan, M. Cartwright, J. P. Bello, and J. Salamon, "Few-shot continual learning for audio classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 321–325.

[8] B. Bayram and G. İnce, "An incremental class-learning approach with acoustic novelty detection for acoustic event recognition," *Sensors*, vol. 21, no. 19, p. 6622, 2021.

[9] J. Berg and K. Drosos, "Continual learning for automated audio captioning using the learning without forgetting approach," in *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2021, pp. 140–144.

[10] H. Ma, J. Yi, J. Tao, Y. Bai, Z. Tian, and C. Wang, "Continual Learning for Fake Audio Detection," in *INTERSPEECH*, 2021, pp. 886–890.

[11] K. Joseph, S. Khan, F. S. Khan, R. M. Anwer, and V. N. Balasubramanian, "Energy-based latent aligner for incremental learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7452–7461.

[12] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "ICaRL: Incremental classifier and representation learning," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2017, pp. 2001–2010.

[13] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.

[14] S. Hou, X. Pan, C. C. Loy, Z. Wang, and D. Lin, "Learning a unified classifier incrementally via rebalancing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 831–839.

[15] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 621–625.

[16] D. A. Krause and A. Mesaros, "Binaural signal representations for joint sound event detection and acoustic scene classification," in *European Signal Processing Conference (EUSIPCO)*, 2022, pp. 399–403.

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

[18] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=Skq89Scxx

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