Discriminative Distillation to Reduce Class Confusion in Continual Learning

Changhong Zhong[1,2], Zhiying Cui[1,2], Wei-Shi Zheng[1,2], Hongmei Liu[1,3(✉)], and Ruixuan Wang[1,2(✉)]

[1] School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
wangruix5@mail.sysu.edu.cn

[2] Key Laboratory of Machine Intelligence and Advanced Computing, MOE, Guangzhou, China

[3] Key Laboratory of Information Security Technology, Guangzhou, Guangdong, China

Abstract. Successful continual learning of new knowledge would enable intelligent systems to recognize more and more classes of objects. However, current intelligent systems often fail to correctly recognize previously learned classes of objects when updated to learn new classes. It is widely believed that such downgraded performance is solely due to the catastrophic forgetting of previously learned knowledge. In this study, we argue that the class confusion phenomena may also play a role in downgrading the classification performance during continual learning, i.e., the high similarity between new classes and any previously learned classes would also cause the classifier to make mistakes in recognizing these old classes, even if the knowledge of these old classes is not forgotten. To alleviate the class confusion issue, we propose a discriminative distillation strategy to help the classifier well learn discriminative features between confusing classes during continual learning. Experiments on multiple datasets support that the proposed distillation strategy, when combined with existing methods, is effective in improving continual learning.

Keywords: Continual learning · Confusing classes · Discriminative distillation

1 Introduction

Continual learning or lifelong learning aims to continually learn and absorb new knowledge over time while retaining previously learned knowledge [21]. With this ability, humans can accumulate knowledge over time and become experts in certain domains. It is desirable for the intelligent system to obtain this ability and recognize more and more objects continually, with the presumption that

This work is supported by NSFCs (No. 62071502, U1811461), the Guangdong Key Research and Development Program (No. 2020B1111190001), and the Meizhou Science and Technology Program (No. 2019A0102005).

S. Yu et al. (Eds.): PRCV 2022, LNCS 13534, pp. 580–592, 2022.
https://doi.org/10.1007/978-3-031-18907-4_45

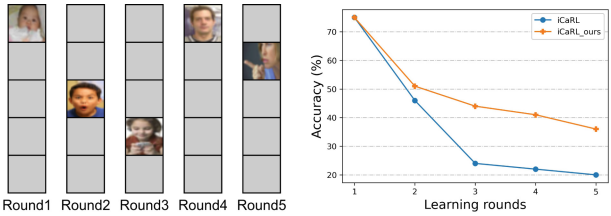

Fig. 1. Continual learning suffers not only from catastrophic forgetting but also from the confusion between old and new classes. Left: five similar classes ('baby', 'boy', 'girl', 'man', 'woman') were learned at different rounds; gray boxes represent certain other classes learned at each round. Right: classification performance on 'baby' class learned at the first round decreases over learning rounds, but the proposed method (orange) can better handle the confusion between the 'baby' class and its similar classes at later rounds compared to baseline iCaRL (blue). (Color figure online)

very limited amount or even no data is stored for the old classes when learning knowledge of new classes. The intelligent system has to update its parameters when acquiring new knowledge and often inevitably causes the downgraded performance on recognizing old classes. It has been widely believed that the downgraded performance is solely due to the *catastrophic forgetting* of old knowledge during learning new knowledge [13,14], and various approaches have been proposed to alleviate the catastrophic forgetting issue, such as by trying to keep important model parameters or outputs at various layers in convolutional neural networks (CNNs) unchanged during learning new knowledge [4,5,8,14,17].

However, sometimes simply keeping old knowledge from forgetting during continual learning may not be enough to keep classification performance from downgrading. At an early round of continual learning, since only a few classes of knowledge needs to be learned, the classifier may easily learn to use part of class knowledge to well discriminate between these classes. When any new class is visually similar to any previously learned class during continual learning, the visual features learned to recognize the old class may not be discriminative enough to discriminate between the new class and the visually similar old class (e.g., 'girl' vs. 'baby', Fig. 1), causing downgraded performance on previously learned class. We call this phenomena the *class confusion issue*. In this study, we propose a novel knowledge distillation strategy to help the classifier learn such discriminative knowledge information between old and new classes during continual learning. The basic idea is to train a temporary expert classifier to learn both the new classes and visually similar old classes during continual learning, and then distill the discriminative knowledge from the temporary expert classifier to the new classifier. To our best knowledge, it is the first time to explore the class confusion issue in continual learning. The main contributions are below:

– It is observed that continual learning is affected not only by catastrophic forgetting, but also by potential class confusion between new classes and visually similar old classes.

- A discriminative knowledge distillation strategy is proposed to help the classifier discriminate confusing classes.
- Initial experiments on multiple image classification datasets support that the proposed discriminative distillation can be flexibly combined with existing methods and is effective in improving continual learning.

2 Related Work

Generally, there are two types of continual learning problems, task-incremental and class-incremental. Task-incremental learning (TIL) presumes that one model is incrementally updated to solve more and more tasks, often with multiple tasks sharing a common feature extractor but having task-specific classification heads. The task identification is available during inference, i.e., users know which model head should be applied when predicting the class label of a new data. In contrast, class-incremental learning (CIL) presumes that one model is incrementally updated to predict more and more classes, with all classes sharing a single model head. This study focuses on the CIL problem.

Existing approaches to the two types of continual learning can be roughly divided into four groups, regularization-based, expansion-based, distillation-based, and regeneration-based. Regularization-based approaches often find the model parameters or components (e.g., kernels in CNNs) crucial for old knowledge, and then try to keep them unchanged with the help of regularization loss terms when learning new classes [1,11,14]. While keeping the parameters unchanged could help models keep old knowledge in a few rounds of continual learning, it is not able to solve the confusion issue because more and more parameters in CNNs are frozen. To make models more flexibly learn new knowledge, expansion-based approaches are developed by adding new kernels, layers, or even sub-networks when learning new knowledge [9,12,16,22,29]. Although expanding the network architecture can potentially alleviate the confusion issue to some extent because the expanded kernels might help extract more discriminative features, most expansion-based approaches are initially proposed for TIL and might not be flexibly extended for CIL. In comparison, distillation-based approaches can be directly applied to CIL by distilling knowledge from the old classifier (for old classes) to the new classifier (for both old and new classes) during learning new classes [2,10,17,19,24]. In addition, regeneration-based approaches have also been proposed particularly when none of old-class data is available during learning new classes. The basic idea is to train an auto-encoder [6,23,25] or generative adversarial network (GAN) [20,28] to synthesize old data for each old class, such that plenty of synthetic but realistic data for each old class are available during learning new classes. All the existing approaches are proposed to alleviate the catastrophic forgetting issue, without aware of the existence of the class confusion issue.

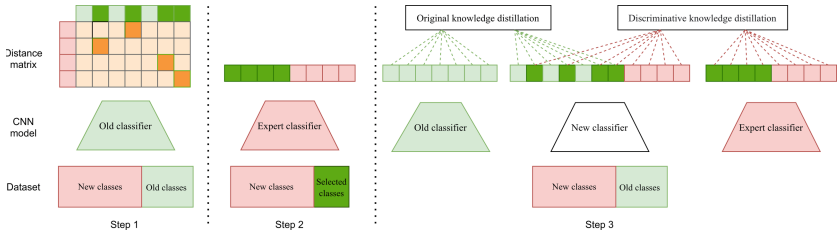

Fig. 2. Discriminative knowledge distillation pipeline. First, the old classifier learned at the previous round is used to identify the similar old class(es) for each new class (Step 1). Then, the temporary expert classifier is trained to recognize both the new classes and their similar old classes (Step 2). Finally, the old classifier and the expert classifier are simultaneously used to teach the new classifier (Step 3). The potential confusion between new and old classes can be alleviated by the distillation from the expert classifier to the new classifier. (Color figure online)

3 Method

In contrast to most continual learning methods which only aim to reduce catastrophic forgetting during learning new classes, this study additionally aims to reduce the potential confusions between new classes and visually similar old classes. As in most distillation-based continual learning methods, only a small subset of training data is stored for each old class and available during continual learning of new classes.

3.1 Overview of the Proposed Framework

We propose a distillation strategy particularly to reduce the class confusion issue during continual learning. At each round of continual learning, besides the knowledge distillation from the old classifier learned at the previous round to the new classifier at the current round, a temporary expert classifier is trained to classify not only the new classes but also those old classes which are visually similar to the new classes (Fig. 2, Step 2), and then the discriminative knowledge of the expert classifier is distilled to the new classifier as well during training the new classifier (Fig. 2, Step 3). The knowledge distillation from the expert classifier to the new classifier would largely reduce the potential confusion between these similar classes during prediction by the new classifier. It is worth noting discriminative knowledge distillation from the expert can be used as a plug-in component for most distillation-based continual learning methods.

3.2 Expert Classifier

The key novelty of the proposed framework is the addition of the expert classifier whose knowledge will be distilled to the new classifier. The expert classifier at each learning round is trained to classify both the new classes at the current

round and those old classes which are similar to and therefore more likely confused with the new classes. In this way, the discriminative knowledge between such similar classes can be explicitly learned, and the distillation of such discriminative knowledge would likely reduce the confusion between each new class and its similar old class(es).

To find the old class(es) similar to each new class, the feature extractor part of the old classifier is used to output the feature representation of each new-class data and stored old-class data, and then the class-centre representation is obtained respectively for each class by averaging the feature representations of all data belonging to the same class. The Euclidean distance from the class-centre representation of the new class to that of each old class is then used to select the most similar (i.e., closest) old class(es) for the new class (Fig. 2, Step 1). While sometimes one new class may have multiple similar old classes and another new class may have no similar old classes, without loss of generality, the same number of similar old classes is selected for each new class in this study and no old class is selected multiple times at each learning round.

Once the old classes similar to the new classes are selected, the expert classifier can be trained using all the training data of the new classes and the stored similar old-class data (Fig. 2, Step 2). Since only very limited number of old data is available for each old class, the training data set is imbalanced across classes, which could make the classifier focus on learning knowledge of the large (i.e., new) classes. To alleviate the imbalance issue, the expert classifier is initially trained (for 80 epochs in this study) using all the available training set and then fine-tuned (for 40 epochs in this study) with balanced dataset across classes by down-sampling the dataset of new classes.

3.3 Knowledge Distillation

The expert classifier, together with the old classifier from the previous round of continual learning, is used to jointly teach the new classifier based on the knowledge distillation strategy. Suppose $D = \{(\mathbf{x}_i, \mathbf{y}_i), i = 1, \ldots, N\}$ is the collection of all new classes of training data at current learning round and the stored small old-class data, where $\mathbf{x}_i$ is an image and the one-hot vector $\mathbf{y}_i$ is the corresponding class label. For image $\mathbf{x}_i$, let $\mathbf{z}_i = [z_{i1}, z_{i2}, \ldots, z_{it}]^\mathsf{T}$ denote the logit output (i.e., the input to the last-layer softmax operation in the CNN classifier) of the expert classifier, and $\hat{\mathbf{z}}_i = [\hat{z}_{i1}, \hat{z}_{i2}, \ldots, \hat{z}_{it}]^\mathsf{T}$ denote the corresponding logit output of the new classifier (Fig. 2, Step 3, outputs of the new classifier with dashed red lines linked), where t is the number of outputs by the expert classifier. Then, the distillation of the knowledge from the expert classifier to the new classifier can be obtained by minimizing the distillation loss $\mathcal{L}_n$,

$$\mathcal{L}_n(\boldsymbol{\theta}) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{t} p_{ij} \log \hat{p}_{ij} \,, \tag{1}$$

where $\boldsymbol{\theta}$ represents the model parameters of the new classifier, and p_{ij} and $\hat{p}_{ij}$ are from the temperature-tuned softmax operation,

$$p_{ij} = \frac{\exp\left(z_{ij}/T_n\right)}{\sum_{k=1}^{t} \exp\left(z_{ik}/T_n\right)}, \quad \hat{p}_{ij} = \frac{\exp\left(\hat{z}_{ij}/T_n\right)}{\sum_{k=1}^{t} \exp\left(\hat{z}_{ik}/T_n\right)}, \tag{2}$$

and $T_n \geq 1$ is the temperature coefficient used to help knowledge distillation [7]. Since the expert has been trained to discriminate new classes from visually similar old classes, the knowledge distillation from the expert classifier to the new classifier is expected to help the new classifier gain similar discriminative power. In other words, with the distillation, the new classifier would become less confused with the new classes and visually similar old classes, resulting in better classification performance after each round of continual learning.

Besides the knowledge distillation from the expert classifier, knowledge from the old classifier can be distilled to the new classifier in a similar way, i.e., by minimizing the distillation loss $\mathcal{L}_o$,

$$\mathcal{L}_o(\boldsymbol{\theta}) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{s} q_{ij} \log \hat{q}_{ij}\,, \tag{3}$$

where s is the number of old classes learned so far, and q_{ij} and $\hat{q}_{ij}$ are respectively from the temperature-tuned softmax over the logit of the old classifier and the corresponding logit part of the new classifier (Fig. 2, Step 3, outputs of the new classifier with dashed green lines linked), with the distillation parameter T_o.

As in general knowledge distillation strategy, besides the two distillation losses, the cross-entropy loss $\mathcal{L}_c$ over the training set D based on the output of the new classifier is also applied to train the new classifier. In combination, the new classifier can be trained by minimizing the loss $\mathcal{L}$,

$$\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}_c(\boldsymbol{\theta}) + \lambda_1 \mathcal{L}_o(\boldsymbol{\theta}) + \lambda_2 \mathcal{L}_n(\boldsymbol{\theta})\,, \tag{4}$$

where λ_1 and λ_2 are trade-off coefficients to balance the loss terms.

The proposed distillation strategy is clearly different from existing distillations for continual learning. Most distillation-based continual learning methods only distill knowledge from the old class to the new class at each learning round. The most relevant work is the dual distillation [18] which reduces catastrophic forgetting with the help of two classifiers (called expert classifier and old classifier respectively), where the expert classifier is trained only for new classes and then, together with the old classifier, distilled to the new classifier. In comparison, the expert classifier in our method is trained to learn not only the new classes but also likely confusing old classes, particularly aiming to alleviate the class confusion issue. Therefore, our method extended the dual distillation but with a brand new motivation. Most importantly, the proposed discriminative distillation can be easily combined with most existing continual learning methods by simply adding the loss term $\mathcal{L}_n(\boldsymbol{\theta})$ during classifier training at each round of continual learning.

Table 1. Statistics of datasets. [75, 2400]: size range of image height and width.

Dataset	#class	Train/class	Test/class	Size
CIFAR100	100	500	100	32×32
mini-ImageNet	100	∼1,200	100	[75, 2400]
ImageNet	1000	∼1,200	100	[75, 2400]

4 Experiments

4.1 Experimental Settings

The proposed method was evaluated on three datasets, CIFAR100 [15], the full ImageNet dataset [3], and a subset of ImageNet which contains randomly selected 100 classes (Table 1). During model training, each CIFAR100 image was randomly flipped horizontally, and each ImageNet image was randomly cropped and then resized to 224×224 pixels. On each dataset, an CNN classifier was first trained for certain number (e.g., 10, 20) of classes, and then a set of new classes' data were provided to update the classifier at each round of continual learning. Ths SGD optimizer (batch size 128) was used with an initial learning rate 0.1. The new classifier at each round of continual learning was trained for up to 100 epochs, with the training convergence consistently observed. ResNet32 and ResNet18 were used as the default CNN backbone for CIFAR100 and ImageNet (including mini-ImageNet) respectively, and $\lambda_1 = \lambda_2 = 1.0$, $T_n = T_o = 2.0$. One similar old class was selected for each new class in the expert classifier. Following iCaRL [24], the herding strategy was adopted to select a small subset of images for each new class with a total memory size K. For CIFAR100 and mini-ImageNet, the memory size is $K = 2000$. And for ImageNet, $K = 20000$.

After training at each round, the average accuracy over all learned classes so far was calculated. Such a training and evaluation process was repeated in next-round continual learning. For each experiment, the average accuracy over three runs were reported, each run with a different and fixed order of classes to be learned. All baseline methods were evaluated on the same orders of continual learning over three runs and with the same herding strategy for testing.

4.2 Effectiveness Evaluation

The proposed discriminative distillation can be plugged into most continual learning methods. Therefore, the effectiveness of the proposed distillation is evaluated by combining it respectively with existing continual learning methods, including LwF [17], iCaRL [24], UCIR [8], and BiC [27]. All the four methods are distillation-based, and therefore the only difference between each baseline and the corresponding proposed method is the inclusion of the discriminative distillation loss term during classifier training. The inference method proposed in the original papers were adopted during testing (nearest-mean-of-exemplars for iCaRL, and softmax output for LwF, UCIR, and BiC). The evaluation was

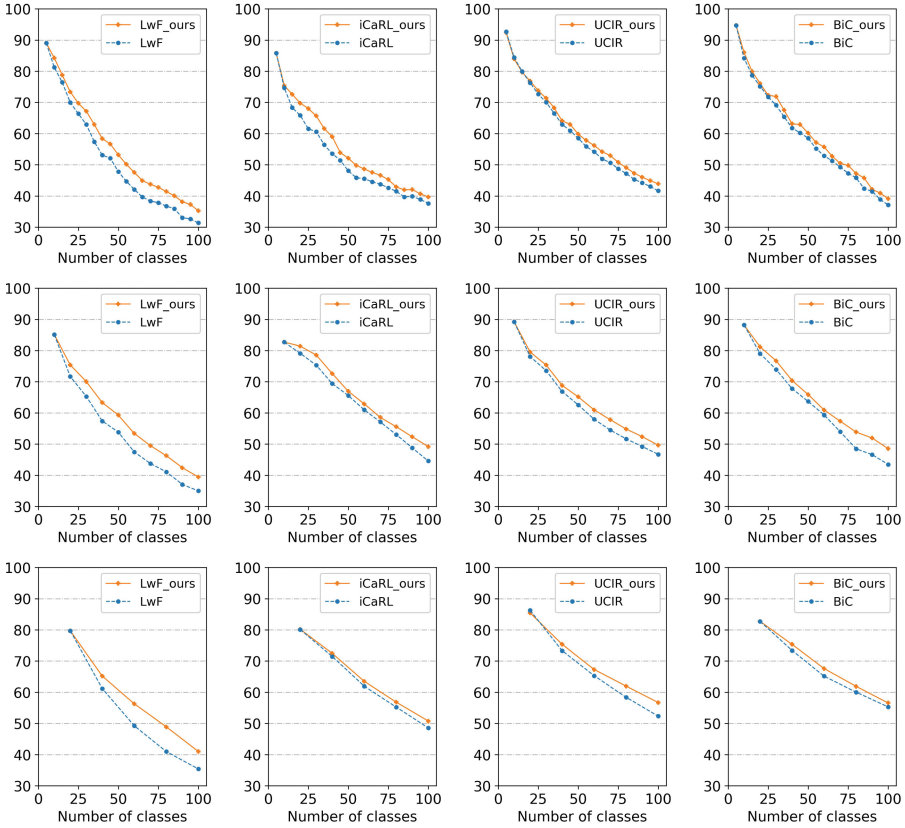

Fig. 3. Continual learning of 5 (first row), 10 (second row), and 20 (third row) new classes at each round with CIFAR100 dataset. Columns 1–4 (blue curve): performance of LwF, iCaRL, UCIR, BiC; Columns 1–4 (orange curve): performance of the proposed method built on the corresponding baseline. (Color figure online)

firstly performed on the CIFAR100 dataset. As shown in Fig. 3, when continually learning 5 classes (first row), 10 classes (second row), and 20 classes (third row) at each round respectively, each baseline method was clearly outperformed by the combination of the proposed discriminative distillation with the baseline, with around absolute 2%–5% better in accuracy at each round of continual learning. The consistent improvement in classification performance built on different continual learning methods supports the effectiveness of the proposed discriminative distillation for continual learning. Similar results were obtained from experiments on mini-ImageNet (Fig. 4) and ImageNet (Fig. 5), suggesting that the proposed discriminative distillation is effective in various continual learning tasks with different scales of new classes at each learning round.

To further investigate the effect of the proposed discriminative distillation on reducing the confusions between similar classes, the total reduction in

Table 2. Effect of the proposed distillation on the reduction of class confusion and catastrophic forgetting. Each value is the number of images incorrectly classified by the new classifier at that learning round.

Types	Methods	Learning rounds			
		2	3	4	5
Confusion	UCIR	193	547	1037	1525
	UCIR+ours	178	524	986	1399
Forgetting	UCIR	328	1002	1813	2706
	UCIR+ours	309	944	1633	2484

classification error compared to the baseline method at each round of learning was divided into two parts, one relevant to class confusion and the other to catastrophic forgetting. For CIFAR100, it is well-known that the dataset contains 20 meta-classes (e.g., human, flowers, vehicles, etc.) and each meta-class contains 5 similar classes (e.g., baby, boy, girl, man, and woman). At any round of continual learning, if the trained new classifier mis-classify one test image into another class which shares the same meta-class, such an classification error is considered partly due to class confusion (Table 2, 'Confusion'). Otherwise, if one test image of any old class is mis-classified to another class belonging to a different meta-class, this classification error is considered partly due to catastrophic forgetting (Table 2, 'Forgetting'). Table 2 (2nd row) shows that the proposed discriminative distillation did help reduce the class confusion error compared to the corresponding baseline (1st row) at various learning rounds. In addition, the discriminative distillation can also help reduce catastrophic forgetting (Table 2, last two rows), consistent with previously reported results based on distillation of only new classes from the expert classifier [18].

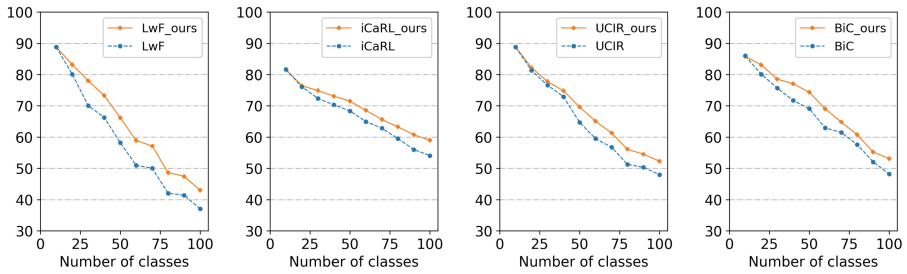

Fig. 4. Continual learning of 10 new classes at each round on mini-ImageNet.

The effect of the discriminative distillation was also visually confirmed with demonstrative examples of attention map changes over learning rounds (Fig. 6). For example, while the classifier trained based on the baseline UCIR can attend

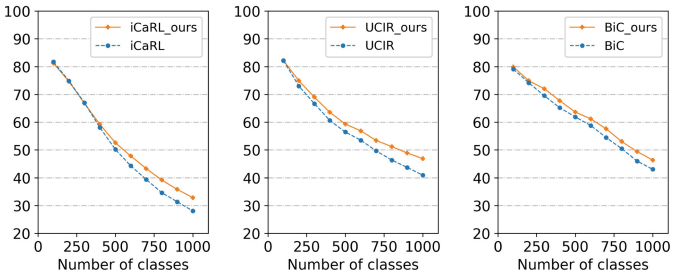

Fig. 5. Continual learning of 100 new classes at each round on ImageNet.

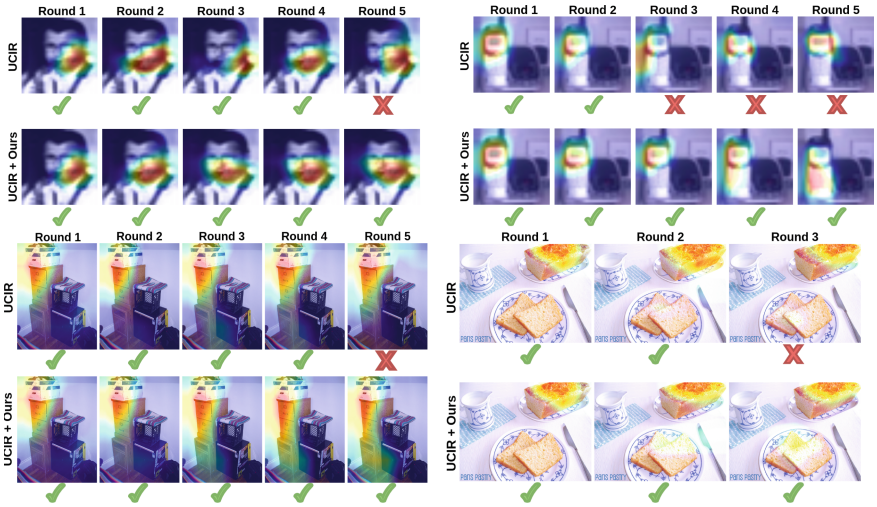

Fig. 6. Demonstrative attention maps over learning rounds from the baseline UCIR (upper row) and the correspondingly proposed method (bottom row). Input images are from CIFAR100 and mini-ImageNet, and each attention map (heatmap overlapped on input) for the ground-truth class was generated by Grad-CAM [26] from the trained classifier at each learning round. The tick or cross under each image represents the classification result.

to some part of the 'man' face over learning rounds, this test image was mis-classified at the last round (Fig. 6, top row, left half). This suggests that the mis-classification is probably not due to forgetting old knowledge (otherwise the attended region at last learning round would be much different from that at the first round). In comparison, the classifier based on the correspondingly proposed method learned to attend to larger face regions and can correctly classify the image over all rounds (Fig. 6, second row, left half), probably because the expert classifier learned to find that more face regions are necessary in order to discriminate different types of human faces (e.g., 'man' vs. 'women') and such discriminative knowledge was distilled from the expert classifier to the new classifier during continual

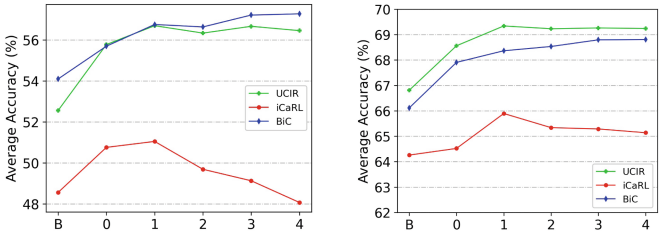

Fig. 7. Ablation study built on different baseline methods, with 20 new classes learned at each round on CIFAR100. X-axis: number of similar old classes selected for each new class during continual learning; '0' means the expert classifier only learns new classes, and 'B' means the expert classifier is not applied during learning. Y-axis: the mean classification accuracy over all the classes at the final round (Left), or the average of mean class accuracy over all rounds (Right).

learning. Similar results can be obtained from the other three examples (Fig. 6, 'phone', 'file cabinet', and 'bread' images).

4.3 Ablation Study

The effect of the discriminative distillation is further evaluated with a series of ablation study built on different baseline methods. As Fig. 7 shows, compared to the baselines ('Baseline' on the X-axis) and the dual distillation which does not learn any old classes in the expert classifier ('0' on the X-axis), learning to classify both new and similar old classes by the expert classifier and then distilling the discriminative knowledge to the new classifier ('1' to '4' on X-axis) often improves the continual learning performance, either at the final round (Left) or over all rounds (Right). Adding more old classes for the expert classifier does not always improve the performance of the new classifier (Fig. 7, Left, red curve), maybe because the inclusion of more old classes distracts the expert classifier from learning the most discriminative features between confusing classes.

5 Conclusions

Continual learning may be affected not only by catastrophic forgetting of old knowledge, but also by the class confusion between old and new knowledge. This study proposes a simple but effective discriminative distillation strategy to help the classifier handle both issues during continual learning. The distillation component can be flexibly embedded into existing approaches to continual learning. Initial experiments on natural image classification datasets shows that explicitly handling the class confusion issue can further improve continual learning performance. This suggests that both catastrophic forgetting and class confusion may need to be considered in future study of continual learning.

References

1. Abati, D., Tomczak, J., Blankevoort, T., Calderara, S., Cucchiara, R., Bejnordi, B.E.: Conditional channel gated networks for task-aware continual learning. In: CVPR (2020)
2. Castro, F.M., Marín-Jiménez, M.J., Guil, N., Schmid, C., Alahari, K.: End-to-end incremental learning. In: ECCV (2018)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: CVPR (2009)
4. Dhar, P., Singh, R.V., Peng, K.C., Wu, Z., Chellappa, R.: Learning without memorizing. In: CVPR (2019)
5. Douillard, A., Cord, M., Ollion, C., Robert, T., Valle, E.: PODNet: pooled outputs distillation for small-tasks incremental learning. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12365, pp. 86–102. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58565-5_6
6. Hayes, T.L., Kafle, K., Shrestha, R., Acharya, M., Kanan, C.: REMIND your neural network to prevent catastrophic forgetting. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12353, pp. 466–483. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58598-3_28
7. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: NIPS Workshop (2015)
8. Hou, S., Pan, X., Loy, C.C., Wang, Z., Lin, D.: Learning a unified classifier incrementally via rebalancing. In: CVPR (2019)
9. Hung, C.Y., Tu, C.H., Wu, C.E., Chen, C.H., Chan, Y.M., Chen, C.S.: Compacting, picking and growing for unforgetting continual learning. In: NIPS (2019)
10. Iscen, A., Zhang, J., Lazebnik, S., Schmid, C.: Memory-efficient incremental learning through feature adaptation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12361, pp. 699–715. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58517-4_41
11. Jung, S., Ahn, H., Cha, S., Moon, T.: Continual learning with node-importance based adaptive group sparse regularization. In: NIPS (2020)
12. Karani, N., Chaitanya, K., Baumgartner, C., Konukoglu, E.: A lifelong learning approach to brain MR segmentation across scanners and protocols. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 476–484. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_54
13. Kemker, R., McClure, M., Abitino, A., Hayes, T.L., Kanan, C.: Measuring catastrophic forgetting in neural networks. In: AAAI (2018)
14. Kirkpatrick, J., et al.: Overcoming catastrophic forgetting in neural networks. In: Proceedings of the National Academy of Sciences (2017)
15. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Technical report, University of Toronto (2009)
16. Li, X., Zhou, Y., Wu, T., Socher, R., Xiong, C.: Learn to grow: a continual structure learning framework for overcoming catastrophic forgetting. In: ICML (2019)
17. Li, Z., Hoiem, D.: Learning without forgetting. IEEE Trans. Pattern Anal. Mach. Intell. **40**(12), 2935–2947 (2017)
18. Li, Z., Zhong, C., Wang, R., Zheng, W.-S.: Continual learning of new diseases with dual distillation and ensemble strategy. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 169–178. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_17

19. Meng, Q., Shin'ichi, S.: ADINet: attribute driven incremental network for retinal image classification. In: CVPR (2020)
20. Ostapenko, O., Puscas, M., Klein, T., Jahnichen, P., Nabi, M.: Learning to remember: a synaptic plasticity driven framework for continual learning. In: CVPR (2019)
21. Parisi, G.I., Kemker, R., Part, J.L., Kanan, C., Wermter, S.: Continual lifelong learning with neural networks: a review. Neural Netw. **113**, 54–71 (2019)
22. Rajasegaran, J., Hayat, M., Khan, S.H., Khan, F.S., Shao, L.: Random path selection for continual learning. In: NIPS (2019)
23. Rao, D., Visin, F., Rusu, A., Pascanu, R., Teh, Y.W., Hadsell, R.: Continual unsupervised representation learning. In: NIPS (2019)
24. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: iCaRL: incremental classifier and representation learning. In: CVPR (2017)
25. Riemer, M., Klinger, T., Bouneffouf, D., Franceschini, M.: Scalable recollections for continual lifelong learning. In: AAAI (2019)
26. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: ICCV (2017)
27. Wu, Y., et al.: Large scale incremental learning. In: CVPR (2019)
28. Xiang, Y., Fu, Y., Ji, P., Huang, H.: Incremental learning using conditional adversarial networks. In: ICCV (2019)
29. Yan, S., Xie, J., He, X.: DER: dynamically expandable representation for class incremental learning. In: CVPR (2021)