Expert with Outlier Exposure for Continual Learning of New Diseases

Zhengjing Xu[1,2], Kanghao Chen[1,2], Wei-Shi Zheng[1,2], Zhijun Tan[1], Xiaobo Yang[3], and Ruixuan Wang✉[1,2,4]

[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2]Key Laboratory of Machine Intelligence and Advanced Computing, MOE, China
[3]State Key Laboratory of Dampness Syndrome of Chinese Medicine, Guangzhou, China
[4]Department of Network Intelligence, Peng Cheng Laboratory, Shenzhen, China

2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-6819-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/BIBM55620.2022.9995352

***Abstract*—Current intelligent diagnosis systems struggle to continually learn to diagnose more and more diseases due to catastrophic forgetting of old knowledge when learning new knowledge. Although storing small old data for subsequent continual learning can effectively help alleviate the forgetting issue, the heavy data imbalance between old classes and to-be-learned new classes in classifier training often causes biased prediction towards the new classes just learned by the updated classifier. In this study, an outlier detection technique is novelly applied to train an additional expert classifier for new classes to help alleviate the class imbalance issue and discriminate the learned new classes from old classes during inference (instead of the training phase). Specially, the stored small data of old classes are considered as outliers during training the expert classifier, such that the output probability distributions from the expert classifier are expected to be obviously different between test data of the old classes and those of the new classes. Such difference between old classes and new classes can be used to fine-tune the original output from the updated classifier which is responsible for prediction of all learned (old and new) classes. During inference, a novel ensemble strategy is proposed to combine the predictions from the updated classifier, the expert classifier, and the previously learned old classifier. The proposed learning and inference framework can be easily combined with existing continual learning strategies. Empirical evaluations on three medical image datasets and one natural image dataset show that the proposed framework can effectively improve continual learning performance.**

***Index Terms*—Continual learning, Outlier exposure, Ensemble model**

I. Introduction

One obstacle of deploying current intelligent diagnosis systems is that each system often can only diagnose a very limited number of diseases and can not cover all diseases for a specific organ or tissue. This is because it is difficult to collect training data for all possible diseases (particularly rare diseases) with limited resource [11], [14]. Thus, enabling an intelligent diagnosis system to have the lifelong or continual learning ability, i.e., incrementally learning to diagnose more and more diseases over time as human clinicians, may be a more practical solution. However, catastrophic forgetting of old knowledge has been widely observed when intelligent systems learn new knowledge [4], [8], [9], [16], [24]

Corresponding author: wangruix5@mail.sysu.edu.cn

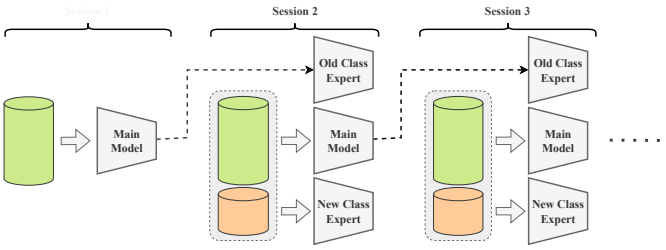

Fig. 1. Overview of the proposed continue learning framework. At the first learning session, a classifier (Main Model) is trained for the first set of classes. At each following session, the classifier (Main Model) for both new and old classes is updated, and an expert classifier (New Class Expert) for new classes is also trained with outlier exposure. The main model from the previous session is used as the expert (Old Class Expert) for old classes. The three models are ensembled for prediction at each session.

There has been much effort attempting to alleviate the catastrophic forgetting issue in continual learning mainly built on deep learning models. Among them, perhaps the most effective strategy is to integrate all the previously learned old models into the new model [12], [22]. However, integrating old models would quickly expand the model scale over multiple sessions of continual learning [5]. Currently, most studies in continual learning assume that model scale is kept from increasing substantially. In this case, keeping a small subset of old data for each class has been proven effective in keeping old knowledge from fast forgetting [1]–[3], [7], [13], [17], [18]. However, the heavy data imbalance between old classes and new classes at each learning session often leads to prediction bias towards new classes during inference [21], [25]. To alleviate the class imbalance issue, BiC [21] adds a bias correction layer to correct the model's output, where the layer is trained on a separate validation set. WA [25] corrects the biased weights in the last fully connected layer by aligning the norms of weight vectors for new classes to those for old classes. These class rebalancing strategies were evaluated only on natural images and their effects on medical image classification are still unclear.

In this study, one outlier detection strategy is novelly applied to the continual learning task to help alleviate the class imbalance issue and discriminate the learned new classes from old classes during inference (rather than during the training phase).

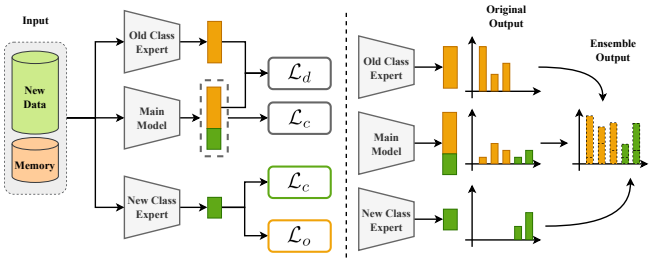

Fig. 2. The proposed framework for training and inference at a single session. Left: training phase. Right: inference phase.

Specially, at each incremental learning session, an additional expert classifier is trained with Outlier Exposure (OE) [6], [15] to recognize the new classes appearing at the current session and detect outliers not belonging to the new classes. During expert classifier training, the kept small subset of old data for each old class were used as outliers. During inference, the expert classifier for the new classes ('New Class Expert' in Figure 1), the updated classifier for both old and new classes ('Main Model' in Figure 1), and the previously trained classifier for old classes ('Old Class Expert' in Figure 1) are novelly ensembled to predict class of any new (test) data. Empirical evaluations on three medical image datasets and one natural image dataset confirmed the effectiveness of the proposed framework for continual learning.

II. Methodology

This study aims to alleviate the bias of incrementally trained classifier towards relatively new classes. Such bias is mainly from the heavy imbalance in training data between new classes and previously learned old classes, where only very limited number of old classes data are allowed to be stored during continual learning. Different from previous class rebalancing strategies, an outlier exposure technique was novelly applied to the training of an additional expert classifier which is part of a new ensemble model for more fairly and discriminative prediction between old and new classes.

A. A baseline framework for class-incremental learning

Class-incremental learning tries to make a classifier incrementally learn more and more classes over sessions. For a new session of learning, suppose the classifier has previously learned m old classes and will learn n new classes with the training dataset $D = \{(\mathbf{x}_i, \mathbf{y}_i), i = 1, \ldots, N\}$, where D is the collection of previously stored small subset of data for each old class and the whole set of training data for each new class. $\mathbf{x}_i$ is the i-th training image and the one-hot vector $\mathbf{y}_i$ represents the corresponding class label. A general continual learning framework is based on the knowledge distillation from the previous old classifier ('Old Class Expert' in Figure 2) to the current classifier ('Main Model' in Figure 2) when the current classifier learns the n new class in addition to the previously learned m old classes. In such a learning framework, the current classifier is trained by minimizing the loss $\mathcal{L}(\boldsymbol{\theta}; D)$,

$$\mathcal{L}(\boldsymbol{\theta}; D) = \mathcal{L}_c(\boldsymbol{\theta}; D) + \lambda \mathcal{L}_d(\boldsymbol{\theta}; D)\,, \tag{1}$$

where $\boldsymbol{\theta}$ denotes the model parameters of the current classifier, and λ is a coefficient constant to balance the two loss terms $\mathcal{L}_c(\boldsymbol{\theta}; D)$ and $\mathcal{L}_d(\boldsymbol{\theta}; D)$. The classification loss $\mathcal{L}_c(\boldsymbol{\theta}; D)$ helps the current classifier learn the n new classes (and the m old classes), and the distillation loss $\mathcal{L}_d(\boldsymbol{\theta})$ helps the new classifier well keep knowledge of old classes [17], respectively with the typical form

$$\mathcal{L}_c(\boldsymbol{\theta}; D) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{m+n} y_{ij} \log \hat{y}_{ij}\,, \tag{2}$$

$$\mathcal{L}_d(\boldsymbol{\theta}; D) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{m} p_{ij} \log \hat{p}_{ij}\,, \tag{3}$$

where y_{ij} is the j-th component of the one-hot class label $\mathbf{y}_i$, and $\hat{y}_{ij}$ is the j-th output element of the current classifier for the input image $\mathbf{x}_i$. p_{ij} is the j-th component of the output $\mathbf{p}_i$ from the temperature-tuned softmax operation on the logit vector of the old classifier, and similarly $\hat{p}_{ij}$ is that from the corresponding logit part of the current classifier. Note that the distillation loss may have more sophisticated forms as in PODNet [2] and UCIR [7], which is also considered in the empirical evaluation below.

Since only very limited number of old training data are available when training the current classifier as in previous studies [2], [7], [17], [21], [25], the training dataset is dominated by the n new classes. Although knowledge distillation can largely help keep knowledge of old classes in the current classifier, the heavy data imbalance between the old classes and the new classes often causes the new classifier to have biased prediction towards new classes during inference, which partly leads to worse classification performance on old classes.

B. Expert of new classes with outlier exposure

To alleviate the biased prediction towards new classes due to the data imbalance, we propose training an additional expert classifier ('New Class Expert' in Figure 2) for new classes by making use of not only the training data of the new classes but also the available limited data of the old classes. The idea is to consider data of the old classes as outliers for the expert classifier, and outlier exposure to the expert classifier during training will help the expert classifier more easily detect whether a new data is an outlier (i.e., from one of the old classes) or not during inference. In training, the objective is to obtain an expert classifier which has confident predictions (i.e., close to one-hot output) for data from the new classes but unconfident predictions (e.g., close to uniform output) for data from the old classes. This can be achieved by minimizing the loss $\mathcal{L}_e(\boldsymbol{\omega}; D)$,

$$\mathcal{L}_e(\boldsymbol{\omega}; D = D_n \cup D_m) = \mathcal{L}_c(\boldsymbol{\omega}; D_n) + \beta \mathcal{L}_o(\boldsymbol{\omega}; D_m)\,, \tag{4}$$

$$\mathcal{L}_c(\boldsymbol{\omega}; D_n) = -\frac{1}{N_n}\sum_{i=1}^{N_n}\sum_{j=1}^{n} y_{ij} \log q_{ij}\,, \tag{5}$$

$$\mathcal{L}_o(\boldsymbol{\omega}; D_m) = -\frac{1}{N_m}\sum_{i=1}^{N_m}\sum_{j=1}^{n} u_j \log q_{ij}\,, \tag{6}$$

where $\boldsymbol{\omega}$ denotes the model parameters of the expert classifier, $\mathcal{L}_c(\boldsymbol{\omega}; D_n)$ is the cross-entropy loss on the training set D_n of the n new classes, and $\mathcal{L}_o(\boldsymbol{\omega}; D_m)$ is the cross-entropy loss on the training set D_m of the m old classes specially with the expected output being uniform (i.e., $u_j = 1/n$) for each old class data. q_{ij} denotes the j-th output element of the expert classifier for the i-th input data from either D_n or D_m. N_n and N_m denote the number of data in D_n and D_m respectively, and β is a coefficient constant to balance the two loss terms.

After the expert classifier is well trained, then during inference, if a test data is from certain old class, the output from the expert classifier would be more likely close to a discrete uniform distribution. Such output information from the expert classifier can obviously help the main model reduce its biased prediction towards new classes for data of old classes, e.g., based on an ensemble strategy (See Section II-C below). Besides helping reduce biased prediction of data from old classes towards the new classes, the expert may also help reduce the possible mis-prediction from the new classes to the old classes. For example, if a test data is from one of the new classes, the prediction from the expert classifier is more likely close to a one-shot vector. The ensemble of such peaky output and the corresponding output from the main model would more likely cause higher output probability for one of the new classes, thus reducing the possibility of predicting the data as one old class.

C. Ensemble model for prediction

During inference, besides the expert classifier for the n new classes, the previous classifier (i.e., 'Old Class Expert' in Figure 2) for the m old classes can also be used to further help the current classifier (i.e., 'Main Model') alleviate the biased prediction. In particular, if a test data is from one old class, the output of the previous classifier would be more likely close to a one-hot vector and the ensemble of such peaky output and the corresponding output part from the main model would more likely cause higher output probability for one of the old classes. With this consideration, we propose an ensemble strategy based on the three models (i.e., main model, old class expert, and new class expert) for the final prediction $\mathbf{o}$ of any new data, i.e.,

$$\mathbf{o} = \gamma_0[\hat{\mathbf{y}}_{1:m}^{\intercal} + \gamma_1 \mathbf{p}^{\intercal}, \hat{\mathbf{y}}_{m+1:m+n}^{\intercal} + \gamma_2 \mathbf{q}^{\intercal}]^{\intercal}, \tag{7}$$

where $\hat{\mathbf{y}}_{1:m}$ denotes the first m output elements from the main model for the m old classes, and similarly $\hat{\mathbf{y}}_{m+1:m+n}$ denotes the last n output elements for the n new classes. $\mathbf{p}$ and $\mathbf{q}$ are the output from the previous classifier and the expert classifier respectively. γ_1 and γ_2 are two coefficient constants to balance the contributions from the three classifiers. Considering that the number of old classes is often larger than that of the new classes (i.e., $m > n$), the influence of each element from $\mathbf{q}$ in average would be larger than that from $\mathbf{p}$ on the final prediction, which would cause biased prediction towards the n new classes if $\gamma_1 = \gamma_2$. To reduce such unsatisfactory bias, γ_1 should be set a relatively larger value (i.e., $\gamma_1 > \gamma_2$). $\gamma_0 = \frac{1}{1+\gamma_1+\gamma_2}$ is a normalization factor to assure the final prediction $\mathbf{o}$ is a discrete probability distribution.

It is worth noting that our approach is different from the related dual distillation ('DDistill') method [13]. In DDistill, the new class expert is trained only with new classes of data and is utilized to help train the main model. In contrast, our approach additionally uses the kept old data to train the New Class Expert for outlier detection, and the New Class Expert is utilized as part of the ensemble model during inference in our approach. The novelty of our approach is three-fold. First, it novelly applies an outlier detection technique to the continual learning task, which can help alleviate the class imbalance issue and discriminate the learned new classes from old classes during inference. Second, partly based on the outlier detection ability of the New Class Expert, a novel ensemble strategy is proposed by combining the Old Class Expert, the New Class Expert, and the Main Model. Third, our approach can be used as a plug-in to easily combine with existing continual learning methods to boost learning performance, as shown in the following experiments.

TABLE I
DATASETS USED IN EXPERIMENTS.

Datasets	Modality	Classes	Train Set	Test Set	Image size	Memory size
Skin40	Dermoscopy	40	2,000	400	[420, 1640]	totally 50
Skin8	Dermoscopy	8	3,555	705	[600, 1024]	totally 50
MedMNIST8	Hybrid	51	483,201	79,444	28×28	40 per class
CIFAR100	Natural	100	50,000	10,000	32×32	20 per class

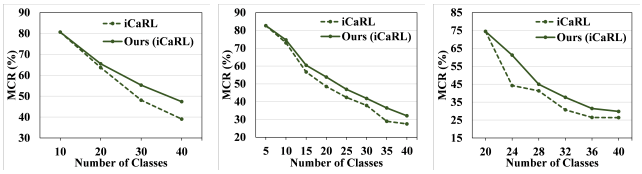

Fig. 3. The classification performance over learning sessions with the protocol B0-4 (Left), B0-8 (Middle) and B20-6 (Right) respectively on Skin40.

III. EXPERIMENT

A. Experimental setting

Datasets: Four public datasets were used to evaluate the performance of the proposed class-incremental learning framework (Table I). While data are balanced across classes in Skin40 [19] and CIFAR100 [10], Skin8 and MedMNIST8 are not. Skin8 is a 8-class dataset which comes from the classification challenge of dermoscopic images held by ISIC'2019 [20]. Due to highly class-imbalanced in the original dataset, 628 images were randomly selected from six larger classes and all images (fewer than 628) were kept for the other two smaller classes. MedMNIST8 contains 8 different datasets from MedMNIST [23], including PathMNIST, DermaMNIST, OCTMNIST, PneumoniaMNIST, BreastMNIST, BloodMNIST, TissueMNIST and OrganAMNIST. Each of the 8 datasets contains different number of images with a different imaging modality, which is challenging for continual learning.

TABLE II
CONTINUAL LEARNING PERFORMANCE ON SKIN8 AND SKIN40. EACH SUBSCRIPT IS THE STANDARD DEVIATION OF MCRS OVER THREE RUNS.

Methods	Skin8 B0-4		Skin40 B0-4		Skin40 B0-8		Skin40 B20-6	
	Avg	Last	Avg	Last	Avg	Last	Avg	Last
iCaRL	$55.2_{\pm 6.39}$	$34.7_{\pm 3.10}$	$57.9_{\pm 0.63}$	$39.1_{\pm 2.01}$	$49.7_{\pm 1.63}$	$27.5_{\pm 1.48}$	$40.6_{\pm 1.61}$	$26.4_{\pm 0.81}$
Ours (iCaRL)	$56.1_{\pm 6.41}$	$36.4_{\pm 2.55}$	$62.2_{\pm 0.73}$	$47.4_{\pm 1.65}$	$53.6_{\pm 1.82}$	$32.1_{\pm 2.11}$	$46.6_{\pm 2.47}$	$29.8_{\pm 1.77}$
UCIR	$58.4_{\pm 5.17}$	$35.6_{\pm 2.40}$	$52.0_{\pm 2.11}$	$30.3_{\pm 2.35}$	$44.5_{\pm 1.53}$	$23.3_{\pm 1.82}$	$38.7_{\pm 2.87}$	$26.6_{\pm 1.15}$
Ours (UCIR)	$61.5_{\pm 6.64}$	$40.7_{\pm 2.21}$	$58.3_{\pm 3.17}$	$41.4_{\pm 2.98}$	$50.2_{\pm 1.23}$	$27.3_{\pm 0.85}$	$46.2_{\pm 1.97}$	$30.8_{\pm 1.71}$
BiC	$54.7_{\pm 4.65}$	$27.1_{\pm 3.00}$	$51.5_{\pm 1.46}$	$29.1_{\pm 1.91}$	$42.4_{\pm 3.76}$	$24.5_{\pm 1.50}$	$36.7_{\pm 2.61}$	$22.5_{\pm 2.36}$
Ours (BiC)	$59.2_{\pm 4.18}$	$37.8_{\pm 1.35}$	$56.8_{\pm 1.21}$	$40.8_{\pm 0.76}$	$46.2_{\pm 2.79}$	$25.7_{\pm 1.90}$	$39.3_{\pm 2.00}$	$22.8_{\pm 0.76}$
PODNet	$58.0_{\pm 6.53}$	$35.2_{\pm 3.07}$	$56.7_{\pm 1.42}$	$34.9_{\pm 2.48}$	$45.0_{\pm 2.68}$	$18.6_{\pm 1.69}$	$38.4_{\pm 1.44}$	$17.6_{\pm 1.15}$
Ours (PODNet)	$59.1_{\pm 5.97}$	$36.1_{\pm 3.49}$	$61.0_{\pm 3.03}$	$43.7_{\pm 2.05}$	$48.7_{\pm 2.06}$	$23.8_{\pm 1.08}$	$44.8_{\pm 1.44}$	$21.4_{\pm 3.76}$
DDistill	$55.8_{\pm 1.78}$	$34.9_{\pm 1.79}$	$52.2_{\pm 1.31}$	$34.2_{\pm 1.26}$	$48.6_{\pm 0.21}$	$29.1_{\pm 2.40}$	$42.5_{\pm 0.45}$	$26.7_{\pm 2.80}$
Ours (DDistill)	$56.2_{\pm 4.96}$	$36.5_{\pm 1.32}$	$57.2_{\pm 1.09}$	$37.9_{\pm 1.39}$	$49.5_{\pm 1.21}$	$29.9_{\pm 2.29}$	$44.2_{\pm 1.70}$	$30.8_{\pm 1.78}$

TABLE III
CLASSIFICATION PERFORMANCE (MCR) FOR THE OLD CLASSES LEARNED IN THE FIRST SESSION (SESSION 1) OVER LEARNING SESSIONS WITH THE SKIN40 B0-4 PROTOCOL.

Mothod	Session 1	Session 2	Session 3	Session 4
iCaRL	$79.0_{\pm 1.00}$	$54.7_{\pm 4.04}$	$47.3_{\pm 4.93}$	$38.0_{\pm 4.00}$
Ours (iCaRL)	$79.0_{\pm 1.00}$	$67.3_{\pm 3.51}$	$48.0_{\pm 4.36}$	$39.0_{\pm 6.00}$
UCIR	$80.7_{\pm 1.53}$	$50.3_{\pm 5.51}$	$24.3_{\pm 1.53}$	$15.7_{\pm 3.06}$
Ours (UCIR)	$80.7_{\pm 1.53}$	$51.3_{\pm 4.73}$	$49.0_{\pm 5.57}$	$31.0_{\pm 2.65}$
BiC	$80.3_{\pm 2.02}$	$61.3_{\pm 6.43}$	$32.7_{\pm 2.89}$	$31.0_{\pm 1.73}$
Ours (BiC)	$80.3_{\pm 2.02}$	$67.3_{\pm 1.53}$	$51.3_{\pm 6.66}$	$40.0_{\pm 2.65}$
PODNet	$81.0_{\pm 3.61}$	$58.7_{\pm 6.81}$	$40.7_{\pm 5.13}$	$22.3_{\pm 4.04}$
Ours (PODNet)	$81.0_{\pm 3.61}$	$72.0_{\pm 8.66}$	$57.3_{\pm 6.11}$	$32.0_{\pm 5.57}$
DDistill	$80.7_{\pm 2.08}$	$46.0_{\pm 5.57}$	$39.3_{\pm 4.04}$	$23.7_{\pm 1.53}$
Ours (DDistill)	$80.7_{\pm 2.08}$	$61.0_{\pm 2.00}$	$51.3_{\pm 5.69}$	$35.3_{\pm 4.62}$

TABLE IV
CONTINUAL LEARNING PERFORMANCE ON CIFAR100 AND MEDMNIST8.

Methods	CIFAR100 B0-10		CIFAR100 B0-5		MedMNIST8	
	Avg	Last	Avg	Last	Avg	Last
iCaRL	$55.5_{\pm 0.71}$	$32.6_{\pm 1.37}$	$61.8_{\pm 1.04}$	$42.9_{\pm 1.22}$	$33.4_{\pm 2.13}$	$12.6_{\pm 1.85}$
Ours (iCaRL)	$58.6_{\pm 0.65}$	$36.9_{\pm 1.30}$	$66.0_{\pm 1.08}$	$49.3_{\pm 1.42}$	$37.4_{\pm 1.98}$	$13.4_{\pm 1.13}$
UCIR	$56.9_{\pm 0.76}$	$43.4_{\pm 0.85}$	$62.9_{\pm 1.23}$	$49.5_{\pm 0.78}$	$42.4_{\pm 1.93}$	$31.9_{\pm 5.73}$
Ours (UCIR)	$60.0_{\pm 0.75}$	$45.6_{\pm 1.10}$	$64.4_{\pm 1.45}$	$51.2_{\pm 1.03}$	$53.5_{\pm 1.75}$	$35.2_{\pm 7.15}$
BiC	$56.2_{\pm 2.36}$	$41.5_{\pm 1.13}$	$61.2_{\pm 1.52}$	$45.4_{\pm 2.78}$	$56.9_{\pm 0.90}$	$49.0_{\pm 4.75}$
Ours (BiC)	$57.6_{\pm 2.31}$	$42.5_{\pm 0.99}$	$64.3_{\pm 1.67}$	$47.1_{\pm 3.37}$	$62.2_{\pm 0.12}$	$58.7_{\pm 2.23}$
PODNet	$55.0_{\pm 1.02}$	$39.4_{\pm 0.17}$	$64.0_{\pm 0.84}$	$48.9_{\pm 0.50}$	$46.7_{\pm 1.28}$	$42.4_{\pm 3.12}$
Ours (PODNet)	$59.6_{\pm 1.04}$	$41.5_{\pm 0.74}$	$68.3_{\pm 0.83}$	$54.4_{m 0.45}$	$54.5_{\pm 0.87}$	$45.1_{\pm 0.87}$
DDistill	$57.9_{\pm 1.80}$	$41.8_{\pm 0.36}$	$59.6_{\pm 1.65}$	$45.6_{\pm 0.92}$	$63.3_{\pm 0.55}$	$57.4_{\pm 0.32}$
Ours (DDistill)	$65.7_{\pm 1.42}$	$46.2_{\pm 1.11}$	$68.7_{\pm 1.09}$	$52.7_{\pm 0.79}$	$64.9_{\pm 0.80}$	$60.2_{\pm 0.46}$

Protocol: Our method was evaluated mainly based on two widely adopted protocols. The first protocol (with format B0-T) is to split the dataset into T subsets, each of which is for one learning session and contains the same number of classes. The second protocol (with format Bα-T) is to split the dataset into a larger subset containing α classes for the first learning session and the other $T-1$ smaller subsets for the remaining $T-1$ learning sessions, with the same (but smaller than α) number of classes contained in each smaller subset. However, for MedMNIST8, each of the 8 pre-defined datasets is used for a learning session. Following iCaRL [17], a small number of images for each old class (Table I, 'Memory') were selected and updated using the herding strategy.

TABLE V
ABLATION STUDY OF THE PROPOSED FRAMEWORK.

Main.	Old.	New.	Skin40 B0-4		Skin40 B0-8		Skin40 B20-6	
			Avg	Last	Avg	Last	Avg	Last
✓			$56.7_{\pm 1.42}$	$34.9_{\pm 2.48}$	$45.0_{\pm 2.68}$	$18.6_{\pm 1.69}$	$38.4_{\pm 1.44}$	$17.6_{\pm 1.15}$
✓	✓		$58.1_{\pm 2.47}$	$40.1_{\pm 0.36}$	$47.6_{\pm 2.79}$	$23.1_{\pm 0.69}$	$42.7_{\pm 1.37}$	$20.3_{\pm 2.17}$
✓		✓	$59.8_{\pm 2.94}$	$38.9_{\pm 2.15}$	$48.5_{\pm 2.18}$	$22.4_{\pm 1.65}$	$42.5_{\pm 0.63}$	$20.3_{\pm 3.33}$
✓	✓	✓	$61.0_{\pm 3.04}$	$43.7_{\pm 2.05}$	$48.7_{\pm 2.06}$	$23.8_{\pm 1.08}$	$44.8_{\pm 1.44}$	$21.4_{\pm 3.76}$

Implementation details: For the classifier backbones, ResNet-18 was adopted on Skin40 and Skin8, and ResNet-32 was adopted on CIFAR100 and MedMNIST8. In training, stochastic gradient descent (SGD) optimizer with initial learning rate 0.01, weight decay (0.0005), and cosine annealing scheduling adopted. Batch size was 16 for two skin datasets and 128 for the other two. All models were trained up to 250 epochs and the convergence of training was consistently observed. Hyper-parameters were empirically set, with $\lambda = 1.0$, $\beta = 0.5$, $\gamma_1 = 1.0$, and $\gamma_2 = 0.7$ for all experiments on all datasets. For evaluation, the mean class recall (MCR) over all learned classes after each session of continual learning, including the MCR at the last learning step ('Last'), and the average MCR over all learning sessions ('Avg') were used as the performance measure. The mean and standard deviation of MCR over three runs were reported for each experiment.

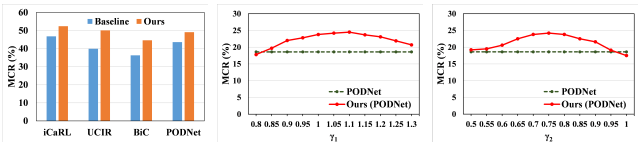

Fig. 4. Sensitivity study on Skin40. Left: with memory size 100 and protocol B0-4. Middle and Right: with protocol B0-8.

B. Evaluation of the proposed framework

The proposed learning and inference framework was evaluated by comparing with state-of-the-art continual learning methods including iCaRL [17], UCIR [7], BiC [21], PODNet [2], and DDistill [13]. All the strong baseline methods

were performed with suggested hyper-parameter settings in the original studies and evaluated on the same orders of continual learning over three runs for each experiment. The same memory buffer setting was adopted as mentioned above. For fair comparison, each baseline was compared to the corresponding version which used the same baseline as part of our method. Note that the traditional multi-class head instead of the neareast class mean was used in iCaRL during inference in order to faily compare with our method. As Table II shows, each version of our method outperforms the corresponding strong baseline with all learning settings on the two skin datasets Skin8 and Skin40. As an example, Figure 3 shows the classification performance over the course of continual learning based on the iCarL baseline, which consistently supports that the proposed framework helps improve the continual learning. Table III summarizes the classification performance for the first set of old classes learned at the first session over the continual learning process. It can be observed that our method always performs better than the corresponding strong baseline method on the old classes, further confirming that our method can help alleviate the catastrophic forgetting of old knowledge. What is more, the boosted performance by our method on the CIFAR100 dataset and the challenging MedMNIST dataset (Table IV) suggests that our method can be well generalized to natural imaging domain and even hybrid imaging domains.

C. Ablation and sensitivity studies

To further confirm the effectiveness of the proposed framework, an ablation study was performed by removing each component of the framework. As demonstrated in Table V, both New Class Expert ('New' in Table V) and Old Class Expert ('Old') help the Main Model ('Main') improve the continual learning performance. In addition, the effectiveness of the proposed framework is not limited to specific hyper-parameter settings. For example, when varying memory size from 50 to 100 (Figure 4, Left) or the hyper-parameters γ_1 and γ_2 (Middle and Right) within a relatively large range respectively, our method always outperforms the corresponding baseline, confirming the stability of the proposed framework.

Conclusion

In this study, an outlier detection technique was novelly applied to help improve the class-incremental learning performance, mainly by training an additional expert classifier with outlier exposure and then ensembled with the main model and the previous old model during inference. Extensive evaluations on two medical image datasets and one natural image dataset consistently support that the proposed class-incremental learning framework is effective and can be easily combined with existing strategies as a plug-in component. Future work includes investigation of alternative outlier detection strategies for continual learning and its application in practical intelligent diagnosis systems.

Acknowledgments. This work is supported by NSFCs (No. 62071502, U1811461, 81673845), the Guangdong Key Research and Development Program (No. 2020B1111190001), Key Project of State Key Laboratory of Dampness Syndrome of Chinese Medicine (No. SZ2021ZZ0303) and the Meizhou Science and Technology Program (No. 2019A0102005).

References

[1] J. Bang, H. Kim, Y. Yoo, J.-W. Ha, and J. Choi, "Rainbow memory: Continual learning with a memory of diverse samples," in *CVPR*, 2021, pp. 8218–8227.

[2] A. Douillard, M. Cord, C. Ollion, T. Robert, and E. Valle, "Podnet: Pooled outputs distillation for small-tasks incremental learning," in *ECCV*, 2020.

[3] R. M. French, "Catastrophic forgetting in connectionist networks," *Trends in Cognitive Sciences*, pp. 128–135, 1999.

[4] R. M. French and N. Chater, "Using noise to compute error surfaces in connectionist networks: A novel means of reducing catastrophic forgetting," *Neural Computation*, pp. 1755–1769, 2002.

[5] S. Golkar, M. Kagan, and K. Cho, "Continual learning via neural pruning," *ArXiv Preprint PrXiv:1903.04476*, 2019.

[6] D. Hendrycks, M. Mazeika, and T. G. Dietterich, "Deep anomaly detection with outlier exposure," in *ICLR*, 2019.

[7] S. Hou, X. Pan, C. C. Loy, Z. Wang, and D. Lin, "Learning a unified classifier incrementally via rebalancing," in *CVPR*, 2019, pp. 831–839.

[8] R. Kemker, M. McClure, A. Abitino, T. Hayes, and C. Kanan, "Measuring catastrophic forgetting in neural networks," in *AAAI*, 2018.

[9] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences*, pp. 3521–3526, 2017.

[10] A. Krizhevsky, "Learning multiple layers of features from tiny images," in *Technical Report*, 2009, p. 9.

[11] Z. Li and D. Hoiem, "Learning without forgetting," *TPAMI*, pp. 2935–2947, 2018.

[12] Z. Li, C. Zhong, S. Liu, R. Wang, and W.-S. Zheng, "Preserving earlier knowledge in continual learning with the help of all previous feature extractors," *ArXiv*, 2021.

[13] Z. Li, C. Zhong, R. Wang, and W.-S. Zheng, "Continual learning of new diseases with dual distillation and ensemble strategy," in *MICCAI*, 2020, pp. 169–178.

[14] V. Lomonaco and D. Maltoni, "Core50: a new dataset and benchmark for continuous object recognition," in *Conference on Robot Learning*, 2017, pp. 17–26.

[15] A.-A. Papadopoulos, M. R. Rajati, N. Shaikh, and J. Wang, "Outlier exposure with confidence control for out-of-distribution detection," *Neurocomputing*, pp. 138–150, 2021.

[16] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural networks*, pp. 54–71, 2019.

[17] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "icarl: Incremental classifier and representation learning," in *CVPR*, 2017, pp. 5533–5542.

[18] Robins, Anthony, "Catastrophic forgetting, rehearsal and pseudorehearsal," *Connection Science*, pp. 123–146, 1995.

[19] X. Sun, J. Yang, M. Sun, and K. Wang, "A benchmark for automatic visual classification of clinical skin disease images," in *ECCV*, 2016.

[20] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, 2018.

[21] Y. Wu, Y. Chen, L. Wang, Y. Ye, Z. Liu, Y. Guo, and Y. R. Fu, "Large scale incremental learning," in *CVPR*, 2019, pp. 374–382.

[22] S. Yan, J. Xie, and X. He, "Der: Dynamically expandable representation for class incremental learning," in *CVPR*, 2021.

[23] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni, "Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification," *ArXiv*, 2021.

[24] Y. Yang, Z. Cui, J. Xu, C. Zhong, R. Wang, and W.-S. Zheng, "Continual learning with bayesian model based on a fixed pre-trained feature extractor," in *MICCAI*, 2021, pp. 397–406.

[25] B. Zhao, X. Xiao, G. Gan, B. Zhang, and S. Xia, "Maintaining discrimination and fairness in class incremental learning," in *CVPR*, 2020, pp. 13 205–13 214.