Continual Learning of New Diseases with Dual Distillation and Ensemble Strategy

Zhuoyun Li[1,2⋆], Changhong Zhong[1,2⋆], Ruixuan Wang[1,2], and Wei-shi Zheng[1,2,3]

[1] School of Data and Computer Science, Sun Yat-sen University, China
[2] Key Laboratory of Machine Intelligence and Advanced Computing, MOE, Guangzhou, China
[3] Pazhou Lab, Guangzhou, China

Abstract. Most intelligent diagnosis systems are developed for one or a few specific diseases, while medical specialists can diagnose all diseases of certain organ or tissue. Since it is often difficult to collect data of all diseases, it would be desirable if an intelligent system can initially diagnose a few diseases, and then continually learn to diagnose more and more diseases with coming data of these new classes in the future. However, current intelligent systems are characterised by catastrophic forgetting of old knowledge when learning new classes. In this paper, we propose a new continual learning framework to alleviate this issue by simultaneously distilling both old knowledge and recently learned new knowledge and by ensembling the class-specific knowledge from the previous classifier and the learned new classifier. Experiments showed that the proposed method outperforms state-of-the-art methods on multiple medical and natural image datasets.

Keywords: Continual learning · Distillation · Ensemble.

1 Introduction

Deep learning has been a common tool for medical image analysis in recent years and achieved human-level performance in diagnosis of various diseases [2, 5, 18]. Most intelligent diagnosis systems would be fixed once developed and deployed. However, it is difficult to collect training data for all diseases of certain organ or tissue. As a result, every current intelligent system can diagnose just one or a few diseases, unable to diagnose all diseases as medical specialists do. Therefore, it would be desirable to enable an intelligent system to continually learn to diagnose more and more diseases, finally becoming a human-like specialist. Such a continual or lifelong learning process often presumes that the intelligent systems can access little or no old data of previously learned diseases due to various factors (e.g., privacy, data not shared across institutes). Unlike humans who can learn new knowledge without forgetting old knowledge, current intelligent

⋆ The authors contribute equally to this paper.

classification systems are characterised by catastrophic forgetting of old classes when learning new classes [7, 8, 13]. This makes it very challenging to develop an intelligent system which can continually learn to diagnose new diseases without sacrificing diagnosis performance on previously learned old diseases [3].

Typically, there always exists a trade-off between learning new knowledge and keeping previously acquired knowledge about old classes. To mitigate the forgetting of previously learned knowledge, one way is to find system units (e.g., kernels in convolutional neural networks) which are crucial for old knowledge, and then try to keep parameters of such units unchanged when learning new knowledge [6, 14, 15, 17]. Although this can help the system keep old knowledge to some extent, it would cause more and more difficulty in continually learning new knowledge, particularly when most kernels in the convolutional neural network (CNNs) become crucial for keeping old knowledge. To solve this dilemma, researchers tried to dynamically add new components to the existing system specifically for new knowledge [1, 12, 20, 24, 25]. However, this approach often assumes that the intelligent system can discriminate between old and new classes in advance, which is impractical in an intelligent diagnosis system where old and new diseases need to be diagnosed together. Inspired by the human learning process which often replays memory of old knowledge during learning new knowledge, another approach is to either store a small subset of original exemplars [4, 10, 11, 16, 19] or regenerate synthetic data [23, 26] for each old class, which has been shown to help reduce catastrophic forgetting more effectively in continual learning of new classes. However, synthetic data may not faithfully contain fine-grained discriminative features for each old class, particularly when applied to different diseases part of which may look similar to each other, and imbalanced small subset of exemplars for each old class compared to relatively large set of data for new classes often makes the intelligent system biased to recently learned (relatively new) class during prediction [22].

Here we propose a more effective framework to keep old knowledge during learning new classes, with a better way to simulate the idea of memory replay. In this framework, an *expert* classifier only for new classes is first trained by fine-tuning from the *original* classifier of old classes, and then the knowledge from both classifiers are distilled to teach a new classifier to recognize both old and new classes. The fine-tuning from the *original* classifier to the *expert* classifier would make the *expert* classifier largely contain old knowledge while learning new classes, therefore distillation from the *expert* classifier to the new classifier would additionally help keep knowledge of old classes. Besides the dual distillation, an ensemble strategy is applied to help keep old knowledge by combining the information of old classes extracted from both the *original* classifier and the new classifier. The proposed approach shows state-of-the-art performance in continual learning on two skin disease datasets and one natural image dataset. To our best knowledge, this is the first time to explore *continual learning* of intelligent diagnosis on relatively large number of (i.e., 40) diseases.

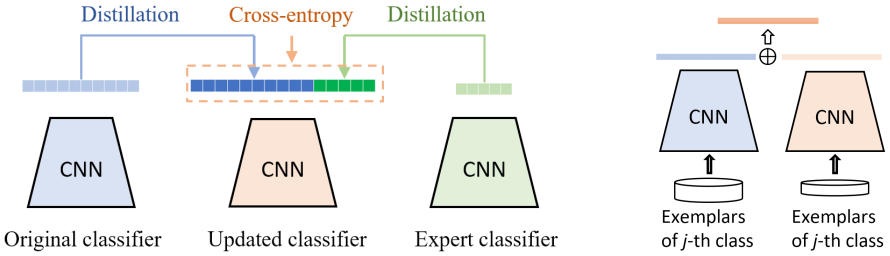

Fig. 1. The proposed continual learning framework. Left: distillation of old and new knowledge from the *original* classifier and the *expert* classifier to the *updated* classifier. Right: ensemble (denote by $\oplus$ above) of mean feature vectors for the j-th old class.

2 Continual Learning

The objective is to make an intelligent diagnosis system continually learn to diagnose new classes of diseases while keeping its diagnosis performance on old classes not reduced largely. One presumption of the continual learning problem is that most, if not all, training data of the old classes are not available when continually learning new classes. The key challenge is to reduce the catastrophic forgetting of old knowledge when continually learning new classes each time. Considering that keeping small subset of data for each old class in continual learning has been confirmed crucial to potentially reduce the catastrophic forgetting of old knowledge in previous studies [19], here we also assume that a small subset of data for each old class is available during continual learning.

2.1 Overview

In this study, we apply a dual distillation strategy to effectively reduce the catastrophic forgetting of old knowledge. The continual learning framework is demonstrated in Figure 1, with the diagnosis system for the old classes represented by the *original* CNN classifier (Figure 1, blue), and the system after continual learning of a set of new classes represented by the *updated* CNN classifier (Figure 1, orange). Different from existing approaches [4, 19], an *expert* CNN classifier composed of the feature extractor part of the *original* CNN classifier and a new fully connected layer is trained specifically for classification of the new classes (Figure 1, green), and then the knowledge from both the *original* classifier and the *expert* classifier is distilled to train the *updated* classifier (Figure 1, left). Since the *expert* classifier is fine-tuned from the *original* classifier, it is expected that knowledge of the old classes would be partly kept in the *expert* classifier, and therefore distillation from the *expert* classifier to the *updated* classifier would distill not only knowledge of new classes, but also partial knowledge of old classes, thus helping the *updated* classifier keep more knowledge of old classes compared to the only distillation from the *original* classifier to the *updated* classifier.

Besides the dual distillation, an ensemble strategy is also proposed to reduce the catastrophic forgetting of old knowledge (Figure 1, right). When training the *updated* classifier, the previously stored small subset of data for each old class and the relatively large set of data for each new class are collected together as the training set. In the testing phase, the *nearest-mean-of-exemplars* method [19] was adopted for category prediction due to its simplicity and effectiveness, where the mean is often calculated by averaging the feature vectors of the stored or selected subset of data in the feature space for each class (see Section 2.3). However, the imbalanced training set could lead to the bias toward the new classes when predicting the category of any test data by the *updated* classifier. To mitigate the prediction bias, for each old class, the mean feature vectors from the *original* classifier and the *updated* classifier were ensembled (i.e., averaged here) for prediction. Such ensemble strategy makes the mean feature vector of each old class deviated less from the mean feature vector previously generated with the *original* classifier, thus helping the knowledge representation of each old class changed less. Such ensemble was shown to reduce the catastrophic forgetting of old knowledge (Section 3.4). It is worth noting that the *updated* classifier would become the *original* classifier in next-round continual learning.

2.2 Dual distillation

The dual distillation is for the training of the *updated* classifier. Formally, let $D = \{(\mathbf{x}_i, \mathbf{y}_i), i = 1, \ldots, N\}$ denote the collection of previously stored small subset of data for each old class and the whole set of data for each new class, with $\mathbf{x}_i$ representing an image and the one-hot vector $\mathbf{y}_i$ representing the corresponding class label. For each image $\mathbf{x}_i$, denote by $\mathbf{z}_i = [z_{i1}, z_{i2}, \ldots, z_{is}]^\mathsf{T}$ the logit output (just before performing the softmax operation) of the *original* classifier, where s is the number of old classes. Similarly, denote by $\hat{\mathbf{z}}_i = [\hat{z}_{i1}, \hat{z}_{i2}, \ldots, \hat{z}_{i,s+t}]^\mathsf{T}$ the logit output of the *updated* classifier, where t is the number of new classes. Then, the distillation of old knowledge from the *original* classifier to the *updated* classifier can be realized by minimization of the distillation loss $\mathcal{L}_o$,

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

where $\boldsymbol{\theta}$ represents the to-be-learned parameters of the *updated* classifier, and p_{ij} and $\hat{p}_{ij}$ are from the modified softmax operation,

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

and $T_o \geq 1$ is the distillation parameter. Larger T_o will force the *updated* classifier to learn a more fine-grained separation between different feature vectors $\mathbf{z}_i$'s [9].

Similarly, the distillation of new knowledge from the *expert* classifier to the *updated* classifier can be realized by minimization of the distillation loss $\mathcal{L}_n$,

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

where q_{ij} and $\hat{q}_{ij}$ are respectively from the modified softmax over the logit of the *expert* classifier and the corresponding logit part of the *updated* classifier, with the distillation parameter T_n.

Besides the two distillation losses, the cross-entropy loss $\mathcal{L}_c$ is also applied to help the *updated* classifier discriminate both old and new classes,

$$\mathcal{L}_c(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{s+t} y_{ij}\log\hat{y}_{ij}\,, \tag{4}$$

where y_{ij} is the j-th component of the one-hot class label $\mathbf{y}_i$, and the $\hat{y}_{ij}$ is the j-th output of the *updated* classifier for the input image $\mathbf{x}_i$. In combination, the *updated* classifier can be trained by minimizing the overall loss $\mathcal{L}$,

$$\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}_c(\boldsymbol{\theta}) + \lambda_1\mathcal{L}_o(\boldsymbol{\theta}) + \lambda_2\mathcal{L}_n(\boldsymbol{\theta})\,, \tag{5}$$

where λ_1 and λ_2 is a coefficient constant balancing different loss terms.

2.3 Ensemble of class means

Considering that the majority of the training set is from new classes, and often only a small subset of data for each old class is stored, directly using output of the *updated* classifier to predict the category of any test data would be inevitably biased to new classes. As adopted in recent work [19], the nearest-mean-of-exemplars method is also used for category prediction of any test data in this study. The basic idea is to extract and average the feature vectors of all available data for each class based on the output of the penultimate layer of the *updated* classifier, and then any test data is categorized as the class whose mean vector is nearest to the test data in the feature space. That means, for each new class, the class mean is the average of feature vectors over the whole training set of this class based on the output of the penultimate layer of the *updated* classifier. However, for each old class, different from previous work, we propose an ensemble strategy for calculating the class mean. Specifically, suppose $\mathbf{m}_j$ is the previously obtained class mean for the j-th old class (in the preivous round of continual learning), and $\hat{\mathbf{m}}_j$ is the average of feature vectors over the stored small subset of data for the j-th class based on output of the penultimate layer of the well-trained *updated* classifier. Then the ensemble class mean of the j-th old class is the average of the two means, i.e., $(\mathbf{m}_j + \hat{\mathbf{m}}_j)/2$. Using the ensemble class mean instead of the updated individual mean $\hat{\mathbf{m}}_j$ for each old class is experimentally shown to improve the prediction performance for old classes.

Before going to the next round of continual learning, due to the limited memory to store exemplar training data for each class, a small subset of training data needs to be selected from the whole training set for each new class, and the previously stored subset of data for each old class needs to be further reduced. Suppose the memory size is K, then only $K/(s+t)$ exemplars would be stored in the memory for each of the $(s+t)$ classes. As in the iCaRL method, the herding selection strategy is applied to generate a sorted list of exemplars for

Table 1. Statistics of three datasets used in experiments.

Datasets	#Classes	Train set	Test set	New classes per time	Size
CIFAR100	100	50,000	10,000	5, 10, 20	32×32
Skin40	40	2,000	400	2, 5, 10	[420, 1640]
Skin8	8	3,555	705	2	[600, 1024]

each (new) class, with exemplars approximating the class mean having higher priorities [19]. Thus, the first $K/(s+t)$ exemplars in the sorted list would be selected and stored for each new class. For each old class, the first $K/(s+t)$ exemplars from the already shortened list would be kept, discarding the others in the memory.

3 Experimental Results

3.1 Experimental setup

Three datasets were used to extensively evaluate the proposed method (Table 1). Among them, the Skin8 dataset is from the classification challenge of dermoscopic images held by ISIC'2019 [21]. The original Skin8 is highly imbalanced between classes. To alleviate the potential effect of imballance in continual learning, 628 images were randomly selected from six classes and all images (fewer than 260) were kept for the other two smaller classes. The dataset Skin40 consists of 40 classes of skin disease images collected from the internet, with 50 training images and 10 test images for each class. In training, each image was randomly cropped within scale range [0.8,1.0] and then resized to 224×224 pixels.

On each dataset, an *original* CNN classifier was first trained for a small number (e.g., 2, 5) of classes, and then a set of (e.g., 2, 5) new classes' data were provided to train the *expert* classifier and the *updated* classifier, finishing the first-round continual learning. SGD optimizer (batch size 16) was used, with initial learning rate 0.01 and then divided by 10 at the 80th, and 160th epoch respectively. Weight decay (0.00001) and momentum (0.9) were also applied. Each model was trained for up to 200 epochs, with training convergence consistently observed. After training the *updated* classifier each time, 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 and standard deviation of accuracy over five runs were reported, each run with a different order of classes to be learned. Unless otherwise mentioned, ResNet-18 was used as backbone, memory size $K = 50$, parameters $T_o = T_n = 4.0$ on Skin40, $T_o = T_n = 2.0$ on Skin8, and $\lambda_1 = 3.0$, $\lambda_2 = 1.0$.

3.2 Performance on continual learning of medical image classes

Our method was first evaluated on skin datasets by comparing with state-of-the-art baselines, including iCaRL [19], End-to-End Incremental Learning

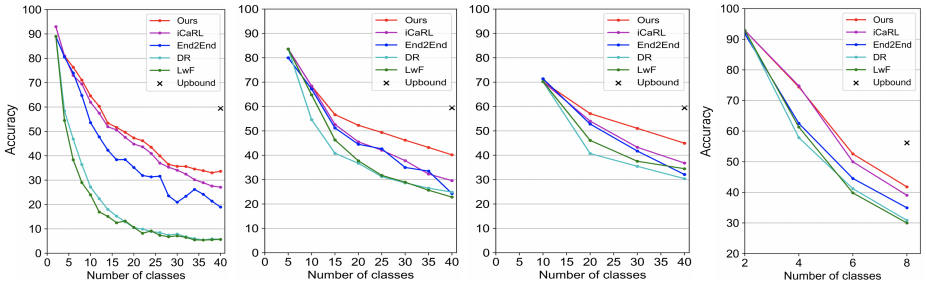

Fig. 2. Performance comparison between the proposed method and baselines on Skin40 (first three) and Skin8 (last). Standard deviation ([0.7, 3.8]) omitted for clearer views.

Table 2. Performance comparison between the proposed method and baselines with different CNN backbones. The average and standard deviation of accuracy at the end of continual learning of 5 classes every time over five runs were reported.

Backbones	ResNet18				AlexNet				VGG19			
Methods	iCaRL	E2E	DR	Ours	iCaRL	E2E	DR	Ours	iCaRL	E2E	DR	Ours
Accuracy	29.6 ±2.1	24.3 ±1.4	24.9 ±3.8	**40.2 ±1.6**	29.3 ±0.4	20.8 ±1.6	20.6 ±3.9	**33.0 ±0.8**	25.5 ±2.5	19.3 ±0.8	20.3 ±3.8	**28.8 ±1.7**

(End2End) [4], Distillation and Retrospection (DR) [10] and LwF [16]. Similar amount of effort was put into tuning each baseline method. The same memory size was used for all methods except the LwF which does not store old exemplars. In addition, an upper-bound result was also reported (Figure 2, **black** cross) by training a non-continual classifier with all classes of training data. Figure 2 shows that our method outperforms all the strong baselines when the classifier continually learn either 2, 5, or 10 classes every time on Skin40 (first three subfigures), and 2 classes every time on Skin8 (last one), all confirming the effectiveness of our method. Smaller improvement on Skin8 than on Skin40 may be due to fewer classes and fewer rounds of continual learning on Skin8.

3.3 Robustness and generalization of continual learning

To show the robustness of our method, multiple CNN backbones were used for continual learning of skin diseases. As shown in Tabel. 2, our method consistently outperforms the strong baselines with all three CNN backbones, no matter whether the classifier continually learn 2, 5, or 10 classes each time on Skin40 (only showing 5-class case). Besides the CNN backbones, different memory sizes ($K = 100, 200$) were also evaluated (not shown due to limited space), all consistently showing our method is superior to the strong baselines.

Considering that the baselines were evaluated only on natural image datasets in previous studies, our method was also evaluated on CIFAR100. Consistently, our method showed better performance in continual learning of 5, 10, or 20

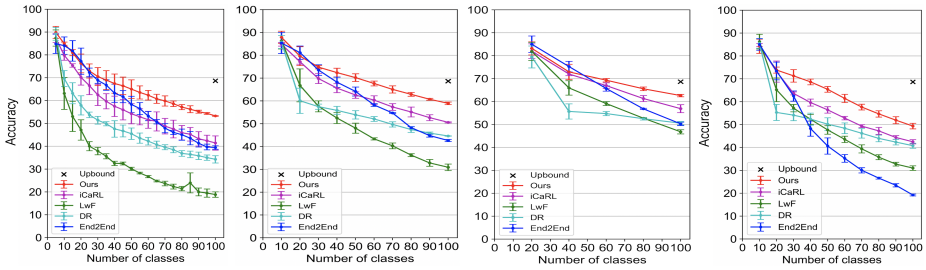

Fig. 3. Performance of continual learning on the the CIFAR100 dataset. From left to right: continual learning of 5, 10, and 20 new classes each time with memory size $K = 2000$, and the continual learning of 10 new classes each time with $K = 500$.

Table 3. Ablation study on CIFAR100. 'Original+CE': training the *updated* classifier with the cross-entropy loss and the single distillation from the *original* classifier. Final-round learning performance was reported, with 10 new classes learned each time.

K	Original+CE (OCE)	OCE+ expert(random)	OCE+ expert(finetune)	OCE+expert(finetune) +ensemble
2000	55.7	55.4	57.9	**59.2**
500	44.2	45.6	47.4	**49.3**
100	27.8	31.9	32.7	**34.8**

new classes each time (Figure 3), supporting that our method is generalizable to various domains. One interesting observation is that the performance of our method is similar to that of iCaRL at first few rounds of continual learning, and then becomes better at subsequent rounds. This might be because the stored old data is sufficient in first few rounds to help the classifier reduce catastrophic forgetting of old classes, leading to the negligible effect of the *expert* classifier and the ensemble strategy in our method. This is confirmed in Figure 3 (right), where our method outperforms iCaRL from earlier rounds of continual learning when memory size is smaller ($K = 500$ here versus $K = 2000$ in 2nd subfigure).

3.4 Ablation Study

This section evaluates the role of the *expert* classifier and the ensemble strategy in continual learning. Table 3 shows that although the *expert* classifier trained from scratch each time improve learning performance little (column 3 vs. column 2), the *expert* classifier fine-tuned from the *original* classifier clearly improves more (column 4). Adding the ensemble strategy further improve performance (column 5). It is worth noting the ensemble strategy would help more if memory size becomes smaller (columns 4-5, $K = 2000$ vs. 500 vs. 100).

4 Conclusion

In this study, we proposed a new continual learning framework to help intelligent systems reduce catastrophic forgetting of old knowledge during learning new classes. Experiments showed that dual distillation from two teachers can help teach a new classifier more effectively during continual learning of new diseases, and the ensemble prediction from old and new classifiers further alleviate forgetting old knowledge.

Acknowledgement

This work is supported in part by the National Key Research and Development Program (grant No. 2018YFC1315402), the Guangdong Key Research and Development Program (grant No. 2019B020228001), the National Natural Science Foundation of China (grant No. U1811461), and the Guangzhou Science and Technology Program (grant No. 201904010260).

References

1. Aljundi, R., Chakravarty, P., Tuytelaars, T.: Expert gate: Lifelong learning with a network of experts. In: Conference on Computer Vision and Pattern Recognition. pp. 3366–3375 (2017)
2. Ardila, D., Kiraly, A.P., Bharadwaj, S., Choi, B., Reicher, J.J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., et al.: End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature medicine **25**(6), 954–961 (2019)
3. Baweja, C., Glocker, B., Kamnitsas, K.: Towards continual learning in medical imaging. arXiv preprint arXiv:1811.02496 (2018)
4. Castro, F.M., Marín-Jiménez, M.J., Guil, N., Schmid, C., Alahari, K.: End-to-end incremental learning. In: European Conference on Computer Vision. pp. 233–248 (2018)
5. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115–118 (2017)
6. Fernando, C., Banarse, D., Blundell, C., Zwols, Y., Ha, D., Rusu, A.A., Pritzel, A., Wierstra, D.: Pathnet: Evolution channels gradient descent in super neural networks. arXiv preprint arXiv:1701.08734 (2017)
7. French, R.M.: Catastrophic forgetting in connectionist networks. Trends in Cognitive Sciences **3**(4), 128–135 (1999)
8. Goodfellow, I.J., Mirza, M., Da Xiao, A.C., Bengio, Y.: An empirical investigation of catastrophic forgeting in gradientbased neural networks. In: International Conference on Learning Representations (2014)
9. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
10. Hou, S., Pan, X., Change Loy, C., Wang, Z., Lin, D.: Lifelong learning via progressive distillation and retrospection. In: European Conference on Computer Vision. pp. 437–452 (2018)

11. Isele, D., Cosgun, A.: Selective experience replay for lifelong learning. In: AAAI Conference on Artificial Intelligence (2018)
12. Karani, N., Chaitanya, K., Baumgartner, C., Konukoglu, E.: A lifelong learning approach to brain MR segmentation across scanners and protocols. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 476–484 (2018)
13. Kemker, R., McClure, M., Abitino, A., Hayes, T.L., Kanan, C.: Measuring catastrophic forgetting in neural networks. In: AAAI Conference on Artificial Intelligence (2018)
14. Kim, H.E., Kim, S., Lee, J.: Keep and learn: Continual learning by constraining the latent space for knowledge preservation in neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 520–528 (2018)
15. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. National Academy of Sciences **114**(13), 3521–3526 (2017)
16. Li, Z., Hoiem, D.: Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(12), 2935–2947 (2017)
17. Mallya, A., Lazebnik, S.: Packnet: Adding multiple tasks to a single network by iterative pruning. In: Conference on Computer Vision and Pattern Recognition. pp. 7765–7773 (2018)
18. McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.C., Darzi, A., et al.: International evaluation of an AI system for breast cancer screening. Nature **577**(7788), 89–94 (2020)
19. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: Conference on Computer Vision and Pattern Recognition. pp. 2001–2010 (2017)
20. Rusu, A.A., Rabinowitz, N.C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., Hadsell, R.: Progressive neural networks. arXiv preprint arXiv:1606.04671 (2016)
21. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**, 180161 (2018)
22. Wu, Y., Chen, Y., Wang, L., Ye, Y., Liu, Z., Guo, Y., Fu, Y.: Large scale incremental learning. In: Conference on Computer Vision and Pattern Recognition. pp. 374–382 (2019)
23. Xiang, Y., Fu, Y., Ji, P., Huang, H.: Incremental learning using conditional adversarial networks. In: International Conference on Computer Vision. pp. 6619–6628 (2019)
24. Xu, J., Zhu, Z.: Reinforced continual learning. In: Advances in Neural Information Processing Systems. pp. 899–908 (2018)
25. Yoon, J., Yang, E., Lee, J., Hwang, S.J.: Lifelong learning with dynamically expandable networks. In: International Conference on Learning Representations (2018)
26. Zhai, M., Chen, L., Tung, F., He, J., Nawhal, M., Mori, G.: Lifelong gan: Continual learning for conditional image generation. In: International Conference on Computer Vision. pp. 2759–2768 (2019)