Continual Learning with Bayesian Model Based on a Fixed Pre-trained Feature Extractor

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

[1] School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2] Key Laboratory of Machine Intelligence and Advanced Computing, MOE, China
wangruix5@mail.sysu.edu.cn
[3] Pazhou Lab, Guangzhou, China

Abstract. Current deep learning models are characterised by catastrophic forgetting of old knowledge when learning new classes. This poses a challenge in intelligent diagnosis systems where initially only training data of a limited number of diseases are available. In this case, updating the intelligent system with data of new diseases would inevitably downgrade its performance on previously learned diseases. Inspired by the process of learning new knowledge in human brains, we propose a Bayesian generative model for continual learning built on a fixed pre-trained feature extractor. In this model, knowledge of each old class can be compactly represented by a collection of statistical distributions, e.g. with Gaussian mixture models, and naturally kept from forgetting in continual learning. Experiments on two skin image sets showed that the proposed approach outperforms state-of-the-art approaches which even keep some images of old classes during continual learning of new classes.

Keywords: Continual learning · Bayesian model · Generative approach

1 Introduction

Deep learning models, particularly convolutional neural networks (CNNs), have shown human-level performance in diagnosis of various diseases [1,6,12]. However, most intelligent diagnosis systems are limited to diagnosis of only one or a few diseases and cannot be easily extended once deployed, therefore cannot diagnose all diseases of certain tissue or organ (e.g., skin or lung) as medical specialists do. Since collecting data of all (e.g., skin or lung) diseases is challenging due to various reasons (e.g., privacy and limited data sharing), it is impractical to train an intelligent system diagnosing all diseases all at once. One possible solution is to make the intelligent system have the continual or lifelong learning ability, such that it can continually learn to diagnose more and more diseases

M. de Bruijne et al. (Eds.): MICCAI 2021, LNCS 12905, pp. 397–406, 2021.
https://doi.org/10.1007/978-3-030-87240-3_38

without resourcing (or resourcing few) original data of previously learned diseases [3]. However, current intelligent models are characterised by catastrophic forgetting of old knowledge when learning new classes.

Researchers have recently tried to reduce the catastrophic forgetting issue mainly in the natural image domain. One approach is to find model components (e.g., kernels in CNNs) crucial for old knowledge, and then try to change them as little as possible when learning new knowledge [9]. However, it would become increasingly more difficult to continually learn new knowledge because more and more kernels in CNNs become crucial for increasingly old knowledge. To make models more flexibly learn new knowledge, another approach tries to modify model structures by add new layers or kernels when learning new knowledge [16]. Knowledge distillation has also been widely used during learning new classes [4,7,8,10,11,14], where the old knowledge is implicitly represented by soft outputs of the old classifier with stored small old images or new classes of images as inputs. An alternative approach is to train a generative model to produce enough number of synthetic training data for each old class when learning new classes [19,20].

However, almost all existing approaches modify the feature extraction part of the classifiers either in parameter values or in structures during continual learning of new classes. In contrast, humans seem to learn new knowledge by adding memory of the learned new information without modifying the (e.g., visual) perceptual pathway. Therefore, one possible cause to catastrophic forgetting in existing models is the change in the feature extraction part (corresponding to the perceptual pathway in human brains) when learning new knowledge. With this consideration, we propose a generative model for continual learning built on a fixed pre-trained feature extractor, which is different from all existing (discriminative) models. The generative model can naturally keep knowledge of each old class from forgetting, without storing original images of old classes or regenerating synthetic images during continual learning. Experiments on two skin disease classification tasks showed the proposed approach outperforms state-of-the-art approaches which even keep some images of old classes during continual learning.

2 A Generative Model for Continual Learning

The proposed method is inspired by two interesting findings in neuroscience. One finding is that most infants cannot form episodic memory before 3 years old [2,17], and the other finding is that humans continually form memory from infants to elderly people [13]. One hypothetical explanation is that the visual pathway in younger infant's brain might be rapidly changing with daily visual stimuli from surroundings and then become firm with little change since 3 years old or so. Humans can continually learn new visual knowledge through their whole lives probably because they formed new memories about the new knowledge, but without changing the visual pathway which works as a visual feature extractor. This could explain why current deep learning models are characterised by catastrophic forgetting of old knowledge, i.e., model parameters or model structures from the the feature extractor part are always changed to

some extent in almost all continual learning approaches. With this consideration, we propose a human-like continual learning framework, i.e., first pre-training a feature extractor, then fixing the feature extractor and forming new memory for each new knowledge. In the following, we will introduce one general way to pre-train the feature extractor, one statistical method to represent the memory, and one Bayesian model to predict class of any new (test) data after continual learning each time.

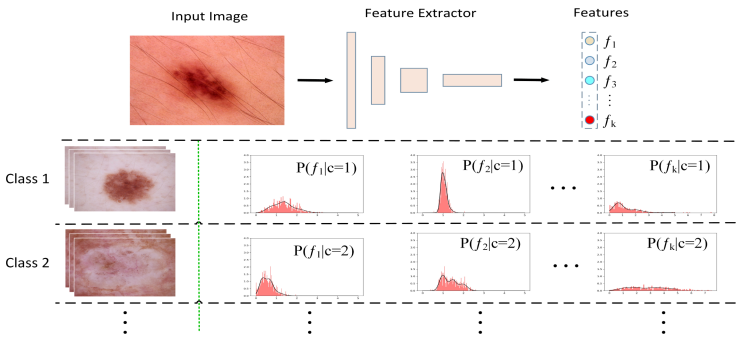

Fig. 1. Fixed pre-trained feature extractor (top) and memory formation (middle to bottom). Feature extractor is pre-trained and fixed during continual learning. Memory of each class is represented by a set of statistical distributions over features.

2.1 Fixed Pre-trained Feature Extractor

An ideal feature extractor should output two different feature vectors if two input data were visually different, meanwhile visually more similar inputs should result in more similar feature vectors from the feature extractor. The visual feature extractor (i.e., visual pathway) in younger infants are probably taught in certain self-supervised way, although the mechanism of self-supervision in infant brain has not been explicitly understood [15]. While it is worth exploring various self-supervised learning approaches (e.g., auto-encoder) to train a feature extractor, here we leave the self-supervision exploration for future work, and adopt a simpler but widely used approach, i.e., pre-training a CNN classifier with relatively large number images whose classes or domains are relevant but different from those in the task of interest, and then using the pre-trained CNN feature extractor (often consisting of all the convolutoinal layers) for the continual learning classification task of interest (Fig. 1, top row). It is expected that the pre-trained feature extractor would probably be powerful enough to discriminate different input images from the task of interest. Experiments showed that even such a simple approach to a fixed pre-trained feature extractor can already significantly help reduce catastrophic forgetting of old knowledge with the proposed generative approach. It is worth noting that, during continual learning of

new classes in the classification task of interest, the pre-trained feature extractor is fixed and not updated. The knowledge of each learned new class is represented and stored as described in the following subsection.

2.2 Memory Formation

Different from the state-of-the-art continual learning approaches which often store a small number of original images for each old class, the proposed approach stores not original images but the statistical information of each class based on the feature extractor outputs of all training images belonging to the class. Here, each element of the output feature vector is assumed to represent certain type of visual feature. Then based on the class of training images, the distribution of each feature is estimated and collected together to form the memory of the knowledge of the specific class (Fig. 1, second to bottom rows). Formally, denote by $D_c = \{\mathbf{x}_i, i = 1, \ldots, N_c\}$ the set of training images for class c, $\mathbf{z}_i = [z_{i1}, z_{i2}, \ldots, z_{ik}, \ldots, z_{iK}]^\mathsf{T}$ the L_2-normalized output feature vector from the feature extractor for the input image $\mathbf{x}_i$, and $\mathbf{f} = [f_1, f_2, \ldots, f_k, \ldots, f_K]^\mathsf{T}$ the vector of random variables representing the output of the feature extractor, then the statistical distribution of the k-th feature f_k for class c can be represented by a probability density distribution $p(f_k|c, D_c)$,

$$p(f_k|c, D_c) = g(\{z_{ik}, i = 1, \ldots, N_c\}), \quad \forall\, k \in \{1, \ldots, K\} \tag{1}$$

where $g(\cdot)$ could be any appropriate distribution estimator. Here a Gaussian mixture model (GMM) with a small number of S components is adopted to represent $g(\cdot)$ for its simplicity. Since each Gaussian component can be compactly represented by its mean and standard deviation, totally only $2 \cdot S \cdot K$ numbers are stored in the memory to represent the knowledge of each class. D_c would be omitted from $p(f_k|c, D_c)$ in the following for simplicity.

2.3 Bayesian Model for Prediction

Based on the statistical distributions of visual features for each class, we propose a generative classification model based on the Bayesian rule for prediction. Given a test image $\mathbf{x}_j$, denote by $\mathbf{z}_j = [z_{j1}, z_{j2}, \ldots, z_{jk}, \ldots, z_{jK}]^\mathsf{T}$ the corresponding output from the feature extractor, and $p(c|\mathbf{z}_j)$ the probability of the test image belonging to class c. Then based on the Bayes rule, we can get

$$p(c|\mathbf{z}_j) = \frac{p(\mathbf{z}_j|c) \cdot p(c)}{\sum_{m=1}^{M} p(\mathbf{z}_j|m) \cdot p(m)}, \tag{2}$$

where M is the number of classes learned so far. Considering that potential correlations between certain feature components are probably caused by co-occurred visual parts of a specific class of objects, it can be assumed that different feature components f_k's are conditionally independent given specific class c. Then, the logarithm of Eq. 2 gives

$$\log p(c|\mathbf{z}_j) = \sum_k \log p(f_k = z_{jk}|c) + \log p(c) - \alpha, \tag{3}$$

where $\alpha = \log \sum_m p(\mathbf{z}_j|m)p(m)$ can be considered a constant for different classes. In Eq. (3), the likelihood function value $p(f_k = z_{jk}|c)$ for each feature element k can be directly obtained based on the previously stored knowledge $p(f_k|c)$ (Eq. 1) in the memory. The prior $p(c)$ for class c can be simply estimated based on the ratio of the number of training images for this class over the total number of training images of all learned classes so far, i.e., $p(c) = N_c / \sum_m N_m$. Note that in this case, the number of training images for each class needs to be stored in the memory such that $p(c)$ can be easily updated when new classes' knowledge is learned as above (Eq. 1). Based on Eq. (3), the class of the test image $\mathbf{x}_j$ would be directly predicted as the one with the highest value of $\log p(c|\mathbf{z}_j)$ over all classes learned so far.

The advantages of the the proposed approach over existing continual learning approaches are obvious. First, the knowledge of each old class is statistically represented by the set of likelihood functions (Eq. 1) and compactly stored in the memory. Therefore, old knowledge will not be forgotten over continual learning of new classes. In comparison, old knowledge will be inevitably and gradually forgotten over multiple rounds of continual learning in existing approaches, either due to the changes in feature extractor or due to the reduced number of original images to be stored in the limited memory. Second, the final performance of the proposed approach over multiple rounds of continual learning is not affected by the number of learning rounds and the number of new classes added in each round. In contrast in existing approaches, more rounds of continual learning with smaller number of new classes added each time would often lead to worse classification performance. Therefore, the proposed approach is more robust to various learning conditions with little forgetting of old knowledge.

3 Experimental Evaluations

3.1 Experimental Setup

The proposed approach was extensively evaluated on two medical skin image datasets. Skin7 [5] is a skin lesion dataset from the challenge of dermoscopic image classification held by the International Skin Imaging Collaboration (ISIC) in 2018. It consists of 7 disease categories, and each image is of size 600×450 pixels. This dataset presents severe class imbalance, with the largest class 60 times larger than the smallest one. Skin40 is a subset of 193 classes of skin disease images collected from the internet [18]. Skin40 contains two types of images, dermoscopic images which have relatively consistent imaging conditions (e.g., similar illumination) and therefore low levels of imaging noise, and clinical images captured mostly with digital cameras or mobile phones. The 40 classes with relatively more number of images (60 images per class) were chosen from the 193 classes to form the Skin40 dataset, while the remaining 153 classes (10 to 40 image per class) were used to train a CNN classifier whose final classification layer was then removed to form the fixed feature extractor for our approach or to be used as tunable feature extractor for baseline methods in most experiments. It is worth noting that there is no overlap between the 153 classes (for training the

feature extractor in advance) and the classes in Skin7 and Skin40 (for continual learning evaluation).

During feature extractor training based on the 153 skin image classes, each image was randomly cropped within the scale range [0.8, 1.0] and then resized to 224×224 pixels, followed by random horizontal and vertical flipping. The mini-batch stochastic gradient descent (batch size 32) was used, with initial learning rate 0.01 and then divided by 10 at the 35th, 70th, and 105th epoch respectively. Weight decay (0.0005) and momentum (0.9) were also applied. The feature extractor was trained for 120 epochs with observed convergence.

In each experiment, multiple rounds of continual learning were performed, with a few (e.g., 2, 5) new classes learned each time. After each round of continual learning, the mean class recall (MCR) over all classes learned so far was calculated. The mean and standard deviation of MCR over five runs were reported, where the five orders of classes to be continually learned were fixed and used in all methods. Unless otherwise mentioned, ResNet-101 was used as the backbone for the feature extractor, and the dimension of feature vector K was 2048 and the number of Gaussian components in each GMM model was empirically set to 2 based on a small validation set for each dataset.

3.2 Effectiveness of the Generative Model

This section evaluates the effectiveness of the proposed approach by comparing with state-of-the-art strong baselines, including iCaRL [14], End-to-End Incremental Learning (End2End) [4], learning a Unified Classifier Incrementally via Rebalancing (UCIR) [8], Distillation and Retrospection (DR) [7], and Learning without Forgetting (LwF) [10]. The suggested hyper-parameter settings in the original work were adopted. In each round of continual learning, for the iCaRL, End2End, DR, and UCIR which need certain number of old data, the number of images stored (i.e., memory size) for all old classes is respectively 50 for Skin7 and 100 for Skin40. The memory size was chosen such that stored number of images for each class was only a small portion of the original training images at the last round of learning. An upper-bound result was also reported (Fig. 2, green star) by training a non-continual classifier with all training data.

Figure 2 shows that, with certain number of new classes to be continually learned at each round, the proposed approach always performs better than all the strong baselines particularly at later round of continual learning, although the same pre-trained feature extractor was used to fine-tune the CNN classifier for each baseline method. Even with more images of old classes stored for the representative strong baseline iCaRL, the proposed approach still performs better (Fig. 2, second row, last). The results also tell us the final-round performance of the proposed approach is not affected by the number of new classes to be learned each time. In comparison, the final performance of each baseline becomes worse with smaller number of new classes to be learned each time. These results clearly support that the proposed approach is more effective in keeping old knowledge from forgetting.

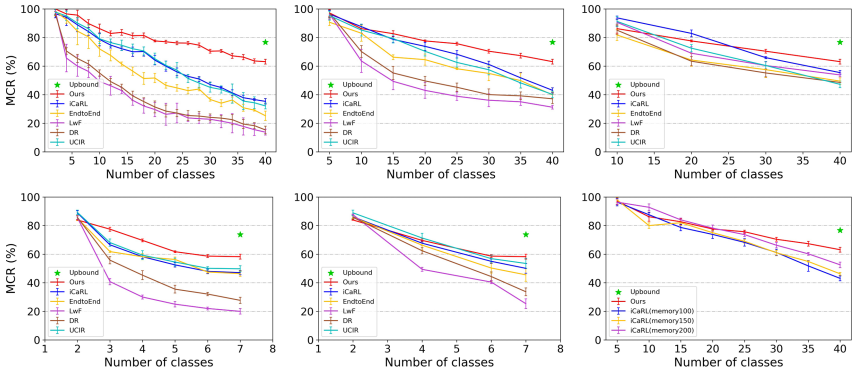

Fig. 2. Performance comparison on Skin40 and Skin7. First row (from left): learning 2, 5, 10 classes each time on Skin40. Second row (from left): learning 1 and 2 classes each time on Skin7, and comparison with iCaRL with varying memory size on Skin40. X-axis in each sub-figure represents the accumulated number of learned classes in the corresponding continual learning task.

Table 1. Performance on various feature extractor backbones. Results after last round of continual learning were reported, with 1 (Skin7) or 5 (Skin40) new classes per round.

Dataset	VGG19				ResNet18				ResNet34				ResNet101			
	LwF	iCaRL	IR	Ours	LwF	iCaRL	IR	Ours	LwF	iCaRL	IR	Ours	LwF	iCaRL	IR	Ours
Skin7	18.9	39.7	38.3	**46.5**	19.8	44.3	46.2	**55.6**	20.1	46.9	48.3	**56.8**	21.0	48.5	50.0	**58.4**
Skin40	27.4	33.6	32.5	**52.8**	30.4	41.8	37.1	**61.9**	31.1	42.3	39.5	**62.8**	31.2	43.1	40.2	**63.1**

3.3 Generalizability and Robustness of the Generative Model

The proposed approach is a general framework and therefore can employ different feature extractor backbones. As Table 1 shows, the proposed approach performs consistently better than strong baselines with different feature extractor backbones, supporting that the proposed approach is not limited to specific feature extractor structures. To evaluate the robustness of the generative model, the GMM with varying number of Gaussian components and different orders of classes to be continually learned were tried during continual learning. Figure 3 clearly shows that the generative model works stably well with different number of Gaussian components in GMM and with different class orders, with little change in performance. In comparison, the performance of the representative iCaRL varied a lot with different class orders. This is because knowledge of each old class is compactly stored and not changed throughout the whole process of continual learning by our approach. In comparison, all the strong baselines modify the feature extractor during continual learning, which would then change the representation of each stored old data and further change the representation of old knowledge, differently with different orders of classes to be learned.

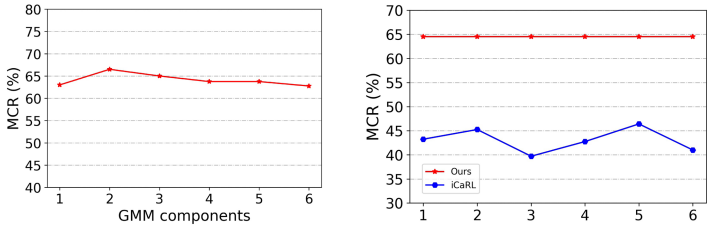

Fig. 3. Robustness of the proposed approach. Left: stable performance with varying GMM components on Skin40. Right: final-round learning performance with different class orders (x-axis indices for different sets of class orders on Skin40). Five new classes were continually learned per round.

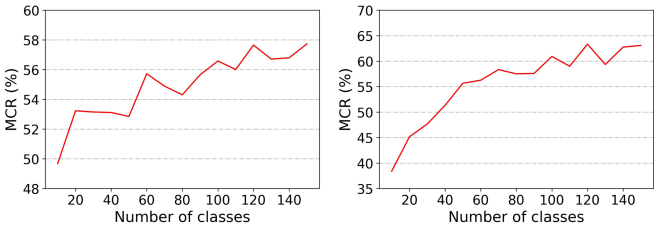

Fig. 4. Effect of feature extractor on continual learning. More classes (x-axis) used to train feature extractor result in better performance on Skin7 (left) and Skin40 (right).

3.4 Effect of Feature Extractor

The proposed approach is based on a fixed pre-trained feature extractor. To confirm that better feature extractors would help the generative model perform better in continual learning, the original 153 classes of skin images used for training the feature extractor were reduced gradually to only 10 classes, each time using such reduced number of classes to train the feature extractor and then the performance of the proposed approach at last round of continual learning on both the Skin7 and Skin40 datasets was calculated. Figure 4 does show that more classes used for training the feature extractor would generally result in better performance of the proposed approach. The feature extractor trained by more classes of data would probably have learned to extract more types of features and therefore could be more generalizable to a new but relevant domain. Consistent with the observation and explanation, when the feature extractor is fixed by random parameter weights (i.e., without any training), the classifier in continual learning showed the worst performance (MCR is 21% on Skin7, 6% on Skin40; not shown in Fig. 4). These results strongly suggest that exploring better ways to obtain a better feature extractor would further improve performance of the generative model in continual learning.

4 Conclusion

In this study, we propose a Bayesian generative model for continual learning of new classes. The model does not update the feature extractor but generates statistical information to represent knowledge of each class. Without storing any original data, the generative model can keep knowledge of each old class from forgetting and outperforms existing state-of-the-art approaches which often store small number of old data. The model is not limited to any specific feature extractor, and the final-round performance is not affected by the process of continual learning such as the number of new classes to be learned each time or the number of rounds of continual learning. Better pre-trained feature extractor could be explored to further improve the performance of the generative approach.

Acknowledgement. This work is supported in part by the National Natural Science Foundation of China (grant No. 62071502, U1811461), the Guangdong Key Research and Development Program (grant No. 2020B1111190001, 2019B020228001), and the Meizhou Science and Technology Program (grant No. 2019A0102005).

References

1. Ardila, D., et al.: End-to-End lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature Medicine (2019)
2. Bauer, P.J.: A complementary processes account of the development of childhood amnesia and a personal past. Psychological Review (2015)
3. Baweja, C., Glocker, B., Kamnitsas, K.: Towards continual learning in medical imaging. In: NIPS Workshop (2018)
4. Castro, F.M., Marín-Jiménez, M.J., Guil, N., Schmid, C., Alahari, K.: End-to-End incremental learning. In: ECCV (2018)
5. Codella, N.C.F., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). CoRR abs/1902.03368 (2019). http://arxiv.org/abs/1902.03368
6. De Fauw, J., et al.: Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature Medicine (2018)
7. Hou, S., Pan, X., Change Loy, C., Wang, Z., Lin, D.: Lifelong learning via progressive distillation and retrospection. In: ECCV (2018)
8. Hou, S., Pan, X., Loy, C.C., Wang, Z., Lin, D.: Learning a unified classifier incrementally via rebalancing. In: CVPR (2019)
9. Kirkpatrick, J., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences (2017)
10. Li, Z., Hoiem, D.: Learning without forgetting. IEEE Trans. Pattern Anal. Mach. Intell. **40**, 2935–2947 (2017)
11. Li, Z., Zhong, C., Wang, R., Zheng, W.-S.: Continual learning of new diseases with dual distillation and ensemble strategy. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 169–178. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_17

12. McKinney, S.M., et al.: International evaluation of an ai system for breast cancer screening. Nature (2020)
13. Nadel, L., Hupbach, A., Gomez, R., Newman-Smith, K.: Memory formation, consolidation and transformation. Neuroscience & Biobehavioral Reviews (2012)
14. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: iCaRL: Incremental classifier and representation learning. In: CVPR (2017)
15. Ribordy, F., Jabès, A., Lavenex, P.B., Lavenex, P.: Development of allocentric spatial memory abilities in children from 18 months to 5 years of age. Cognitive Psychology (2013)
16. Rusu, A.A., et al.: Progressive neural networks. arXiv preprint arXiv:1606.04671 (2016)
17. Scarf, D., Gross, J., Colombo, M., Hayne, H.: To have and to hold: Episodic memory in 3-and 4-year-old children. Developmental Psychobiology (2013)
18. Sun, X., Yang, J., Sun, M., Wang, K.: A benchmark for automatic visual classification of clinical skin disease images. In: ECCV (2016)
19. Xiang, Y., Fu, Y., Ji, P., Huang, H.: Incremental learning using conditional adversarial networks. In: ICCV (2019)
20. Zhai, M., Chen, L., Tung, F., He, J., Nawhal, M., Mori, G.: Lifelong GAN: Continual learning for conditional image generation. In: ICCV (2019)