ADINet: Attribute driven incremental network for retinal image classification

Qier Meng
Research Center for Medical Bigdata,
National Institute of Informatics
qiermeng@nii.ac.jp

Satoh Shin'ichi
Research Center for Medical Bigdata,
National Institute of Informatics
satoh@nii.ac.jp

Abstract

Retinal diseases encompass a variety of types, including different diseases and severity levels. Training a model with all possible types of disease is impractical. Dynamically training a model is necessary when a patient with a new disease appears. Deep learning techniques have stood out in recent years, but they suffer from catastrophic forgetting, i.e., a dramatic decrease in performance when new training classes appear. We found that keeping the feature distribution of a teacher model helps maintain the performance of incremental learning. In this paper, we design a framework named "Attribute Driven Incremental Network" (ADINet), a new architecture that integrates class label prediction and attribute prediction into an incremental learning framework to boost the classification performance. With image-level classification, we apply knowledge distillation (KD) to retain the knowledge of base classes. With attribute prediction, we calculate the weight of each attribute of an image and use these weights for more precise attribute prediction. We designed attribute distillation (AD) loss to retain the information of base class attributes as new classes appear. This incremental learning can be performed multiple times with a moderate drop in performance. The results of an experiment on our private retinal fundus image dataset demonstrate that our proposed method outperforms existing state-of-the-art methods. For demonstrating the generalization of our proposed method, we test it on the ImageNet-150K-sub dataset and show good performance.

1. Introduction

Retinal diseases encompass a variety of types, including different diseases and severities. Usually, there are several decades of types of retinal diseases [29]. As the diseases change in different stages, it is difficult to collect all of the disease types at the same time to train a model, especially in the case of some usual disease like retinal vein occlusion or central serous chorioretinopathy. Dynamically training a model is necessary when a patient with a new disease appears. In the medical imaging application, there is an increasing demand for systems that can implement incremental learning over a series of tasks. Because it is difficult to obtain all of the old dataset due to the privacy of the disease dataset. Deep convolutional networks have achieved great performance in classification tasks in computer vision. However, the incremental learning paradigm still suffers from catastrophic forgetting, i.e., a performance decrease for base classes when datasets for new classes appear [27]. In real-world object classification, systems must be continuously upgraded by examining new knowledge. However, retraining a modal always with old data and new data together is impractical [30]. When a new type of data appears, a natural way of performing incremental learning is to fine-tune a pre-trained model on new data. Figure 1 shows the basic idea of our proposed method and example images of a retinal disease dataset.

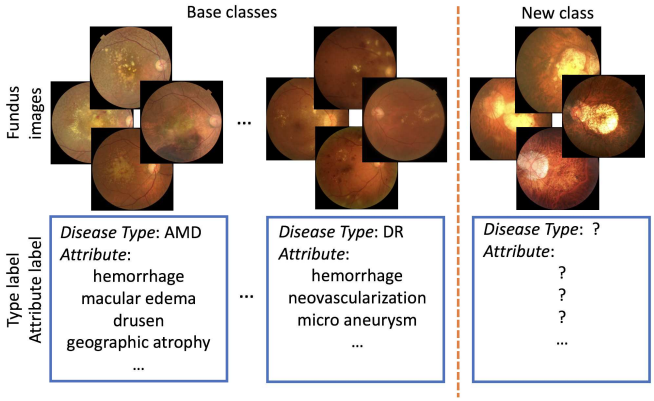


Figure 1. Examples of several retinal diseases. Figure illustrates our basic idea. We study images with image-level labels and corresponding attributes of base classes. We predict labels and attributes of new classes with teacher model.

Visual attributes are an important research area in computer vision because they can be a powerful mid-level representation that can bridge low-level features and high-level human recognition. Attributes used for mid-level representation have been investigated in a variety of computer vision tasks, including recognition, classification, and retrieval, for

many years. In fundus imaging, attributes are summarized from the case histories of the disease symptoms of each patient. For image classification, we found that good attribute prediction helps in boosting classification performance.

Given the scenarios described above, there are two straightforward solutions to learning new classes without forgetting the base classes: (1) preserving the parameters of an original model, namely, adding the initialized output layers to an original model and tuning the whole network [34, 14], and (2) preserving the knowledge of the base class in an original model with technology like knowledge distillation (KD) [11, 25]. However, while these methods can alleviate catastrophic forgetting to some extent, the overall classification performance remains significantly worse than classical joint learning.

The main contribution of our work is to provide an attribute-based incremental learning approach. We hypothesize that attribute prediction for each class can be used to encode representations of models, and attribute prediction for teacher and student models can be constrained by using attribute distillation (AD) loss, as explained in Sec. 3.5. AD loss helps a model remember some visual knowledge. By integrating attribute prediction, we boost the performance of image classification.

Another contribution is that our model can classify different classes and predict the attributes of each image. We asked ophthalmologists to provide attribute annotations for each class instead of each image. For predicting attributes precisely, we calculate the weight of each attribute in each image based on the information entropy of each attribute. Then, the weight of each attribute is integrated into a fully-connected layer to predict the attributes. We integrate the predicted attribute information into an incremental learning framework. Our paper is the first to use feature distribution to retain the knowledge of base classes to boost incremental learning performance. The resulting framework outperforms the existing state-of-the-art methods on our private retinal fundus image dataset. We found that the proposed method can be generalized in other domains, so we also experiment on a public dataset, ImageNet-150K [22], which contains attribute annotations made by experts.

2. Related work

Work related to the proposed method can be summarized into two categories: incremental learning and attribute learning. The following is an explanation of the connections and differences between our work and these methods in terms of corresponding aspects.

2.1. Incremental learning

Incremental learning has been a long-term problem in machine learning [3, 15]. Because the manner of the training procedure is incremental, the main problem is overcoming catastrophic forgetting. On the basis of the success of deep learning, the existing works can be categorized into two types: parameter-based and distillation-based. Parameter-based methods estimate the weight parameters of a teacher model and student model according to the importance of network weights. MAS [1] also focuses on studying the importance of the weights of a network in an unsupervised and continuous manner. When new data appears, changes to important parameters can be penalized to prevent the forgetting of the previous knowledge. Distillation-based methods mainly rely on knowledge distillation. Knowledge distillation [11] is an effective way to transfer knowledge from one network to another. The first application of KD in the incremental learning is in *Learning without Forgetting (LwF)* [20], where a modified cross-entropy loss is used to preserve the knowledge in a teacher model. Hou *et al*. [12] propose a framework for distilling previous knowledge from a base class via distillation and retrospection. M. Castro *et al*. [2] proposed an end-to-end incremental framework by using KD loss to retain the knowledge, while cross-entropy loss is used to classify the new type. S. Rebuffi *et al*. [26] selects some exemplars near the mean exemplars and uses KD loss to distill the knowledge from them.

2.2. Attribute learning

Attribute learning has attracted much attention for image classification in large-scale datasets [16, 7]. Learning visual attributes is beneficial for boosting classification performance [17]. The attribute descriptions of an instance or category are useful as a semantically meaningful intermediate representation to bridge the gap between low-level features and high-level class concepts. [31] proposed a joint learning architecture that is for face recognition and attribute prediction. [32] proposed a multi-task learning mechanism for increasing the discrepancy between different classes. [19] addresses the large-scale content-based face image-retrieval problem by learning a binary code that is comprised of different attributes. There have been several applications of incremental learning used on the study of attribute like [13, 5, 33].

3. Proposed method

3.1. Motivation

Our approach is motivated by the recent works on KD. In the incremental learning procedure, we use KD to retain the knowledge of base classes. We also designed the AD loss for preserving the knowledge of the attributes in base classes. For attribute prediction, we estimate the weight of each attribute in each image. The weight is used to calculate the representations used for predicting the attributes of each image. It helps predict the attributes precisely and boost the

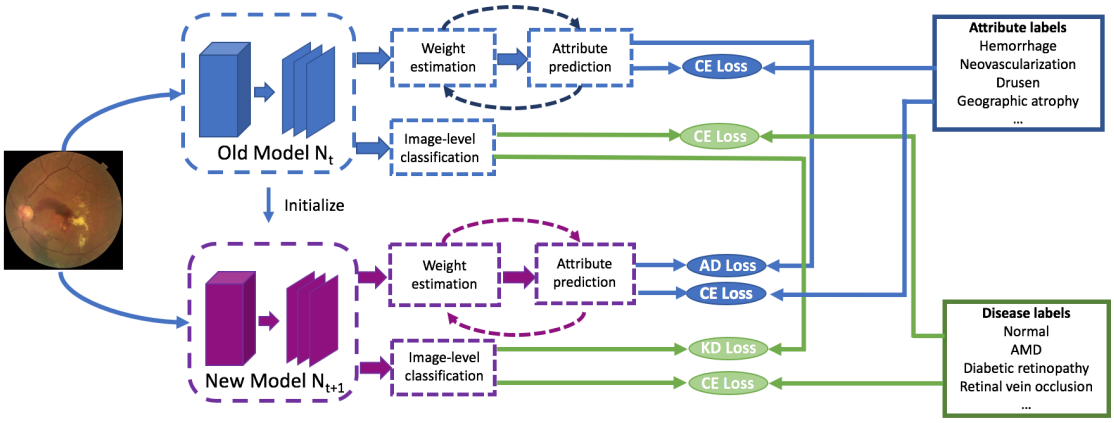


Figure 2. Framework of proposed method. We perform image-level classification and attribute prediction at same time. In attribute prediction, we estimate weight of each attribute for predicting attributes precisely. For distillation, we use KD loss for image-level classification and AD loss for attribute prediction.

classification performance. Our loss function is specially designed to make use of partially labelled attributes, which is more general in the real world. All of the procedures proceed incrementally. Our proposed method aims to improve the performance of incremental learning classification and predict the attributes of each image by using attributes only via class annotation instead of attribute annotation via each image. A flowchart is shown in Fig. 2.

3.2. Problem description

In this section, we explain each symbol used in our proposed method. Assume that we have N training images expressed as $X = [x_1, x_2, ..., x_N] \in \mathbb{R}^{d \times N}$, where $x_i \in \mathbb{R}^d$ is the i^{th} image with a d-dimension representation. We also have the ground-truth class label as $l \in [1, 2, ..., P]$, where P indicates the number of labels. We also have the ground-truth annotation of the attributes of all categories in the form of a class-attribute annotation matrix, denoted by $A \in \{0, 1, 2\}^{P \times T}$, where T indicates the number of attributes per class. $a_{l,j}$ is an element in matrix A that indicates the j^{th} attribute in category l^{th}. $a_{l,j} = 1$ and 0 indicates whether this attribute is present/absent. We use $a_{l,j} = 2$ to denote that the j^{th} attribute is unrelated to category l^{th}. Thus, the attribute label is missing, i.e., this attribute cannot provide useful information for classifying this category. In our proposed method, we need to perform image-level classification and attribute prediction at the same time. We also calculate the weight of each attribute $w_{l,j}$ to specify the contribution of the j^{th} attribute to the l^{th} class. We denote p_l as the prediction of each class label l, $p_{l,j}$ as the prediction of each attribute label, and $w_{l,j}$ as the weight of each attribute.

3.3. Image-level classification

We used ResNet50 [10] as the backbone of our proposed method since the ResNet architecture performs well in image classification. The global feature, i.e., a fully-connected layer fc after avg-pooling of $ResidualBlock4$, is fed to $T+1$ classifiers as $\left[fc^1, fc^2, \ldots, fc^{T+1}\right]$. fc^1 is used for image-level classification. fc^2 to fc^{T+1} are used for classifying the attributes of each category.

In image-level classification, we use a softmax layer to obtain the class prediction after the layer of fc^1. This classification is multi-class classification.

3.4. Weight estimation and attribute prediction

Even in the same class, the same attribute contributes differently in different images. Treating all attributes as equally informative will degrade the prediction performance. For precisely reflecting the information amount of each attribute in each image, we estimate the weight of each attribute and perform attribute prediction in this section. We adopt ResNet50 as the feature extractor. Then, we apply T attribute predictors, which consist of a fully-connected layer and a sigmoid layer, to proceed with attribute prediction. Figure 3 shows the procedure of weight estimation.

We send an image into ResNet50 and obtain the initial prediction result of the j^{th} attribute in category l^{th} as $p_{l,j}$ first. Then, we calculate the entropy of this attribute to represent the information amount of this attribute. The entropy is calculated as:

$$Entropy(p_{l,j}) = -\frac{1}{2}(p_{l,j}log(p_{l,j}) + (1-p_{l,j})log(1-p_{l,j})). \quad (1)$$

After we calculate the entropy of each attribute in each class, we calculate the exponent of each entropy as

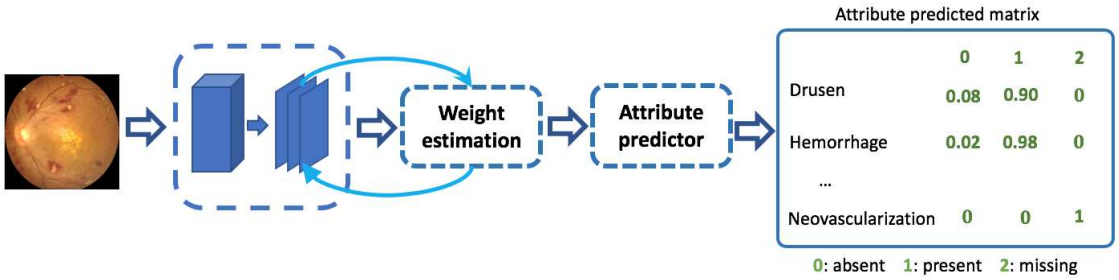


Figure 3. Framework of weight estimation and attribute prediction

$Conf(p_{l,j})$ and then normalize these exponents to obtain the weights:

$$Conf(p_{l,j}) = \begin{cases} e^{\frac{Entropy(p_{l,j})}{\sigma^2}}, & a_{l,j} \neq 2, \\ 0, & a_{l,j} = 2, \end{cases} \tag{2}$$

$$w_{l,j} = \frac{Conf(p_{l,j})}{\sum_{j=1}^{T} Conf(p_{l,j})}, \tag{3}$$

this weight is examined to show how much of a contribution an attribute has to distinguish classes. When $a_{l,j} = 2$, it means that this attribute has no contribution to distinguish this class, so we set $Conf(p_{l,j})$ as 0. We multiply these weights with the fully-connected layer output to obtain the new fully-connected layer output $fc_{l,j}^{new}$ of attribute j for class l:

$$fc_{l,j}^{new} = w_{l,j} fc_{l,j}, \tag{4}$$

$fc_{l,j}^{new}$ is sent to the attribute predictor, which consists of a fully-connected layer and a sigmoid layer. The attribute predictor is trained by a modified binary cross-entropy loss with an attribute label.

3.5. Loss function

The whole framework is trained in an incremental manner; namely, at incremental learning step t, we define the teacher model as N_t and the student model as N_{t+1}. The N_t is trained on base classes n, and N_{t+1} is trained on base classes n and new classes m by adding m neurons to the network's output layer of N_t. The weight parameters of the student model are initialized by using the parameters from the teacher model except for the newly added neurons in the output layer, which are randomly initialized.

To alleviate the catastrophic forgetting of base classes while training the data of new classes, we leverage KD in the loss function [11]. Instead of using hard labels to train the loss function, KD loss uses the teacher model's output as the ground-truth labels to train the student model. For the image-level classification, we jointly optimize KD loss on the base classes and cross-entropy loss on the new classes to achieve good classification performance.

For the classification part, we back-propagate the loss of image-level classification and attribute classification. The loss function for classification L_{cls} is defined as:

$$L_{cls} = L_{category} + \alpha L_{attribute}, \tag{5}$$

$$L_{category} = -\sum_{l=1}^{n+m} l_l log(p_l), \tag{6}$$

$$L_{l,j} = -\mathbb{I}(a_{l,j} \neq 2)(a_{l,j} log(p_{l,j}) + (1 - a_{l,j}) log(1 - p_{l,j})), \tag{7}$$

$$L_{attribute} = \sum_{l=1}^{n+m} \sum_{j=1}^{T} L_{l,j}, \tag{8}$$

where $L_{category}$ is the loss function for image-level classification, and l_l is the ground-truth label of a disease label. $L_{attribute}$ is the loss function for attribute classification. $L_{l,j}$ is a modified cross entropy loss function, and $a_{l,j}$ is the ground-truth label of an attribute label. In $L_{l,j}$, $\mathbb{I}$(cond.) is 1 when the condition is true and 0 otherwise. When an attribute label is missing, we have $\mathbb{I}(a_{l,j} = 2) = 0$. α is a trade-off parameter defined as 0.5. L_{cls} is the loss function for the classification part in our work.

After we obtain the teacher model, we want to keep the knowledge of the base classes. We use the knowledge distillation loss in image-level classification [20]. The loss function L_{dis} for distillation is defined as:

$$L_{dis} = L_D + \alpha L_{AD}, \tag{9}$$

$$L_D = -\sum_{l=1}^{n} \hat{l}_l^t log(p_l^{t+1}), \tag{10}$$

$$\hat{l}_l^t = softmax(fc_1^t / T_{dis}), \tag{11}$$

where T_{dis} is a temperature scalar. L_D is a distillation function for image-level classification (the KD loss in Fig. 2), and $\hat{l}_l^t$ is the distilled version of output probability of the l^{th}

class of the teacher model N_t. And p_l^{t+1} is the prediction probability of the l^{th} class of the student model N_{t+1} [20].

For keeping the attributes knowledge for base classes, we designed the attribute distillation loss L_{AD} to preserve the knowledge of old attributes. For any given input image x, let b be the top base class predicted by N_t, and we denote the attribute prediction vectors of models N_t and N_{t+1} as $A_t^{x,b}$ and $A_{t+1}^{x,b}$. We use the sum of the element wise L_1 difference of these two attribute prediction vectors to calculate the L_{AD} as:

$$L_{AD} = \sum_{j=1}^{T} \left|\left| A_{t,j}^{x,b} - A_{t+1,j}^{x,b} \right|\right|_1 . \tag{12}$$

Essentially, the attribute prediction of image x represents the feature distribution, which reflects the teacher model's study of base classes. If N_t and N_{t+1} have equivalent knowledge of base classes, they should predict attributes similarly. Therefore, $A_t^{x,b}$ and $A_{t+1}^{x,b}$ should be similar.

The overall loss combines the distillation loss and the classification loss:

$$L = \lambda L_{dis} + (1 - \lambda) L_{cls}, \tag{13}$$

where the scalar λ is used to balance between the two terms. The scalar λ is set to $\frac{n}{n+m}$, where n and m are the number of base and new classes.

3.6. Using exemplars of base classes

In our application, we choose a pipeline to apply a small number of exemplars from the base classes to our training dataset. That is because only using a new class for the next iteration of training will cause a large part of the information on base classes to be lost. For reducing the training time and the storage cost for incremental learning, we select a part of the dataset of base classes instead of all of the dataset.

Exemplar image selection is usually performed in two ways. The first, random selection, randomly selects a fixed number of images from each base class. The second, the exemplar management strategy proposed by iCaRL [26], selects images such that the average feature vector of exemplars will be closest to the mean value. In our pipeline, we select the second strategy.

4. Experiment

4.1. Experimental settings

Dataset We conducted an experiment on two datasets. The first, the **fundus image dataset**, is our private fundus image dataset. It contains 6,000 images consisting of 20 different types of diseases. Twenty-four attributes were annotated in this dataset by multiple expert ophthalmologists for each class. Table 1 shows disease-label and attribute-label parts [1]. Thirteen universities compiled this dataset, and multiple expert ophthalmologists annotated the images with image-level labels. 60% of the dataset was used for training, 20% of it was used for validation, and the remaining 20% was used for testing. The second is **ImageNet-150K-sub**, which is a subset of ImageNet-150K [22]. ImageNet-150K is a subset of the ILSVRC2012 dataset [28] with 150,000 images. In ImageNet-150K, 148 images from the training set and 2 images from the validation set were selected from the dataset for each of the 1,000 categories. Twenty-five attributes for each image were annotated in this dataset by multiple experts for each image. Each attribute has three kinds of labels, in which -1 means absent, 0 means uncertain, and is treated as missing in [22], and 1 means present. We treat an attribute with 1 as present and -1 as absent. Because "uncertain" cannot provide useful information on an attribute, we treat an attribute with 0 as missing. We randomly select a subset of 100 classes to conduct our experiment. We call this dataset "ImageNet-150K-sub."

Table 1. Part of disease labels and of semantic attributes in fundus image dataset

Disease label	Attributes
Normal	central reflux of macula
Age-related macular degeneration-early [AMD (early)]	hemorrhage,..., macular edema
Age-related macular degeneration-atrophic [AMD (atrophic)]	hemorrhage,..., macular edema
Age-related macular degeneration-exudative [AMD (exudative)]	hemorrhage,...,drusen, atrophy
Central serous chorioretinopathy	macular edema,..., intraretinal fluid
Retinal vein occlusion branch (RVO(B))	hemorrhage,...,vitreous hemorrhage
Diabetic retinopathy (DR)	hemorrhage,..., neovascularization
Glaucoma	pale optic disc, enlarged cupping
Myopic maculopathy	disc change, macular hole
Myopic choroidal neovascularization	geographic atrophy,..., hemorrhage

Here, we experimented with the dataset with attribute annotation for each class and the dataset with attribute annotation for each image respectively. We wanted to compare the performance and see whether the simple annotation could improve the experiment performance or not.

4.2. Implementation details

Before the training, we preprocessed the fundus images by cropping them to remove the black regions because these regions contain no information. After that, we resized the images to 512 $\times$ 512 pixels. During the training phase, we augmented the dataset with randomly rotated and horizontally and vertically flipped images from the dataset. The CNN model was implemented by using PyTorch trained on a Quadro GV100. We initially fine-tuned the ResNet50 [10] with ImageNet [4]. A gradient descent optimizer was used with a momentum of 0.9. We trained our CNN model

[1] The full disease labels and corresponding attribute labels can be found in supplementary meterials.

with a batch size of 16 and an initial learning rate of 0.01. We chose iCarL [26] as the baseline of our work. For the experiment with fundus images, 10 images for each base class were stored as exemplars. For the experiment with the ImageNet-150K-sub, 20 images for each base class were stored as exemplars.

4.3. Evaluation on incremental learning

Incremental learning was evaluated by the curve of the classification accuracies after each phase. We also calculated the average of all of the accuracies, i.e., average incremental accuracy.

We compared our proposed method ADINet with the state-of-the-art on our private fundus image dataset. We compared the results of ADINet with the results from learning without forgetting (LwF) [20], incremental classifier and representation learning (iCaRL) [26], and end-to-end incremental learning (EEIL) [2]. Figure 4 shows the results with an incremental setting of 10, 5, and 2 phases. The average incremental accuracies are shown in the bracket next to each method. As shown in Fig. 4, ADINet outperformed the state-of-the-art either in terms of the trend of the classification accuracy curve or average incremental accuracy. Particularly, with 5 phases, the increase of ADINet was more than 10% over iCaRL [26]. According to Fig. 4, it can be seen that ADINet could retain the knowledge of the base classes, which solved the imbalance between the base and new classes. The average incremental accuracies of ADINet with 10, 5, and 2 phases are 82.7%, 83.2%, and 83.1%, respectively. According to Fig. 4, it can be seen that as the incremental phase increased, the performance dropped due to catastrophic forgetting on our fundus image dataset.

We also compared ADINet for the dataset of ImageNet-150K-sub. We show the performance of the proposed method on the validation dataset. Figure 5 shows the results of comparison with 10, 5, and 2 phases. The average incremental accuracies of ADINet with 10, 5, and 2 phases were 87.4%, 87.3%, and 82.5% respectively.

The average incremental accuracy of ADINet for the fundus dataset was not as good as the average incremental accuracy for the ImageNet-150K-sub due to the dataset in ImageNet-150K-sub having more significant variance, which can be easier to classify. ImageNet-150K-sub provides attribute labels per image, but our fundus dataset provides only attribute labels per class. However, observing the results curves of these two datasets, we found that only using attribute labels per class can also alleviate the catastrophic forgetting of incremental learning effectively in each phase. On the basis of the results, our method still can boost the performance of incremental learning with only attribute labels per class.

Table 2. Classification results on fundus image dataset

Method	top-1 accuracy
Bilinear-CNN [21]	76.1%
PDFR [35]	77.0%
FV-CNN [8]	77.2%
FCAN [23]	78.5%
RA-CNN [6]	78.1%
Boost-CNN [24]	79.5%
MA-CNN [36]	81.2%
A^3M [9]	80.5%
ADINet	**82.7%**

Table 3. Classification results for ImageNet-150K-sub

Method	top-1 accuracy
Bilinear-CNN [21]	81.2%
PDFR [35]	83.0%
FV-CNN [8]	83.4%
FCAN [23]	83.9%
RA-CNN [6]	84.2%
Boost-CNN [24]	84.9%
MA-CNN [36]	85.4%
A^3M [9]	85.7%
ADINet	**87.4%**

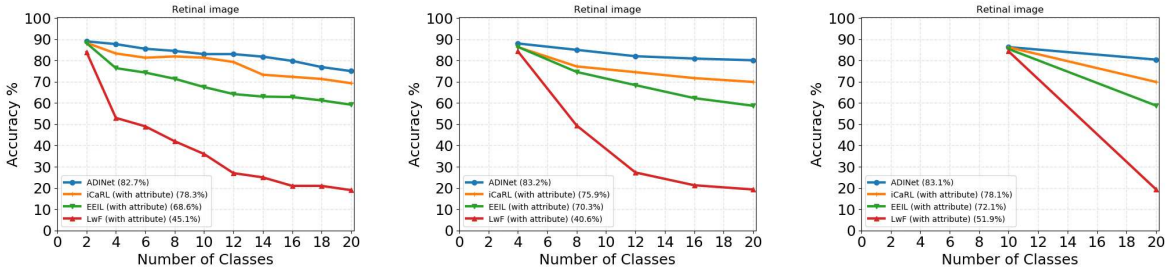
4.4. Evaluation on image classification

We compared ADINet with some classical image classification methods. We conducted the experiment on our private fundus image dataset and ImageNet-150K-sub. We used the classification accuracy (top-1 accuracy) for measurement. The results are summarized in Table 2 and Table 3. We used the average incremental accuracy of 10 phases as the classification performance for comparison with the state-of-the-art. From the results shown, our model showed competitive performance. This indicates that integrating with attribute and weight estimation boosts the incremental learning performance, which is effective when performing image classification.

4.5. Evaluation on attribute prediction

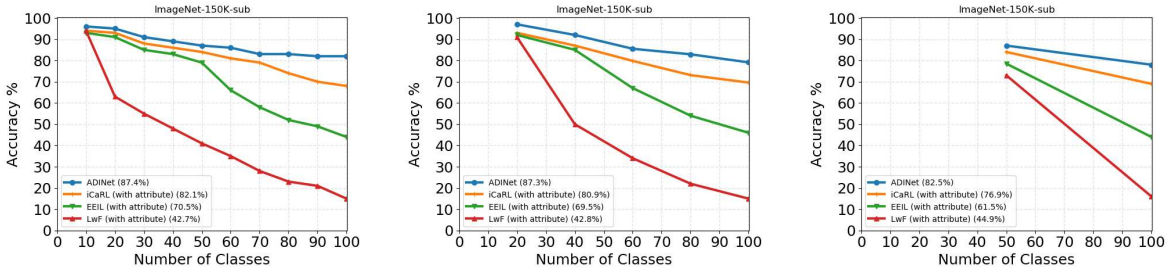
We compared our proposed method in terms of attribute prediction. We conducted our experiment on the dataset of ImageNet-150K-sub. We used the classification accuracy (top-1 accuracy) for measurement. Table 4 compares the results of 10 attributes in the ImageNet-150K-sub dataset with two previous methods [2]. One is only using ResNet50 to predict attribute recognition, and the other is DeepMAR [18], which performs attribute recognition by using the correlations between human attributes to improve the overall recognition performance further. ADINet had a more significant improvement in terms of low ratio attributes than the other two previous methods, according to Table 4. By using only ResNet50, it can be seen that the overall performance of attribute recognition was relatively low. By considering the correlations between attributes, the overall performance is increased. Calculating the weight prediction and integrating the weights with attribute recognition can

[2]The results of all attribute recognition in ImageNet-150K-sub dataset can be found in supplementary material.



(a) (b) (c)

Figure 4. Performance on our private fundus image dataset with incremental setting of 10, 5, and 2 phases.



(a) (b) (c)

Figure 5. Performance on ImageNet-150K-sub with incremental setting of 10, 5, and 2 phases.

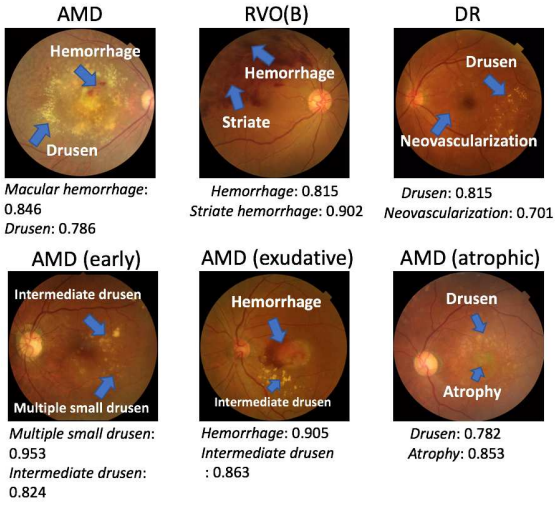


Figure 6. Attribute recognition on fundus image dataset. First row shows attribute prediction results of three different diseases, and second row shows attribute prediction results of three different severities of AMD. Blue arrows show corresponding regions of attributes. We show attributes with top-2 prediction scores.

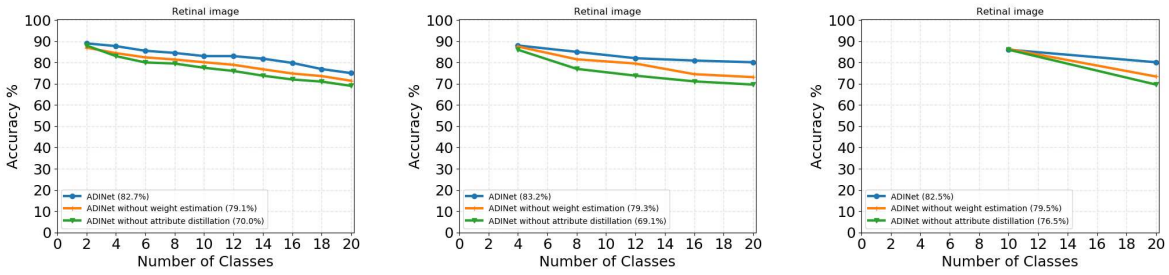
boost the performance significantly. Figure 6 shows examples of attribute prediction with the fundus image dataset. It can be seen that our approach predicted the attributes of each fundus image effectively, which would help ophthalmologists in making diagnoses.

Table 4. Attribute recognition comparison for ImageNet-150K-sub

Attribute	Ratio	ResNet50	DeepMAR	ADINet
Black	0.12	57.2	65.2	**75.2**
Blue	0.0235	66.3	76.1	**81.4**
Brown	0.0895	62.8	71.6	**78.2**
Gray	0.0265	61.8	67.7	**73.6**
Green	0.0315	52.9	65.2	**78.9**
Orange	0.012	57.3	69.0	**73.4**
Pink	0.0075	57.4	67.8	**72.1**
Red	0.0435	59.3	**76.2**	75.1
Purple	0.003	64.1	81.7	**83.2**
White	0.111	67.6	78.7	**82.3**
Average	*	60.7	71.9	**77.3**

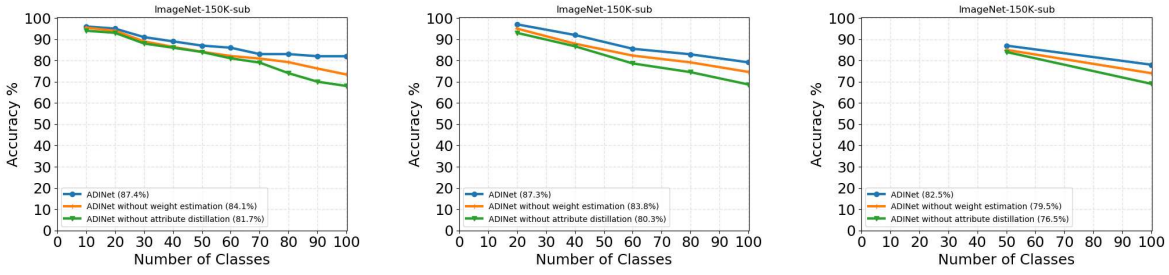
4.6. Ablation study

We now analyze the components of our approach and demonstrate their contribution to the overall performance. We evaluated our approach with an incremental setting of 10 phases. We performed two different experiments: ADINet without attribute distillation and ADINet without weight estimation. In the first experiment, we conducted our method on the fundus image dataset and ImageNet-150K-sub. We report the classification accuracy (top-1 accuracy) for each incremental step. We also compare the average incremental accuracy with different methods. In the second experiment, we also conducted our method on the both datasets. We compared the classification accuracy (top-1 accuracy) of each incremental step on our private fundus image dataset and ImageNet-150K-sub. And we compared the average attribute recognition accuracy on the ImageNet-150K-sub.



(a) (b) (c)

Figure 7. Ablation study with fundus images. Results for (a), (b), (c) are for 10, 5, and 2 phases, respectively.



(a) (b) (c)

Figure 8. Ablation study with ImageNet-150K-sub. Results for (a), (b), (c) are for 10, 5, and 2 phases, respectively.

4.6.1 Evaluation on attribute distillation

As aforementioned, for estimating the effect of the attribute in the image classification, we add attribute distillation for boosting the classification performance. Comprehensive experiments were performed, and the results are displayed in Fig. 7 and Fig. 8. We compared ADINet and ADINet without attribute distillation. According to the figures, without attribute distillation, the classification accuracy of each incremental step dropped significantly. In particular, as the incremental step increased, the discrepancy between ADINet and ADINet without attribute distillation increased. This demonstrates that distilling the attribute information from the base classes helps retain the feature preservation and boosts the classification performance.

4.6.2 Evaluation on weight estimation

We added weight estimation for predicting the attributes precisely for each class. According to Fig. 7 and Fig. 8, after removing the weight estimation, the classification accuracy of each incremental step degraded. That is because, in each image, different attributes contribute differently. A more informative attribute should be given more weight. We also compared the average accuracy of attribute recognition between ADINet and ADINet without estimation in Table 5. According to this table, it can be seen that our proposed method produced more precise recognition results for the dataset of ImageNet-150K-sub. We only show the comparison results of iCaRL here. The comparison of other baselines can be found in supplementary material.

Table 5. Comparison of attribute recognition accuracy for two datasets

Dataset	ADINet w/o weight estimation	ADINet
ImageNet-150K-sub	73.6	**76.6**

5. Conclusion

We explored the incremental learning problem for the task of image classification, and we proposed a method: attribute distillation and attribute weight estimation. By integrating the attribute information to transfer the knowledge of a base class from a teacher to student model, the proposed method boosts the performance of classification. At the same time, our proposed method can also investigate the the predicted attributes of each image. This approach outperforms the state-of-the-art. Regarding future work, the proposed method could be applied to a scenario in which there are a few attribute labels even without attribute labels. Incremental attribute recognition is a challenging problem due to the absence of ground-truth attributes for each image. We intend to extend our work in this direction.

Acknowledgement The fundus image dataset is provided by Japan Ocular Imaging Registry Research Group. This research is supported by the ICT infrastructure establishment and implementation of artificial intelligence for clinical and medical research from Japan Agency for Medical Research and development, AMED.

References

[1] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. *Proceedings of the European Conference on Computer Vision (ECCV)*, 11 2017. 2

[2] Francisco M. Castro, Manuel J. Marin-Jimenez, Nicolas Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 2, 6

[3] Gert Cauwenberghs and Tomaso Poggio. Incremental and decremental support vector machine learning. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, NIPS'00, pages 388–394, 2000. 2

[4] J. Deng, W. Dong, R. Socher, and L. Li. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. 5

[5] Emrah Ergul. Relative attribute based incremental learning for image recognition. *CAAI Transactions on Intelligence Technology*, 2(1):1 – 11, 2017. 2

[6] J. Fu, H. Zheng, and T. Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4476–4484, July 2017. 6

[7] Gang Wang and D. Forsyth. Joint learning of visual attributes, object classes and visual saliency. In *2009 IEEE 12th International Conference on Computer Vision*, pages 537–544, Sep. 2009. 2

[8] Philippe Henri Gosselin, Naila Murray, Hervé Jégou, and Florent Perronnin. Revisiting the fisher vector for fine-grained classification. *Pattern Recognition Letters*, 49:92–98, 2014. 6

[9] Kai Han, Jianyuan Guo, Chao Zhang, and Mingjian Zhu. Attribute-aware attention model for fine-grained representation learning. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, pages 2040–2048, New York, NY, USA, 2018. ACM. 6

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016. 3, 5

[11] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. 2, 4

[12] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Lifelong learning via progressive distillation and retrospection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[13] P. Kankuekul, A. Kawewong, S. Tangruamsub, and O. Hasegawa. Online incremental attribute-based zero-shot learning. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3657–3664, June 2012. 2

[14] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114, 12 2016. 2

[15] Ilja Kuzborskij, Francesco Orabona, and Barbara Caputo. From N to N+1: Multiclass transfer incremental learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013. 2

[16] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958, June 2009. 2

[17] Ryan Layne, Timothy Hospedales, and Shaogang Gong. Person re-identification by attributes. In *Proceedings of the British Machine Vision Conference (BMVC)*, volume 2, 01 2012. 2

[18] D. Li, X. Chen, and K. Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 111–115, Nov 2015. 6

[19] Yan Li, Ruiping Wang, Haomiao Liu, Huajie Jiang, Shiguang Shan, and Xilin Chen. Two birds, one stone: Jointly learning binary code for large-scale face image retrieval and attributes prediction. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. 2

[20] Zhizhong Li and Derek Hoiem. Learning without forgetting. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 40, pages 2935–2947, 2016. 2, 4, 5, 6

[21] T. Lin, A. RoyChowdhury, and S. Maji. Bilinear CNN models for fine-grained visual recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1449–1457, Dec 2015. 6

[22] H. Liu, R. Wang, S. Shan, and X. Chen. Learning multi-functional binary codes for both category and attribute oriented retrieval tasks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6259–6268, July 2017. 2, 5

[23] Xiao Liu, Tian Xia, Jiang Wang, and Yuanqing Lin. Fully convolutional attention localization networks: Efficient attention localization for fine-grained recognition. *CoRR*, abs/1603.06765, 2016. 6

[24] Mohammad Saberian Jian Yang Nuno Vasconcelos Mohammad Moghimi, Serge Belongie and Li-Jia Li. Boosted convolutional neural networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 24.1–24.13. BMVA Press, September 2016. 6

[25] Amal Rannen, Rahaf Aljundi, Matthew B. Blaschko, and Tinne Tuytelaars. Encoder based lifelong learning. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2

[26] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2, 5, 6

[27] ANTHONY ROBINS. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146, 1995. 1

[28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 5

[29] Steffen Schmitz-Valckenberg, Frank Holz, Alan Bird, and Richard Spaide. Fundus autofluorescence imaging: Review and perspectives. *Retina (Philadelphia, Pa.)*, 28:385–409, 04 2008. 1

[30] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. Incremental learning of object detectors without catastrophic forgetting. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1

[31] Fariborz Taherkhani, Nasser M. Nasrabadi, and Jeremy Dawson. A deep face identification network enhanced by facial attributes prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018. 2

[32] Zhanxiong Wang, Keke He, Yanwei Fu, Rui Feng, Yu-Gang Jiang, and Xiangyang Xue. Multi-task deep neural network for joint face recognition and facial attribute prediction. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, ICMR '17, pages 365–374, New York, NY, USA, 2017. ACM. 2

[33] Liuyu Xiang, Xiaoming Jin, Guiguang Ding, Jungong Han, and Leida Li. Incremental few-shot learning for pedestrian attribute recognition. pages 3912–3918, 08 2019. 2

[34] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *ICML*, 2017. 2

[35] X. Zhang, H. Xiong, W. Zhou, W. Lin, and Q. Tian. Picking deep filter responses for fine-grained image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1134–1142, June 2016. 6

[36] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 6