Exemplar-supported representation for effective class-incremental learning.

GUO, L., XIE, G., XU, X. and REN, J.

2020

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 21, 2020, accepted March 9, 2020, date of publication March 12, 2020, date of current version March 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2980386

Exemplar-Supported Representation for Effective Class-Incremental Learning

LEI GUO[1], GANG XIE[2,3,4], (Member, IEEE), XINYING XU[2], AND JINCHANG REN[2,5], (Senior Member, IEEE)

[1]College of Information and Computer, Taiyuan University of Technology, Taiyuan 030024, China
[2]College of Electrical and Power Engineering, Taiyuan University of Technology, Taiyuan 030024, China
[3]School of Electronic Information Engineering, Taiyuan University of Science and Technology, Taiyuan 030024, China
[4]Shanxi Key Laboratory of Advanced Control and Intelligent Information System, Taiyuan University of Science and Technology, Taiyuan 030024, China
[5]Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow G11XW, U.K.

Corresponding authors: Gang Xie (xiegang@tyut.edu.cn) and Jinchang Ren (jinchang.ren@strath.ac.uk)

This work was supported in part by the Key Research and Development Plan of Shanxi Province under Grant 201703D111023, in part by the Shanxi International Cooperation Project under Grant 201803D421039, in part by the Hundred Talents Program of Shanxi, and in part by the Natural Science Foundation of Shanxi Province under Grant 201801D121144 and Grant 201801D221190.

ABSTRACT Catastrophic forgetting is a key challenge for class-incremental learning with deep neural networks, where the performance decreases considerably while dealing with long sequences of new classes. To tackle this issue, in this paper, we propose a new exemplar-supported representation for incremental learning (ESRIL) approach that consists of three components. First, we use memory aware synapses (MAS) pre-trained on the ImageNet to retain the ability of robust representation learning and classification for old classes from the perspective of the model. Second, exemplar-based subspace clustering (ESC) is utilized to construct the exemplar set, which can keep the performance from various views of the data. Third, the nearest class multiple centroids (NCMC) is used as the classifier to save the training cost of the fully connected layer of MAS when the criterion is met. Intensive experiments and analyses are presented to show the influence of various backbone structures and the effectiveness of different components in our model. Experiments on several general-purpose and fine-grained image recognition datasets have fully demonstrated the efficacy of the proposed methodology.

INDEX TERMS Exemplar-based subspace clustering, incremental learning, memory aware synapses, image recognition.

I. INTRODUCTION

In real-world applications, most of the image recognition systems are incremental [1], thus they should be updated continuously to adapt to the new data that are different from the existing ones. To save the computation cost and storage requirement, the model to be obtained need adapt to or extendable to the new data, rather than retraining from scratch. This is the motivation of incremental learning, a kind of method for learning models to cope with new classes or tasks with less catastrophic forgetting [2].

From the perspective of goal, incremental learning can be divided into two categories, i.e. task-incremental learning [3]–[12] and class-incremental learning [13]–[17]. Task-incremental learning is trained with multiple classifiers to handle old and new tasks. Regarding class-incremental learning, a unified classifier is utilized to process mixed data of the old and new classes, which is more realistic, but more difficult to train [16]. In this paper, we focus on class-incremental learning.

The dominating issue for incremental learning is catastrophic forgetting, where the model performance for old classes or tasks will be disrupted by the training of new data. Currently, deep-learning-based incremental learning methods are the main paradigm to alleviate catastrophic forgetting. From the perspective of the employed strategies, these methods can be divided into memory-based and model-based methods. For memory-based methods, exemplars are selected to preserve the performance of the old classes or tasks, where the model is trained with the new data and exemplar set in the subsequent stage. Obviously, selecting enough and diverse exemplars is the key to preserving the performance of old classes or tasks. In [13], [14], herding selection is utilized to extract the exemplars, which is a sampling method with replacement thus cannot properly rank the exemplars. In [15], exemplars are selected according to their scores associated

The associate editor coordinating the review of this manuscript and approving it for publication was Md. Moinul Hossain.

with all current classes, yet disregarding the diversity of the exemplars. Furthermore, in [13]–[15], the pre-trained DCNN is not employed as the backbone due to the improper resolution of the images. For model-based methods [3]–[12], they update the weight parameters by using specific learning algorithms or certain defined loss functions. However, these methods cannot work well in long sequences of new classes or tasks, due mainly to the omitting of as the old data [16].

To tackle the aforementioned issues, in this paper, we propose an exemplar-supported representation for incremental learning (ESRIL) approach for both general-purpose and fine-grained image recognition. The main contributions of our paper are highlighted as follows:

1) We propose a novel incremental learning approach that incorporates three key components, i.e. i) memory aware synapses (MAS) [8] for representation learning and classification, ii) scalable exemplar-based subspace clustering (ESC) for selecting and ranking exemplars [18] to guarantee sufficient and diverse exemplars from each subspace, and iii) the nearest class multiple centroids (NCMC) classifier for effective classification to save the training time and to reduce the impact of class imbalance between old and new classes.

2) Specifically, MAS based representation learning and classification can help to retain the ability to extract distinctive features for old classes. By identifying the important network weight parameters via a learned function MAS can make use of these more robust important parameters and perform better class-incremental learning for both general-purpose and fine-grained image recognition, compared to those using knowledge distillation loss. As we resize the image to a proper resolution, DCNN pre-trained on ImageNet can be adopted as the backbone in our MAS module, boosting the representation learning and classification performance.

3) Comprehensive experiments have demonstrated the superior performance of our proposed class-incremental learning approach for various image recognition tasks. We first explore the effect of various backbone structures on three datasets. By isolated experiments, we further verify the efficacy of ESC and MAS. Experiments on five general-purpose and fine-grained image recognition datasets have fully demonstrated the effectiveness of our proposed approach as it has significantly outperformed the baselines.

The rest of this paper is organized as follows. In Section II the related work is introduced. Section III discusses in detail the proposed ESRIL approach. In Section IV, we conduct experiments to show the effect of different backbone structures and isolated components and compare the proposed approach with the baselines on five datasets. Finally, some concluding remarks and future work are summarized in Section V.

II. RELATED WORK

Incremental learning has been studied for a long time in varied areas (e.g., image recognition [14], [19], [20], object detection [21], visual tracking [22], genetic search [23]). Catastrophic forgetting is the main challenge for incremental learning, indicating the stability-plasticity dilemma. In this work, we focus on image recognition. We now depict works related to our approach. The related works can be divided into conventional methods and deep learning methods.

A. INCREMENTAL LEARNING WITH CONVENTIONAL METHODS

Incremental learning with conventional methods consists of neural-network-based methods and others. From the perspective of the strategies used, the neural-network-based methods have two types, i.e. memory-based and model-based methods. Memory-based methods use the exemplars of old data to maintain performance. In the 1990s, a fair amount of memory-based methods were proposed to address the issue of catastrophic forgetting [24], [25]. Robins first employs pseudo exemplars to alleviate catastrophic forgetting rather than the original old data [26], [27]. Model-based methods utilize specific learning algorithms or certain defined loss functions to keep the performance. In [28], a variation of the backpropagation algorithm is proposed, which just punishes those activations related to the errors. Kruschke presents a new architecture called ALCOVE [29]. By the specific architecture, the representation of new tasks or classes will have less overlap with the representation of old data. French uses dual-network architectures including the early-processing area and the final-storage area to deal with the old and new data [30]. Learn++ algorithm is an early attempt that uses the Ensemble Learning scheme in incremental Learning [31]. In [32], Coop *et al.* also employ the Ensemble Learning scheme, and they embed a sparsely encoding layer to alleviate the change of prior learned representations.

Several other conventional methods also focus on incremental learning, including support vector machine (SVM), Random Forests and nearest class mean (NCM). In [33], SVM is trained with the new data and support vectors to adapt to incremental training. In [34], Cauwenberghs *et al.* propose a method to boost performance by retaining the Kuhn-Tucker conditions on the old classes or tasks, while adding new data to the solution. In [35], the NCM classifier utilizes one or multiple class centroids to classify, which can incorporate data of new classes with less computational cost and outperforms standard parametric classifiers. In [36], Ristin *et al.* combine Random Forests with NCM or SVM for incrementally learning the new data.

B. INCREMENTAL LEARNING WITH DEEP LEARNING

Deep learning is a powerful data-driven tool for representation learning, promoting the development of incremental learning. Incremental learning with deep learning also can be categorized into memory-based methods and model-based methods.

First, we review the memory-based methods with deep learning [13]–[17]. These methods classify old classes or tasks relatively well in long sequences, due to the

existence of true or pseudo exemplars of old data. In [13], Rebuffi *et al.* present a herding selection approach for exemplar selection, and employ exemplar set and knowledge distillation loss to train the model. In the End-to-End incremental learning [14], herding selection and knowledge distillation loss are adopted, and balanced fine-tuning is performed to tackle the imbalance between the old and new classes. However, herding selection cannot rank the exemplars properly. In [15], Chen *et al.* adopt Generative Adversarial Networks (GAN) to synthesize data of old classes and train the model with the synthesized data and exemplars for reducing the impact of the imbalance. The method can have good performance only if GAN learns the data of old classes well. And the exemplar selection method in [15] needs to be improved. As shown in [16], the imbalance between old and new classes is a crucial reason for catastrophic forgetting, cosine normalization, and inter-class separation are used to address the problem. In [17], a linear model is added after the FC layer to alleviate the impact of class imbalance.

Second, we summarize model-based methods with deep learning [3]–[12]. Learning without Forgetting (LwF) uses knowledge distillation loss and the outputs of new data on the original network to reduce catastrophic forgetting [3], which is an early attempt. In [4], meta-learning is used to boost domain generalization ability and to learn a robust feature representation. Regarding the Elastic Weight Consolidation [5], important weight parameters are identified by the Fisher information metric and the corresponding loss function is employed to preserve the performance for previous tasks. In [6], the unimportant weight parameters are pruned by network pruning methods, and multiple tasks can be packed in a network. In [7], the past and current parameters are recorded by intelligent synapses, and the parameters' importance is estimated online. Nonetheless, as the importance is computed according to the batch gradient descent, it can be overestimated and underestimated. In terms of MAS [8], importance is estimated by the sensitivity of the learned function. In this way, the network is no longer stuck to a local minimum, thus it obtains a fine performance. These have contributed to advances in this area. In [5], [6], [8], the important weights are identified by pruning technique or learned binary masks, and kept fixed in the learning process of new tasks. The methods in [10], [11] build universal parametric families of networks that share large numbers of parameters among different tasks, employ residual adapter to switch the network for the new target task. In [12], transfer learning is utilized to adapt the model to new tasks. The works mentioned above are task-incremental methods. As the classifier should be assigned in advance for the task [10]–[12], it is hard to generalize the methods for class-incremental learning. The important network parameters are more robust in [5]–[8], and these works can be used for reference. But the main drawback of these methods is that they cannot be kept well in a long sequence of classes or tasks [16]. Based on the considerations above, memory-based methods, and model-based methods are combined in our work, and MAS is adopted for representation learning and classification.

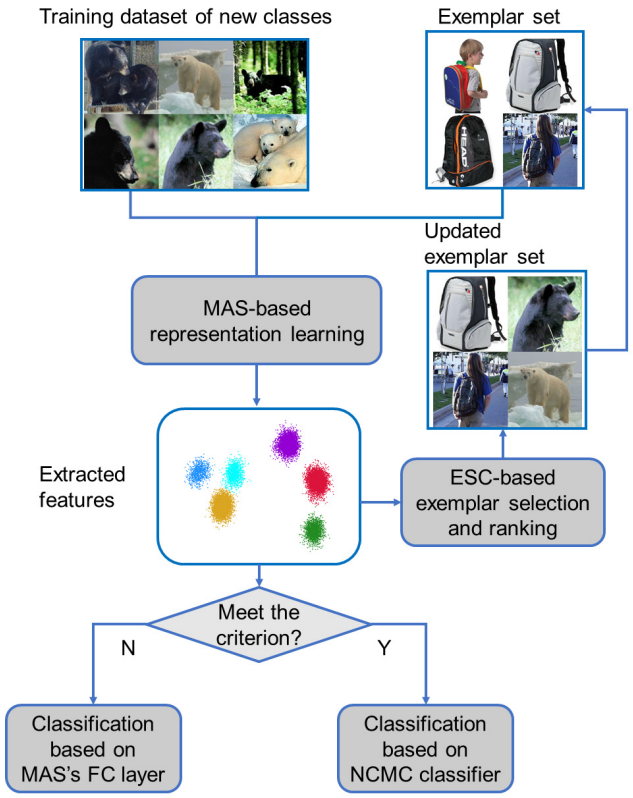

FIGURE 1. Illustration for the proposed approach ESRIL. The criterion is that samples are not quite different from samples in the ImageNet dataset.

III. THE PROPOSED APPROACH

In this paper, we propose a novel incremental learning approach ESRIL that incorporates three key components. First, the memory aware synapses (MAS) module [8] is employed as a representation learning and classification module to retain the ability to extract distinctive features for old classes. Compared to the networks in [13]–[15], important network parameters in [8] are more robust. As we resize the image to a proper resolution, DCNN pre-trained on the ImageNet can be adopted as the backbone in the MAS module. Resizing the image and applying a pre-trained backbone is an early attempt in class-incremental learning. Second, when samples are not quite different from those in the ImageNet dataset, the NCMC is used as a classifier to save the training time of the FC layer of MAS whilst preserving the high accuracy in both general-purpose and fine-grained recognition tasks. Third, scalable exemplar-based subspace clustering (ESC) is used to select exemplars [18], which can rank exemplars and is guaranteed to select sufficient and diverse exemplars from each subspace.

A. MAS-BASED REPRESENTATION LEARNING AND CLASSIFICATION

First, we use MAS to alleviate catastrophic forgetting from different views of the model. MAS is a model-based method for representation learning and classification. By identifying the important network weight parameters via a learned function, the feature extraction and classification ability for the old classes can be retained. The cost of changing the important parameters is large, which can mitigate the influence of class imbalance between the new and old classes slightly. And in the MAS module, DCNN pre-trained on ImageNet is used as a backbone. When the image is not quite different from images in the ImageNet dataset, MAS will have excellent representational and generalization ability after the module is trained for one or a few epochs.

After the initial model is trained on data $\boldsymbol{X}^{1,p-1}$ of old classes $1, \ldots, p-1$, the importance of model parameters should be estimated to penalize changes of the important parameters and to retain feature extraction and classification ability on the old classes. As a result, we have a mapping function $G(\boldsymbol{x};\theta)$ for classification, where parameters $\theta = \{\theta_{ij}\}$. According to [8], the importance $\alpha_{ij}^{1,p-1}$ for parameter $\theta_{ij}^{1,p-1}$ on the training data of classes $1, \ldots, p-1$, can be defined as,

$$\alpha_{ij}^{1,p-1} = \frac{1}{M}\sum_{k=1}^{M}\left\|f_{ij}(\boldsymbol{x}_k)\right\|, \tag{1}$$

where M is the number of data points in the training dataset, $f_{ij}(\boldsymbol{x}_k) = \partial(l_2^2(G(\boldsymbol{x}_k;\theta)))/\partial\theta_{ij}^{1,p-1}$ is the gradient of the squared L_2 norm of mapping function for the parameter $\theta_{ij}^{1,p-1}$. Consequently, the importance $\alpha^{1,p-1}$ is obtained.

When data $\boldsymbol{X}^{p,q}$ of the new classes $p, \ldots, q$ arrive, we construct an augmented training set $\boldsymbol{D}^{1,q}$ that contains exemplar set $\boldsymbol{R}^{1,p-1}$ for old classes $1, \ldots, p-1$ and $\boldsymbol{X}^{p,q}$ for new classes $p, \ldots, q$, and adapt the model to the new classes. The MAS-based representation learning and classification algorithm is given in Table 2. The loss is calculated as,

$$L^{1,q} = L_{\mathrm{CE}} + \lambda\sum_{i,j}\alpha_{ij}^{1,p-1}\left(\theta_{ij}^{1,q} - \theta_{ij}^{1,p-1}\right)^2, \tag{2}$$

where λ a balance parameter, $\theta^{1,q}$ the new parameters for all the classes $1, \ldots, p-1, p, \ldots, q$, $\theta^{1,p-1}$ the old parameters for old classes $1, 2, \ldots, p-1$, L_{CE} the cross-entropy loss function. In the loss function, $\theta^{1,q}$ need to be optimized by stochastic gradient descent, and the importance $\alpha^{1,p-1}$ and parameters $\theta^{1,p-1}$ are fixed. After the learning process, the sub importance $\alpha\_sub^{1,q}$ is computed according to the augmented training dataset of classes $1, \ldots, q$, whilst $\alpha^{1,q}$ is obtained by,

$$\alpha^{1,q} = \alpha^{1,p-1} + \alpha\_sub^{1,q}, \tag{3}$$

which can be used for future training stage.

The mapping function G can be decomposed as $C \circ F$, where F is the representation function and C is the mapping function of the FC layer of MAS, namely the classification function. When a data point $\boldsymbol{x}$ arrives, the feature $\boldsymbol{z}$ can be represented by,

$$\boldsymbol{z} = F(\boldsymbol{x}). \tag{4}$$

Finally, the classification by the FC layer can be performed by,

$$\widehat{y} = C(\boldsymbol{z}). \tag{5}$$

TABLE 1. MAS-based representation learning algorithm.

Algorithm 1

Input: data $\boldsymbol{X}^{p,q}$ of the new classes $p, \ldots, q$, exemplar set $\boldsymbol{R}^{1,p-1}$ of the old classes $1, \ldots, p$-1, the model parameters $\theta^{1,p-1}$ for the old classes $1, \ldots, p$-1, the importance $\alpha^{1,p-1}$ for $\theta^{1,p-1}$

Construct an augmented training set, $\boldsymbol{D}^{1,q} = \boldsymbol{R}^{1,p-1} \cup \boldsymbol{X}^{p,q}$.

Train the model with the loss function,

$$L^{1,q} = L_{\mathrm{CE}} + \lambda\sum_{i,j}\alpha_{ij}^{1,p-1}\left(\theta_{ij}^{1,q} - \theta_{ij}^{1,p-1}\right)^2.$$

Calculate the sub importance of $\theta^{1,q}$, $\alpha\_sub_{ij}^{1,q} = \frac{1}{M}\sum_{k=1}^{M}\left\|f_{ij}(\boldsymbol{x}_k)\right\|$,

where $f_{ij}(\boldsymbol{x}_k) = \frac{\partial(l_2^2(G(\boldsymbol{x}_k;\theta^{1,q})))}{\partial\theta_{ij}^{1,q}}$.

The total importance is computed by, $\alpha^{1,q} = \alpha^{1,p-1} + \alpha\_sub^{1,q}$.

Output: model parameters $\theta^{1,q}$ for the classes $1, \ldots, p$-1, $p, \ldots, q$, importance $\alpha^{1,q}$ for $\theta^{1,q}$

B. NEAREST CLASS MULTIPLE CENTROIDS CLASSIFIER

After the feature z is extracted by the MAS module, the nearest class multiple centroids (NCMC) classifier is used for classification [35], when the image is not quite different from those in the ImageNet dataset. NCMC is utilized to replace the FC layer of MAS, as the training of the FC layer is more time-consuming. And NCMC classifier employs a set of centroids to represent a class, resulting in more flexible class representations. In this work, considering the high representation ability of the pre-trained MAS, we use Euclidean distance rather than the Mahalanobis distance for simplicity. As a result, in this paper, the NCMC classifier is a non-parametric method, reducing the influence of class imbalance of old and new classes [37].

As for the augmented training dataset of classes $1, \ldots, p-1, p, \ldots, q$, we can acquire a set of c feature centroids $\left\{\boldsymbol{\mu}_{ij} | i = 1, \ldots, p-1, p, \ldots, q; j = 1, \ldots, c\right\}$ with k-means clustering method. For the class i, the posterior probability can be determined as:

$$p(i|\boldsymbol{z}) = \sum_{j=1}^{c} p\left(\boldsymbol{\mu}_{ij}|\boldsymbol{z}\right), \tag{6}$$

$$p\left(\boldsymbol{\mu}_{ij}|\boldsymbol{z}\right) = \frac{1}{V}\exp\left(-0.5 \times d\left(\boldsymbol{z}, \boldsymbol{\mu}_{ij}\right)\right), \tag{7}$$

TABLE 2. ESC-based exemplar selection and ranking algorithm.

Algorithm 2

Input: exemplar set $\boldsymbol{R}^{1,p-1}=\{\boldsymbol{R}_k^{1,p-1}|\boldsymbol{R}_k^{1,p-1}=(\boldsymbol{r}_1,...,\boldsymbol{r}_m),\ k=1,...,p-1\}$, data $\boldsymbol{X}^{p,q}$ of classes $p, \ldots, q$, the number of extracted exemplars n per class, hyperparameter $\eta>1$, the representation function F

Remove the unimportant exemplars.

For k=1, …, p-1

$\boldsymbol{R}_{1,p-1}^{k} \leftarrow (\boldsymbol{r}_1,...,\boldsymbol{r}_n)$ //Select the first n exemplars.

End for

Select exemplars.

For k=p, …, q

$\boldsymbol{Z}_k^{p,q}=F\left(\boldsymbol{X}_k^{p,q}\right)$

Random select $z\in\boldsymbol{Z}_k^{p,q}$, and initialize $\boldsymbol{Z}_k^{\mathrm{Ex}}=\{z\}$.

For i=1, …, n-1 do

$\boldsymbol{Z}_k^{\mathrm{Ex}}=\boldsymbol{Z}_k^{\mathrm{Ex}}\bigcup\arg\max_{z\in Z} f_\eta(z,\boldsymbol{Z}_k^{\mathrm{Ex}})$

End for

According to the $\boldsymbol{Z}_k^{\mathrm{Ex}}$, select the corresponding exemplars $\boldsymbol{R}_k^{p,q}$ of class k.

End for

$\boldsymbol{R}^{p,q}=\bigcup_{k=p,...,q}\boldsymbol{R}_k^{p,q}$

Update exemplar set.

$\boldsymbol{R}^{1,q}=\boldsymbol{R}^{1,p-1}\bigcup\boldsymbol{R}^{p,q}$

Output: exemplar set $\boldsymbol{R}^{1,q}$

where $V=\sum_{i=1}^{q}\sum_{j=1}^{c}\exp(-0.5\times d(z,\boldsymbol{\mu}_{ij}))$ and $p(i|z)$ is the posterior probability for class i, $d(z,\boldsymbol{\mu}_{ij})$ is the Euclidean distance between feature z and a feature centroid. Afterward, the class label $\widehat{y}$ is predicted by,

$$\widehat{y}=\underset{i=1,...,q}{\operatorname{argmax}}\, p\,(i|z)\,. \tag{8}$$

C. ESC-BASED EXEMPLAR SELECTION AND RANKING ALGORITHM

Third, exemplar selection is employed to mitigate catastrophic forgetting from the perspective of data. It aims to select a small amount of exemplar data points that represent the whole data set, which can be categorized into two types. The first type of methods assumes that the data points lie around centers [13], [38]–[40]. And these centers are served as exemplars. The second type of methods assumes that the data points are distributed in one or several low-dimensional spaces [18], [41]–[43]. ESC [18] is an excellent exemplar selection method of recent years. The method is guaranteed to find sufficient and diverse data points for each subspace and can rank the exemplars.

After the MAS module for exemplar set $\boldsymbol{R}^{1,p-1}$ and new data $\boldsymbol{X}^{p,q}$ is trained, we perform ESC-based exemplar selection and ranking. The corresponding algorithm is given in Table 2. We use the strategy of fixed exemplar set size of K, the number of exemplars per class is $n=K/q$. First, we remove the relatively unimportant exemplars to allocate space for the new classes. Second, exemplar selection is performed by minimizing the objective function below. For images $\boldsymbol{X}_k^{p,q}$ of class k, the features $\boldsymbol{Z}_k^{p,q}$ are extracted by the representation learning module.

$$f_\eta(\boldsymbol{z}_j,\boldsymbol{Z}_k^{p,q})=\min_{\boldsymbol{h}_j\in R^N}\left(\|\boldsymbol{h}_j\|_1+\frac{\eta}{2}\left\|\boldsymbol{z}_j-\sum_{z_i\in Z_k^{p,q}}h_{ij}\boldsymbol{z}_i\right\|_2^2\right), \tag{9}$$

where $\boldsymbol{h}_j=[h_{1j},h_{2j},\ldots,h_{Nj}]^{\mathrm{T}}$, N is the number of data points for classes k, $\boldsymbol{z}_j$ is the extracted feature of a data point $\boldsymbol{x}_j\in\boldsymbol{X}_k^{p,q}$, $\boldsymbol{Z}_k^{p,q}$ are the extracted features of exemplars $\boldsymbol{R}_k^{p,q}$, $\eta>1$ is a hyperparameter. As a result, the exemplars are extracted iteratively, and the acquired set is a prioritized list. Third, the exemplars of the new classes are added to the exemplar set.

IV. EXPERIMENTS

A. EXPERIMENTAL SETTINGS AND DATASETS

Our experiments are conducted on five publicly available datasets, Caltech-256 dataset [44], CIFAR 100 dataset [45] and Oxford Flowers 102 dataset [46], Describable Textures Dataset [47], and Stanford Dogs Dataset [48] for general-purpose and fine-grained image recognition.

Caltech-256 dataset [44] and CIFAR 100 dataset [45] are two general-purpose image recognition datasets. The Caltech-256 dataset contains 30607 images with 256 object classes and a background class. For each class, it has 80 images at least. In this paper, 2/3 of images for a class are used for training, and the remaining images are employed for testing. The CIFAR 100 dataset has 60000 32×32 color images in 100 classes totally. For each class, 500 samples are used for training, and 100 samples are utilized for testing.

To demonstrate the performance for find-grained image recognition, Oxford Flowers 102 dataset [46], Describable Textures Dataset [47], and Stanford Dogs Dataset [48] are introduced. The Oxford Flowers 102 dataset contains 8189 images and 102 classes. For each class, 20 images are utilized for training and the rest are used for testing. The Describable Textures Dataset is a texture dataset and has 47 classes. As for each class, 40 images are utilized for training, and 40 images are employed for testing. The Stanford Dogs Dataset contains 120 breeds of dogs. For the training dataset, each class has 100 samples. And for the testing dataset, the average number per class is 71.5.

Regarding the Caltech-256 dataset, experiments are conducted with incremental intervals of about 5 and 50 classes. The total memory size is 2570, and the upper bound per class is 30. In terms of CIFAR 100 dataset, the model is trained in batches of 20 classes, the maximal size of exemplar set is 2000, and the resolution of images is resized to 224. With respect to Oxford Flowers 102 dataset, the incremental intervals are 1 and 20 classes, the maximal memory size is 408, and the upper bound per class is 6. For the Describable Textures Dataset, the incremental interval is about 10 classes, the maximal memory size is 235, and the upper bound per

class is 15. For the Stanford Dogs Dataset, the incremental interval is 10 classes, the maximal memory size is 1200, and the upper bound per class is 30.

For performance assessment, the proposed ESRIL approach is implemented in two different runs, i.e. ESRIL-FC and ESRIL-NCMC. The first one uses the FC layer of MAS as the classifier, and the second one employs NCMC as the classifier. For the CIFAR 100 dataset, the images are pre-processed. The Describable Textures Dataset is a texture attribute dataset. Regarding the two datasets, images feeding to the network have a large difference with images in the ImageNet dataset. It takes more epochs to train the network for obtaining good feature representation ability. Consequently, we use the FC layer of MAS for classification directly. And for the other three datasets, the NCMC classifier is used. We run the experiments three times with different class orders and report the average accuracy curve and corresponding standard deviation except for the experiments of studying the effect of the backbone structure.

First, differential analysis is performed. The impact of the different backbone structures is investigated, and the structures include AlexNet [49], Resnet [50], and EfficientNet [51]. To study the effect of components, we conduct experiments on the Caltech-256 dataset, in which ESC and MAS are isolated for experimental evaluation. For exemplar selection, AP and herding methods are used for comparing. In terms of representation learning, we compare with Piggyback and LwF. Second, we compare our proposed approach with baselines on five datasets. The baselines include LwF [3] and iCaRL [13].

To simulate the scenario in the real world, the class-incremental learning methods are assessed from models trained on the data of old classes, and then the data of new classes arrive in batches. The experiments are conducted on a workstation with 2 Intel Xeon Processor E5-2620V4 CPUs 2 NVIDIA Titan Xp GPUs, and 128GB RAM.

B. EVALUATION METRICS

In this work, accuracy at each incremental step is utilized as the main metric,

$$Acc = \frac{tp + tn}{tp + fp + tn + fn}, \tag{10}$$

where tp, fp, tn and fn are the true positive, false positive, true negative and false negative samples. The performance can be reflected by accuracy at each incremental step in detail.

And for simplicity, the mean incremental accuracy and final incremental accuracy are employed as evaluation metrics. The mean incremental accuracy is defined as,

$$MeanAcc = \left(\sum_{i=1}^{K} Acc_i\right) \Big/ K, \tag{11}$$

where K is the number of incremental class batches. And the final incremental accuracy is,

$$FinAcc = Acc_K. \tag{12}$$

TABLE 3. Mean incremental accuracy (%) of ESRIL for different backbone structures on three datasets. The experiments are conducted in incremental intervals of about 50, 20 and 20 classes.

Dataset	AlexNet	Resnet	EfficientNet
Caltech-256 dataset	62.47	78.94	81.74
CIFAR 100 dataset	69.52	67.53	80.03
Oxford Flowers 102 dataset	72.93	72.71	69.89

TABLE 4. Final incremental accuracy (%) of ESRIL for different backbone structures on three datasets. The experiments are conducted in incremental intervals of about 50, 20 and 20 classes.

Dataset	AlexNet	Resnet	EfficientNet
Caltech-256 dataset	52.77	73.57	74.28
CIFAR 100 dataset	59.16	51.66	68.51
Oxford Flowers 102 dataset	56.95	56.20	45.32

As we can see, the mean incremental accuracy is used to measure the total performance of the incremental learning process. The final incremental accuracy is employed to evaluate the residual performance after the incremental learning process.

C. DIFFERENTIAL ANALYSIS

In the incremental learning process, the backbone structure is a key factor for classification performance. Tables 3 and 4 summarize the experimental results of ESRIL for different backbone structures on three datasets. According to Table 3 and 4, EfficientNet on Caltech-256 and CIFAR 100 dataset has excellent performances, while AlexNet and Resnet predict better than EfficientNet on Oxford Flowers 102 dataset. The reason is that EfficientNet has a more complex structure and needs more samples of a class to train. Therefore, we choose EfficientNet as the backbone for the dataset with more samples and use AlexNet or Resnet for the dataset with fewer samples.

To study the effect of components of exemplar selection and representation learning, we perform isolated experiments. The accuracy curves of isolated experiments and the corresponding average selection rates on the Caltech-256 dataset are shown in Fig. 2. As we use all the training data in the first run, the average selection rate is 100%. Generally, our method outperforms other methods according to Fig. 2. AP-MAS-NCMC suffers from catastrophic forgetting. The reason is that Affinity Propagation cannot extract enough exemplars. By herding, enough exemplars can be extracted. As a result, the classification performance is improved. Furthermore, the final incremental accuracy of ESRIL-NCMC is 8.6% better than that of herding-MAS-NCMC. The reason is that ESC can extract enough exemplars appropriately covering the whole dataset and rank the exemplars. ESRIL-NCMC outperforms ESC-Piggyback-NCMC, indicating MAS can identify the important weights, and adapt to the new classes properly. ESRIL-NCMC performs

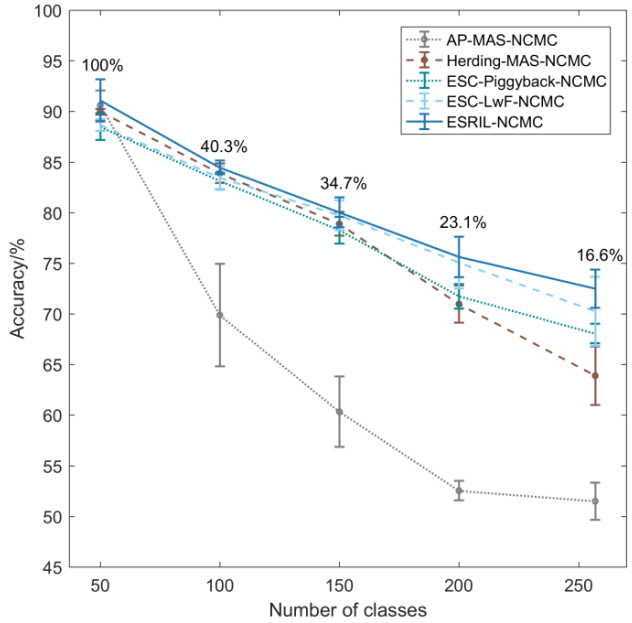

FIGURE 2. Accuracy curves of isolated experiments on the Caltech-256 dataset (incremental interval: ∼50 classes). The numbers above the curves are the average selection rates for old classes.

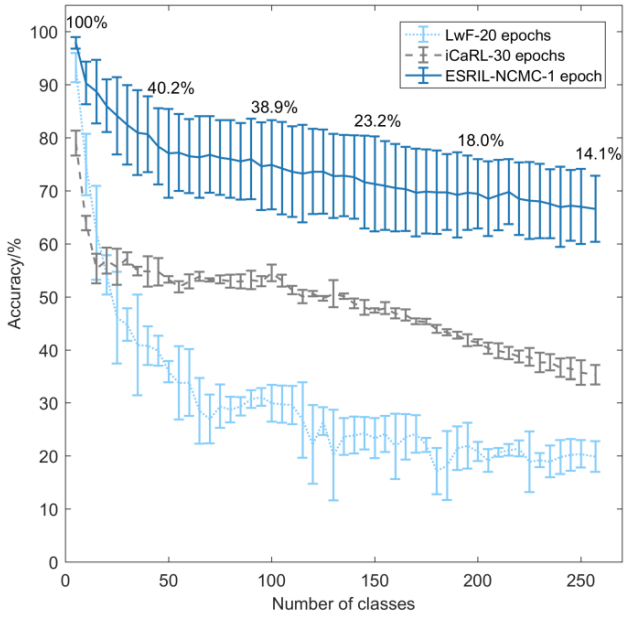

FIGURE 3. Accuracy curves on the Caltech-256 dataset (incremental interval: ∼5 classes). The numbers above the curves are the average selection rates for old classes.

better than ESC-LwF-NCMC. A possible reason could be that MAS is more suitable for the Caltech-256 dataset, a general-purpose image recognition dataset.

D. EXPERIMENTS FOR GENERAL-PURPOSE IMAGE RECOGNITION

To demonstrate the approach's effectiveness for general-purpose image recognition, experiments are performed on the Caltech-256 dataset and CIFAR 100 dataset. General-purpose classification is used to classify various classes, and the classes usually have a large visual difference. The experimental results and training epochs of the models are given in Figs. 3 and 4.

As shown in Figs. 3 and 4, performances of our proposed methods are better than those comparing methods in the trend of whole the accuracy curve. LwF has a nice performance in the early stage of the incremental learning process. However, the accuracy drops quickly in the later stage. And obviously, iCaRL outperforms LwF. The reason is that, by selecting exemplars, iCaRL can predict the previous classes more accurately. It suggests that exemplars are vital for the prediction performance of the model. On the Caltech-256 dataset, the training epochs for iCaRL and our method are 30 and 1, and on the CIFAR 100 dataset, the training epochs are 50 and 20. The correct training epoch of ESRIL-FC is 20. Thus our method has a lower computational complexity. This is because we use the pre-trained DCNN as the backbone. The pre-trained DCNN has excellent feature extraction ability and can enhance the training efficacy. On the Caltech-256 dataset, the accuracy has a large variation. The possible reason is that the class order has slightly high influence with a small incremental interval for ESRIL-NCMC.

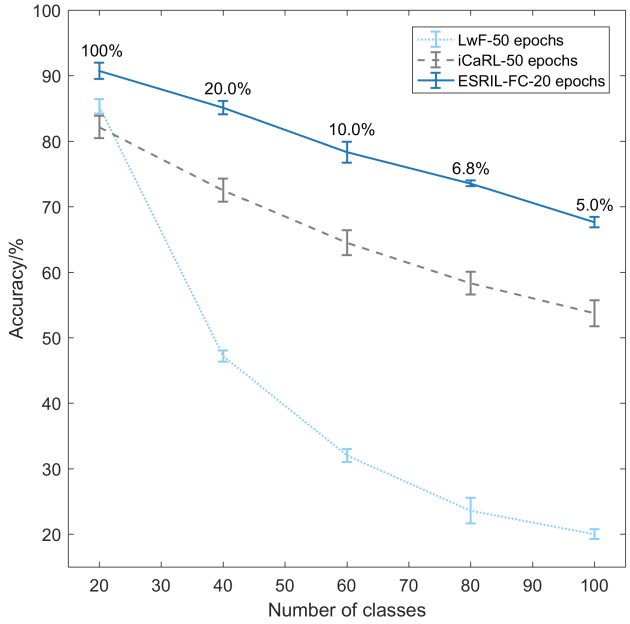

FIGURE 4. Accuracy curves on CIFAR 100 dataset (incremental interval: 20 classes). The numbers above the curves are the selection rates for old classes.

E. RESULTS ON FINE-GRAINED IMAGE RECOGNITION

Finally, experiments are performed on three fine-grained image recognition datasets, Stanford Dogs Dataset, Oxford Flowers 102 dataset, and Describable Textures Dataset. The objective of fine-grained classification is to classify subclasses of a superior class, such as dog breeds. Visually, these subclasses only have subtle differences in particular parts. Figs. 5, 6, and 7 show the corresponding experimental results of the three datasets and the training epochs of the models.

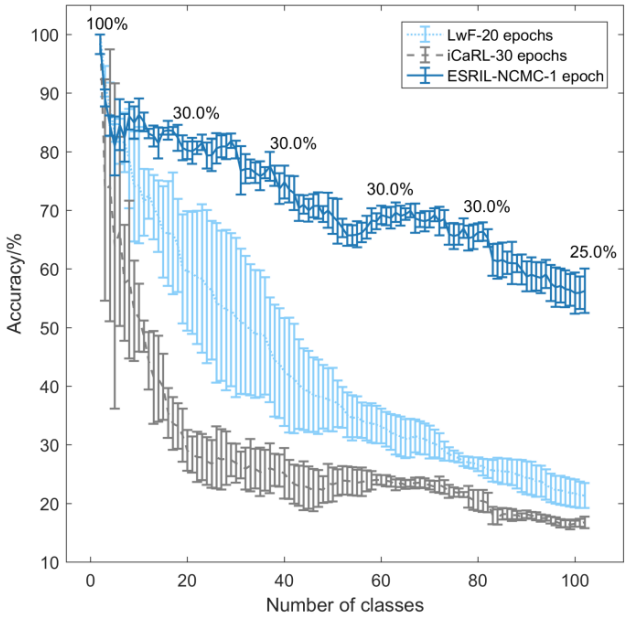

FIGURE 5. Accuracy curves on Oxford Flowers 102 dataset (incremental interval: 1 class). The numbers above the curves are the selection rates for old classes.

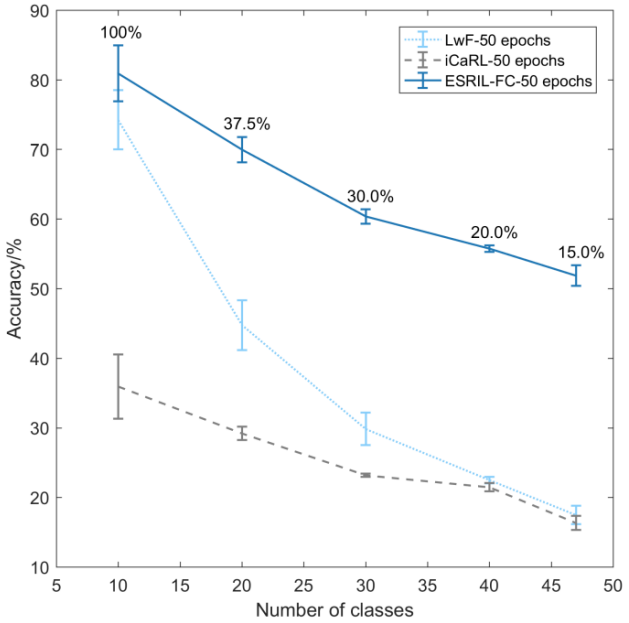

FIGURE 7. Accuracy curves on the Describable Textures Dataset (incremental interval: ∼10 classes). The numbers above the curves are the selection rates for old classes.

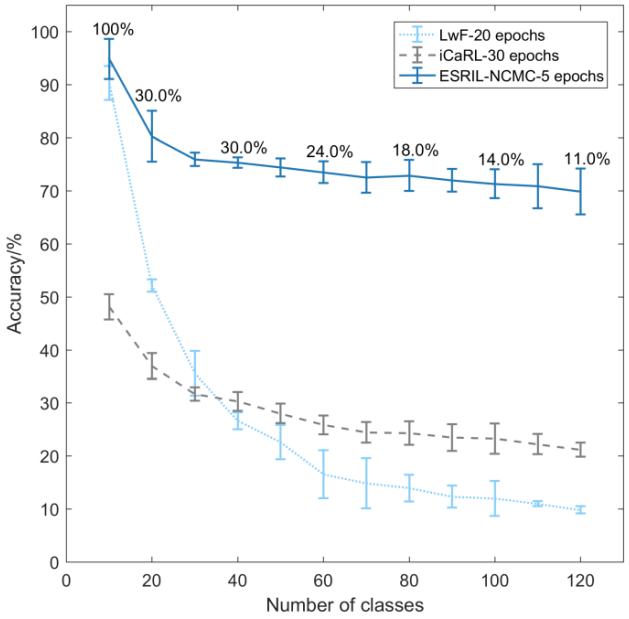

FIGURE 6. Accuracy curves on the Stanford Dogs Dataset (incremental interval: 10 classes). The numbers above the curves are the selection rates for old classes.

For all these datasets, our method beats iCaRL and LwF. On the Oxford Flowers 102 dataset, though the incremental learning has been performed for 101 runs, ESRIL-NCMC has a superior performance. The classification task on Stanford Dogs Dataset is relatively complex. Therefore, we train ESRIL-NCMC for five epochs. And in the final incremental run, only 11.0% of the samples of the old classes are preserved. Despite that, ESRIL-NCMC still performs well, indicating the abilities to mitigate class imbalance and catastrophic forgetting for MAS and NCMC. Regarding the Describable Textures Dataset, we train ESRIL-Classifier for 50 epochs to obtain a good performance. On the Describable Textures Dataset and Stanford Dogs Dataset, iCaRL has not the desired performance, possibly due to that the pre-trained backbone is not adopted in iCaRL.

V. CONCLUSION

In this work, a class-incremental learning approach ESRIL is proposed, which contains three key components, i.e. MAS, ESC, and NCMC, and have achieved a significantly improved performance. Specifically, MAS with a pre-trained backbone has made the model have robust network parameters and excellent representation learning ability. ESC is a crucial module for keeping the performance of previous classes, which can extract and also rank sufficient and diverse exemplars for each class. If the samples are not much different from those in ImageNet, we can employ the NCMC classifier instead of the FC layer of MAS to save the training cost, and to alleviate the effect of class imbalance. Experiments on five datasets for various applications have fully demonstrated the effectiveness of our proposed class-incremental approach for both general-purpose and fine-grained image recognition tasks.

In the future, we will continue working on the end-to-end training algorithm for the three components. Moreover, the exemplar selection algorithm will also be investigated, considering the representativeness, and other factors.

REFERENCES

[1] S. Hou, X. Pan, C. Change Loy, Z. Wang, and D. Lin, "Lifelong learning via progressive distillation and retrospection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 437–452.

[2] L. Fu, H.-H. Hsu, and J. C. Principe, "Incremental backpropagation learning networks," *IEEE Trans. Neural Netw.*, vol. 7, no. 3, pp. 757–761, May 1996.

[3] Z. Li and H. Derek, "Learning without forgetting," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2016, pp. 614–629.

[4] D. Li, Y. Yang, Y. Song, and T. M. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *Proc. 32nd AAAI Conf. Artif. Intell.*, Apr. 2018, pp. 3490–3497.

[5] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, "Overcoming catastrophic forgetting in neural networks," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 13, pp. 3521–3526, Mar. 2017.

[6] A. Mallya and S. Lazebnik, "PackNet: Adding multiple tasks to a single network by iterative pruning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7765–7773.

[7] F. Zenke, P. Ben, and G. Surya, "Continual learning through synaptic intelligence," in *Proc. Int. Conf. Mach. Learn (ICML)*, Aug. 2017, pp. 3987–3995.

[8] R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars, "Memory aware synapses: Learning what (not) to forget," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 139–154.

[9] A. Mallya, D. Davis, and S. Lazebnik, "Piggyback: Adapting a single network to multiple tasks by learning to mask weights," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 67–82.

[10] S. A. Rebuffi, H. Bilen, and A. Vedaldi, "Learning multiple visual domains with residual adapters," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2017, pp. 506–516.

[11] S.-A. Rebuffi, A. Vedaldi, and H. Bilen, "Efficient parametrization of multi-domain deep neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8119–8127.

[12] A. Rosenfeld and J. K. Tsotsos, "Incremental learning through deep adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 3, pp. 651–663, Mar. 2020.

[13] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "ICaRL: Incremental classifier and representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2001–2010.

[14] F. M. M. J. Castro Marín-Jiménez and N. C. K. Guil Schmid Alahari, "End-to-end incremental learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 233–248.

[15] H. Chen, R. Wang, S. Shan, and X. Chen, "Exemplar-supported generative reproduction for class incremental learning," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Sep. 2018, pp. 98–100.

[16] S. Hou, X. Pan, C. C. Loy, Z. Wang, and D. Lin, "Learning a unified classifier incrementally via rebalancing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 831–839.

[17] Y. Wu, Y. Chen, L. Wang, Y. Ye, Z. Liu, Y. Guo, and Y. Fu, "Large scale incremental learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 374–382.

[18] C. You, C. Li, D. P. Robinson, and R. Vidal, "Scalable exemplar-based subspace clustering on class-imbalanced data," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 67–83.

[19] M. Artac, M. Jogan, and A. Leonardis, "Incremental PCA for on-line visual learning and recognition," in *Proc. Object Recognit. Supported Interact. Service Robots*, Aug. 2002, pp. 781–784.

[20] S. S. Sarwar, A. Ankit, and K. Roy, "Incremental learning in deep convolutional neural networks using partial network sharing," *IEEE Access*, vol. 8, pp. 4615–4628, 2020.

[21] K. Shmelkov, C. Schmid, and K. Alahari, "Incremental learning of object detectors without catastrophic forgetting," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3400–3409.

[22] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, "Incremental learning for robust visual tracking," *Int. J. Comput. Vis.*, vol. 77, nos. 1–3, pp. 125–141, May 2008.

[23] S. Baluja, "Population-based incremental learning. A method for integrating genetic search based function optimization and competitive learning," Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-CS-94-163, 1994.

[24] P. Hetherington, "Is there 'catastrophic interference' in connectionist networks?" in *Proc. 11th Annu. Conf. Cognit. Sci. Soc.*, 1989, pp. 26–33.

[25] R. Ratcliff, "Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions," *Psychol. Rev.*, vol. 97, no. 2, pp. 285–308.

[26] A. Robins, "Catastrophic forgetting, rehearsal and pseudorehearsal," *Connection Sci.*, vol. 7, no. 2, pp. 123–146, Jun. 1995.

[27] R. French, "Catastrophic forgetting in connectionist networks," *Trends Cognit. Sci.*, vol. 3, no. 4, pp. 128–135, Apr. 1999.

[28] C. A. Kortge, "Episodic memory in connectionist networks," in *Proc. 12th Annu. Conf. Cogn. Sci. Soc.*, 1990, pp. 764–771.

[29] J. K. Kruschke, "ALCOVE: An exemplar-based connectionist model of category learning," *Psychol. Rev.*, vol. 99, no. 1, pp. 22–29, 1992.

[30] R. M. French, "Pseudo-recurrent connectionist networks: An approach to the 'Sensitivity-Stability' dilemma," *Connection Sci.*, vol. 9, no. 4, pp. 353–380, Dec. 1997.

[31] R. Polikar, L. Upda, S. S. Upda, and V. Honavar, "Learn++: An incremental learning algorithm for supervised neural networks," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 31, no. 4, pp. 497–508, Nov. 2001.

[32] R. Coop, A. Mishtal, and I. Arel, "Ensemble learning in fixed expansion layer networks for mitigating catastrophic forgetting," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 10, pp. 1623–1634, Oct. 2013.

[33] S. Ruping, "Incremental learning with support vector machines," in *Proc. IEEE Int. Conf. Data Mining*, Jul. 1999, pp. 1–6.

[34] G. Cauwenberghs and T. Poggio, "Incremental and decremental support vector machine learning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Nov. 2001, pp. 409–415.

[35] T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka, "Distance-based image classification: Generalizing to new classes at near-zero cost," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2624–2637, Nov. 2013.

[36] M. Ristin, M. Guillaumin, J. Gall, and L. Van Gool, "Incremental learning of random forests for large-scale image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 490–503, Mar. 2016.

[37] S. Qiao, C. Liu, W. Shen, and A. Yuille, "Few-shot image recognition by predicting parameters from activations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7229–7238.

[38] L. Kaufman and J. R. Peter, "Clustering by means of medoids," in *Statistical Data Analysis based on the L1 Norm*, Y. Dodge, Ed. Basel, Switzerland: Birkhäuser, 1987, pp. 405–416.

[39] B. J. Frey and D. Dueck, "Mixture modeling by affinity propagation," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2006, pp. 379–386.

[40] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, Feb. 2007.

[41] C. Boutsidis, M. W. Mahoney, and P. Drineas, "An improved approximation algorithm for the column subset selection problem," in *Proc. 20th Annu. ACM-SIAM Symp. Discrete Algorithms*, Jan. 2009, pp. 968–977.

[42] E. Elhamifar, G. Sapiro, and R. Vidal, "See all by looking at a few: Sparse modeling for finding representative objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1600–1607.

[43] H. Liu, Y. Liu, and F. Sun, "Robust exemplar extraction using structured sparse coding," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 8, pp. 1816–1821, Aug. 2015.

[44] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. 7694, 2007.

[45] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, Tech. Rep., 2009, vol. 1, no. 4.

[46] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. 6th Indian Conf. Comput. Vis., Graph. Image Process.*, Dec. 2008, pp. 722–729.

[47] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3606–3613.

[48] A. Khosla, N. Jayadevaprakash, B. Yao, and F. F. Li, "Novel dataset for fine-grained image categorization: Stanford dogs," in *Proc. CVPR Workshop Fine-Grained Vis. Categorization (FGVC)*, Jun. 2011, pp. 1–2.

[49] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2012, pp. 1097–1105.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[51] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn (ICML)*, Jun. 2019, pp. 6105–6114.