Few-Shot Class-Incremental Learning via Relation Knowledge Distillation

Songlin Dong[1 #], Xiaopeng Hong[2 #], Xiaoyu Tao[1], Xinyuan Chang[3], Xing Wei[3], Yihong Gong[3*]
[1]College of Artificial Intelligence, Xi'an Jiaotong University
[2]School of Cyber Science and Engineering, Xi'an Jiaotong University
[3]School of Software Engineering, Xi'an Jiaotong University
{dsl972731417,txy666793,cxy19960919}@stu.xjtu.edu.cn
{hongxiaopeng,weixing,ygong}@mail.xjtu.edu.cn

Abstract

In this paper, we focus on the challenging *few-shot class-incremental learning* (FSCIL) problem, which requires to transfer knowledge from old tasks to new ones and solves catastrophic forgetting. We propose the *exemplar relation distillation incremental learning* framework to balance the tasks of old-knowledge preserving and new-knowledge adaptation. First, we construct an exemplar relation graph to represent the knowledge learned by the original network and update gradually for new tasks learning. Then an exemplar relation loss function for discovering the relation knowledge between different classes is introduced to learn and transfer the structural information in relation graph. A large number of experiments demonstrate that relation knowledge does exist in the exemplars and our approach outperforms other state-of-the-art class-incremental learning methods on the CIFAR100, miniImageNet, and CUB200 datasets.

Introduction

To date, deep Convolutional Neural Networks (CNNs) have achieved superior performances in a large number of computer vision and pattern recognition tasks (He et al. 2020; Tao et al. 2020; Krizhevsky, Sutskever, and Hinton 2012; He et al. 2015; Deng et al. 2009). Currently, the universal approach is to learn a model on a large amount of data, which is usually fixed and can not change in line with the needs of users. Practical artificial intelligence models are supposed to adapt to the ever-changing world. For example, they can continuously learn new tasks without forgetting tasks that they have learned before. Thus, incremental learning (*i.e.*, continuous learning, or lifelong learning) has attracted much attention due to the ability to perform continuous model learning in a wide range of practical applications. The scenarios of incremental learning can be briefly divided into the class-incremental (Rebuffi et al. 2017; Tao et al. 2020a; Chang et al. 2021) and the task-incremental (Rajasegaran et al. 2020) ones. This paper mainly focuses on the former one.

Most of the class-incremental learning (CIL) approaches (Rebuffi et al. 2017; Castro et al. 2018; Hou et al. 2019) learn new tasks from large scale training samples with annotations. However, in practice, it is expensive and infeasible to continuously label a steady stream of data. As a result, it is more realistic that only very small amounts of samples with annotations are available. Therefore, in this paper, we focus on this *few-shot class-incremental learning* (FSCIL) (Tao et al. 2020b) problem which is challenging but has great application prospects.

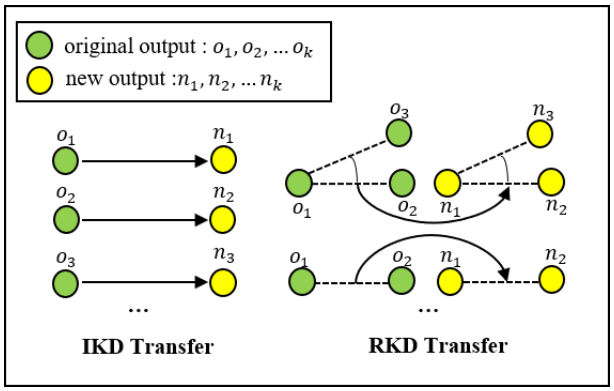

Figure 1: Comparisons of two different knowledge transfer techniques: The individual knowledge distillation (IKD) transfers the knowledge point to point (original output to new output) while the relation knowledge distillation (RKD) transfers the *structural* relations (e.g., adjacency) of the output.

Knowledge distillation is a common and effective approach to transfer knowledge from an *old* network to a *new* one in CIL (Gou et al. 2020). For example, some early CIL approaches (Rebuffi et al. 2017; Castro et al. 2018) distill the original model's soft targets directly into a new one on the output layer to preserve old knowledge. In addition, the responses of the penultimate feature layer are also used as the knowledge to guide the training of the new task model (Hou et al. 2019; Tao et al. 2020b). All these methods belong to the *individual knowledge distillation* (IKD) approaches (Park et al. 2019), which only focus on preserving isolated points in the feature (or output) space. Concretely, the IKD method only constrains the differences between the

*Yihong Gong is the corresponding author. # Songlin Dong and Xiaopeng Hong are co-first authors.

responses of the original network and the one learned in the new task for individual exemplars. As a result, the mutual relations of points, which are important to many artificial intelligence applications such as image classification and class incremental learning, are inevitably ignored.

Clearly, there is a wealth of relational information between exemplars. For example, a cat is usually more similar to a dog, than a car. Such relations can serve as useful priors for learning in sequential tasks. Thus, it provides a more effective way to constrain not only the locations of exemplars in the feature space but also their relations for the few-shot CIL problem.

Motivated by this observation, we propose a novel few-shot CIL framework, termed the *exemplar relation distillation incremental learning* (ERDIL). ERDIL uses an *exemplar relation graph* (ERG) to comprehensively explore the relations information of exemplars from the old tasks and leverages the graph-based relational knowledge distillation to effectively transfer old knowledge to the CNN model for new tasks learning. More specifically, ERG is made up of selected typical exemplars of old tasks as vertices and the edges linking vertices. The edges of ERG are weighted by the angles between the chosen exemplars. We first propose a novel degree-based exemplars selection mechanism to construct the directed ERG. Then an exemplar relation loss function is designed to transfer the comprehensive relational information embedded in the ERG for new task learning. Finally, we also introduce a metric learning loss term to avoid the ambiguities problem between old and new class data (Hou et al. 2019).

We perform extensive experiments on three commonly used FSCIL datasets including CIFAR100 (Krizhevsky and Hinton 2009), miniImageNet (Vinyals et al. 2016a), and CUB200 (Wah et al. 2011). Our ERDIL framework greatly improves the performance in terms of the accuracy. The experimental results successfully demonstrate that utilizing the relations between exemplars in the ERG is helpful for better preserving old knowledge. To summarize, our main contributions include:

- We propose to constrain the relation of exemplars, rather than their absolute positions for few-shot incremental learning and put forward a novel relation knowledge distillation based FSCIL framework.
- We propose a degree-based graph construction algorithm to model the relation of the exemplars.
- We make comprehensive comparisons between the proposed method with the state-of-the-art FSCIL methods and also regular CIL methods.

Related Work

Multi-task Incremental Learning

Multi-task incremental learning has explored a variety of strategies to prevent networks from forgetting previously learned tasks. Generally, the literature deals with this problem from the following three perspectives: (1) architectural approaches: improving the network architecture, such as dynamic expansion, network pruning, and parameter masking methods to reduce forgetting (Mallya and Lazebnik 2018; Mallya, Davis, and Lazebnik 2018; Serrà et al. 2018; Yoon et al. 2017). (2) Rehearsal approaches which periodically replay the memory for the past experiences of the old tasks and constrain their losses during training the new tasks (Lopez-Paz et al. 2017; Chaudhry et al. 2018; Shin et al. 2017; Wu et al. 2018). (3) Regularization approaches that constrain the network parameters, loss, and output logits update by the learned knowledge (Li and Hoiem 2018; Kirkpatrick et al. 2017; Zenke, Poole, and Ganguli 2017; Lee et al. 2017).

However, the *multi-task* methods are based on task-level where the network's neural resources are allocated to each task correspondingly while in the scenario where the FSCIL setting is single-task and multi-class. As a consequence, we have to exclude them for comparison in the experiment.

Class-Incremental Learning

Class-Incremental Learning: Class-incremental learning aims to learn a unified classifier for all the classes. Knowledge distillation is a popular technique to solve the catastrophic forgetting problem. Those approaches usually store the old class exemplars to compute the distillation loss. For example, iCaRL (Rebuffi et al. 2017) maintains an 'episodic memory' to mitigate forgetting and uses the nearest-neighbor classifier to learn the new classes. EEIL (Castro et al. 2018) combines the cross-entropy and distillation loss to end-to-end learning. CIL problem exists *critical bias* issue (Hou et al. 2019) because of the imbalanced number of training samples of old and new classes. LUCIR (Hou et al. 2019) adopts a cosine distance metric to reduce classification layer bias and applies distillation to the feature space rather than the output logits.

In short, different from the *individual knowledge distillation* CIL approaches, the proposed ERDIL framework alleviates forgetting by maintaining the relations of feature space.

Few Shot Class-Incremental Learning: FSCIL setting is first proposed in TOPIC (Tao et al. 2020b) which not only focuses on overcoming the catastrophic forgetting problem but also focuses on incrementally learning new classes from very few labeled samples. TOPIC proposed a single *neural gas* (NG) network to learn feature space typologies for knowledge representation, and preserves the stabilization and enhances the adaptation by adjusting NG. Later, FSC (Zhao et al. 2020) proposed a novel multi-model FSCIL method. They make the base knowledge space and new task knowledge space integrate into one space, namely composite representation space and achieves a very good performance though NCM (Mensink et al. 2013) classifier training the new tasks.

In short, our ERDIL framework can be applied to single-model and multi-model FSCIL scenarios and achieve relatively better performance.

Few-Shot Learning

Few-shot learning (FSL) requires the model to be very adaptable to a small number of unknown novel samples. At present, there are mainly two categories. Meta-learning based approaches (Finn, Abbeel, and Levine 2017; Nichol

and Schulman 2018): it typically involves an individual meta-learner model that is given a few novel training samples of a new task and tries to quickly learn a learner model that "solves" this new task. Metric-learning based approaches (Snell, Swersky, and Zemel 2017; Sung et al. 2018): Metric learning approaches try to learn feature representations that preserve the class neighborhood structure. Recently, some FSL works (Gidaris and Komodakis 2018; Ren et al. 2019) attempt to learn a unified model capable of recognizing both the base and novel classes while those are not able to keep learning new tasks because they rely on the base class for sampling meta-learning tasks. As a consequence, these few-shot learning works can not be directly applied to FSCIL.

Knowledge Distillation

Knowledge distillation, as a typical model compression and acceleration method, has attracted much attention from the community by learning small student models from large teacher models (Gou et al. 2020). Knowledge distillation can be divided into three categories according to 'knowledge'. (1) Logits-based knowledge utilizes the final output layer as supervision information to generate student network (Hinton, Vinyals, and Dean 2015; Guo et al. 2020). (2) Feature-based knowledge approaches: Feature layer as intermediate representations is used to train the student network (Romero et al. 2014; Zagoruyko and Komodakis 2016; Heo et al. 2019). (3) Relation-based knowledge further explores the structures between different layers and data samples (Yim et al. 2017; Park et al. 2019; Ramakrishnan et al. 2020). Compared with these works, we focus on a more difficult *few-shot incremental learning* problem where what knowledge to transfer and how to effectively transfer knowledge is equally important.

Methodology

In *few-shot class-incremental-learning*, a model learns tasks continually. Each task contains a batch of new classes and each class only contains a few samples with labels. Concretely, the FSCIL setting is defined as follows. Initially, assume that there is a batch of labeled training tasks $X^{(1)}, X^{(2)}, \cdots$, where $X^{(t)} = \{(\mathbf{x}_i^{(t)}, y_i^{(t)})\}_{i=1}^{|X^{(t)}|}$, and $C^{(t)}$ is the set of classes of the t-th training set. There is no overlap between the categories of different tasks so that $\forall i, j,\ C^{(i)} \cap C^{(j)} = \varnothing$. Only the training set of the first task $X^{(1)}$ has large-scale training data (*a.k.a.*, the *base task*), while other subsequent tasks $X^{(t)}$ (*a.k.a.*, the *new task*) just contains a few samples. The model is continually trained on $X^{(1)}, X^{(2)}, \cdots$ with a *unified* classification layer. Only $X^{(t)}$ is available at the t-th training session. After the training on $X^{(t)}$ is done, the model is tested to recognize all encountered classes $C^{(1)}, \cdots, C^{(t)}$. In the *C-way K-shot* FSCIL setting (Tao et al. 2020b; Zhao et al. 2020), the number of classes in $C^{(t)}$ is C and the number of training samples per class is K for each $X^{(t)}, t > 1$.

Overall Framework

A deep neural network can be seen as a composition of a feature extractor $f(\cdot;\theta)$ with parameter set θ and a classification head with parameter set ϕ, which produces the output logits $o(x;\theta,\phi) = \phi^T f(\cdot;\theta)$. Let $\Theta = \{\theta, \phi\}$ denotes the entire parameter set. Firstly, we train $\Theta^{(1)}$ on $X^{(1)}$ with the cross-entropy loss. Then we incrementally finetune the model on $X^{(2)}, X^{(3)}, \cdots$, and get $\Theta^{(2)}, \Theta^{(3)}, \cdots$. At the t-th task ($t > 1$), the output layer is expanded for new classes by adding $|C^{(t)}|$ output neurons.

Our goal is to learn a new set of tasks while preserving the model performance on the old tasks. To address this purpose, we proposed the ERDIL framework. First, a novel selection mechanism based on the degree is proposed to construct the *exemplar relation graph* (ERG). ERG is made up of those selected exemplars, which are the most typical exemplars for characterizing the feature space of old tasks, and the edges, which are weighted by the direction between the chosen exemplars. Moreover, we design an exemplar relation loss function to transfer the comprehensive relational information embedded in the ERG for new task learning. Finally, the feature map's plasticity is maintained by the metric learning loss term. Let G^t denote the exemplar relation graph constructed by old class exemplars. The overall loss function at the new task $(t+1)$ is defined as:

$$\ell(X^{(t+1)}, G^{(t)}; \Theta^{(t+1)}) = \ell_{CE}(X^{(t+1)}; \Theta^{(t+1)}) + \lambda_1 \ell_{ERL}(G^{(t)}; \Theta^t, \Theta^{(t+1)}) + \lambda_2 \ell_{ML}(X^{(t+1)}, G^{(t)}; \Theta^{(t+1)}), \quad (1)$$

The ℓ_{CE} is the standard cross-entropy loss in the classification task. ℓ_{ERL} is the exemplar relation loss term applied to $G^{(t)}$ and ℓ_{ML} is applied for $X^{(t+1)}$ and $G^{(t)}$ to enhance the plasticity of model for training new task. The hyperparameter λ_1 and λ_2 are used for balancing the strength of two loss terms. We elaborate our approach in the following subsections.

Exemplar Selection for Exemplar Relation Graph

Generally, a main component for knowledge distillation methods is randomly selected a set of exemplars from the old classes to store and compute the distillation loss by these exemplars. However, the randomly-selected exemplars can not well represent data of different classes from the old network in the FSCIL scenarios. Instead, we represent knowledge by constructing the exemplar relation graph (ERG).

To construct the ERG, we first select the exemplars from the base class training set as the vertices. Therefore, we propose a selection mechanism based on the degree. For each class of the base class, we generate a nxn two-dimensional matrix P where n is the number of exemplars and the P is initialized to 0 at first.

Given the number of selected exemplars K. First, update matrix P: The $P(p,q)$ is set to 1 if vector p,q are k-nearest neighbors to each other, or 0 otherwise. Second, degree selection: Calculate the degree of all vertexes in the matrix, $d_p = \sum_{q=1}^{w} P(p,q)$ where w is current total number of exemplars. We assume that the neighbor exemplars with the

same degree are similar so that if $d_p = d_q$ and $P(p,q) = 1$, delete the vertex p. The remaining exemplars form a new matrix and repeat the above steps several times until the number of exemplars is K.

As a result, each class only contains the K most representative exemplars to represent the class feature space. Compared with randomly selection, we also removed similar redundant exemplars. Then all these exemplars construct the directed exemplar relation graph $G = \langle V, E \rangle$ where each vertex $v_p \in V$ is the feature vector of an exemplar and the edge e_{pq} is define as:

$$e_{pq} = \frac{v_p - v_q}{\|v_p - v_q\|_2}. \tag{2}$$

For the new class of exemplars, directly become the ERG's vertexes, and calculate the edges for them. Then we use G^t in computing the relation loss as described in the following sections.

Distilling Knowledge in Exemplar Relation Graph

In popular incremental learning approaches (Rebuffi et al. 2017; Castro et al. 2018; Hou et al. 2019; Tao et al. 2020b), often store a certain amount of old class exemplars to constrain the training of Θ. There is a certain relation between these old exemplars, even between the new task's data. For example, after learning about cats' knowledge, it should be easier to learn about dogs.

Given an exemplar relation graph G^t, we believe that the relation knowledge distillation can transfer structural knowledge using mutual relations of exemplars in the original's output presentation. The structural knowledge in G^t can be divided into two categories, absolute relations, and relative relations. The absolute relations is the distance formed by the double vertexes in G^t and the relative relations is the angle formed by the triplet vertexes (a pair edges) in G^t. In incremental learning, we tend to maintain the relationship between angle more than distance relations. First, the angle relations are more flexible because when the angle relationship remains unchanged, the overall scaling of the distance will not destroy the integral feature space. Second, excessive constraints on distance relations can influence the plasticity of feature space which will limit the learning of new classes. As a result, In our ERDIL framework, we will explore the angle relations knowledge in G^t.

Exemplar Relation Loss: In the incremental problem set, we have the feature extractor $f(\cdot;\theta)$ and the original model Θ^t. At each new task, we train the new task model Θ^{t+1} on X^{t+1}. Given a triplet of vertexes in G^t, exemplar relation measures the angle formed by the three exemplars in the output representation space. Those are used to compose an exemplar relation function, $A(\Theta^t)$ for the original model and $A(\Theta^{t+1})$ for the new task model.

The original model $A(\Theta^t)$ is defined as:

$$A(p,q,z;\Theta^t) = \langle e_{pq}, e_{zq} \rangle,\ p,q,z \subset G^t$$
$$\text{where}\ \ e_{pq} = \frac{v_p - v_q}{\|v_p - v_q\|_2},\ e_{zq} = \frac{v_z - v_q}{\|v_z - v_q\|_2}, \tag{3}$$

and similarly for $A(\Theta^{t+1})$ as:

$$A(p,q,z;\Theta^{t+1}) = \langle e_{pq}, e_{zq} \rangle,\ p,q,z \subset G^t$$
$$\text{where}\ \ e_{pq} = \frac{v_p - v_q}{\|v_p - v_q\|_2},\ e_{zq} = \frac{v_z - v_q}{\|v_z - v_q\|_2}, \tag{4}$$

$\langle \cdot \rangle$ is the inner product of two vectors. Exemplar relation function reflects the angular relationship of the feature space, we hope this will remain unchanged in the new task of model learning. For this we define a loss, called *exemplar relation Loss* (ℓ_{ERL}) which is the ℓ_p norm between two relation function. The loss ℓ_{ERL} is given by

$$\ell_{ERL}(G^t;\Theta^t,\Theta^{t+1}) = |A(\Theta^t) - A(\Theta^{t+1})|_p. \tag{5}$$

The exemplar relation loss and exemplar relation graph provide a strong learning signal for training the new network without forgetting the original classes.

Metric Learning for Plasticity

As a commonly used technique, metric learning is widely used in incremental learning methods (Tao et al. 2020b; Hou et al. 2019; Yu et al. 2020). In this paper, we apply the modified margin ranking loss for FSCIL problem to better distinguish between old exemplars $G^{(t)}$ and new exemplars $X^{(t+1)}$:

$$\ell_{ML}(X^{(t+1)}, G^{(t)};\Theta^{(t+1)}) = \sum_{(\mathbf{x},y)\in X\cup G} \sum_{k=1}^{K} \max(0, -f(x) + f_k(x) + \kappa), \tag{6}$$

where κ is the margin threshold, $f(x)$ is the feature of x, $f_k(x)$ is one of top-K the old class features chosen as hard negatives for x.

Optimization

ERDIL integrates a CNN model and an exemplars relation graph G^t, where G^t is used to preserve the CNN's feature space manifold. Our CNN model is trained with the minibatch stochastic gradient descent (minibatch SGD) algorithm. It is less efficient to update the vertices of G^t at each iteration, as the features obtained at intermediate training sessions have not been fully optimized. Therefore, we update G^t after the training of CNN's parameters Θ^t. G^t is then used for the next new task $(t+1)$.

Comparison with Individual Knowledge Distillation Methods

In general, most of CIL (Li and Hoiem 2018; Rebuffi et al. 2017; Wu et al. 2019; Castro et al. 2018; Hou et al. 2019; Tao et al. 2020b) works are to mitigate forgetting through individual knowledge distillation.

For example, the work of iCaRL (Rebuffi et al. 2017; Castro et al. 2018) uses pre-softmax outputs (logits) for o^t and o^{t+1}, which is the logits-based Knowledge, and puts softmax (with temperature τ) to penalize the difference:

$$\ell_{IKD_1}(X,M) = \sum_{(\mathbf{x},y)\in X\cup M}\sum_{k=1}^{n} -softmax(\frac{o^t(\mathbf{x})}{\tau}) \log(softmax(\frac{o^{t+1}(\mathbf{x})}{\tau}). \quad (7)$$

CIL methods also utilize feature-based knowledge for knowledge transfer (Hou et al. 2019). These approaches use feature maps as the knowledge for f^t and f^{t+1} to supervise the training of the new task. For example, the feature-based distillation loss can be summarized as:

$$\ell_{IKD_2}(X,M) = \sum_{(\mathbf{x},y)\in X\cup M} (1 - f^t(x)^T f^{t+1}(x)). \quad (8)$$

In the formulas 7 and 8, X is the training data of the current t-th task and M is the stored exemplars come form original tasks. The $n = \sum_{i=1}^{t-1}|C^{(i)}|$ is number of the old classes. The t is the original model training on the parameter Θ^t while $t+1$ is the new task model training on the parameter Θ^{t+1}, and τ is the distillation temperature (e.g., τ= 2).

The individual knowledge approach faces critical problems when applied to FSCIL. In addition to ignoring relational knowledge, another problem is the contradiction to balance the contribution between ℓ_{DL} and ℓ_{CE}, which may result in a self-contradictory trade-off effect. Generally, learning few-shot new tasks requires sufficient learning rate and iteration times to minimize the ℓ_{CE} term, while a large learning rate finetuning will damage the stability of the pre-softmax outputs(logits) and cause the difficulty to minimize ℓ_{DL}.

Based on the above considerations, our ERDIL abandons the last output layer distillation term and adopt relation-based knowledge methods to manipulate the knowledge contained CNN's feature space. The detailed experimental comparisons are described in the following section.

Experiments

Experimental Setups

We conduct experiments on there image classification datasets CIFAR100 (Krizhevsky and Hinton 2009),miniImageNet (Vinyals et al. 2016a) and CUB200 (Wah et al. 2011). These datasets are very popular in incremental learning (Rebuffi et al. 2017; Tao et al. 2020b) and *few-shot learning* (Vinyals et al. 2016b).

CIFAR100 dataset. CIFAR100 is labeled as a subset of 80 million tiny image datasets collected by Alex Krizhevsky (Krizhevsky and Hinton 2009). It contains 60,000 RGB images over 100 classes, with 500 images per class for training and 100 images per class for testing. Each image has a size of 32×32.

miniImageNet dataset. The miniImageNet dataset is the subset of ImageNet-1k (Deng et al. 2009) which contains 60,000 color images in 100 categories. Each class has 500 images for training and 100 images for testing. Each image has a size of 84×84. Compared with CIFAR100, the Mini-Imagenet dataset is more complex and more suitable for prototyping so that is often used by *few-shot learning* (Vinyals et al. 2016b).

CUB200 dataset. CUB200 is a fine-grained dataset proposed by California Institute of Technology in 2010 (Wah et al. 2011), and is also the current benchmark dataset for fine-grained classification and recognition research. It contains 11788 bird images, including 200 bird categories, among which there are 5994 images for training and 5794 images for testing. During training, the images are resized to 256×256 and then random cropped to 244×224.

Evaluation protocol: We follow the evaluation protocols in (Tao et al. 2020b) to process these datasets. For CIFAR100 and miniImageNet datasets, 60 classes as the base classes, and the other 40 classes are equally divided for incremental learning. We adopt the *5-way 5-shot* setting so that we have 9 training tasks in total. While for the CUB200 dataset, we adopt the *10-way 5-shot* setting, by picking 100 classes as a base class and choosing the other classes into 10 new learning tasks. For all datasets, we randomly pick 5 samples per class from the original dataset for training set. At the same time, the testing set still uses the original one, which is large enough to ensure that generalization performance is evaluated to prevent over-fitting.

Training details: All our models are implemented through PyTorch and use ResNet18 or ResNet20 as our backbone network. For CIFAR100 and miniImageNet, we trained the basic model for 160 epochs using minibatch SGD with a minibatch size of 128. The learning rate is initialized to 0.1, and decreases to 0.01 and 0.001 at the 80 and 120 periods, respectively. For the CUB200 dataset, we use pre-trained ResNet18 and train the basic model with an initial learning rate of 0.05. After 15 epochs, we reduce the learning rate to 0.005 and stop training in the period 20. For each new task, we finetune our model with a learning rate of $1e^{-4}$ for 50 epochs for all three datasets. In the FSCIL setting, since the new task contains very few training samples, we use them all to construct mini-batches for incremental learning. For data augmentation, we perform standard random cropping and flipping as in (He et al. 2015; Hou et al. 2019) for all the methods and add ColorJitter on miniImageNet.

Exemplars details: As for the strategy to preserve the samples for base classes, there are two usually ways. The first way is considering a memory space with a fixed capacity. For example, BOCL (Tao et al. 2020c) learns a fixed som network notes for the base class and incrementally updates it during the new tasks. The second way is storing a constant number of samples for each class, and thus the size of memory space grows with the number of the classes. In this paper, we adopt the latter way for our experiments (e.g $N_{per} = 20\ vs.\ N_{total} = 1400$ on CIFAR100 and miniImageNet,$N_{per} = 5\ vs.\ N_{total} = 1000$ on CUB200).

Comparison details: For comparative experiments, we run the representative CIL and FSCIL approaches in our FSCIL setting, including the classics KD method: iCaRL (Rebuffi et al. 2017) and state of the art FKD methods: LUCIR (Hou et al. 2019) and TOPIC (Tao et al. 2020b). We compare our method with them and report these results of CNN predictions. While for FSC (Zhao et al. 2020), we found that is a multi-model method and uses nearest-mean-of-exemplars classification. We will compare their methods separately after the ablation study. To show the effectiveness of alleviat-

ing forgetting, we directly finetune the CNN model for new tasks without any less-forgetting techniques and we denote it as "Ft-CNN". For the "Joint-CNN", we retrain the CNN model at each task on a joint set of all encountered classes. We use the abbreviation "ours-ERL","ours-ERL++" to indicate the applied loss terms during the incremental learning.

Comparison Results

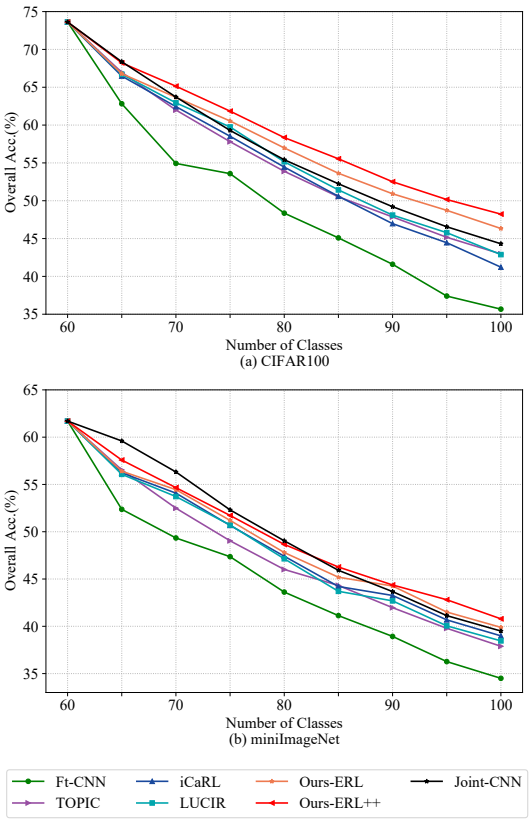

Figure 2: Comparison of the accuracies on CIFAR100 with ResNet20 and miniImageNet with ResNet18.

We compare the proposed ERDIL framework with the state of the arts on CIFAR100, miniImageNet, and CUB200 on the FSCIL setting. As the *5-shot* training samples are randomly picked, we run all methods for 5 times and report the average accuracies. The CIFR100 and miniImageNet datasets' results are shown in Figure 2. Besides, Table 1 reports the test accuracies on CUB200. It can be summarized as follows:

- On three datasets, our ERDIL outperforms all other methods and even surpasses the "Joint-CNN" methods, which is defined as the upper bound in (Tao et al. 2020b). Our method is very stable and has a continuous superiority on each new task.

[1]Noting that the results of comparative methods are higher than those reported in (Tao et al. 2020b), as we use more exemplars and different hyper-parameters settings.

- Our EKDIL methods based on relation knowledge distillation better maintains the substructure in the ERG. "our-ERL" method is average **4.86%** and **4.88%** better than the KD and FKD methods on the three datasets at end.
- On CIFAR100, EKDIL achieves the final accuracy of **48.23%**. In comparison, the iCaRL* and LUCIR* achieves the accuracy of **41.22%** and **42.88%**. ERDIL outperforms the two IKD methods by **7.01%** and of **5.35%**.
- On miniImageNet, EKDIL achieves the accuracy of **40.79%** at end of the task while the IKD methods achieves the accuracy of **38.99%** and **38.46%**. ERDIL outperforms the second best one, iCaRL*, by up to **1.80%**.
- On CUB200, ERDIL achieves the accuracy of **52.28%** at end, which outperforms the KD methods iCaRL*(**41.43%**) and FKD method LUCIR*(**40.26%**) by up to **10.85%** and **12.02%**.

Ablation Study

The effect of different loss terms: We conduct ablation studies to investigate each term's contribution to the end result gain. The experiments are performed on CIFAR100. We research the impact brought by *distillation loss* (DL), *feature distillation loss* (FDL) and our *exemplars relation loss* (ERL). Besides, for "ERL++", as it consist of the *metric learning* (ML) and *new class exemplars* (NCE) terms, we evaluate the each terms' performance separately. Table 2 reports the comparison results of different terms. We summarize it as follows:

- The "ERL" terms achieves the final accuracy, exceeding "FDL" by up to **3.45%** and "DL" by up to **5.11%**.
- For "NCE" term, which uses all the few-shot exemplars that have been trained in the previous tasks for training the current task.
- Both "ERL-ML" and "ERL-NCE" improve the performance of "ERL", and the combined form "ERL++" achieves our best accuracy **48.23%**, which exceeds "ERL" by up to **1.90%**.

Feature Space Composition with Nearest Class Mean Classification: SDC (Yu et al. 2020) proposed that softmax classifier has several fundamental drawbacks which might limit its application to *class-incremental learning*. For example, whenever new classes are added, the classifier's structural require to change. They adopted the embedding networks with a semantic drift compensation method and evaluate it for image classification by using the *nearest class mean* (NCM) (Mensink et al. 2013) which gets a considerable performance on CIL problem. Base on that work, FSC (Zhao et al. 2020) proposed a novel multi-model FSCIL method based on a composite representation space. The composite representation space is generated by integrating two space components, base knowledge space and new task knowledge space. Base the idea of multi-model feature space composition and we replace the regularization term

Method	sessions											Average	F. Acc.
	1	2	3	4	5	6	7	8	9	10	11	Acc	impro.
Ft-CNN	73.52	57.89	58.44	49.67	48.78	44.99	43.02	40.50	37.51	35.19	30.90	47.31	**+21.38**
Joint-CNN	73.52	69.04	64.34	59.88	55.68	51.57	49.04	45.98	43.39	41.25	40.20	53.99	**+12.08**
LUCIR*	73.52	63.01	53.29	47.78	47.35	47.46	45.93	44.12	43.95	42.13	40.26	49.89	**+12.02**
iCaRL*	73.52	63.32	54.93	48.85	46.89	47.74	44.99	43.67	44.17	42.84	41.43	50.21	**+10.85**
TOPIC*	73.52	66.88	61.25	56.61	52.38	48.76	45.58	42.94	41.54	39.35	37.44	51.48	**+14.84**
Ours-ERL	**73.52**	**70.12**	**65.12**	**62.01**	**58.56**	**57.99**	**56.77**	**56.52**	**55.01**	**53.68**	**50.01**	**59.93**	+2.27
Ours-ERL++	**73.52**	**71.09**	**66.13**	**63.25**	**59.49**	**59.89**	**58.64**	**57.72**	**56.15**	**54.75**	**52.28**	**61.18**	

Table 1: Comparison results[1] on CUB200 with ResNet18 using the *10-way 5-shot* FSCIL setting.

Method	CE	DL	RL	ML	sessions									Average
					1	2	3	4	5	6	7	8	9	Acc
DL	✓	✓			73.62	66.48	62.44	58.52	54.45	50.59	46.97	44.45	41.22	55.41
FDL	✓	✓			73.62	66.72	62.94	59.75	55.15	51.43	48.10	45.78	42.88	56.26
ERL	✓		✓		73.62	66.79	63.67	60.54	56.98	53.63	50.92	48.73	46.33	57.91
ERL-ML	✓		✓	✓	73.62	67.40	64.98	61.01	57.66	54.87	52.33	49.59	47.43	58.76
ERL-NCE	✓		✓		73.62	66.82	64.12	60.76	57.89	55.21	51.89	49.49	47.62	58.60
ERL++	✓		✓	✓	**73.62**	**68.22**	**65.14**	**61.84**	**58.35**	**55.54**	**52.51**	**50.16**	**48.23**	**59.29**

Table 2: Comparison results of combining different loss terms on cifar100 with ResNet20.

with our ERDIL framework. Figure 3 shows the comparative result on CIFAR100 and our ERDIL* outperforms their work on average **1.36%**.

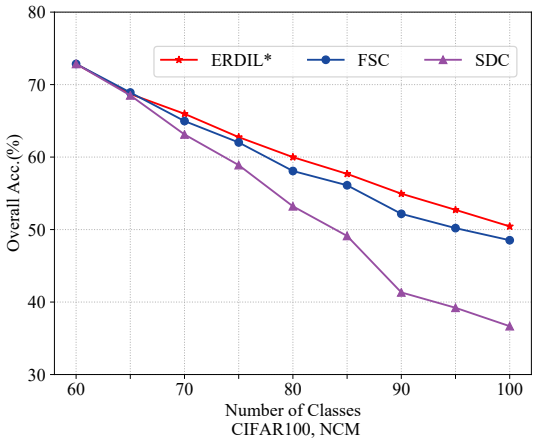

Figure 3: Comparison results under NCM classifier, evaluated on CIFAR100 dateset with ResNet18.

Comparison of Different Methods for Constructing Relation Graph: We compare different methods for constructing a relational graph on CIFAR100, Table 3 reports the accuracy achieved by using different exemplars selection methods, the random exemplars, SOM methods (Tao et al. 2020c) and our exemplar relation graph. We can obverse that constructing a relational graph affects the efficiency of knowledge transfer. Exemplars relation graph method outperforms random selection by up to **0.47%** after the final task.

Name. of *methods*	Random	SOM	ERG
Final Acc. (%)	47.76	47.96	48.23

Table 3: Classification accuracies w.r.t. different methods for constructing relational graph after learning all tasks

Conclusion

We focus on the *few-shot class-incremental learning* task and propose a framework, called ERDIL, to leverage the relation knowledge contained in CNN's feature space. ERDIL constructs the relation graph formed by different classes. We design an exemplars relation loss function to preserve and transfer the relation knowledge between different classes. A large number of experiments show that our method significantly outperforms other CIL and FSCIL methods on CIFAR100, miniImageNet, and CUB200 datasets.

It is somehow surprising to find that our ERDIL exceeded the *Joint-CNN*, which was alleged to be an empirical *upper bound* of few-shot incremental learning approaches (Tao et al. 2020b). The reason probably lies in that *Joint-CNN* is not balanced enough to treat the old and the new tasks. The models trained by the *Joint-CNN* protocol are inclined to overfit to the old class training samples and can hardly learn new ones in FSCIL. As a result, it is more appropriate to regard *Joint-CNN* as an upper bound for regular CIL, as there are enough and relatively balanced samples for both old and new classes, rather than FSCIL. Thus what the upper bound is will be a fruitful topic for future FSCIL research.

Acknowledgements

This work is funded by National Key Research and Development Project of China under Grant No. 2020AAA0105600 and 2019YFB1312000, National Natural Science Foundation of China under Grant No. 62076195 and 62006183, and China Postdoctoral Science Foundation under Grant No. 2020M683489.

References

Castro, F. M.; Marín-Jiménez, M. J.; Guil, N.; Schmid, C.; and Alahari, K. 2018. End-to-end incremental learning. In *ECCV*, 233–248.

Chang, X.; Tao, X.; Hong, X.; Wei, X.; Wei, K.; and Gong, Y. 2021. Class-Incremental Learning with Topological Schemas of Memory Spaces. In *Proceedings of the ICPR Conference on Artificial Intelligence*.

Chaudhry, A.; Ranzato, M.; Rohrbach, M.; and Elhoseiny, M. 2018. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420* .

Deng, J.; Dong, W.; Socher, R.; Li, L. J.; Li, K.; and Li, F. F. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1126–1135. JMLR. org.

Gidaris, S.; and Komodakis, N. 2018. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4367–4375.

Gou, J.; Yu, B.; Maybank, S. J.; and Tao, D. 2020. Knowledge Distillation: A Survey. *arXiv preprint arXiv:2006.05525* .

Guo, Q.; Wang, X.; Wu, Y.; Yu, Z.; Liang, D.; Hu, X.; and Luo, P. 2020. Online Knowledge Distillation via Collaborative Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11020–11029.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385* .

He, Y.; Wei, X.; Hong, X.; Shi, W.; and Gong, Y. 2020. Multi-Target Multi-Camera Tracking by Tracklet-to-Target Assignment. *IEEE Transactions on Image Processing* 29: 5191–5205.

Heo, B.; Lee, M.; Yun, S.; and Choi, J. Y. 2019. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3779–3787.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. *Computer Science* 14(7): 38–39.

Hou, S.; Pan, X.; Loy, C. C.; Wang, Z.; and Lin, D. 2019. Learning a Unified Classifier Incrementally via Rebalancing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 831–839.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences* 114(13): 3521–3526.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. Technical report, Citeseer.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*, 1097–1105.

Lee, S.-W.; Kim, J.-H.; Jun, J.; Ha, J.-W.; and Zhang, B.-T. 2017. Overcoming catastrophic forgetting by incremental moment matching. In *NIPS*, 4652–4662.

Li, Z.; and Hoiem, D. 2018. Learning without forgetting. *T-PAMI* 40(12): 2935–2947.

Lopez-Paz, D.; et al. 2017. Gradient episodic memory for continual learning. In *NIPS*, 6467–6476.

Mallya, A.; Davis, D.; and Lazebnik, S. 2018. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 67–82.

Mallya, A.; and Lazebnik, S. 2018. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7765–7773.

Mensink, T.; Verbeek, J.; Perronnin, F.; and Csurka, G. 2013. Distance-based image classification: Generalizing to new classes at near-zero cost. *IEEE transactions on pattern analysis and machine intelligence* 35(11): 2624–2637.

Nichol, A.; and Schulman, J. 2018. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999* 2(3): 4.

Park, W.; Kim, D.; Lu, Y.; and Cho, M. 2019. Relational knowledge distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3967–3976.

Rajasegaran, J.; Khan, S.; Hayat, M.; Khan, F. S.; and Shah, M. 2020. iTAML: An Incremental Task-Agnostic Meta-learning Approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13588–13597.

Ramakrishnan, K.; Panda, R.; Fan, Q.; Henning, J.; Oliva, A.; and Feris, R. 2020. Relationship Matters: Relation Guided Knowledge Transfer for Incremental Learning of Object Detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 250–251.

Rebuffi, S.-A.; Kolesnikov, A.; Sperl, G.; and Lampert, C. H. 2017. icarl: Incremental classifier and representation learning. In *CVPR*, 2001–2010.

Ren, M.; Liao, R.; Fetaya, E.; and Zemel, R. 2019. Incremental few-shot learning with attention attractor networks. In *Advances in Neural Information Processing Systems*, 5276–5286.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2014. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550* .

Serrà, J.; Suris, D.; Miron, M.; and Karatzoglou, A. 2018. Overcoming catastrophic forgetting with hard attention to the task. *arXiv preprint arXiv:1801.01423* .

Shin, H.; Lee, J. K.; Kim, J.; and Kim, J. 2017. Continual learning with deep generative replay. In *Advances in Neural Information Processing Systems*, 2990–2999.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, 4077–4087.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1199–1208.

Tao, X.; Chang, X.; Hong, X.; Wei, X.; and Gong, Y. 2020a. Topology-preserving class-incremental learning. In *European Conference on Computer Vision*, 254–270. Springer.

Tao, X.; Hong, X.; Chang, X.; Dong, S.; Wei, X.; and Gong, Y. 2020b. Few-Shot Class-Incremental Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12183–12192.

Tao, X.; Hong, X.; Chang, X.; and Gong, Y. 2020c. Bi-objective Continual Learning: Learning 'New' while Consolidating 'Known'. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Tao, X.; Hong, X.; Shi, W.; Chang, X.; and Gong, Y. 2020. Analogy-Detail Networks for Object Recognition. *IEEE Transactions on Neural Networks and Learning Systems* 1–15. doi: 10.1109/TNNLS.2020.3017692.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Kavukcuoglu, K.; and Wierstra, D. 2016a. Matching Networks for One Shot Learning. *arXiv preprint arXiv:1606.04080* .

Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016b. Matching networks for one shot learning. In *Advances in neural information processing systems*, 3630–3638.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Wu, C.; Herranz, L.; Liu, X.; van de Weijer, J.; Raducanu, B.; et al. 2018. Memory replay gans: Learning to generate new categories without forgetting. In *Advances In Neural Information Processing Systems*, 5962–5972.

Wu, Y.; Chen, Y.; Wang, L.; Ye, Y.; Liu, Z.; Guo, Y.; and Fu, Y. 2019. Large Scale Incremental Learning. *arXiv preprint arXiv:1905.13260* .

Yim, J.; Joo, D.; Bae, J.; and Kim, J. 2017. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4133–4141.

Yoon, J.; Yang, E.; Lee, J.; and Hwang, S. J. 2017. Lifelong learning with dynamically expandable networks. *arXiv preprint arXiv:1708.01547* .

Yu, L.; Twardowski, B.; Liu, X.; Herranz, L.; Wang, K.; Cheng, Y.; Jui, S.; and Weijer, J. v. d. 2020. Semantic drift compensation for class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6982–6991.

Zagoruyko, S.; and Komodakis, N. 2016. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928* .

Zenke, F.; Poole, B.; and Ganguli, S. 2017. Continual learning through synaptic intelligence. In *ICML*, 3987–3995. JMLR. org.

Zhao, H.; Fu, Y.; Li, X.; Li, S.; Omar, B.; and Li, X. 2020. Few-Shot Class-Incremental Learning via Feature Space Composition. *arXiv preprint arXiv:2006.15524* .