SEMI-SUPERVISED FEW-SHOT CLASS-INCREMENTAL LEARNING

Yawen Cui[1] *Wuti Xiong*[1*] *Mohammad Tavakolian*[1] *Li Liu*[1,2†]

Univeristy of Oulu[1], Finland; National University of Defense Technology[2], China
{yawen.cui, wuti.xiong, mohammad.tavakolian, li.liu}@oulu.fi

ABSTRACT

The capability of incrementally learning new classes and learning from a few examples is one of the hallmarks of human intelligence. It is crucial to endow a practical recognition system with such ability. Therefore, in this paper, we conduct pioneering work and focus on a challenging yet practical Semi-Supervised Few-Shot Class-Incremental Learning (SSFSCIL) problem, which requires CNN models incrementally learn new classes from very few labeled samples and a large number of unlabeled samples, without forgetting the previously learned ones. To address this problem, a simple and efficient solution for SSFSCIL is proposed to learn novel categories using a self-training strategy in a semi-supervised manner and avoid catastrophic forgetting by distillation-based methods. Our extensive experiments on CIFAR100, *mini*ImageNet and CUB200 datasets demonstrate the promising performance of our proposed method, and define baselines in this new research direction.

Index Terms— Few-shot learning, incremental learning, semi-supervised learning, image classification.

1. INTRODUCTION

Human beings can understand their surroundings by gradually learning new notions throughout their lives. For instance, a kid recognizes a cat by showing a few examples of the cat. He/she can further distinguish cats from dogs by seeing a few examples of the dog. Therefore, our brain constantly receives and learns new concepts based on a few samples and updates the boundaries between the learned ones. This competence is termed as few-shot class-incremental learning (FSCIL) [1]. It is also crucial for deep models to acquire this ability in some real applications, such as medical image analysis [2, 3, 4] and autonomous driving vehicles [5, 6, 7]. Collecting labeled data for such applications is laborious due to several challenges, including timely and expensive processes, privacy issues, and expert knowledge demanded.

FSCIL encompasses two challenging problems: few-shot learning [8, 9, 10, 11, 12, 13] and incremental learning [14, 15, 16, 17, 18, 19]. It aims to continuously learn

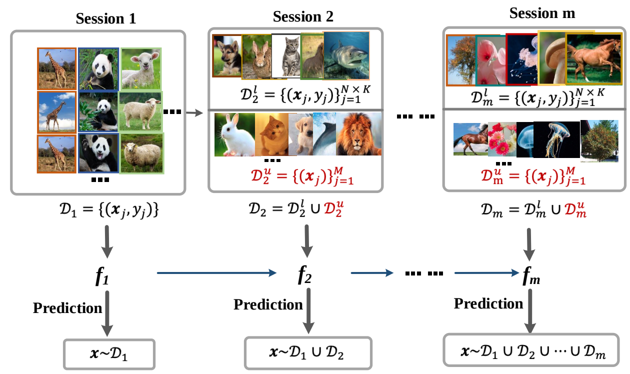

Fig. 1. The task configuration. The first session's training set is a large-scale labeled training set $\mathcal{D}_1$. The sets of following sessions are all N-way K-shot semi-supervised few-shot task settings with labeled data $\mathcal{D}^l$ and unlabeled data $\mathcal{D}^u$.

new categories with limited labeled data arriving in the current session while avoid catastrophic forgetting the old categories of previous sessions. FSCIL has been rarely explored [1]. In this paper, we propose a novel Semi-Supervised Few-Shot Class-Incremental Learning (SSFSCIL) to incrementally learn novel categories with limited labeled and a larger number of unlabeled samples. The task configuration is presented in Fig. 1.

Due to the availability of large-scale unlabeled data, it is possible to only focus on annotations of a few unlabeled samples, and leverage a large portion of unlabeled training examples together with a smaller number of labeled ones for efficient training of learning models. To make the training process of FSCIL more effective, SSFSCIL incrementally presents training data (labeled and unlabeled examples) to the learning model through training sessions. In each training session, we use self-training [20, 21] as the semi-supervised learning technique to train the model using novel labeled and unlabeled data.

In our proposed framework for SSFSCIL, we first initialize parameters of the learning model on a large-scale labeled dataset, which consists of base classes. Then, we incrementally introduce new classes to the learned model by presenting labeled and unlabeled examples of those classes based on self-train in each session. To the best of our knowledge, this

*indicates the equal contribution with the first author.
†indicates the corresponding author.

is the first time few-shot class incremental learning is studied in the context of semi-supervised learning. To evaluate the performance of SSFSCIL, we conduct extensive experiments on three datasets, namely CIFAR100 [22], *mini*ImageNet [10] and CUB200 [23] for image classification. The obtained preliminary experimental results show promising performance in this newly explored research direction.

The contributions of our work are three folds: (1) For the first time, we put forward a challenging and practical learning problem called SSFSCIL, which uses unlabeled data along with labeled training data in FSCIL to improve the efficiency and performance; (2) We incorporate a simple but efficient self-training strategy in SSFSCIL for more effective semi-supervised learning; (3) We perform extensive experiments on three benchmark datasets to evaluate the performance of our proposed method for image classification and introduce new baselines in this new research direction.

2. THE PROPOSED METHOD

We first formulate the problem of semi-supervised few-shot class-incremental learning. The goal is to incrementally learn novel classes in a semi-supervised manner on top of a base session initializing the model. Once the training is completed, the model $f(\cdot)$ will be able to classify samples from all the seen classes, *i.e.,* $y = f(\boldsymbol{x})$, where $\boldsymbol{x}$ and y are the input and the prediction, respectively. In each incremental learning session, we use a small number of labeled training examples together with a large number of unlabeled training examples. Both the labeled and unlabeled examples belong to the same domain in each session. Figure 1 shows an illustration of the task configuration. Specifically, we first conduct a base session which contains a large-scale labeled base dataset $\mathcal{D}_1 = \{(\boldsymbol{x}_j, y_j)\}$.

After training on the base dataset in the first session, we incrementally present novel classes to the model and continue training on the novel data arriving in the following incremental sessions. In the i-th session, we train the model on the dataset $\mathcal{D}_i = \mathcal{D}_i^l \cup \mathcal{D}_i^u$, which contains labeled data $\mathcal{D}_i^l$ and unlabeled data $\mathcal{D}_i^u$. The labeled training data $\mathcal{D}_i^l = \{(\boldsymbol{x}_j, y_j)\}_{j=1}^{N\times K}$ consists of N classes with K labeled examples per class, *i.e.,* a N-way K-shot problem. The unlabeled training data $\mathcal{D}_i^u = \{\boldsymbol{x}_j\}_{j=1}^{M}$ only comprises unlabeled samples, where $M \gg K$. It should be noted that there is not any overlap between the training data of different sessions, *i.e.,* $\mathcal{D}_i \cap \mathcal{D}_j = \emptyset$, where $i \neq j$. The model should be effectively trained so that it does not forget the previously learned classes. Moreover, we should avoid overfitting to newly learned classes in semi-supervised learning.

2.1. Class-Incremental Learning

In our framework, we adopt a knowledge distillation-based [24] incremental learning technique to deal with catastrophic forgetting. To this end, we include a distillation loss to the cross-entropy loss of the model to ensure that the model does not forget the representations of the previously learned classes while new classes are introduced. Hence, we compute the incremental learning loss function as:

$$\mathcal{L}(\mathcal{D}, \mathcal{P}, f) = \mathcal{L}_{ce}(\mathcal{D}, \mathcal{P}, f) + \lambda \mathcal{L}_{dl}(\mathcal{D}, \mathcal{P}, f), \qquad (1)$$

where $\mathcal{L}_{ce}$ is the cross-entropy loss and $\mathcal{L}_{dl}$ is the distillation loss, and $\mathcal{P}$ is old class exemplars drawn from previous sets $\{\mathcal{D}_1, \ldots, \mathcal{D}_{i-1}\}$. $\mathcal{L}_{dl}$ implements in various ways, while it follows the form generally:

$$\mathcal{L}_{dl}(\mathcal{P}, f) = \sum_{j=0}^{|\mathcal{P}|} \left\| f_i(\boldsymbol{x}_j) - f_{i-1}(\boldsymbol{x}_j) \right\|, \qquad (2)$$

where f_i is the model obtained in the i^{th} session and f_{i-1} is the model obtained by the previous session. In order to remember the performance of old classes, the distillation loss mainly measures variations of the predictions obtained by the models in two neighboring sessions.

2.2. Conducting SSFSCIL by Self-Training

The aim of SSFSCIL is to memorize all seen old categories and continuously learn novel categories with labeled and unlabeled instances. The learning process contains two parts: learning the categories arriving at every session and avoiding catastrophic forgetting. We implement semi-supervised learning in the context of incremental learning. We use distillation-based class-incremental learning approaches to address the catastrophic forgetting problem as well. From the semi-supervised perspective, including the unlabeled data in each novel category can significantly improve the performance of the learning algorithm.

In this paper, we propose a simple and efficient solution for SSFSCIL based on self-training. Self-training was first proposed in [20]. The implementation of self-training does not involve any hypothesis and simple. Therefore, we choose self-training to conduct SSFSCIL. We use self-training by selecting unlabeled data based on prediction credibility.

The SSFSCIL process based on self-training is presented in Algorithm 1. Accumulated training set $\mathcal{D}^a$ contains the training data drawn from previous sessions. Total set $\mathcal{D}^t$ contains all the seen data in all the i sessions. For the first session, $\mathcal{D}_1$ is a large-scale training set, and $\mathcal{D}_1$ is added to $\mathcal{D}^t$. The model is trained by $\mathcal{D}_1$, and we obtain f_1.

Regarding the second session, $\mathcal{D}_2^a$ is sampled from the $\mathcal{D}^t$. Then, $\mathcal{D}_2^l$ and $\mathcal{D}_2^a$ train the model f_1 first. After that, the updated f_1 model is used to predict the pseudo labels of $\mathcal{D}_2^u$. We choose S unlabeled samples with higher prediction credibility in $\mathcal{D}_u^2$. The selected unlabeled data together with the pseudo labels is added into $\mathcal{D}_2^l$, and removed from $\mathcal{D}_2^u$. The model is updated again by learning $\mathcal{D}_2^l$ and $\mathcal{D}_2^a$. This described process is iterated till all unlabeled data is added to $\mathcal{D}_2^l$. If the iteration number is not exceed the max epoch number, the model

Algorithm 1 SSFSCIL with self-train.

Input: $\mathcal{D}_1$, $\mathcal{D}_2$, ..., total session number m, $f(\cdot)$, S, $\mathcal{D}^a$, $\mathcal{D}^t = \phi$, the iteration number of n session I^n

Output: $\mathcal{F}$ that could classifies all seen categories

for n in m **do**
  if n==1 **then**
    $f(\cdot)$ updates by learning $\mathcal{D}_1^l$ to obtain f_1;
    Add $\mathcal{D}_1^l$ to $\mathcal{D}^t$;
  else
    Sampling instances from $\mathcal{D}^t$ to form $\mathcal{D}_n^a$,
    while D_n^u is not ϕ **do**
      f_{n-1} updates by learning $\mathcal{D}_n^a \cup \mathcal{D}_n^l$;
      Predict D_n^u using f_{n-1};
      Select S samples from D_n^u;
      Add the selected unlabeled samples into D_n^l;
      Remove them from D_n^u;
      Do the distillation process;
    while iteration number not achieve I_n **do**
      F_{n-1} updates by learning $\mathcal{D}_n^a \cup \mathcal{D}_n^l$;
      Do the distillation process;
  Add D_n^l into $\mathcal{D}^t$ and $f_n = f_{n-1}$;
return $f(\cdot)$ obtained after m sessions.

would update again by learning $\mathcal{D}_2^l$ and $\mathcal{D}_2^a$ till completing all the epochs. Some strategies are also processed in every epoch to memory the categories in previous sessions. The strategy depends on the class-incremental method we use to implement this task. We add $\mathcal{D}_2^l$ to $\mathcal{D}^t$. After this, the session is completed, and we obtain f_2.

When a new incremental session begins, the instances sampled from the total set of the previous session form the accumulated training set first. The sampling process is used for maintaining the performance of the model on old categories. Next, when a dataset in one of the following sessions arrives, the described process of the second session would be repeated until completing all sessions. Finally, we obtain the model endowed the ability that could classify all seen categories with the higher accuracy.

3. EXPERIMENTS

We conducted extensive experiments on three popular image classification databases, *i.e.,* CIFAR100 [22], *mini*ImageNet [10], and CUB200 [23]. For a direct comparison, we followed the same split protocol as in [1] which is the baseline of FSCIL. We implemented SSFSCIL by self-training based on two incremental learning methods, *i.e.,* iCaRL [17] and NCM [16].

3.1. Experimental Setup

CIFAR100 [22] is widely used in class-incremental learning. It includes 100 classes with 600 RGB images per class. For each category, 500 images are used for training and 100 images for testing. The size of the image is 32×32.

***mini*ImageNet** [10] is a subset of the ImageNet with less number of classes. It includes 600 images for each of 100 classes. These images are in the size of 84×84. This dataset is very popular in few-shot learning.

CUB200 [23] contains about $6,000$ training images and $6,000$ test images of over 200 bird categories. The images are resized to 256×256 and then cropped to 224×224 for training.

For CIFAR100 and *mini*ImageNet, we set 60 and 40 classes as the base and novel categories, respectively, and chose a 5-way 5-shot setting. In total, we had 9 training sessions, *i.e.,* one session for base classes and 8 sessions for novel classes. While for CUB200, we adopted the 10-way 5-shot setting by choosing 100 classes as base classes and splitting the remaining 100 classes into 10 incremental learning sessions. For the sessions of learning novel categories, each session's training set was constructed by randomly choosing 5 training instances per class from the original training set to construct 5/10-way 5-shot tasks, and some instances were also picked from the rest of the training set, and their labels were discarded to construct the unlabeled set. We used the whole test set for the evaluation purpose, which was enough to evaluate the generalization ability of the model.

We adopted the same model settings as [16]. We used ResNet-32 for CIFAR100 and ResNet-18 for *mini*ImageNet and CUB200. The total number of unlabeled data in each incremental session was 50. For the first session of CIFAR100 and *mini*ImageNet, the learning rate started from 0.1 and was divided by 10 after 80 and 120 epochs (160 epochs in total). For the rest sessions of CIFAR100, the learning rate was 0.1, and we used early stopping to avoid overfitting. For the rest sessions of *mini*ImageNet, the learning rate was 0.001, and the epoch number was 40. For CUB200, the base learning rate in the first session was 0.001, and divided by 10 after 80 and 120 epochs (160 epochs in total). The learning rate of the following session was 0.001 used in total of 40 epochs. If the unlabeled set was not empty, after every epoch, we selected three unlabeled samples, then added them to the labeled dataset. The models were trained by SGD [25] with the batch size of 128. As for the strategy to preserve the samples for old classes, we considered a memory with a fixed capacity of 500 images.

3.2. Experimental Results

We implemented SSFSCIL using self-training based on two class-incremental learning methods, *i.e.,* iCaRL [17] and NCM [16]. The implemented algorithms are denoted as SS-iCaRL and SS-NCM. Additionally, we conducted the experiments by replacing the nearest-mean-of-exemplar classification in NCM with CNN predictions [26], *i.e.,* NCM-CNN and SS-NCM-CNN (semi-supervised NCM-CNN). The few-shot learning setting results of supervised iCaRL and NCM are also reported. We also compare our results with Ft-CNN [1] and Joint-CNN [1]. In Joint-CNN, all labeled examples

Table 1. The classification accuracies (%) of CUB200. Our results exceed FSCIL method of TOPIC in all sessions.

Method	Session ID										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Ft-CNN [1]	68.68	44.81	32.26	25.83	25.62	25.22	20.84	16.77	18.82	18.25	17.18
Joint-CNN [1]	68.68	62.43	57.23	52.80	49.50	46.10	42.80	40.10	38.70	37.10	35.60
TOPIC-AL [1]	68.68	61.01	55.35	50.01	42.42	39.07	35.47	32.87	30.04	25.91	24.85
TOPIC-AL-MML [1]	68.68	62.49	54.81	49.99	45.25	41.40	38.35	35.36	32.22	28.31	26.28
iCaRL [17]	68.68	52.65	48.61	44.16	36.62	29.52	27.83	26.26	24.01	23.89	21.16
SS-iCaRL (ours)	69.89	61.24 (↑ 8.59)	55.81(↑ 7.20)	50.99(↑ 6.83)	48.18(↑ 11.56)	46.91(↑ 17.39)	43.99(↑ 16.16)	39.78(↑ 13.52)	37.50(↑ 13.49)	34.54(↑ 10.65)	31.33(↑ 10.17)
NCM [16]	68.68	57.12	44.21	28.78	26.71	25.66	24.62	21.52	20.12	20.06	19.87
SS-NCM (ours)	69.89	61.91(↑ 4.79)	55.51(↑ 11.3)	51.71(↑ 22.93)	49.68(↑ 22.97)	46.11(↑ 20.45)	42.19(↑ 17.57)	39.03(↑ 17.51)	37.96(↑ 17.84)	34.05(↑ 13.99)	32.65(↑ 12.78)
NCM-CNN [16]	69.89	62.49	57.68	52.79	50.13	47.44	45.53	42.87	38.97	36.45	33.49
SS-NCM-CNN (ours)	**69.89**	**64.87**(↑ 2.38)	**59.82**(↑ 2.14)	**55.14**(↑ 2.35)	**52.48**(↑ 2.35)	**49.60**(↑ 2.16)	**47.87**(↑ 2.34)	**45.10**(↑ 2.23)	**40.47**(↑ 1.50)	**38.10**(↑ 1.65)	**35.25**(↑ 1.76)

Table 2. Compare the accuracies (%) of *mini*ImageNet obtained by using different numbers of unlabeled instances in each incremental session.

Method	Num_U	Session ID							
		2	**3**	**4**	**5**	**6**	**7**	**8**	**9**
SS-NCM-CNN	25	59.94	56.27	52.84	49.68	46.99	43.50	41.93	39.76
	50	**60.88**	57.63	52.8	**50.66**	**48.28**	**45.27**	41.65	**41.21**
	75	60.25	**57.67**	**53.63**	50.17	47.46	44.91	**41.86**	40.44
	100	58.20	56.27	52.27	48.75	45.89	44.24	42.16	39.80
SS-iCaRL	25	50.15	45.30	42.31	39.86	37.91	33.99	32.59	32.01
	50	**51.64**	**47.43**	**43.92**	**41.69**	**38.74**	**36.67**	**34.54**	**33.92**
	75	50.38	46.56	43.52	40.55	38.26	35.57	33.92	32.79
	100	50.58	46.27	43.56	40.39	38.00	35.38	34.14	32.92
SS-NCM	25	51.55	46.51	43.36	40.95	37.82	34.51	32.89	32.14
	50	**52.60**	**47.97**	44.61	**41.89**	**38.95**	**36.76**	**34.63**	**33.66**
	75	51.48	47.74	**44.73**	41.38	38.91	36.4	34.39	33.49
	100	51.71	47.8	44.59	41.11	38.84	36.02	34.54	33.03

from previous sessions take part in the training process in the current session. In Ft-CNN, the model is simply fine-tuned using a few samples in the current session. This setting usually results in catastrophic forgetting, which drastically deteriorates the model performance.

Fig. 2 presents the results of CIFAR100 and *mini*ImageNet. In the first session, all methods have similar accuracy as they have been trained on the base set. The similar accuracy of the first session is also fair when we compare the deterioration of the accuracy in the following sessions. Our all three implementations achieve remarkable performance on CIFAR100 and *mini*ImageNet compared with the results obtained by FSCIL methods. For CIFAR100, we get 42.62% by using the SS-NCM-CNN, and it outperforms Joint-CNN in nearly all sessions. The performance of SSFSCIL exceeds TOPIC by 13.25%. For *mini*ImageNet, the classification accuracy on all seen categories in the final session is 41.21%. It also outperforms Joint-CNN in all sessions. The results on CUB200 are summarized in Table 1. Our proposed method performs better in comparison with the supervised setting and outperforms TOPIC by 8.97%.

From the obtained results, we can conclude that including unlabeled data in incremental learning sessions improves the performance of few-shot class-incremental classification. Our proposed solution for SSFSCIL paves the way for solving complex and challenging problems in real applications, such as medical image analysis and autonomous driving vehicles.

3.3. Ablation Study

We explore the effect caused by different numbers of unlabeled data added to the labeled set in each session. Table 2 compares the results obtained by adding 25, 50, 75 and 100 unlabeled samples into each incremental session. We obtain most of the best results when the number of unlabeled samples is 50. In general, more data for training would result in better performance. However, for the incremental learning setting, models learn new categories while do not forget previous ones. In order to avoid forgetting, distillation-based methods sample instances of old categories to take part in the training process of the next session. If we put more unlabeled samples of the novel categories into the training set, the proportion of samples belonging to old categories in the training set of a specific session would be lower, so the classification is prone to new categories. For this reason, the accuracies of old categories may affect the classification performance on all the seen categories.

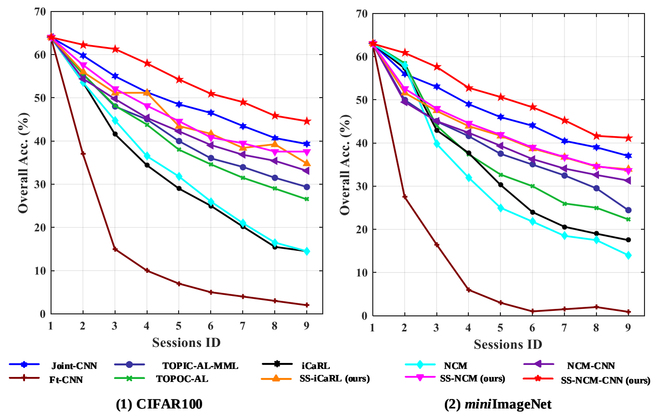

Fig. 2. Results of CIFAR100 and *mini*ImageNet. Our methods perform better for the classification task.

4. CONCLUSION

In this paper, we proposed a Semi-Supervised Few-Shot Class-Incremental Learning (SSFSCIL) for image classification. SSFSCIL is a challenging and practical learning problem that aims to improve the performance of few-shot class-incremental learning by incorporating unlabeled data in the training process together with labeled data. To this end, we used a simple but efficient self-training technique in each incremental session to train the model in a semi-supervised manner. By conducting experiments on three benchmark datasets, we achieved significant results and defined baselines for this novel research direction.

5. REFERENCES

[1] Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong, "Few-shot class-incremental learning," in *CVPR*, 2020.

[2] Mohammad Tavakolian and Abdenour Hadid, "A spatiotemporal convolutional neural network for automatic pain intensity estimation from facial dynamics," *IJCV*, vol. 127(10): 1413–1425, 2019.

[3] Dinggang Shen, Guorong Wu, and Heung-Il Suk, "Deep learning in medical image analysis," *Annual review of biomedical engineering*, vol. 19: 221–248, 2017.

[4] Muhammad Imran Razzak, Saeeda Naz, and Ahmad Zaib, "Deep learning for medical image processing: Overview, challenges and the future," in *Classification in BioApps*, pp. 323–350. Springer, 2018.

[5] Junsung Kim, Ragunathan Rajkumar, and Markus Jochim, "Towards dependable autonomous driving vehicles: a system-level approach," *ACM SIGBED Review*, vol. 10(1): 29–32, 2013.

[6] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *CVPR*, 2020.

[7] Markus Maurer, J Christian Gerdes, Barbara Lenz, and Hermann Winner, *Autonomous driving: technical, legal and social aspects*, Springer Nature, 2016.

[8] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *ICML*, 2017.

[9] K. Hsu, S. Levine, and C. Finn, "Unsupervised learning via meta-learning," in *ICLR*, 2019.

[10] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al., "Matching networks for one shot learning," in *NeurIPS*, 2016.

[11] Zhongjie Yu, Lin Chen, Zhongwei Cheng, and Jiebo Luo, "Transmatch: A transfer-learning scheme for semi-supervised few-shot learning," in *CVPR*, 2020.

[12] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales, "Learning to compare: Relation network for few-shot learning," in *CVPR*, 2018.

[13] Sachin Ravi and Hugo Larochelle, "Optimization as a model for few-shot learning," 2016.

[14] Zhizhong Li and Derek Hoiem, "Learning without forgetting," *IEEE TPAMI*, vol. 40 (12): 2935–2947, 2017.

[15] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari, "End-to-end incremental learning," in *ECCV*, 2018.

[16] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin, "Learning a unified classifier incrementally via rebalancing," in *CVPR*, 2019.

[17] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert, "icarl: Incremental classifier and representation learning," in *CVPR*, 2017.

[18] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars, "Memory aware synapses: Learning what (not) to forget," in *ECCV*, 2018.

[19] Friedemann Zenke, Ben Poole, and Surya Ganguli, "Continual learning through synaptic intelligence," *Proceedings of machine learning research*, vol. 70, pp. 3987, 2017.

[20] H Scudder, "Probability of error of some adaptive pattern-recognition machines," *IEEE Transactions on Information Theory*, vol. 11(3): 363–371, 1965.

[21] Chuck Rosenberg, Martial Hebert, and Henry Schneiderman, "Semi-supervised self-training of object detection models," 2005.

[22] Alex Krizhevsky, Geoffrey Hinton, et al., "Learning multiple layers of features from tiny images," 2009.

[23] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny, "Efficient lifelong learning with a-gem," *arXiv preprint arXiv:1812.00420*, 2018.

[24] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," in *NeurIPS Deep Learning Workshop*, 2014.

[25] Herbert Robbins and Sutton Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60(6): 84–90, 2017.