Incremental Learning with Unlabeled Data in the Wild

Kibok Lee* Kimin Lee† Jinwoo Shin† Honglak Lee*
*University of Michigan, Ann Arbor, MI, USA
†Korea Advanced Institute of Science and Technology, Daejeon, Korea

Abstract

We propose to leverage a continuous and large stream of unlabeled data in the wild to alleviate catastrophic forgetting in class-incremental learning. Our experimental results on CIFAR and ImageNet datasets demonstrate the superiority of the proposed methods over prior methods: compared to the state-of-the-art method, our proposed method shows up to 14.9% higher accuracy and 45.9% less forgetting.

1. Introduction

Class-incremental learning [17] simulates real-world scenarios where the number of tasks continues to grow; the entire tasks are given at once but as a sequence.[1] Deep neural networks (DNNs) tend to forget previous tasks easily when learning new tasks, which is a phenomenon called catastrophic forgetting [4, 15]. The main reason of catastrophic forgetting is the limited resources for scalability: all training data of previous tasks cannot be stored in a limited size of memory as the number of tasks increases. As we live with a continuous and large stream of data, a number of unlabeled data is easily obtainable on the fly or transiently, for example, by data mining on social media [14] and web data [8]. Motivated by this, we propose to leverage such a large stream of unlabeled external data.

Contribution. Under the new class-incremental setup, our contribution is three-fold (see Figure 1 for an overview):

- $\mathcal{A}$. We propose a new training loss, termed global distillation, which utilizes data to distill the knowledge of previous tasks effectively.
- $\mathcal{B}$. We design a 3-step learning scheme to improve the effectiveness of global distillation: (i) training a teacher specialized for the current task, (ii) training a model by distilling the knowledge of the previous model and the teacher learned in (i), and (iii) fine-tuning to avoid overfitting to the current task.
- $\mathcal{C}$. We propose a sampling scheme with a confidence-calibrated model to effectively leverage a large stream of unlabeled data.

[1] In class-incremental learning, a set of classes is given in each task, and we aim to classify data in any class learned so far without task boundaries.

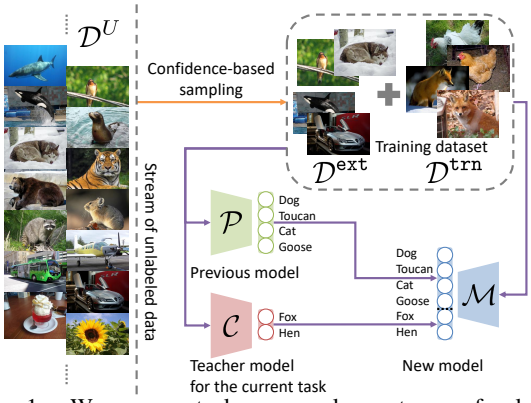


Figure 1. We propose to leverage a large stream of unlabeled data in the wild for class-incremental learning. At each stage, a confidence-based sampling strategy is applied to build an external dataset. Under the combination of the labeled training dataset and the unlabeled external dataset, a teacher model $\mathcal{C}$ first learns the current task, and then the new model $\mathcal{M}$ learns both the previous and the current tasks by distilling the knowledge of $\mathcal{P}$ and $\mathcal{C}$.

2. Approach

2.1. Preliminaries: Class-incremental Learning

Formally, let $(x, y) \in \mathcal{D}$ be a data x and its label y in a dataset $\mathcal{D}$, and let $\mathcal{T}$ be a supervised task mapping x to y. We denote $y \in \mathcal{T}$ if y is in the range of $\mathcal{T}$ such that $|\mathcal{T}|$ is the number of class labels in $\mathcal{T}$. For the t-th task $\mathcal{T}_t$, let $\mathcal{D}_t$ be the corresponding training dataset, and $\mathcal{D}^{\text{cor}}_{t-1} \subseteq \mathcal{D}_{t-1} \cup \mathcal{D}^{\text{cor}}_{t-2}$ be a small coreset containing representative data of previous tasks $\mathcal{T}_{1:(t-1)} = \{\mathcal{T}_1, \ldots, \mathcal{T}_{t-1}\}$, such that $\mathcal{D}^{\text{trn}}_t = \mathcal{D}_t \cup \mathcal{D}^{\text{cor}}_{t-1}$ is the labeled training dataset available at the t-th stage. Let $\mathcal{M}_t = \{\theta, \phi_{1:t}\}$ be the set of learnable parameters of a model, where θ and $\phi_{1:t} = \{\phi_1, \ldots, \phi_t\}$ indicate shared and task-specific parameters, respectively (subscription indicates the task index).[2]

The goal at the t-th stage is to train a model $\mathcal{M}_t$ to perform the current task $\mathcal{T}_t$ as well as the previous tasks $\mathcal{T}_{1:(t-1)}$ without task boundaries, i.e., all class labels in $\mathcal{T}_{1:t}$ are candidates at test time. To this end, a small coreset $\mathcal{D}^{\text{cor}}_{t-1}$ and the previous model $\mathcal{M}_{t-1}$ are transferred from the previous stage. We also assume that a large stream of unlabeled

[2] If task-specific parameters of multiple tasks are given, logits of all learned classes are concatenated for the prediction without task boundaries.

data is accessible, and we would like to sample an external dataset denoted by $\mathcal{D}_t^{\texttt{ext}}$. We do not assume any correlation between the stream of unlabeled data and the tasks.

Learning objectives. With a labeled dataset $\mathcal{D}$, a model $\mathcal{M} = \{\theta, \phi\}$ learns by minimizing a classification loss:

$$\mathcal{L}_{\texttt{cls}}(\theta, \phi; \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} [-\log p(y|x; \theta, \phi)].$$

The following distillation loss is useful when an unlabeled dataset and a reference model $\mathcal{Q} = \{\theta^{\mathcal{Q}}, \phi^{\mathcal{Q}}\}$ is given:

$$\mathcal{L}_{\texttt{dst}}(\theta, \phi; \mathcal{Q}, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{x\in\mathcal{D}} \sum_{y\in\mathcal{T}} [-p(y|x; \theta^{\mathcal{Q}}, \phi^{\mathcal{Q}}) \log p(y|x; \theta, \phi)],$$

where the probabilities can be smoothed for better distillation [6]. For confidence calibration, we also consider a confidence loss to make the model confidence-calibrated, such that its prediction $p(y|x)$ is uniformly distributed if data is from out-of-distribution (OOD) [5, 10, 11]:

$$\mathcal{L}_{\texttt{cnf}}(\theta, \phi; \mathcal{D}) = \frac{1}{|\mathcal{D}||\mathcal{T}|} \sum_{x\in\mathcal{D}} \sum_{y\in\mathcal{T}} [-\log p(y|x; \theta, \phi)].$$

2.2. Global Distillation with 3-step Learning

We propose a novel training method which leverages a large stream of unlabeled external data for class-incremental learning effectively. Intuitively, the previous model $\mathcal{P}_t = \mathcal{M}_{t-1}$ can only produce a prediction on the previous tasks $\mathcal{T}_{1:(t-1)}$, such that unlabeled data are not usable for learning the current task $\mathcal{T}_t$. To compensate for this, another teacher model $\mathcal{C}_t = \{\theta^{\mathcal{C}}, \phi_t^{\mathcal{C}}\}$ learns to be specialized for $\mathcal{T}_t$ by optimizing the following:

$$\min_{\theta^{\mathcal{C}}, \phi_t^{\mathcal{C}}} \mathcal{L}_{\texttt{cls}}(\theta^{\mathcal{C}}, \phi_t^{\mathcal{C}}; \mathcal{D}_t) + \mathcal{L}_{\texttt{cnf}}(\theta^{\mathcal{C}}, \phi_t^{\mathcal{C}}; \mathcal{D}_{t-1}^{\texttt{cor}} \cup \mathcal{D}_t^{\texttt{ext}}), \quad (1)$$

where the confidence loss is jointly minimized with the classification loss to make the model confidence-calibrated for sampling purpose in Section 2.3. However, we note that $\mathcal{P}_t$ and $\mathcal{C}_t$ are not able to distinguish between $\mathcal{T}_{1:(t-1)}$ and $\mathcal{T}_t$, i.e., unlabeled data can only be used to learn either $\mathcal{T}_{1:(t-1)}$ or $\mathcal{T}_t$, not all tasks $\mathcal{T}_{1:t}$ at once. To fully leverage unlabeled data, we define $\mathcal{Q}_t$ as an ensemble of $\mathcal{P}_t$ and $\mathcal{C}_t$: let

$$p_{\max} = \max_y p(y|x, \theta^{\mathcal{P}}, \phi_{1:(t-1)}^{\mathcal{P}}),$$
$$y_{\max} = \arg\max_y p(y|x, \theta^{\mathcal{P}}, \phi_{1:(t-1)}^{\mathcal{P}}).$$

Then, the output of $\mathcal{Q}_t$ can be defined as:

$$p(y|x, \theta^{\mathcal{Q}}, \phi_{1:t}^{\mathcal{Q}}) = \begin{cases} p_{\max} & \text{if } y = y_{\max} \\ \frac{1-p_{\max}-\varepsilon}{1-p_{\max}} p(y|x, \theta^{\mathcal{P}}, \phi_{1:(t-1)}^{\mathcal{P}}) & \text{elif } y \in \mathcal{T}_{1:(t-1)} \\ \varepsilon p(y|x, \theta^{\mathcal{C}}, \phi_t^{\mathcal{C}}) & \text{elif } y \in \mathcal{T}_t, \end{cases} \quad (2)$$

such that $\sum_y p(y|x, \theta^{\mathcal{Q}}, \phi_{1:t}^{\mathcal{Q}}) = 1$. With an assumption that the expected predicted probability is the same over all negative classes $\forall y \notin y_{\max}$, we get

$$\varepsilon = \frac{(1 - p_{\max})|\mathcal{T}_t|}{|\mathcal{T}_{1:t}| - 1}. \quad (3)$$

Now, we define the learning objective of our global distillation (GD) method:

$$\begin{aligned} \min_{\theta, \phi_{1:t}} \quad & \mathcal{L}_{\texttt{cls}}(\theta, \phi_{1:t}; \mathcal{D}_t^{\texttt{trn}}) \\ & + \mathcal{L}_{\texttt{dst}}(\theta, \phi_{1:(t-1)}; \mathcal{P}_t, \mathcal{D}_t^{\texttt{trn}} \cup \mathcal{D}_t^{\texttt{ext}}) \\ & + \mathcal{L}_{\texttt{dst}}(\theta, \phi_t; \mathcal{C}_t, \mathcal{D}_t^{\texttt{trn}} \cup \mathcal{D}_t^{\texttt{ext}}) \\ & + \mathcal{L}_{\texttt{dst}}(\theta, \phi_{1:t}; \mathcal{Q}_t, \mathcal{D}_t^{\texttt{ext}}). \end{aligned} \quad (4)$$

Finally, to eliminate the bias learned from the imbalanced training dataset, we fine-tune the task-specific parameters with the same learning objective. Specifically, for each data in a class k, we normalize the gradient by the portion of data in the class k in the labeled training dataset.

For coreset management, we build a balanced coreset by randomly selecting data for each class. We note that other more sophisticated selection algorithms like herding [16, 17] do not perform significantly better than random selection, as reported in prior works [1, 19].

2.3. Sampling External Dataset

Learning from a large number of data is expensive and most of the data in the wild would be irrelevant to the tasks in interest. To leverage them effectively, we propose to sample an essential external dataset from a large stream of unlabeled data. To alleviate catastrophic forgetting, sampling external data that are expected to be in previous tasks is desired to make the training dataset balanced. Also, to make the model confidence-calibrated, a certain amount of OOD data should also be sampled. Thus, at the beginning of each stage, from a stream of unlabeled data, we randomly sample unlabeled data as OOD[3], and sample most probable data for each class in previous tasks based on the prediction of $\mathcal{P}$.

3. Experiments

3.1. Experimental Setup

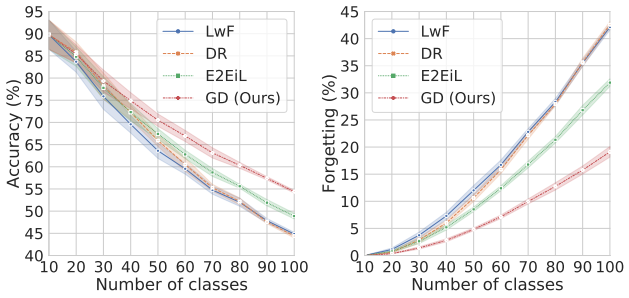
Compared algorithms. *Oracle* provides an upper bound of the performance, which stores all training data of previous tasks and replays them during training. *Baseline* is trained without knowledge distillation. Among prior works, three state-of-the-art methods are compared: *learning without forgetting (LwF)* [12], *distillation and retrospection (DR)* [7], and *end-to-end incremental learning (E2EiL)* [1].

Datasets. CIFAR-100 [9] and downsampled ImageNet ILSVRC 2012 [2, 3] are used. For CIFAR-100, similar to

[3] Since OOD is widely distributed over the data space, randomly sampled data are more useful than the most probable OOD data.

Table 1. Performance of compared methods on CIFAR-100 and ImageNet. We report the mean and the standard deviation of seven trials with different random seeds in %. ↑ (↓) indicates that the higher (lower) number is the better.

Dataset	CIFAR-100						ImageNet					
Task size	5		10		20		5		10		20	
Metric	ACC (↑)	FGT (↓)	ACC (↑)	FGT (↓)	ACC (↑)	FGT (↓)	ACC (↑)	FGT (↓)	ACC (↑)	FGT (↓)	ACC (↑)	FGT (↓)
Oracle	78.7 ± 0.8	3.3 ± 0.2	77.7 ± 0.8	3.2 ± 0.2	75.8 ± 0.7	2.9 ± 0.2	67.3 ± 1.5	3.4 ± 0.4	66.2 ± 1.5	3.2 ± 0.5	64.5 ± 1.2	2.8 ± 0.4
Without an external dataset												
Baseline	57.6 ± 1.1	20.9 ± 0.5	57.0 ± 1.0	19.7 ± 0.4	56.2 ± 1.1	18.0 ± 0.4	43.6 ± 1.1	23.7 ± 0.4	43.5 ± 1.2	21.7 ± 0.6	44.0 ± 0.8	18.7 ± 0.8
LwF [12]	58.7 ± 1.1	19.3 ± 0.5	59.7 ± 1.1	16.9 ± 0.4	60.3 ± 0.9	14.6 ± 0.4	45.0 ± 1.6	21.6 ± 0.4	46.7 ± 1.0	18.6 ± 0.5	48.1 ± 0.8	15.5 ± 0.5
DR [7]	59.4 ± 1.1	19.6 ± 0.4	61.0 ± 1.1	17.1 ± 0.3	62.1 ± 0.8	14.4 ± 0.4	45.9 ± 1.2	22.1 ± 0.6	48.0 ± 1.1	19.0 ± 0.6	50.1 ± 0.9	15.5 ± 0.6
E2EiL [1]	60.6 ± 1.2	16.5 ± 0.5	62.8 ± 1.0	12.8 ± 0.4	65.3 ± 0.7	8.9 ± 0.2	47.1 ± 1.7	17.9 ± 0.5	50.2 ± 1.1	13.5 ± 0.3	53.5 ± 1.1	9.0 ± 0.3
GD (Ours)	**62.4 ± 1.0**	**15.4 ± 0.4**	**65.3 ± 0.9**	**12.1 ± 0.3**	**67.4 ± 0.9**	**8.6 ± 0.4**	**49.4 ± 1.3**	**16.8 ± 0.4**	**53.1 ± 1.2**	**12.9 ± 0.4**	**55.9 ± 1.0**	**8.6 ± 0.5**
With an external dataset												
LwF [12]	60.0 ± 0.8	19.5 ± 0.4	61.3 ± 0.9	17.0 ± 0.4	61.1 ± 1.2	14.7 ± 0.5	46.6 ± 1.1	21.7 ± 0.5	48.6 ± 1.0	18.7 ± 0.5	49.2 ± 0.8	15.9 ± 0.5
DR [7]	60.0 ± 0.9	19.5 ± 0.5	62.5 ± 0.9	16.4 ± 0.3	63.7 ± 1.1	13.4 ± 0.4	46.8 ± 1.2	21.8 ± 0.6	50.0 ± 1.1	18.3 ± 0.5	51.9 ± 1.0	14.6 ± 0.6
E2EiL [1]	61.9 ± 1.0	16.4 ± 0.5	64.5 ± 0.9	12.6 ± 0.4	66.5 ± 1.0	9.0 ± 0.3	48.6 ± 1.3	17.6 ± 0.6	52.2 ± 1.0	13.2 ± 0.3	54.9 ± 0.9	9.1 ± 0.3
GD (Ours)	**66.2 ± 0.9**	**9.6 ± 0.2**	**68.0 ± 0.9**	**7.4 ± 0.3**	**68.9 ± 1.0**	**5.2 ± 0.3**	**54.1 ± 1.5**	**9.7 ± 0.4**	**56.8 ± 1.4**	**7.2 ± 0.4**	**57.9 ± 0.9**	**5.2 ± 0.4**



(a) Average accuracy (ACC) (b) Average forgetting (FGT)

Figure 2. Experimental results on CIFAR-100 with an external data when the task size is 10. We compare (a) the average incremental accuracy, and (b) the average forgetting. We report the mean of the accuracy of seven trials and the standard deviation.

prior works [1, 17], we shuffle the classes uniformly at random and split the classes to build a sequence of tasks. For ImageNet, we first sample 500 images per 100 randomly chosen classes for each trial, and then split the classes. Following the prior works, we divide the classes into splits of 5, 10, and 20 classes. To simulate a large stream of unlabeled data, we take two large datasets: TinyImages [18] with 80M images and ImageNet 2011 with 14M images. The classes appeared in CIFAR-100 and ILSVRC 2012 are excluded to avoid any potential advantage from them. At each stage, our sampling algorithm gets unlabeled data from them uniformly at random to form an external dataset, until the number of retrieved samples is 1M.

Hyperparameters. Our model is based on wide residual networks [20] with 16 layers, a widen factor of 2, and a dropout rate of 0.3. The last fully connected layer is considered to be a task-specific layer, and whenever a task with new classes comes in, the layer is extended by adding more parameters to produce a prediction for the classes. The size of the coreset is set to 2000. Due to the scalability issue, the size of the sampled external dataset is set to the size of the labeled dataset. The temperature for smoothing softmax probabilities [6] is set to 2 for distillation from $\mathcal{P}$ and $\mathcal{C}$ and 1 for distillation from $\mathcal{Q}$ in Eq. (4).

Evaluation metric. We report the performance of the compared methods in two metrics: the average incremental accuracy (ACC) [1, 17] and the average forgetting (FGT). ACC measures the overall performance directly by averaging the accuracy, and FGT measures the amount of catastrophic forgetting, by averaging the accuracy decay, which is essentially the negative of the backward transfer [13].

3.2. Evaluation

Comparison of methods. Table 1 and Figure 2 compare our proposed methods with the state-of-the-art methods. GD outperforms the methods in prior works, LwF, DR, and E2EiL, which shows the effectiveness of the proposed loss function and the 3-step learning scheme. Learning with an external dataset improves the performance consistently, but the improvement is more significant in GD. For example, in the case of ImageNet with a task size of 5, the relative performance gain by learning with an external dataset in E2EiL is 2.8% (from 47.1% to 48.6%) while it is 9.4% (from 49.4% to 54.1%) in GD. Overall, with our proposed learning scheme and the usage of external data, GD shows 14.9% (from 47.1% to 54.1%) of the relative performance improvement from E2EiL, which shows the best performance among the state-of-the-art methods. In terms of forgetting, unlike the other methods, GD shows significantly less forgetting when an external dataset is available: the amount of forgetting in GD is 45.9% (from 17.9% to 9.7%) less than E2EiL in the case above.

4. Conclusion

We propose to leverage a large stream of unlabeled data in the wild for class-incremental learning. The proposed loss encourages a model to distill the knowledge of the reference models without task boundaries, and it is particularly effective when unlabeled data is available. Our 3-step learning scheme effectively leverages the external dataset sampled with the proposed sampling strategy from the stream of unlabeled data.

References

[1] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *ECCV*, 2018. 2, 3

[2] Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819*, 2017. 2

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 2

[4] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999. 1

[5] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, 2017. 2

[6] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2, 3

[7] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Lifelong learning via progressive distillation and retrospection. In *ECCV*, 2018. 2, 3

[8] Jonathan Krause, Benjamin Sapp, Andrew Howard, Howard Zhou, Alexander Toshev, Tom Duerig, James Philbin, and Li Fei-Fei. The unreasonable effectiveness of noisy data for fine-grained recognition. In *ECCV*, 2016. 1

[9] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 2

[10] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *ICLR*, 2018. 2

[11] Kibok Lee, Kimin Lee, Kyle Min, Yuting Zhang, Jinwoo Shin, and Honglak Lee. Hierarchical novelty detection for visual object recognition. In *CVPR*, 2018. 2

[12] Zhizhong Li and Derek Hoiem. Learning without forgetting. In *ECCV*, 2016. 2, 3

[13] David Lopez-Paz et al. Gradient episodic memory for continual learning. In *NeurIPS*, 2017. 3

[14] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *ECCV*, 2018. 1

[15] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of learning and motivation*, 24:109–165, 1989. 1

[16] Cuong V Nguyen, Yingzhen Li, Thang D Bui, and Richard E Turner. Variational continual learning. In *ICLR*, 2018. 2

[17] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, 2017. 1, 2, 3

[18] Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *PAMI*, 30(11):1958–1970, 2008. 3

[19] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, Zhengyou Zhang, and Yun Fu. Incremental classifier learning with generative adversarial networks. *arXiv preprint arXiv:1802.00853*, 2018. 2

[20] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016. 3