Thesis for Bachelor's Degree

Consideration of compatibility between semi-supervised visual representation learning and incremental learning

Lee, Donggun (이 동건 李 동건)

Department of Electronic Engineering and Computer Sciences

Gwangju Institute of Science and Technology

2020

Consideration of compatibility between semi-supervised visual representation learning and incremental learning

준 지도 이미지 표현 학습과 증분학습 사이의 호환성에 대한 고찰

Consideration of compatibility between semi-supervised visual representation learning and incremental learning

Advisor : Professor Jonghyun Choi

by

Lee, Donggun

Department of Electronic Engineering and Computer Sciences
Gwangju Institute of Science and Technology

A thesis submitted to the faculty of Gwangju Institute of Science and Technology in partial fulfillment of the requirements for the degree of Bachelor of Science in Engineering in the Department of Electronic Engineering and Computer Sciences

Gwangju, Republic of Korea

2020. 05. 29.

Approved by

Professor Jonghyun Choi

Committee Chair

Consideration of compatibility between semi-supervised visual representation learning and incremental learning

Lee, Donggun

Accepted in partial fulfillment of requirements
for the degree of Bachelor of Science in Engineering

May. 29. 2020.

Committee Chair ____________________________

Prof. Jonghyun Choi

Committee Member ____________________________

Prof. Haegon Jeon

BS/EC
20165120

Lee, Donggun (이 동 건). Consideration of compatibility between semi-supervised visual representation learning and incremental learning (준 지도 이미지 표현 학습과 증분학습 사이의 호환성에 대한 고찰). Department of Electronic Engineering and Computer Sciences (Artificial Neural Network). 2020. 19p. Advisor Prof. Jonghyun Choi.

Abstract

Incremental Learning is a study of how networks learn the ever-increasing number of tasks over time. Over time, data on past tasks are inaccessible, and direct relearning is impossible, causing catastrophic forgetting to occur essential, and there are numerous prior studies to prevent such forgetting. Among them, a study was conducted on whether improved performance could be obtained when the semi-supervised visual representation learning method was applied to DMC [38], one of the state-of-the-art methods in incremental learning. The evaluation was conducted at the CIFAR10 dataset.

Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1 Introduction

Despite much progress in Deep learning, Incremental learning is still not easily solved. Unlike humans, who naturally learn more new knowledge while preserving knowledge of the past over time in real world, networks need a lot of sacrifice to learn new knowledge. From a network perspective, parameters in a network that are already well-learned by past tasks inevitably lead to a serious drop in performance on past tasks as parameters change as they learn new tasks. This is because there is no change in parameters considering past tasks, and after a certain period of time after learning them, no samples related to those tasks can be accessed at all. In addition to the limitation that 1) data on classes in the past task cannot be accessed, 2)The size of the network should no longer be increased and 3) Both past and present tasks should be performed well in situations where no prior information on test data is required. This is an example of single-headed classification, a method of evaluating a task without any information in class incremental learning. In many cases, numerous methods that address incremental learning problem often fail to fully comply with all three constraints. In some cases, the size of the network may increase over time [6, 20, 27, 29, 36] or need memory for saving samples for past tasks [2, 5, 22, 25, 35].

DMC [38], a method covered in this paper, faithfully observes the three limitations above. No separate sample memory exists, nor does the size of the network increase. In addition, performance is evaluated for single-head classification, an evaluation method that does not require prior information. DMC uses labeled data to learn each model about past and new tasks, and then uses unlabeled data to teach their knowledge into one network. DMC eliminates the intrinsic bias caused by the information asymmetry or over-regularization in the training, as the proposed double distillation objective allows the final student model to learn from two teacher models simultaneously. To further improve DMC's performance, we thought about how we could use unlabeled auxiliary data more effectively. Basically, learning the visual representation of unlabeled data in advance, learning in a way that reduces the difference of distribution between labeled data and unlabeled data on embedded space so that it is not distinguishable from the pre-learned data, and finally, when using unlabeled samples, we set each contribution of sample as the maximum of network's response along to class dimension.

Contributions. The contribution of this paper are as follows:

- The method of effectively using unlabeled data, which was only used in the existing DMC's distillation process, was proposed
- Some improvements were made by modifying the loss used in the existing DMC's distillation process.

Chapter 2. Related Works

2.1 Related Work

2.1.1 Incremental Learning

Incremental learning is the process of the model's adaptation in case of arriving data stream incrementally. The problem of learning sequentially has been emerging since the past [3, 23, 28, 32, 33]. Various strategies like regularization, dynamic architecture, rehearsal based and pseudo-rehearsal based have been proposed to solve the forgetting problems inevitably caused by the incremental learning. Regularization methods [12, 15–17, 19, 37, 38] impose some constraints mitigates severe forgetfulness while the model training. Ensure that model parameters do not deviate too much from previously learned parameters, depending on their importance to old task while training [12, 37]. Especially, focusing on matching of the moment of the posterior distribution of the network which trained in consecutive task respectively [15]. Using distillation loss [9] to maintain the performance of old tasks while the model update with new task [16]. If the strategies we talked about were based on a fixed network architecture, Dynamic architecture methods [6, 20, 27, 29, 36] learn new task after expanding the network dynamically while keeping the previous architecture unchanging. Progressive network [27] increases the capacity of the network with new layers hence, the network size scales quadratically as the number of tasks increasing. Dynamically Expandable Network(DEN) [36] determine expandable capacity of neural networks automatically and efficient training scheme by performing selective retraining. Looking back at past samples later is the quite intuitive way to solve the forgotten. The rehearsal based methods [2, 5, 10, 18, 22, 25, 35] review the past information in current task to recall forgotten knowledge of past tasks. In this case, how to manage the past information is important. Pseudo-rehearsal methods [11, 30, 34, 35] use generative model to generate pseudopatterns [26] instead of reviewing real examplars that used to retrain the model.

2.1.2 Semi-supervised Learning

Semi-supervised learning is a special case of supervised learning. Because the labeling cost is expensive, use a little labeled data and a large amount of unlabeled data to solve visual recognition problems. The several semi-supervised methods are based on the fact that the model's prediction should be consistent for the perturbation. For efficient semi-supervised learning of the ladder network [24], the autoencoder network reflecting the characteristics of the hierarchical latent variable model was established and denoising techniques are actively utilized in the process for more effective learning. Π model [14] defines unsupervised loss to keep the consistency between two different outputs obtained from a network with stochastic augmentation and dropout. Virtual Adversarial Training [21] is similar to Π model, but it used adversarial perturbation instead of independent noise. Mean-Teacher [31] used unsupervised loss to keep the consistency between two different output that comes from exponential moving average of network and current network.

Chapter 3. Approach

We need to formulate class incremental learning problem. There is M sample sets $\{X^1 \ldots, X^M\}$ according to class. The network learns N tasks $T^1, \ldots, T^N$ in turn. When learning a task, they can access $D_{new} = \{X^{a+1}, \ldots, X^b\}$ only. There is no access to the data stream for past tasks $D_{old} = \{X^1, \ldots, X^a\}$. In the end, the goal is learning a network that does not lose classification knowledge about D_{old} at the same time as learning D_{new}.

3.1 Deep Model Consolidation

As Figure 3.1, DMC solve the catastrophic forgetting through the distillation using unlabeled auxiliary data. DMC proposes a way to learn D_{new} and not to forget the knowledge of D_{old}, which cannot be accessed at the same time. DMC uses network $f_{old}(x; \theta_{old})$, which has learned $D_{old} = \{X_1, \ldots X_a\}$ in the past, and network $f_{new}(x; \theta_{new})$, which has learned new data $D_{new} = \{X_{a+1}, \ldots X_b\}$, to pretrained networks on two different tasks, and to conduct knowledge distillation with new networks to learn both past and new task well. This process is called consolidation step, and the double distillation loss used in this process is as follows.

$$L_{dd}(\mathbf{y}, \dot{\mathbf{y}}) = \frac{1}{b} \sum_{j=1}^{b} (y^j - \dot{y}^j), \tag{3.1}$$

y^j is the logit of the jth class generated from the consolidated model, and

$$\dot{y}^j = \begin{cases} \hat{y}^j - \frac{1}{a} \sum_{k=1}^{a} \hat{y}^k, & 1 \le j \le a \\ \hat{y}^j - \frac{1}{b-a} \sum_{k=a+1}^{b} \hat{y}^k, & a < j \le b \end{cases} \tag{3.2}$$

Here, $\hat{y}$ is the normalized version of $\hat{y}_{old}$ and $\hat{y}_{new}$ respectively. Therefore, the final training objective for consolidation step is as follows:

$$min_{\theta}\frac{1}{|N|}\sum_{x_i \in N} L_{dd}(\mathbf{y_i}, \dot{\mathbf{y}}_\mathbf{i}), \tag{3.3}$$

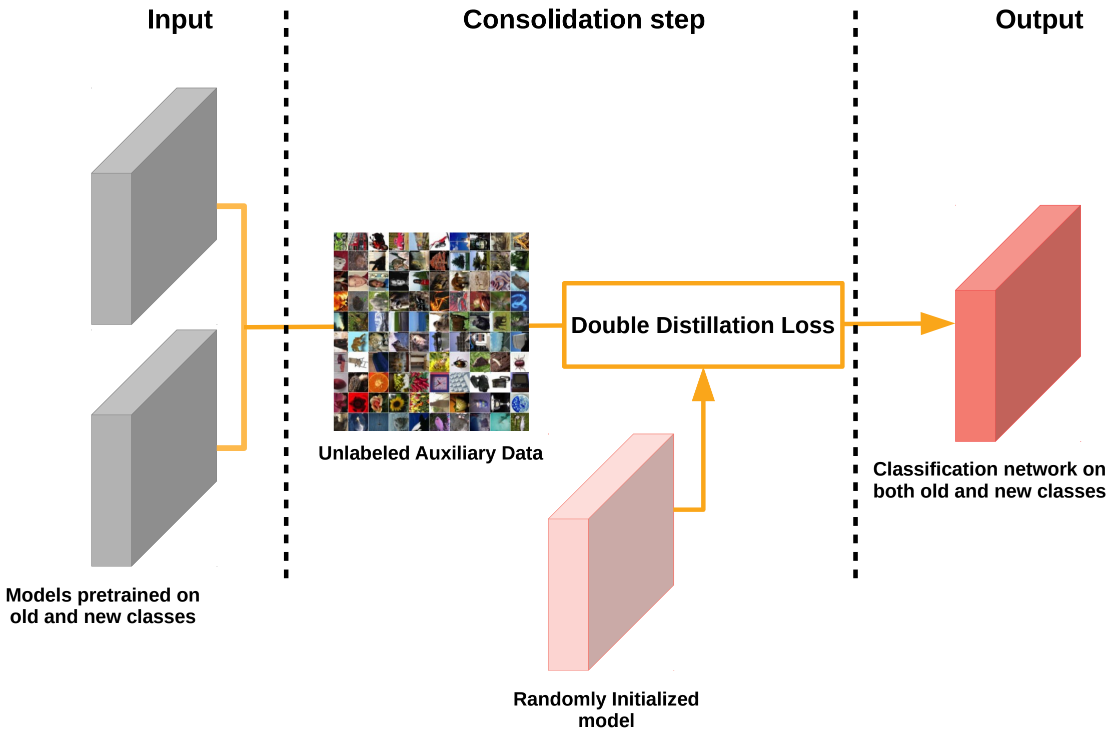


Figure 3.1: **Overview of the DMC**. DMC uses double distillation with numerous unlabeled data to learn incremental tasks from the network that has learned the tasks of the past and the current ones separately from two pretrained networks.

To sum up, DMC can be seen as a repetition of a cycle of training steps for each specialized model and a consolidation step that combines them into a single model. The algorithm for Consolidation step in DMC is organized in Algorithm 1. Based on this method, we have proposed various methods in which semi-supervised visual representation learning has been added.

Algorithm 1 Consolidation step in DMC

Require: θ_{old} ▷ specified model parameters on old task
Require: θ_{new} ▷ specified model parameters on new task
Require: θ ▷ random initialized model parameters
Require: $D_{unlabel}$ ▷ unlabeled auxiliary data
  for t in [1, epochs] **do**
    for each minibatch B in $D_{unlabel}$ **do**
      $\hat{y}_{old} = f_{old}(x; \theta_{old})$ ▷ logit from old model
      $\hat{y}_{new} = f_{new}(x; \theta_{new})$ ▷ logit from new model
      $\hat{y} = Concatenate(\hat{y}_{old}, \hat{y}_{new})$
      $\dot{y} = Normalize(\hat{y})$ ▷ normalize logit by eq 3.2
      $y = f(x; \theta)$
      loss = $\frac{1}{|B|} \sum_{i \in B} L_{dd}(\mathbf{y_i}, \dot{\mathbf{y}}_\mathbf{i})$ ▷ loss by eq 3.1
      update θ using, e.g., ADAM ▷ update network parameters
    end for
  end for
  return θ

3.2 Overview

What this paper is suggesting is that unlabeled auxiliary data cannot be used more effectively by DMC. As described earlier, in existing DMC, unlabeled data is used only for the distilling process within the consolidation step, which is used instead of the labeled data, which is always inaccessible at consolidation step, and thus has no choice but to bear the loss of information loss because of using unlabeled data instead of labeled data. So, using the unlabeled data that can be collected anytime and anywhere, we can learn the visual representation of unlabeled data in advance along with the labelled data in the training process before the consolidation step.It starts from the intuition that it will prevent further loss of information during the distilling process using unlabeled data in the future. If these semi-supervised resetting leads are added first, and performance improvements exist in the distilling process, this will be compatible with all of the incremental learning methods that use unlabeled auxiliary data as well as DMC, which will be expected to improve performance. A study was conducted to verify the compatibility between these two technologies. Thus, depending on how or by what method or standard the representation of unlabeled data is learned, methods of Temporal Enembling or Domain Adversarial Training were applied.

3.2.1 DMC with Temporal Ensembling, TE-DMC

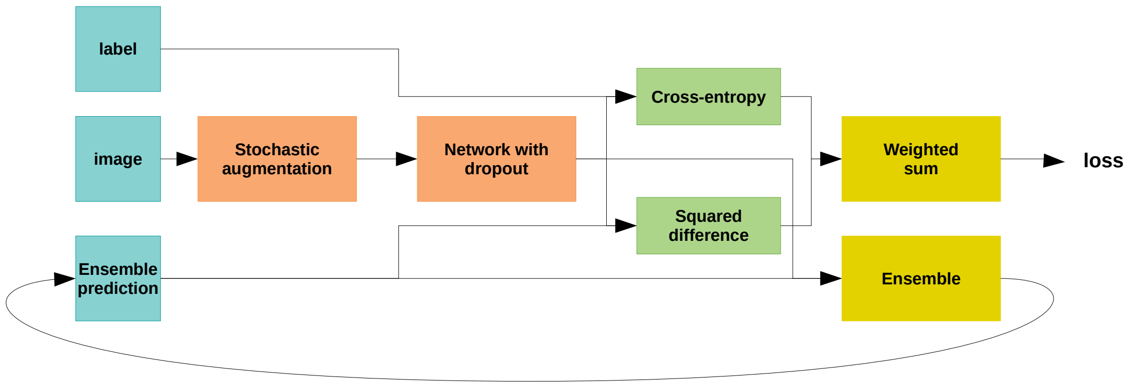


Figure 3.2: **Overview of the Temporal Ensembling**. Temporal enembling uses the prediction of the network, which is different by stochastic aggregation and dropout, as ground truth of unsupervised loss over time to conduct the learning with supervised loss.

Temporal Ensembling [14] is one of the methods of semi-supervised learning that can overcome the limitations of the number of labeled data and successfully proceed with learning in situations where small and unlabeled data are mixed. In particular, they take advantage of the fact that in a neural network, they can generally get better prediction from the ensemble of multi-network than a single network. After applying stochastic augmentation to input image, the network with dropout produces different outputs for the same input. They defined unsupervised loss as the squared loss with the ensembled output along with epoch and current output. Total loss is defined by the weighted sum of existing cross entropy loss and unsupervised loss. Through Temporal Ensembling, instead of learning θ_{old} and θ_{new} in the existing DMC process only in the labeled data, add unlabeled auxiliary data to proceed with semi-supervised setting. When this happens, θ_{old} and θ_{new} pre-learn the visual representation on unlabeled data before consolidation step, creating a better logit for that unlabeled data in the consolidation step so that we look forward to a more quality distilling process in consolidation step. More detail of algorithm is expressed in Algorithm 2. The algorithm for the method is as follows, and we will name this method TE-DMC.

Algorithm 2 Temporal Ensembling in DMC's training step

Require: x_i = training stimuli
Require: L = set of training input indices with known labels
Require: y_i = labels for labeled input $i \in L$
Require: α = ensembling momentum, $0 \leq \alpha < 1$
Require: $w(t)$ = unsupervised weight ramp-up function
Require: $f_\theta(x)$ = stochastic neural network with trainable parameters θ
Require: $g(x)$ = stochastic input augmentation function

```
Z ← 0_[N×C]                                  ▷ initialize ensemble predictions
z̃ ← 0_[N×C]                                  ▷ initialize target vectors
for t in [1, epochs] do
    for each minibatch B in D_unlabel do
        z_{i∈B} ← f_θ(g(x_{i∈B}, t))          ▷ evaluate network outputs for augmented inputs
        loss ← −(1/|B|) Σ_{i∈(B∩L)} log z_i[y_i]          ▷ supervised loss component
               + w(t) (1/(C|B|)) Σ_{i∈B} ||z_i − z̃_i||^2   ▷ unsupervised loss component
        update θ using, e.g., ADAM            ▷ update network parameters
    end for
    Z ← αZ + (1 − α)z                         ▷ accumulate ensemble predictions
    z̃ ← Z/(1 − α^t)                           ▷ construct target vectors by bias correction
end for
return θ
```

3.2.2 DMC with Domain adversarial training, DAT-DMC

Domain-Adversarial Training [7] suggests a way to solve unsupervised domain adaptation with domain adversarial training as shown in Figure 3.3. They used H-divergence proposed by [1], calculate the distance between domain and conduct domain adversarial learning in order to reduce this H-divergence. As shown in the picture(refer image) below, you can learn the domain classifier together with the existing classifier, but in the feature extractor section, visual retention can be learned that makes it difficult to distinguish between target and source domain by backward pass using the gradient value of the opposite symbol from the domain classifier. I thought this could also be applied to DMC cases. What's different from TE-DMC is that the distribution of labeled data and unlabeled auxiliary data, by sufficiently different premises, proceeds with domain adversarial training so that the network

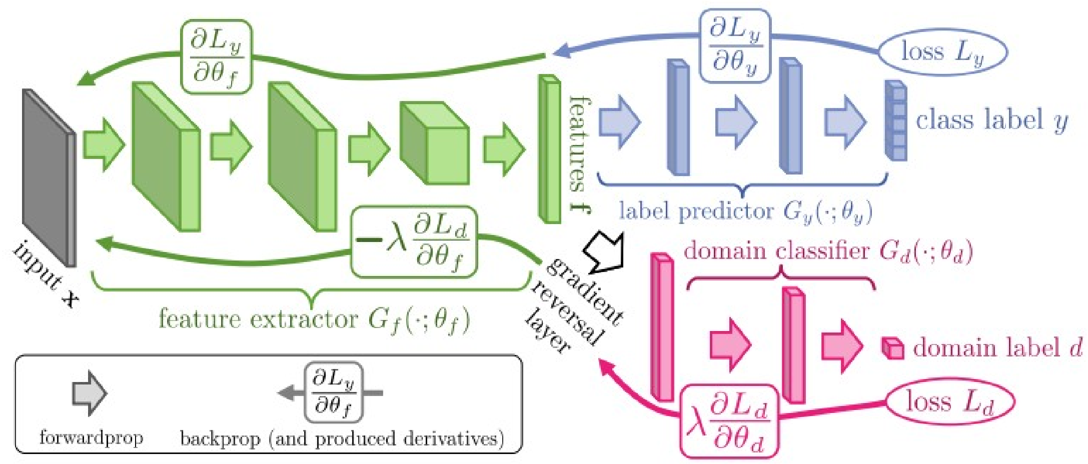


Figure 3.3: **Overview of the Domain Adversarial Training**. Training with domain classifier and use gradient reversal layer at backward pass to make feature dustributions over the two domains are made similar.

does not recognize the difference between labeled data and unlabeled data while learning visual representation. After going through the domain adversarial learning, we make sure that the feature extractor makes the distribution of labeled data and unlabeled data similar on the embedded space, so that when carrying out DMC's consolidation step, we expects the strength to proceed with the process with unlabeled data sampled in distribution similar to the distribution of already learned labeled data on embedding space. DMC using that method will be named DAT-DMC.

3.2.3 DMC with weighting distillation, WD-DMC

The method is not about visual representation learning, such as the previous two methods. This starts with a problem in which all samples are considered to contribute the same amount to the existing distillation loss when distilling using all unlabeled auxiliary data in DMC's consolidation step. In proceeding with the distillation, as described earlier, DMC defined the loss as the mean squared error between the concatenated normalized value of logit from two specified models $\theta_{old}, \theta_{new}$ and the logit from new model θ for unlabeled data to learn θ. At this time, we cannot guarantee the reliability of the network's response and logit obtained from the existing DMC since the unlabeled data is sampled from the distribution completely different from the labeled data distribution that the network have seen during the training process. Therefore, we propose a way to multiply the weight, which means the reliability of logit, by each logit, so that each sample can be more focused on learning the more reliable response. We use the maximum value on the class dimension axis of the corresponding logit as a measure of this reliability. DMC using that method will be named WD-DMC.

Chapter 4. Experiments

4.1 Experiments

4.1.1 Experimental Setup

Datasets. We used CIFAR10 [13] dataset for training and evaluation. We defined a single task as classifying two different classes and make 5 multiple disjoint tasks that are learned in turn. Let's name the curriculum the information about the order in which you are learning the class. We trained and evaluated five different curriculum for all experiments. In addition, cifar100 was used, excluding classes associated with cifar10 classes, as an unlabeled auxiliary dataset required to learn DMC.

Architectures. we used resnet32 for all experiments. Since there is no dropout in Resnet32 [8], specifically for TE-DMC, we used Resnet32 with added dropout layer. In addition, the DAT-DMC added an fc layer for the domain classifier, because it requires additional domain classifier. In all experiments, the batch size was 128 and SGD optimizer was used in the consolidation step. For training steps, however, we used Adam optimizer for TE-DMC and SGD optimizer for all others.

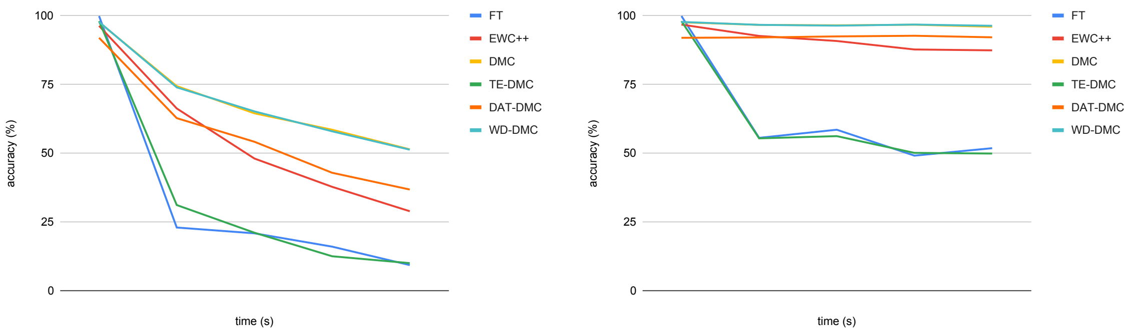
Baselines. We compare TE-DMC, DAT-DMC and WD-DMC against the following baselines: 1) FT : fine tuning with classification loss only, 2) EWC++ [4] : advanced version of Deep neural networks regularized with Elastic Weight Consolidation , 3) DMC : original version of DMC [38].

4.1.2 Results on CIFAR10

We now quantitatively compare our methods with other methods on the CIFAR10 dataset in Table 4.1. FT corresponds to the lower bound in the experiment in such a way that no essential instruction method has been applied. The accuracy of TE-DMC is similar to that of FT without any method, so it can be seen that visual representation learned by Temporal Ensembling destroys DMC's consolidation step to the point where the effects of DMC's method are lost. DAT-DMC records higher performance than FT, indicating that it helps to some extent to solve the catastrophic forgetting problem. However, lower than the existing DMC performance, it can be seen that although the pre-learned visual representation extracts some meaningful logits from the consolidation step, it is not as useful in terms of distillation as in DMC, which has learned visual representation only through labeled data. Finally, WD-DMC's performance has improved slightly compared to DMC's, which also shows that focusing on samples that contribute to the samples to produce reliable logits is of fine help.

Method	Single-Head Acc. (%)	Multi-Head Acc. (%)
FT	9.29	51.7
EWC++ [4]	28.82	87.34
DMC [38]	51.54	95.94
TE-DMC	10	50.11
DAT-DMC	36.76	92.08
WD-DMC	51.88	96.28

Table 4.1: **Classification accuracy (single-head and multi-head) on Cifar10.** These are the results of the accuracy of multi-head and single-head, which is evaluated according to whether or not there is prior information about task.



(a) Single-Head Accuracy. (%). (b) Multi-Head Accuracy. (%)

Figure 4.1: **Test accuracy change along to time.** Except for TE-DMC, all methods perform better than FT.

Chapter 5. Conclusion

5.1 Conclusion

In order to improve DMC, which is one of the latest ways to address necessarily existing catastrophic forgetting, DMC has added several semi-supervised visual representation learning, or proposed slightly improved loss. TE-DMC, DAT-DMC, and WD-DMC are those. E-DMC and DAT-DMC were about how to pre-learn the unlabeled auxiliary data used by DMC directly from the training step. However, as a result, the visual representation of unlabeled data learned through these methods has led to a significant drop in performance over DMC in the consolidation step. Through this, the visual representation learning proposed in this paper was not successfully compatible with the existing DMC. On the other hand, in the case of WD-DMC, different degrees of contribution by each sample to loss were used actively for distillation, which showed little performance improvement over the existing DMC at CIFAR-10.

References

[1] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando C Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2009.

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

[3] Gert Cauwenberghs and Tomaso A. Poggio. Incremental and decremental support vector machine learning. In *NIPS*, 2000.

[4] Arslan Chaudhry, Puneet Kumar Dokania, Thalaiyasingam Ajanthan, and Philip H. S. Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *ECCV*, 2018.

[5] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. *ArXiv*, 2018.

[6] Robert Coop, Aaron Mishtal, and Itamar Arel. Ensemble learning in fixed expansion layer networks for mitigating catastrophic forgetting. *IEEE Transactions on Neural Networks and Learning Systems*, 2013.

[7] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. Domain-adversarial training of neural networks. *ArXiv*, abs/1505.07818, 2016.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[9] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *ArXiv*, 2015.

[10] Khurram Javed and Faisal Shafait. Revisiting distillation and incremental classifier learning. In *ACCV*, 2018.

[11] Nitin Kamra, Umang Gupta, and Yan Liu. Deep generative dual memory network for continual learning. *ArXiv*, 2018.

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

[13] A. Krizhevsky. Learning Multiple Layers of Features from Tiny Images. Technical report, 2009.

[14] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *ArXiv*, abs/1610.02242, 2017.

[15] Sang-Woo Lee, Jin-Hwa Kim, JungWoo Ha, and Byoung-Tak Zhang. Overcoming catastrophic forgetting by incremental moment matching. *ArXiv*, 2017.

[16] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[17] Xialei Liu, Marc Masana, Luis Herranz, Joost van de Weijer, Antonio M. López, and Andrew D. Bagdanov. Rotate your networks: Better weight consolidation and less catastrophic forgetting. *2018 24th International Conference on Pattern Recognition (ICPR)*, 2018.

[18] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continuum learning. In *NIPS*, 2017.

[19] Arun Mallya, Dillon Davis, and Svetlana Lazebnik. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In *ECCV*, 2018.

[20] Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[21] Takeru Miyato, Shin ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:1979–1993, 2019.

[22] Cuong V. Nguyen, Yingzhen Li, Thang D. Bui, and Richard E. Turner. Variational continual learning. *ArXiv*, 2017.

[23] Robi Polikar, L. Upda, S. S. Upda, and Vasant Honavar. Learn++: an incremental learning algorithm for supervised neural networks. *IEEE Trans. Systems, Man, and Cybernetics, Part C*, 2001.

[24] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In *NIPS*, 2015.

[25] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[26] Anthony V. Robins. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connect. Sci.*, 1995.

[27] Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *ArXiv*, 2016.

[28] Jeffrey C. Schlimmer and Douglas H. Fisher. A case study of incremental concept induction. In *AAAI*, 1986.

[29] Joan Serrà, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. *ArXiv*, 2018.

[30] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *NIPS*, 2017.

[31] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NIPS*, 2017.

[32] Sebastian Thrun. Is learning the n-th thing any easier than learning the first? In *NIPS*, 1995.

[33] Sebastian Thrun and Tom M. Mitchell. Lifelong robot learning. *Robotics and Autonomous Systems*, 1995.

[34] Chenshen Wu, Luis Herranz, Xialei Liu, Yaxing Wang, Joost van de Weijer, and Bogdan Raducanu. Memory replay gans: learning to generate images from new categories without forgetting. *ArXiv*, 2018.

[35] 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*, 2018.

[36] Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. *ArXiv*, abs/1708.01547, 2017.

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

[38] Junting Zhang, Jie Zhang, Shalini Ghosh, Dawei Li, Serafettin Tasci, Larry Heck, Heming Zhang, and C.-C. Jay Kuo. Class-incremental learning via deep model consolidation. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1120–1129, 2020.

Summary

Consideration of compatibility between semi-supervised visual representation learning and incremental learning

본 논문에서는 증분 학습에서의 성능 향상을 하는 방법으로써 준 지도 이미지 표현 학습을 도입하는 방식에 대해서 논하고 있다. 구체적으로는 증분 학습에서 필연적으로 발생하는 치명적 망각 현상을 해결하는 최신의 알고리즘 DMC를 개선하고자 하는 방안에 대해 치중되어있다. 본 논문에서는 기존의 증분 학습 알고리즘 DMC를 개선하고자 하는 방안으로 TE-DMC, DAT-DMC, WD-DMC 3가지를 제안한다. TE-DMC는 준 지도 학습의 대표적인 방안을 적용하여 DMC의 훈련과정에서 라벨링이 되어있지 않은 이미지에 대한 표현 방식을 네트워크가 미리 배울 수 있도록 한다. DAT-DMC는 TE-DMC와 다르게 적대적 훈련을 통해 라벨링이 되어있지 않은 이미지와 라벨링이 되어있는 이미지 간의 구별을 할 수 없게끔 하면서 라벨링이 되어 있지 않은 이미지에 대한 표현 방식을 네트워크가 미리 배울 수 있도록 한다. 이 두방법 모두 DMC의 작동 과정 중 두개의 모델의 지식을 하나의 모델로 전수하는 과정에서 사용되는 라벨링이 되어있지 않은 이미지에 대한 표현 방식을 두개의 모델이 미리 잘 알고 있을 경우, 지식 전수를 더 효율적으로 할 수 있기를 기대하는 목적에서 제안하였다. 마지 막으로 WD-DMC는 지식 전수 과정에서 모든 라벨링이 되어있지 않은 데이터들을 똑같은 신뢰도로 수용하는 기존의 손실 함수를 조금 더 유연하게 하기 위하여, 각 샘플들의 신뢰도를 모델의 로짓의 최대값으로 정의하여 유연한 지식 전수 과정을 기대하는 목적으로 제안하였다. 본 논문에서는 라벨링이 되어 있는 CIFAR10 데이터셋과 라벨링이 되어 있지 않으며 CIFAR10과 관련된 클래스들을 모두 제거한 CIFAR100 데이터셋을 훈련 및 평가에 사용하였다. 제안하는 방법을 통해 성공적으로 이미지 표현 학습은 수행하였지만, 지식 전수 과정에서 호환성의 문제가 발생하여

성능 하락이 일어나게 되었다. 이를 통해 준 지도 학습을 통한 이미지 표현 학습이 DMC의 지식 전수 과정과 호환성이 떨어진다는 것을 알게되었다.

감 사 의 글

학사학위논문 연구에 도움을 주신 모든 분들께 감사드립니다.

Curriculum Vitae

Name	:	**Donggun Lee**
Date of Birth	:	July 29, 1998
Gender	:	Male
Nationality	:	Korea
Affiliation	:	Department of Electronic Engineering and Computer Sciences Gwangju Institute of Science and Technology (GIST)
Office Address	:	Dasan Bldg. 504, GIST, 123 Cheomdangwagi-ro, Buk-gu, Gwangju 500-712, Republic of Korea
Home Address	:	204 Dormitory A for undergraduate students, GIST
Phone	:	+82-10-8636-5629
E-mail	:	unlimitlife@gist.ac.kr

Educations

B.S. Department of Electronic Engineering and Computer Science, **Gwangju Institute of Science and Technology (GIST)**, Gwangju, Korea (2016.3 $\sim$ 2020.6)