TCP: Triplet Contrastive-relationship Preserving for Class-Incremental Learning

Shiyao Li[1], Xuefei Ning[1]*, Shanghang Zhang[2], Lidong Guo[1], Tianchen Zhao[1],
Huazhong Yang[1] and Yu Wang[1]*
[1]Tsinghua University, China, [2]Peking University, China
lishiyao20@mails.tsinghua.edu.cn, foxdoraame@gmail.com, yu-wang@tsinghua.edu.cn

A. Experimental Setups

A.1. Datasets

We employ three datasets, which are extensively used in the literature of class-incremental learning, for our experiments: CIFAR-100, ImageNet-100 and ImageNet-1000. CIFAR-100 contains 60,000 images of the size 32×32 over 100 classes, including 50,000 training images and 10,000 test images, respectively. ImageNet-100 is a subset of ImageNet-1000 with only 100 classes, randomly sampled from the original 1000 classes and it contains about 130,000 training images and 5,000 testing images. For both datasets, we select 50 classes as the base classes, and the rest 50 classes are equally divided for incremental learning phases. ImageNet-1000 has 1000 classes, we use the first 500 classes to train the base model and the rest 5000 classes are used for incremental learning.

A.2. Training Details

All models are trained on RTX 3090 GPUs. We use ResNet-32 and ResNet-18 for CIFAR-100 and ImageNet, respectively. We add a nonlinear projection head after the ResNet [2], and remove the ReLU in the penultimate layer to allow the features to take both positive and negative values for the cosine classifier [3]. We train the base 50% classes model for 200 epochs on CIFAR-100 and ImageNet using SGD with a batch size of 256. The learning rate is initialized to 0.1 and follows a cosine annealing schedule. At each incremental learning phase, we finetune the model for 160 epochs with the memory bank $\mathcal{M}$, batch size of 128, and new data batch size of 128. The learning rate is initially set to 0.005 for CIFAR-100 and ImageNet with the cosine annealing strategy. At the end of each incremental phase, we apply the herding sampling strategy proposed in iCaRL [9] and use the data in memory bank to train a unified classifier with a learning rate 0.1 and batch size 256. Then, we evaluate the model on the union of all the encountered test datasets.

*Corresponding authors

	margin of old model	constant σ			
		0.05	0.1	0.15	0.2
Ave. Acc.	**66.54**	63.15	64.67	63.26	61.28

Table 1. Average Accuracy of different margin applied in the triplet contrastive-relationship preserving loss

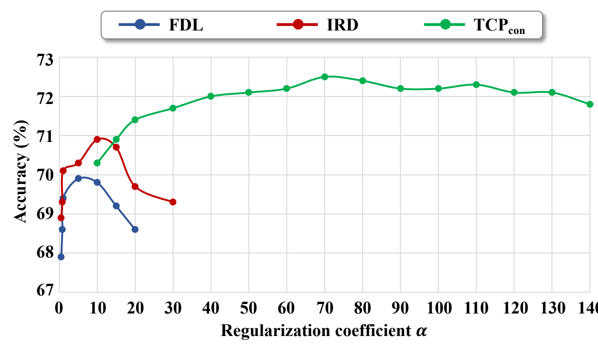

Figure 1. The accuracy of the first incremental step under the 5-phase setting on CIFAR100.

B. Hyperparameter Study

In this section, we report the ablation studies on two different hyperparameters, including the margin coefficient σ in TCP loss, and the regularization coefficient α in the overall loss function. We perform all experiments on CIFAR100 under 5-phase setting.

B.1. The Effect of the Margin Coefficient

In our TCP loss, we need to choose a proper margin coefficient σ. In triplet loss [11], they regard σ as a constant number for each triplet to learn new knowledge. However, TCP loss aims to distill old knowledge from a pretrained old model. Thus, the pretrained model can provide more in-

Model	backbone	CIFAR-FS 5-way		FC100 5-way	
		1-shot	5-shot	1-shot	5-shot
TADAM [5]	ResNet-12	-	-	40.1	56.1
Shot-Free [8]	ResNet-12	69.2	84.7	-	-
TEWAM [7]	ResNet-12	70.4	81.3	-	-
ProtoNet [12]	ResNet-12	72.2	83.5	37.5	52.5
MetaOptNet [10]	ResNet-12	72.6	84.3	41.1	55.5
RethinkDistill [13]	ResNet-12	73.9	86.9	44.6	60.9
RethinkDistill [13] + TCP	ResNet-12	**74.1**	**87.6**	**44.7**	**62.1**

Table 2. Average few-shot classification accuracies (%) on CIFAR-FS and FC100 datasets.

formative margins that indicate the sample similarities in each old triplet. For instance, if the negative sample is highly similar to the anchor sample, then the margin will be small; otherwise, the margin will be large. **This additional similarity information in old model's margin will help TCP loss to preserve old knowledge**, so we use the old model's margins as σ instead of a constant σ. Note that, when we compute the contrastive relationship $D^{'}_{i,jk}$ of new model, we simultaneously compute the contrastive relationship $D_{i,jk}$ od old model as the old margin σ. As shown in Table 1, the results indicate that the average accuracy achieved by using the old model's margins outperforms the best constant σ choice by 1.87%. This experiment demonstrates that incorporating the more informative margins of the old model is an effective approach to improve performance.

B.2. Regularization Coefficient Stability

The overall loss function of the distillation-based incremental learning methods can be described as Equation 1:

$$\mathcal{L} = \mathcal{L}_{\rm A2CL} + \alpha \mathcal{L}_{\rm distill}, \tag{1}$$

where $\mathcal{L}_{\rm A2CL}$ and $\mathcal{L}_{\rm TCP}$ denote asymmetrical augmented contrastive loss and the distillation loss, α denotes the regularization coefficient.

To investigate the impact of the regularization coefficient α, we employed the proposed A2CL to learn new data and evaluated the effectiveness of three different distillation losses in preserving old knowledge: point-wise FDL [3], pair-wise IRD [1], and our proposed $\text{TCP}_{\rm con}$.

Figure 1 displays the first phase accuracy of three distillation losses with varying values of the regularization coefficient α. FDL (blue line) and IRD (red line) exhibit sensitivity to the value of α, with appropriate values for α only found within a narrow range. In contrast, **TCP is less sensitive to the regularization coefficient α than FDL and IRD**. When using the TCP (green line) loss, the appropriate α value can be chosen from a wide range (from 40 to 130), as shown in Figure 1. This outcome is expected since TCP offers greater flexibility to allow for changes in the feature space when learning new tasks. When α is very large, the distillation loss carries greater weight, and the optimizer endeavors to minimize the distillation loss. In this case, FDL and IRD aim to maintain the exact value of old feature positions or similarities, leading the model to sacrifice the learning of new classes. However, using the TCP loss with a large α enables the model to easily learn new classes since the TCP loss only preserves the contrastive relationship of features instead of any exact value.

C. TCP on Few-shot Learning

Few-shot scenarios are very common in real-world applications due to the long-tail distribution of data [6, 14]. In such scenarios, the model requires fine-tuning using a limited number of samples from new tasks [5, 12, 13]. We investigate whether the TCP loss can improve performance in the important few-shot scenarios by plugging it into existing few-shot learning algorithms. We report the few-shot learning ability of the proposed TCP loss on CIFAR-FS and FC100 datasets in Table 2.

The CIFAR-FS (CIFAR100 Few-shots) dataset is derived from the original CIFAR-100 by splitting 100 classes into 64, 16, and 20 classes for training, validation, and testing. The FC100 (Few-shot CIFAR100) dataset is also derived from CIFAR-100. Different from CIFAR-FS, it first groups the original 100 classes into 20 high-level classes, then it splits the 20 high-level classes into 12, 4, and 4 classes for training, validation, and testing. This results in 60 classes for training, 20 classes for validation, and 20 classes for testing.

TCP can easily be plugged into few-shot learning methods and boost their performance. As shown in Table 2, we plug the proposed TCP loss into RethinkDistill [13] and denote it as RethinkDistill+TCP. TCP can effectively boost the performance of the original RethinkDistill on both CIFAR-FS and FC100 datasets. RethinkDis-

ID	new image / class	old image / class	Avg. Acc
A	500	500	77.01%
B	500	100	73.45%
C	300	300	75.96%

Table 3. Average Accuracy of the 50 old classes with different number of new images and old images per class in CIL.

till+TCP outperforms the original RethinkDistill on both datasets. Notably, we observe that TCP can bring more significant improvements in the 5-way 5-shot setting than in the 5-way 1-shot setting. This result is expected because in the 5-shot setting, the feature space needs to be more flexible to learn new data than in the 1-shot setting. Thus, TCP can provide even more performance improvements in the 5-shot setting.

In fact, the benefits of TCP loss can be leveraged in various scenarios where the model needs to effectively learn new knowledge while distilling relevant old knowledge from a teacher model. These scenarios include few-shot [12, 13], incremental [4, 9], and others. By simply integrating TCP loss into the original algorithms, we can enhance their performance.

D. The effect of class imbalance problem

As we mentioned in the abstract and introduction, the imbalance problem in CIL makes it difficult to preserve the feature relation of old classes and hard to learn the feature relation between old and new classes. To empirically substantiate our assertion, we devised a meticulous experiment. Specifically, we continually learn ten new classes based on a pretrained 50-class classifier on CIFAR-100 with the contrastive distillation loss [1] and evaluate the average accuracy on the 50 old classes. The findings, presented in Tab. 3, reveal the following: In case A, where all training data is utilized during continual learning (serving as our baseline), the mean accuracy of the old 50 classes stands at 77.01%. In case B, which limits the storage to merely 100 images for each old class, there's a noticeable decline in accuracy from 77.01% to 73.45%. This case shows that the imbalance problem between old and new classes can truly make it difficult for the model to preserve the accuracy of the old classes. However, in case C, all classes have 300 training images, which means that the data in each class is balanced. In this case, the average accuracy of old classes is much higher than case B and only 1.05% lower than the baseline. Consequently, it becomes evident that the imbalance indeed impedes the model's capacity to preserve the feature relation of old classes, making it hard for the model to learn the feature relation between old and new classes.

References

[1] Hyuntak Cha, Jaeho Lee, and Jinwoo Shin. Co2l: Contrastive continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9516–9525, 2021.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[3] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 831–839, 2019.

[4] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

[5] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. *Advances in neural information processing systems*, 31, 2018.

[6] Archit Parnami and Minwoo Lee. Learning from few examples: A summary of approaches to few-shot learning. *arXiv preprint arXiv:2203.04291*, 2022.

[7] Limeng Qiao, Yemin Shi, Jia Li, Yaowei Wang, Tiejun Huang, and Yonghong Tian. Transductive episodic-wise adaptive metric for few-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3603–3612, 2019.

[8] Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto. Few-shot learning with embedded class models and shot-free meta training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 331–339, 2019.

[9] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

[10] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. *arXiv preprint arXiv:1807.05960*, 2018.

[11] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[12] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[13] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 266–282. Springer, 2020.

[14] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.