Supplementary Material of Adapt without Forgetting: Distill Proximity from Dual Teachers in Vision-Language Models

Mengyu Zheng[1,2], Yehui Tang[2], Zhiwei Hao[2,3], Kai Han[2], Yunhe Wang[2], and Chang Xu[1⋆]

[1] School of Computer Science, Faculty of Engineering, The University of Sydney
[2] Huawei Noah's Ark Lab
[3] School of information and Electronics, Beijing Institute of Technology
mzhe4259@uni.sydney.edu.au, {yehui.tang,kai.han,yunhe.wang}@huawei.com, haozhw@bit.edu.cn, c.xu@sydney.edu.au

A Appendix

This appendix includes

- Description of datasets in MTIL benchmark(Sec. A.1);
- Details of experimental results(Sec. A.2);
- Additional experiments(Sec. A.3)
- The algorithm of implementing graph distillation in CLIP(Sec. A.4).

A.1 Description of Datasets in MTIL Benchmark.

The detailed characteristics of the 11 datasets applied in the MTIL benchmark are shown in Table 1.

A.2 Details of Experimental Results.

Table 2 shows the accuracy results on all datasets after training on each task in the MTIL benchmark under task order setting I. Each row represents the accuracy results of our model on each dataset after completing training for the current task.

Meanwhile, we provide additional experiments details about order setting II. Firstly, the sequence of order setting II is StanfordCars, Food, MNIST, Oxford-Pet, Flowers, SUN397, Aircraft, Caltech101, DTD, EuroSAT, and CIFAR100. Moverover, the complete experiment result of our method are shown in Table 3. Lastly, Table 4 displays the experimental results, showcasing a comparative results of our method against others using the Transfer, Avg., and Last metrics under task order setting II. Despite varying order settings, our method consistently demonstrates significant improvements across nearly all metrics.

⋆ Corresponding author

Table 1: Discription of dataset in MTIL benchmark.

Dataset	# classes	# train	# test	Description
Aircraft [11]	100	3334	3333	aircraft model
Caltech101 [5]	101	6941	1736	real-life object
CIFAR100 [9]	100	50000	10000	real-life object
DTD [2]	47	1880	1880	texture database
EuroSAT [6]	10	21600	5300	satellite images
Flowers [12]	102	1020	6149	images of flowers
Food [1]	101	75750	25250	images of food
MNIST [3]	10	60000	10000	handwritten digits
OxfordPet [14]	37	3680	3669	pet image dataset
Cars [8]	196	7144	8041	car images
SUN397 [17]	397	87003	21751	scene categories
Total	1201	319352	97109	

Table 2: Accuracy(%) on every dataset after training on each task in the MTIL benchmark under task order setting I.

	Aircraft	Caltech101	CIFAR100	DTD	EuroSAT	Flowers	Food	MNIST	OxfordPet	Cars	SUN397
Aircraft	59.3	88.8	67.7	46.0	54.6	70.6	88.5	56.4	88.6	62.8	65.8
Caltech101	50.0	94.8	69.1	45.8	59.0	68.9	87.7	58.7	86.2	62.3	66.0
CIFAR100	48.3	94.1	87.5	46.6	55.2	70.3	88.0	63.0	87.7	61.8	67.0
DTD	48.2	94.5	86.6	77.6	55.8	71.5	87.8	64.7	89.1	62.1	67.2
EuroSAT	47.8	94.4	85.6	76.9	98.4	71.5	87.9	67.0	89.1	62.3	67.5
Flowers	45.5	94.1	85.3	76.8	98.3	97.2	87.3	63.1	88.0	61.6	67.5
Food	44.8	93.7	84.8	75.1	97.9	95.5	92.6	63.6	88.2	62.6	67.8
MNIST	43.0	93.4	84.0	74.6	97.7	94.8	92.5	99.2	88.1	62.1	66.7
OxfordPet	42.9	93.7	83.9	74.6	97.5	94.6	92.5	99.2	95.3	62.0	67.0
Cars	40.0	93.3	84.0	73.8	97.3	94.1	92.3	99.1	94.5	88.1	67.0
SUN397	42.4	92.7	83.2	73.2	97.0	91.9	92.2	99.1	94.0	87.4	82.6
Transfer		88.8	68.4	46.1	56.2	70.6	87.9	62.4	88.1	62.2	67.0
Avg.	46.6	93.4	82.0	67.4	82.6	83.7	90.0	75.7	89.9	66.8	68.4
Last	42.4	92.7	83.2	73.2	97.0	91.8	92.2	99.1	93.9	87.4	82.6

Table 3: Accuracy(%) on every dataset after training on each task in the MTIL benchmark under task order setting II.

	Cars	Food	MNIST	OxfordPet	Flowers	SUN397	Aircraft	Caltech101	DTD	EuroSAT	CIFAR100
Cars	88.7	88.6	59.3	87.7	71.2	65.6	22.6	89.0	45.7	54.4	68.6
Food	88.4	92.9	64.1	87.6	70.6	67.2	22.0	89.3	45.0	55.5	68.8
MNIST	88.0	92.8	99.2	87.1	70.3	66.8	22.2	89.3	45.0	55.6	69.0
OxfordPet	87.5	92.6	99.1	95.2	70.7	66.6	22.7	89.2	45.1	54.1	68.3
Flowers	86.9	92.4	99.0	94.5	97.0	66.4	22.0	89.6	45.3	51.9	68.5
SUN397	86.1	92.2	99.0	93.9	95.1	82.8	22.3	89.8	47.1	53.6	69.0
Aircraft	84.3	91.9	98.9	93.2	94.1	81.9	60.3	88.7	45.8	49.8	67.8
Caltech101	84.4	91.9	98.8	93.0	92.7	81.4	53.4	94.0	45.4	52.9	68.2
DTD	84.5	91.8	98.8	92.6	93.2	81.3	52.9	93.8	77.9	55.1	69.0
EuroSAT	84.7	91.7	98.8	92.7	92.8	81.2	51.8	93.8	77.2	98.5	70.2
CIFAR100	83.9	91.5	97.6	92.7	89.2	80.7	50.6	92.9	75.3	97.0	87.8
Transfer		88.6	61.7	87.5	70.7	66.5	22.3	89.3	45.5	53.6	68.7
Avg.	86.1	91.8	92.0	91.8	85.2	74.7	36.6	90.9	54.1	61.7	70.5
Last	83.9	91.5	97.6	92.7	89.2	80.7	50.6	92.9	75.3	97.0	87.8

A.3 Additional experiments.

Results on MTIL with single teacher. To highlight the effect of 1st- and 2nd-order proximities. We present the results of our proposed method and four compared methods with a single teacher in Table 5. The results show that the proposed method outperforms all other methods under every metric with a single teacher. Besides, the Transfer of the two classic knowledge distillation methods, traditional KD [7] and RKD [13], are lower than the original CLIP model, which means that these two knowledge distillation methods are not able to keep the zero-shot transfer ability.

Compared with AttriClip under CIL benchamark. We present the average for every two steps of the proposed method and AttriClip, under the CIL setting on CIFAR100 dataset in Table 6. As the results show, the proposed method has a better average on every step.

A.4 Algorithm of Method.

To help understand the training process of our method, we now provide the whole training algorithm of our method in Algorithm 1.

Table 4: Performance of trained model on each task in the MTIL benchmark under task order setting II.

Method	Cars	Food	MNIST	OxfordPet	Flowers	SUN397	Aircraft	Caltech101	DTD	EuroSAT	CIFAR100
Zero-shot	64.7	88.5	59.4	89.0	71.0	65.2	24.3	88.4	44.6	54.9	68.2
Fine-tuning	89.6	92.7	99.6	94.7	97.5	81.8	62.0	95.1	79.5	98.9	89.6
Transfer											
Continual FT		85.9	59.6	57.9	40.0	46.7	11.1	70.0	30.5	26.6	37.7
LwF [10]		87.8	58.5	71.9	46.6	57.3	12.8	81.4	34.5	33.7	46.8
iCaRL [15]		86.1	51.8	67.6	50.4	57.9	11.0	72.3	31.2	32.7	48.1
LwF-VR [4]		88.2	57.0	71.4	50.0	58.0	13.0	82.0	34.4	29.3	47.6
WiSE-FT [16]		87.2	57.6	67.0	45.0	54.0	12.9	78.6	35.5	28.4	44.3
ZSCL [18]		88.3	57.5	84.7	68.1	64.8	21.1	88.2	45.3	**55.2**	68.2
Ours		**88.6**	**61.7**	**87.5**	**70.7**	**66.5**	**22.3**	**89.3**	**45.5**	53.6	**68.7**
Avg.											
Continual FT	42.1	70.5	92.2	80.1	54.5	59.1	19.8	78.3	41.0	38.1	42.3
LwF [10]	49.0	77.0	92.1	85.9	66.5	67.2	20.9	84.7	44.6	45.5	50.5
iCaRL [15]	52.0	75.9	77.4	74.6	58.4	59.3	11.7	79.6	42.1	43.2	51.7
LwF-VR [4]	44.9	75.8	91.8	85.3	63.5	67.6	16.9	84.9	44.0	40.6	51.3
WiSE-FT [16]	52.6	79.3	91.9	83.9	63.4	65.2	23.3	83.7	45.4	40.0	48.2
ZSCL [18]	81.7	91.3	91.1	91.0	82.9	72.5	33.6	89.7	53.3	**62.8**	69.9
Ours	**86.1**	**91.8**	**92.0**	**91.8**	**85.2**	**74.7**	**36.6**	**90.9**	**54.1**	61.7	**70.5**
Last											
Continual FT	24.0	67.3	99.1	87.4	44.3	67.0	29.5	92.3	61.3	81.0	**88.1**
LwF [10]	34.6	69.6	99.3	88.7	61.1	72.5	32.5	88.1	65.6	90.9	87.9
iCaRL [15]	46.0	81.5	91.3	82.8	66.5	72.2	16.3	91.6	68.1	83.2	87.8
LwF-VR [4]	27.4	61.2	99.4	86.3	60.6	70.7	23.4	88.0	61.3	84.3	**88.1**
WiSE-FT [16]	35.6	76.9	99.5	89.1	62.1	71.8	27.8	90.8	67.0	85.6	87.6
ZSCL [18]	78.2	91.1	**97.6**	92.5	87.4	78.2	45.0	92.3	72.7	96.2	86.3
Ours	**83.9**	**91.5**	**97.6**	**92.7**	**89.2**	**80.7**	**50.6**	**92.9**	**75.3**	**97.0**	87.8

Table 5: Comparison results on the MTIL benchmark under task order setting I with singe teacher.

Metric	Zero-shot	ZSCL	Ours only C_{i-1}	KD	RKD
Transfer.	69.4	68.1	**69.2**	65.4	65.2
Avg.	65.3	75.4	**76.7**	73.0	73.2
Last.	65.3	83.6	**84.8**	83.2	83.0

Table 6: Comparison results of each step on the CIL benchmark.

Method	**Step 2**	**Step 4**	**Step 6**	**Step 8**	**Step 10**
AttriClip	93.7	87.5	82.5	81.9	81.4
Ours	**97.1**	**92.8**	**89.9**	**87.8**	**86.2**

Algorithm 1 Implementing Graph Distillation in CLIP

Input: A series of task datasets $\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_N$, wild images and text datasets $p_{\mathrm{w}}(\boldsymbol{x})$ and $p_{\mathrm{w}}(\boldsymbol{t})$, pretrained CLIP model C_0.

Input: Initializing the model fine-tuned on previous dataset C_{i-1} by C_0, the number of iterations *iter* trained on each tasks, the batch size m for task dataset, and n for wild data, temperature τ.

for $i = 1, ..., N$ **do**

  while iterations $<$ *iter* **do**

    Randomly sample $\{\boldsymbol{x}^{(k)}, \boldsymbol{y}^{(k)}\}_{k=1}^{m}$ from $\mathcal{T}_i$

    Randomly sample $\{\boldsymbol{x}_w^{(j)}\}_{j=1}^{m} \sim p_{\mathrm{w}}(\boldsymbol{x})$, $\{\boldsymbol{t}_w^{(j)}\}_{j=1}^{n} \sim p_{\mathrm{w}}(\boldsymbol{t})$, donate as $\{v^k\}_{k=1}^{m+n}$

    Calculate $L_{\mathrm{CE}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} CE(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}, \tau)$

    Calculate 1st-order proximity $p_1(\cdot|v^k)$, $\tilde{p}_1(\cdot|v^k)$, $\overline{p}_1(\cdot|v^k)$.

    Calculate 2nd-order proximity $p_2(\cdot|v^k)$, $\tilde{p}_2(\cdot|v^k)$, $\overline{p}_2(\cdot|v^k)$.

    Calculate sample-wise distillate weights $\overline{w}_k$ and $\tilde{w}_k$ of each vertex for C_0 and C_{i-1}.

    Calculate $L \leftarrow L_{\mathrm{CE}} + \frac{1}{m+n}\sum_{k=1}^{m+n}[\overline{w}_k \cdot \bar{\ell}_{\mathrm{pd}}(v_k) + \tilde{w}_k \cdot \tilde{\ell}_{\mathrm{pd}}(v_k)$

    Update the parameters θ of C_i with optimizer.

  end while

  Update the previous model $C_{i-1} = C_i$.

end for

Output: A continual learning CLIP model C_N.

References

1. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101–mining discriminative components with random forests. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13. pp. 446–461. Springer (2014) 2
2. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3606–3613 (2014) 2
3. Deng, L.: The mnist database of handwritten digit images for machine learning research [best of the web]. IEEE signal processing magazine **29**(6), 141–142 (2012) 2
4. Ding, Y., Liu, L., Tian, C., Yang, J., Ding, H.: Don't stop learning: Towards continual learning for the clip model. arXiv preprint arXiv:2207.09248 (2022) 4
5. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: 2004 conference on computer vision and pattern recognition workshop. pp. 178–178. IEEE (2004) 2
6. Helber, P., Bischke, B., Dengel, A., Borth, D.: Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **12**(7), 2217–2226 (2019) 2
7. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015) 3
8. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 554–561 (2013) 2
9. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images (2009) 2
10. Li, Z., Hoiem, D.: Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence **40**(12), 2935–2947 (2017) 4
11. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013) 2
12. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian conference on computer vision, graphics & image processing. pp. 722–729. IEEE (2008) 2
13. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3967–3976 (2019) 3
14. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3498–3505. IEEE (2012) 2
15. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 2001–2010 (2017) 4
16. Wortsman, M., Ilharco, G., Kim, J.W., Li, M., Kornblith, S., Roelofs, R., Lopes, R.G., Hajishirzi, H., Farhadi, A., Namkoong, H., et al.: Robust fine-tuning of zero-shot models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7959–7971 (2022) 4

17. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: 2010 IEEE computer society conference on computer vision and pattern recognition. pp. 3485–3492. IEEE (2010) 2
18. Zheng, Z., Ma, M., Wang, K., Qin, Z., Yue, X., You, Y.: Preventing zero-shot transfer degradation in continual learning of vision-language models. arXiv preprint arXiv:2303.06628 (2023) 4