Tip-Adapter: Training-free Adaption of CLIP for Few-shot Classification

Renrui Zhang*[1,2], Wei Zhang*[1], Rongyao Fang[2], Peng Gao†[1], Kunchang Li[1], Jifeng Dai[3], Yu Qiao[1], and Hongsheng Li[2,4]

[1] Shanghai AI Laboratory
[2] The Chinese University of Hong Kong
[3] SenseTime Research
[4] Centre for Perceptual and Interactive Intelligence (CPII)
{zhangrenrui, gaopeng, qiaoyu}@pjlab.org.cn, hsli@ee.cuhk.edu.hk

1 Fine-tuning Settings

Compared to Tip-Adapter without training, Tip-Adapter-F fine-tunes the keys $\mathbf{F}_{\text{train}}$ in the cache model, but freezes values $\mathbf{L}_{\text{train}}$, CLIP's [12] visual encoder and textual encoder. Here, we explore whether other modules in Tip-Adapter could be fine-tuned for performance improvement. In Table 1, we conduct 7 fine-tuning experiments for unfreezing different modules of Tip-Adapter. Note that we set the learning rates of two CLIP's encoders as 1/1000 of the $\mathbf{F}_{\text{train}}$ and $\mathbf{L}_{\text{train}}$'s for training stability, and train every settings for 20 epochs on ImageNet [3] with 16-shot training set. As shown, the first two rows denote the performance for Tip-Adapter's 62.03% and Tip-Adapter-F's 65.51%. The third row by fine-tuning the cached values $\mathbf{L}_{\text{train}}$ decreases the performance to 60.90%, and fine-tuning all cache model even leads to collapse during training, which accords with our assumption that the one-hot ground-truth labels shall not be updated to preserve the few-shot knowledge. Furthermore, we experiment to fix all parameters in the cache model and fine-tune the pre-trained CLIP's weights. If the visual encoder or textual encoder is independently tuned, the performance could be improved to

Table 1. Fine-tuning different modules for Tip-Adapter. '✓' denotes fine-tuning and the symbol '-' denotes freezing. 'Vis.' and 'Tex.' stand for visual encoder and textual encoder of CLIP. The accuracy (%) and training time are tested on 16-shot ImageNet [3] and a single NVIDIA GeForce RTX 3090 GPU.

Vis.	Tex.	$\mathbf{F}_{\text{train}}$	$\mathbf{L}_{\text{train}}$	Accuracy	Time
-	-	-	-	62.03	**0**
-	-	✓	-	**65.51**	5min
-	-	-	✓	60.90	5min
-	-	✓	✓	Collapsed	-
✓	-	-	-	62.84	8min
-	✓	-	-	63.15	1h 20min
✓	✓	-	-	51.22	1h 27min

62.84% and 63.15%, respectively, but when both encoders are jointly fine-tuned, the classification accuracy would significantly drop to 51.22%. This is because of the severe over-fitting for such a huge-parameter model learning from the few-shot training set. Compared to unfreezing CLIP's encoders, only fine-tuning $\mathbf{F}_{\text{train}}$ brings larger performance improvement but less time consumption, which fully demonstrates the superiority of our Tip-Adapter-F.

2 Performance Gain without Training

In Figure 1, we show the absolute accuracy improvement brought by Tip-Adapter over Zero-shot CLIP [12] on 11 classification datasets under 16-shot settings: EuroSAT [7], Flowers102 [10], DTD [2], SUN397 [15], StandfordCars [8], FGVCAircraft [9], UCF101 [13], Caltech101 [5], OxfordPets [11], ImageNet [3] and Food101 [1]. Without any training, Tip-Adapter greatly boosts Zero-shot CLIP on EuroSAT by 33.02% and Fowers102 by 23.87%. Now that the CLIP is pre-trained on large-scale web-collected image-text pairs for daily scenarios, when the domain gap between downstream dataset and the pre-trained data is larger, the performance gain by Tip-Adapter would be normally higher. Taking EuroSAT and DTD as examples, they respectively contain land cover and detailed texture pictures with distinctive semantics, which thus require more few-shot knowledge memorized in the cache model to update the pre-trained CLIP's knowledge for better performance.

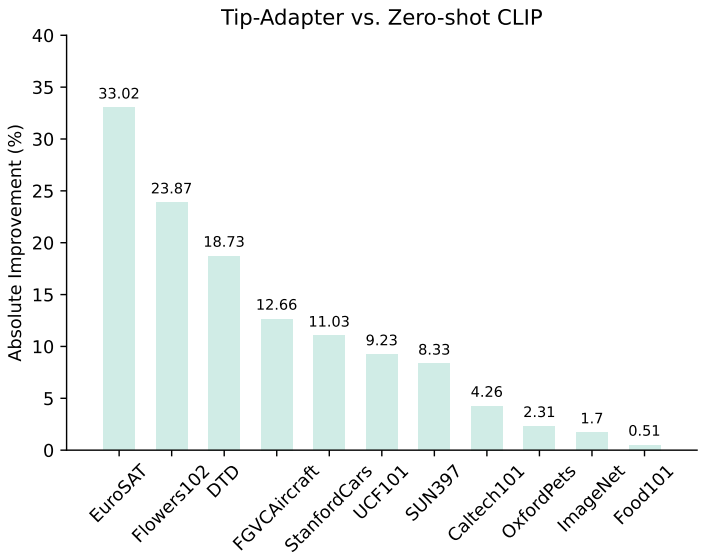

Fig. 1. Performance gain contributed from the proposed training-free cache model, which is constructed by the 16-shot training set on 11 classification datasets.

3 Compared to Fully-trained Methods

Although our Tip-Adapter and Tip-Adapter-F are based on the few-shot training sets, they are evaluated by the full test sets, the same as conventional methods [6, 4] trained by full training sets. In Table 2, we compare the learnable parameters and training settings between ours and the series of ResNet [6] and DeiT [14]. We adopt ViT-Large [4] as the visual backbone of Tip-Adapter and Tip-Adapter-F. As shown, only by 16-shot training set, Tip-Adapter without parameters or training outperforms ResNet-50 and DeiT-T by +1.9% and +3.9%, respectively. Tip-Adapter-F further achieves higher performance by the efficient fine-tuning of 6 minutes. This demonstrates the superiority of our approach in low-data and resource-limited regimes.

Table 2. Comparison between Tip-Adapter, Tip-Adapter-F and conventional methods trained by full training set on ImageNet [3]. The training time is tested on a single NVIDIA GeForce RTX 3090 GPU.

Method	Acc. (%)	Param. (M)	Train. Set	Train. Time
ResNet-50 [6]	74.2	25.6	full set	>1 day
ResNet-101 [6]	77.4	44.5	full set	>1 day
DeiT-T [14]	72.2	6.0	full set	>1 day
DeiT-S [14]	**79.9**	22.1	full set	>1 day
Tip-Adapter	76.1	**0**	**16-shot**	**0**
Tip-Adapter-F	79.4	6.2	**16-shot**	6 min

References

1. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101–mining discriminative components with random forests. In: European conference on computer vision. pp. 446–461. Springer (2014)
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)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
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)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
7. 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)
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)
9. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
10. 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)
11. 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)
12. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021)
13. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
14. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. pp. 10347–10357. PMLR (2021)
15. 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)