

Training-Free Feature Reconstruction with Sparse Optimization for Vision-Language Models

Anonymous Author(s)

In Supplementary Materials, we provide additional experimental results and implementation details for further comprehension of our proposed method.

1 METHODS

In this section, we give a detailed derivation for the selection of λ in Equation (4) of the main text. We introduce a regularization term, $\lambda\|\mathbf{m}_q - \mathbf{S}_d\mathbf{w}\|_2^2$, to constrain the reconstruction error. However, note that, if λ is too large, the optimal solution would ignore the sparse optimization of weight \mathbf{w} . Conversely, if we choose a λ that is too small, the optimization process would not sufficiently consider the reconstruction error penalty. Following [9], we utilize the binary penalty search (BPS) algorithm to address this issue. First, we rewrite Equation (4) as a penalty problem of λ :

$$\Gamma(\lambda) : \min_{\mathbf{w}} \|\mathbf{w}\|_1 + \lambda\|\mathbf{m}_q - \mathbf{S}_d\mathbf{w}\|_2^2.$$

We establish the upper and lower bonds of λ , i.e., $\lambda_r > \lambda_l \geq 0$. By solving $\Gamma(\lambda_r)$, we obtain a solution \mathbf{w}_r that satisfies the reconstruction constraint $\|\mathbf{m}_q - \mathbf{S}_d\mathbf{w}\|_2^2 \leq \sigma$, while solving $\Gamma(\lambda_l)$ yields a solution \mathbf{w}_l that exceeds the error constraint. Our objective is to find a λ in (λ_l, λ_r) with which solving $\Gamma(\lambda)$ yields a solution with sparser weights than that of $\Gamma(\lambda_r)$ while still adhering to the reconstruction constraint.

To achieve this, we employ a binary search for the optimal λ . At the start of each iteration, we set $\lambda = (\lambda_l + \lambda_r)/2$ and assess the solution of $\Gamma(\lambda)$. If $\|\mathbf{m}_q - \mathbf{S}_d\mathbf{w}\|_2^2 > \delta$, we update the lower bound λ_l to λ ; otherwise, we decrease the upper bound λ_r . The interval range is halved after each iteration, progressively narrowing down to a sufficiently small search range smaller than ϵ . The most appropriate λ is then determined as λ_r . The details of the BPS algorithm are outlined below.

Algorithm 1 Binary Penalty Search

Require: $\mathbf{m}_q, \mathbf{S}_d, \delta$, stop threshold ϵ .

Ensure: optimal solution \mathbf{w}_\star

Initialize $\lambda_l = 0, \lambda_r = \lambda_{\max}$.

repeat

$\lambda = (\lambda_l + \lambda_r)/2$;

 Solve Problem (4) by ADM to obtain \mathbf{w}_\star ;

if $\|\mathbf{m}_q - \mathbf{S}_d\mathbf{w}_\star\|_2^2 \leq \delta$ **then**

$\lambda_r = \lambda$;

else

$\lambda_l = \lambda$;

end if

until $\lambda_r - \lambda_l \leq \epsilon$

2 MORE IMPLEMENTATION DETAILS

2.1 Dataset Details

We adopt the few-shot evaluation protocol to assess our method on 11 widely-used image classification datasets, spanning the breadth of generic object classification (ImageNet [14], Caltech101 [5]), fine-grained object classification (OxfordPets [12], StanfordCars [8], Flowers102 [11], Food-101 [1], FGVCircraft [10]), texture classification (DTD [3]), remote sensing recognition (EuroSAT [7]), scene recognition (SUN397 [18]), and action recognition (UCF101 [15]). We evaluate the domain generalization performance of ImageNet [4], ImageNet-V2 [14] and ImageNet-Sketch [17]. The details of each dataset are shown in Table 1, including the number of classes, the sizes of training and testing sets, and the original tasks.

3 MORE EXPERIMENTAL RESULTS

3.1 Few-Shot Recognition Accuracy

In Section 4.2, we provide line charts to exhibit the performance of our proposed TaCo and other baseline methods, including zero-shot CLIP and [13] and CALIP [6], along with other state-of-the-art training-free methods, encompassing APE [21], Tip-X [16] and Tip-Adapter [20]. Here we provide detailed per-dataset results on all 11 recognition datasets in Table 2. We include results from those existing works for easier comparison and bold the best result for each shot and each dataset in the table. All the few-shot recognition results are based on the ResNet-50 backbone. According to the table, our proposed method demonstrates strong consistency across all types of recognition tasks and remarkably outperforms APE on Average by +1.72% across all few-shot settings.

Meanwhile, to test the additional performance improvement of TaCo when incorporated with other training-methods, we present the per-dataset accuracy result when combined with PLOT [2] and [19], respectively. As shown in Table 3, we compare the incorporated methods with PLOT [2], Tip-Adapter-F [20] and APE-T [21]. It can be observed that with the aid of TaCo, both prompt-based and adapter-based method surpasses baseline methods and on action recognition task (UCF101), our method remarkably surpasses APE-T by 4.8% with PLOT on the 8-shot setting. All the experiment results manifest the consistency and portability of our method.

REFERENCES

- [1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101-mining discriminative components with random forests. In *European Conference on Computer Vision*. 446–461.
- [2] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. 2023. PLOT: Prompt learning with optimal transport for vision-language models. In *International Conference on Learning Representations*.
- [3] Mircea Cimpoi, Subhansu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. Describing textures in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3606–3613.
- [4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 248–255.

Table 1: The detailed statistics of datasets used in experiments.

Dataset	Classes	Training size	Testing size	Task
Caltech101 [5]	100	4,128	2,465	Object recognition
DTD [3]	47	2,820	1,692	Texture recognition
EuroSAT [7]	10	13,500	8,100	Satellite image recognition
FGVCAircraft [10]	100	3,334	3,333	Fine-grained aircraft recognition
Flowers102 [11]	102	4,093	2,463	Fine-grained flowers recognition
Food101 [1]	101	50,500	30,300	Fine-grained food recognition
ImageNet [14]	1,000	1.28M	50,000	Object recognition
OxfordPets [12]	37	2,944	3,669	Fine-grained pets recognition
StanfordCars [8]	196	6,509	8,041	Fine-grained car recognition
SUN397 [18]	397	15,880	19,850	Scene recognition
UCF101 [15]	101	7,639	3,783	Action recognition
ImageNet-V2 [14]	1,000	-	10,000	Robustness of collocation
ImageNet-Sketch [17]	1,000	-	50,889	Robustness of sketch domain

Table 2: Training-free few-shot recognition accuracy on 11 datasets.

Method	Shots	Dataset											
		ImageNet	Caltech	Pets	Car	Flowers	Food	Aircraft	DTD	SUN397	EuroSAT	UCF101	Average
Zero-Shot CLIP	0	60.3	84.5	85.5	54.3	65.5	77.3	17.0	41.2	58.6	41.8	61.4	58.8
CALIP	0	60.6	87.7	86.2	56.4	66.5	77.4	17.8	42.4	58.6	38.9	61.7	59.4
Tip-Adapter	1	60.7	87.2	86.0	57.4	73.1	77.4	18.8	46.2	61.3	54.4	62.7	62.3
	2	60.9	88.3	87.0	58.5	79.1	77.5	21.2	49.6	62.7	61.5	64.7	64.6
	4	61.0	89.2	86.5	61.5	83.8	77.5	22.1	53.9	64.2	65.5	66.2	66.5
	8	61.5	89.6	86.9	63.1	87.9	77.7	25.5	58.3	65.6	67.9	68.4	68.4
	16	62.0	90.2	88.5	66.7	89.8	77.9	29.9	60.6	66.8	70.6	70.7	70.3
Tip-X	1	60.7	88.4	85.4	58.8	73.4	77.4	20.1	46.9	61.7	55.8	62.8	62.8
	2	61.0	88.6	87.9	60.3	79.4	77.5	22.0	50.1	63.4	61.5	66.4	65.3
	4	61.1	89.3	88.3	63.8	85.9	77.5	22.9	55.2	64.9	68.1	66.8	67.6
	8	61.5	89.8	89.1	65.8	88.5	77.9	26.9	59.3	66.3	68.8	68.9	69.3
	16	62.1	90.7	89.9	67.3	90.2	77.9	30.1	63.5	68.0	73.1	71.9	71.3
APE	1	62.0	90.5	86.3	59.6	79.6	77.6	20.8	52.6	64.4	59.7	63.2	65.1
	2	62.3	91.0	87.1	61.4	83.6	77.6	22.9	58.6	65.9	62.8	65.7	67.2
	4	62.5	91.8	87.3	65.1	87.9	77.6	24.4	60.7	66.6	70.3	69.9	69.4
	8	62.8	91.9	87.5	66.7	91.1	78.3	28.3	66.0	68.4	74.4	71.7	71.6
	16	63.4	92.5	88.9	70.3	91.9	78.5	31.2	67.7	69.7	78.1	74.5	73.3
TaCo (Ours)	1	62.5	91.3	87.1	61.0	81.0	78.6	23.2	53.6	65.3	61.3	65.3	66.4
	2	63.0	92.4	88.0	63.6	84.9	79.0	25.3	60.4	68.0	64.8	69.0	68.9
	4	63.6	93.2	88.4	66.3	89.4	79.0	28.7	62.6	68.7	72.2	71.6	71.2
	8	64.2	93.8	89.0	69.7	92.5	79.4	32.4	66.9	69.9	77.1	76.8	73.8
	16	65.2	94.0	90.7	73.2	94.1	80.6	35.6	70.0	71.9	81.9	78.0	75.9

Table 3: Training-based few-shot recognition accuracy on 11 datasets.

Method	Shots	Dataset											
		ImageNet	Caltech	Pets	Car	Flowers	Food	Aircraft	DTD	SUN397	EuroSAT	UCF101	Average
Zero-Shot CLIP	0	60.3	84.5	85.5	54.3	65.5	77.3	17.0	41.2	58.6	41.8	61.4	58.8
PLOT	1	59.5	89.7	87.5	56.6	71.7	77.7	17.9	46.6	62.6	54.1	64.5	62.6
	2	60.6	90.8	86.6	57.5	81.2	77.7	18.9	51.2	61.7	64.2	66.8	65.2
	4	61.5	90.8	88.6	63.4	87.8	77.2	22.4	56.0	65.1	72.4	69.6	68.6
	8	61.9	91.5	87.4	67.0	92.4	75.3	26.5	61.7	67.7	78.2	74.7	71.3
	16	63.0	92.2	87.2	72.8	94.8	77.1	31.5	65.6	70.0	82.2	77.3	74.0
Tip-Adapter-F	1	61.3	89.4	86.9	58.5	80.1	77.6	20.8	50.3	62.5	59.2	64.9	64.7
	2	61.7	89.8	87.1	62.1	82.5	77.8	23.5	54.0	63.6	65.8	66.2	66.7
	4	62.5	90.6	87.7	64.8	89.0	78.3	26.0	57.8	66.1	73.9	70.9	69.8
	8	64.0	91.5	88.1	69.5	91.9	78.7	30.2	62.7	68.8	77.8	74.5	72.5
	16	65.5	92.9	89.7	75.5	95.0	79.5	35.5	67.3	71.3	83.8	78.0	75.8
APE-T	1	62.5	90.5	87.1	60.8	83.3	77.6	24.6	54.6	66.0	65.9	66.5	67.2
	2	63.3	90.8	87.3	63.9	87.9	78.1	25.5	58.8	67.3	71.7	69.2	69.4
	4	63.7	91.7	88.5	68.4	92.1	78.6	29.1	65.4	69.0	74.8	73.4	72.2
	8	64.8	93.0	88.8	71.4	94.7	78.9	33.3	67.0	71.1	80.8	75.4	74.5
	16	66.1	93.2	90.2	76.5	96.2	79.5	39.5	69.6	72.7	86.8	79.7	77.3
TaskRes+TaCo	1	63.3	91.5	87.5	61.8	84.1	78.3	26.1	55.0	66.5	66.2	67.8	68.0
	2	64.0	92.0	87.9	64.9	88.8	79.1	27.3	58.8	68.0	72.9	71.6	70.5
	4	64.9	92.8	89.6	70.1	94.1	79.3	31.2	66.3	70.5	76.8	75.9	73.8
	8	65.9	94.5	89.9	72.9	95.7	80.3	34.6	67.8	71.7	81.0	78.6	75.7
	16	67.3	95.0	91.6	78.6	97.2	82.2	40.6	71.3	73.1	87.9	81.3	78.7
PLOT+TaCo	1	63.2	91.3	87.9	63.1	84.8	78.1	26.3	57.5	66.8	68.1	68.8	68.7
	2	64.0	92.4	88.1	65.6	89.1	78.7	28.3	62.4	68.1	74.9	71.8	71.2
	4	64.8	92.8	90.1	70.7	94.9	78.8	31.4	67.3	70.0	77.8	77.4	74.2
	8	66.1	93.6	91.5	74.6	95.3	79.6	35.1	69.6	71.6	81.2	81.2	76.3
	16	67.1	94.2	92.3	79.0	97.6	81.3	40.9	72.4	73.3	88.8	83.6	79.1

- [5] Li Fei-Fei, Rob Fergus, and Pietro Perona. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 178–178.
- [6] Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzheng Ma, Xupeng Miao, Xuming He, and Bin Cui. 2023. Calip: Zero-shot enhancement of clip with parameter-free attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 746–754.
- [7] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 2019. 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 (2019), 2217–2226.
- [8] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3D object representations for fine-grained categorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 554–561.
- [9] Zihan Liao, Fan Liu, Ang Li, and Christos Masouros. 2023. Faster-Than-Nyquist Symbol-Level Precoding for Wideband Integrated Sensing and Communications. *arXiv preprint arXiv:2306.14509* (2023).
- [10] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151* (2013).
- [11] Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*.
- [12] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. 2012. Cats and dogs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3498–3505.
- [13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.
- [14] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do imagenet classifiers generalize to imagenet?. In *International Conference on Machine Learning*.
- [15] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).
- [16] Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. 2023. Sus-x: Training-free name-only transfer of vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2725–2736.
- [17] Haoan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. 2019. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, Vol. 32.
- [18] Jianxiang Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3485–3492.
- [19] Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. 2023. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10899–10909.
- [20] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2022. Tip-adapter: Training-free adaption of clip for

[21] Xiangyang Zhu, Renrui Zhang, Bowei He, Aojun Zhou, Dong Wang, Bin Zhao, and Peng Gao. 2023. Not all features matter: Enhancing few-shot clip with

adaptive prior refinement. *arXiv preprint arXiv:2304.01195* (2023).

349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406

407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464