Noise-Tolerant Few-Shot Unsupervised Adapter for Vision-Language Models

Eman Ali[1, 2]
eman.ali@mbzuai.ac.ae
Muhammad Haris Khan[1]
muhammad.haris@mbzuai.ac.ae

[1] Mohamed Bin Zayed University of Artificial Intelligence,
Abu Dhabi, UAE
[2] Alexandria University
Alexandria, Egypt

1 More Dataset Details

The detailed statistics of each dataset and the corresponding prompt engineering are shown in Table 1. We follow [4, 5] and remove the "BACKGROUND_Google" and "Faces_easy" classes from Caltech101. For UCF101, we only take the middle frame of each video for the image encoder.

2 More Implementation Details

To generate pseudo-labels for the unlabeled target data, we adhere to the data pre-processing pipeline established by CLIP. This pipeline involves random cropping, resizing, and horizontal flipping of images. However, when constructing and fine-tuning the weighted cache model, we employ a more extensive set of data-specific augmentations, as outlined in Table 2. The use of a comprehensive set of data-specific augmentations, in addition to the standard CLIP pre-processing pipeline, is a critical factor in enhancing the effectiveness of the weighted cache model.

In instances where pseudo-labels cannot be generated for image data points, we utilize the weights of the CLIP classifier itself as image features. This strategic approach ensures that all available data is effectively incorporated into the model's training process, maximizing the utilization of the dataset. Furthermore, we draw inspiration from [4] and implement prompt ensembling for ImageNet using CLIP. This involves combining the outputs from multiple prompts to generate a more robust and accurate representation of the data. In contrast, for the remaining datasets, we employ a single handcrafted prompt specifically designed to capture the unique characteristics of each dataset.

In the context of few-shot unlabeled selection, we adopt the methodology outlined in UPL [1]. This method involves generating pseudo-labels for the entire dataset and selecting the top-k confidence samples per class to enrich the training set. To ensure a fair comparison, we utilize a large CLIP model (ViT-L-14) for generating pseudo-labels in MaPLe [2] and PromptSRC [3], maintaining consistency in the experimental setup across

Dataset	Abbreviation	Class Number	Train Set	Test Set	Prompt Engineering
ImageNet	ImgNet	1,000	1.28M	50,000	“itap of a [class].", “a bad photo of the [class]." “an origami [class].", “a photo of the large [class].", “a [class] in a video game.", “art of the [class].", “a photo of the small [class]."
Caltech101	Caltech	100	4,128	2,465	“a photo of a [class].”
DTD	DTD	47	2,820	1,692	“[class] texture.”
EuroSAT	ESAT	10	13,500	8,100	“a centred satellite photo of [class].”
FGVCAircraf	FGVCA	100	3,334	3,333	“a photo of a [class], a type of aircraft.”
Food101	Food	101	50,500	30,300	“a photo of [class], a type of food.”
Flowers102	Flower	102	4,093	2,463	“a photo of a [class], a type of flower.”
OxfordPets	OxPets	37	2,944	3,669	“a photo of a [class], a type of pet.”
SUN397	SUN	397	15,880	19,850	“a photo of a [class].”
StandfordCars	StCars	196	6,509	8,041	“a photo of a [class].”
UCF101	UCF	101	7,639	3,783	“a photo of a person doing [class].”

Table 1: The detailed dataset statistics and the corresponding handcraft prompts.

Dataset	Abbreviation	Data Augmentation
ImageNet	ImgNet	Random Horizontal Flipping
Caltech101	Caltech	Random Horizontal Flipping + Random Affine
DTD	DTD	Random Horizontal Flipping
EuroSAT	ESAT	ColorJitter
FGVCAircraf	FGVCA	Random Horizontal Flipping + Random Affine
Food101	Food	Random Affine
Flowers102	Flower	Random Horizontal Flipping
OxfordPets	OxPets	Random Horizontal Flipping
SUN397	SUN	Random Horizontal Flipping + Random Affine
StandfordCars	StCars	Random Affine
UCF101	UCF	Random Horizontal Flipping + Random Affine

Table 2: The detailed data augmentations utilized for each dataset.

different methodologies. Finally, The hyperparameters α and β are set following the values specified in [4] for consistency and comparability.

3 More Experimental Results

Different number of shots: To assess the effectiveness of NtUA, we manipulated the quantity of unlabeled data. Remarkably, even when utilizing a minimal amount of unlabeled data, specifically 2, 4, or 8 samples per class (refer to Tables 3, 4, and 5), NtUA consistently outperformed alternative methodologies.

Methods	ImgNet	Caltech	DTD	ESAT	FGVCA	Food	Flower	OxPets	SUN	StCars	UCF	Average
CLIP-ViT-B/32	63.77	91.48	44.09	45.27	19.17	80.40	66.59	87.44	62.08	60.12	63.47	62.17
UPL	59.06	91.48	46.22	**54.67**	6.93	78.70	64.68	82.69	**64.33**	52.92	63.52	60.47
UPL*	59.46	92.17	**46.87**	48.67	5.22	78.55	66.71	84.87	63.76	54.57	66.93	60.71
LaFTer	56.90	85.11	44.33	33.89	16.95	75.84	66.42	80.08	58.26	47.26	60.67	56.88
MaPLe	61.26	90.22	41.43	18.21	17.94	79.22	63.05	79.50	64.00	53.77	63.73	57.48
PromptSRC	58.03	86.90	43.68	34.27	17.37	76.43	67.32	82.37	60.07	49.87	61.38	57.97
NtUA (ours)	**64.49**	**92.90**	46.81	52.21	**20.70**	**80.66**	**72.76**	**89.62**	63.13	**61.56**	**67.86**	**64.79**
Supervised	64.95	93.75	56.38	72.26	25.8	80.93	84.94	89.26	66.48	65.03	71.27	70.10

Table 3: Comparison of NtUA with Five SOTA adaptation methods over 10 widely adopted image classification benchmarks. We leverage CLIP-ViT-B/32 as the backbone model and evaluate performance in a 2-shot setting.

Methods	ImgNet	Caltech	DTD	ESAT	FGVCA	Food	Flower	OxPets	SUN	StCars	UCF	Average
CLIP-ViT-B/32	63.77	91.48	44.09	45.27	19.17	80.40	66.59	87.44	62.08	60.12	63.47	62.17
UPL	59.68	92.41	48.29	55.89	16.77	79.34	67.24	82.61	64.08	54.06	64.39	62.25
UPL*	60.93	92.41	48.17	50.26	15.24	78.58	72.55	84.66	64.15	57.37	**69.13**	63.04
LaFTer	58.67	88.97	47.22	51.73	17.22	76.35	67.97	84.16	60.97	51.24	63.71	60.75
MaPLe	62.55	91.32	37.29	39.01	3.15	79.85	64.15	83.84	63.82	51.24	64.47	58.24
PromptSRC	60.08	91.12	45.15	39.19	17.37	77.16	68.98	83.62	61.92	53.54	64.76	60.26
NtUA (ours)	**65.11**	**94.12**	**49.76**	**61.40**	**18.51**	**80.85**	**73.20**	**89.21**	**64.30**	**62.13**	67.38	**66.00**
Supervised	65.83	94.73	60.87	77.2	28.05	81.23	90.26	89.29	68.78	68.25	75.6	62.64

Table 4: Comparison of NtUA with Five SOTA adaptation methods over 10 widely adopted image classification benchmarks. We leverage CLIP-ViT-B/32 as the backbone model and evaluate performance in a 4-shot setting.

Methods	ImgNet	Caltech	DTD	ESAT	FGVCA	Food	Flower	OxPets	SUN	StCars	UCF	Average
CLIP-ViT-B/32	63.77	91.48	44.09	45.27	19.17	80.40	66.59	87.44	62.08	60.12	63.47	62.17
UPL	61.07	92.37	48.46	58.67	17.55	79.61	67.93	84.55	64.94	54.98	64.18	63.12
UPL*	61.86	92.09	**52.48**	52.44	**21.42**	79.19	75.11	86.24	65.53	60.85	69.23	65.13
LaFTer	59.69	90.99	45.98	50.65	18.30	77.76	69.10	82.45	61.72	53.74	64.74	61.37
MaPLe	62.47	91.81	43.74	28.58	18.96	80.37	65.25	85.12	63.72	55.15	62.44	59.78
PromptSRC	61.34	91.24	46.22	48.25	20.61	78.79	71.05	83.59	62.91	55.09	65.05	62.19
NtUA (ours)	**65.82**	**93.71**	51.95	**59.99**	20.25	**81.39**	**76.53**	**89.59**	**65.85**	**65.56**	**69.60**	**67.29**
Supervised	67.23	94.85	65.07	80.35	33.09	81.66	93.18	89.67	71.56	73.10	79.20	75.36

Table 5: Comparison of NtUA with Five SOTA adaptation methods over 10 widely adopted image classification benchmarks. We leverage CLIP-ViT-B/32 as the backbone model and evaluate performance in an 8-shot setting.

References

[1] Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*, 2022.

[2] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023.

[3] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational

model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15190–15200, 2023.

[4] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European Conference on Computer Vision*, pages 493–510. Springer, 2022.

[5] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.