Supplementary to Visual-Language Prompt Tuning with Knowledge-guided Context Optimization

Hantao Yao[1],Rui Zhang[2], Changsheng Xu[1,3]

[1] State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, CAS

[2] State Key Lab of Processors, Institute of Computing Technology, CAS; [3] University of Chinese Academy of Sciences(CAS),

{hantao.yao,csxu}@nlpr.ia.ac.cn; zhangrui@ict.ac.cn

1. Comparison for Cross-Dataset Transfer

Similar to CoOp [4] and CoCoOp [3], we also evaluate the generalizability of the KgCoOp by applying the learnable prompts inferred from the source dataset (ImageNet) on the other downstream dataset. The related results are shown in Table 1. As shown in Table 1, CoCoOp obtains the best average performance of all existing methods. The reason is that the prompts in CoCoOp are a combination of textual prompts and visual descriptions, leading to CoCoOp having high generalizability on unseen datasets. However, CoCoOp is a time-consuming method. Different from CoCoOp, CoOp, ProGrad [5] and the proposed KgCoOp only use textual-based prompts. Compared to CoOp and ProGrad, KgCoOp obtains a higher performance on almost all datasets except EuroSAT. The superior performance proves that the proposed KgCoOp has a high generalizability for cross-dataset transfer.

Table 1. Comparison in the cross-dataset transfer learning by learning the prompts from ImageNet(16-shot samples) with ViT-16/B, and evaluating on the other 10 datasets. "tp" denotes the "textual prompt", and "v" denotes the visual information of each instance.

	Methods	CoCoOp	CoOp	ProGrad	KgCoOp
	Prompts	tp+v		tp	
Source	ImageNet	71.02	71.51	**72.24**	70.66
Targets	Caltech101	94.43	93.70	91.52	**93.92**
	OxfordPets	90.14	89.14	89.64	**89.83**
	StandfordCars	65.32	64.51	62.39	**65.41**
	Flowers	71.88	68.71	67.87	**70.01**
	Food101	86.06	85.30	85.40	**86.36**
	FGVCAircraft	22.94	18.47	20.61	**22.51**
	SUN397	67.36	64.15	62.47	**66.16**
	DTD	45.73	41.92	39.42	**46.35**
	EuroSAT	45.37	**46.39**	43.46	46.04
	UCF101	68.21	66.55	64.29	**68.50**
	Avg.	65.74	63.88	62.71	**65.51**

2. Effect of Context Length

For the learnable prompts, the context length is a critical aspect. We thus analyze the effect of the context length in the base-to-new generalization setting with the backbone of ViT-16/B. Similar to CoOp [4], we study 4, 8, and 16 context tokens. For the context length of 8 and 16, the prompt is initialized with "X X ... X a photo of a [Class]". The averaging performance on 11 datasets is summarized in Figure 1. We can observe that setting the context length as 8 obtains a higher performance than the other two settings on all three metric terms. Furthermore, the learning prompt with lengths of 4 and 16 obtain similar performance. However, for making a fair comparison with CoOp and CoCoOp, the context length is set as 4 in our final model.

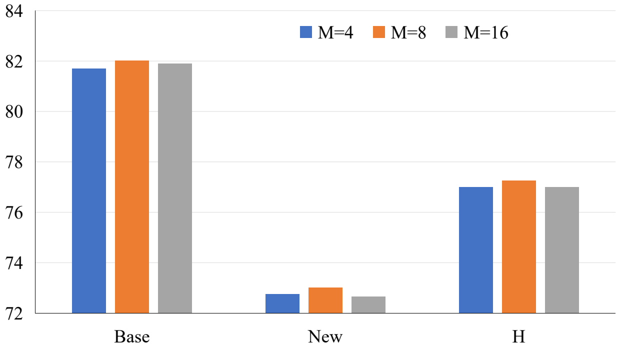

Figure 1. Effect of context length.

3. Effect of Initialization

To verify the impact of initialization for prompt tuning, we conduct a comparison based on the word embeddings-based initialization('w/ init') and random initialization('w/o init'). The random initialization applies a zero-mean Gaussian distribution with 0.02 standard deviation to initialize the prompt tokens, and the word embeddings-based initialization uses the "a photo of a" to initialize the prompt tokens. The averaging performance on 11 datasets is summarized in Figure 2. We can observe that using the word embedding-based initialization obtains a higher per-

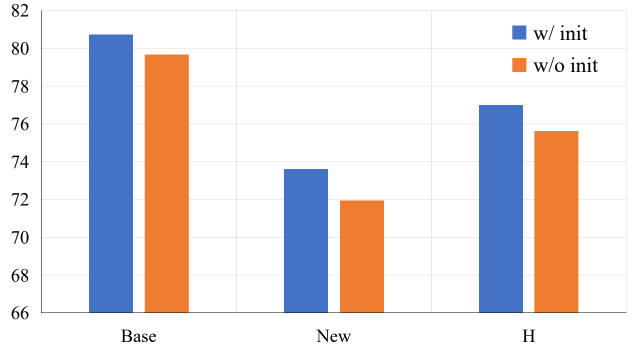

Figure 2. Effect of initialization.

formance in all three terms than random initialization.

4. Effect of hand-crafted prompts

As different hand-crafted prompts would provide different knowledge to constrain the prompt tuning, we thus evaluate the effect of different hand-crafted prompts. Evaluation on six hand-crafted prompts shows in Table 3, *i.e.,* T1:'{}'; T2:'a photo of a {}'; T3:'itap of a {}'; T4:'a photo of the large {}'; T5:'a {} in a video game'; T6:'a photo of a {}, a type of {}'. Although different hand-crafted prompots have achieved different performnce, we observe that T1 without using any prompts obtains the performance of 76.02%. Furthermore, the more information given by the hand-crafted prompts, the higher performance, *e.g.,* T6 obtains the highest performance.

5. How to reduce the discrepancy between special knowledge and general knowledge?

The key insight of our work is to reduce the discrepancy between special knowledge and general knowledge for improving the generability of unseen datasets. In KgCoOp, $\mathcal{L}_{kg}$ is used to minimize the distance between the general textual embeddings and specific textual embeddings for reducing the discrepancy. For the CoOp-based methods, they exist other two ways to measure the discrepancy between special knowledge and general knowledge besides $\mathcal{L}_{kg}$: 1) $\mathcal{L}_{pt}$:reducing the distance between the tokens of the learnable prompts and the fixed prompts; 2) $\mathcal{L}_{kl}$: using the Kullback-Leibler divergence measure the consistency between the predictions generated by the general textual embeddings and specific textual embeddings. We thus conduct a comparison among all three methods and summarize the results in Table 4. As shown in Table 4, using $\mathcal{L}_{pt}$ obtains a worse performance of H than CoOp, demonstrating the direct constrain of the similarity between prompts is not a reasonable way. Different from $\mathcal{L}_{pt}$, $\mathcal{L}_{kg}$ and $\mathcal{L}_{kl}$ both obtain a higher performance than CoOp. Furthermore, the proposed KgCoOp using $\mathcal{L}_{kg}$ obtains the best performance in the terms of *New* and *H*. The superior performance proves that it is reasonable to mitigate knowledge forgetting by minimizing the distance between embeddings.

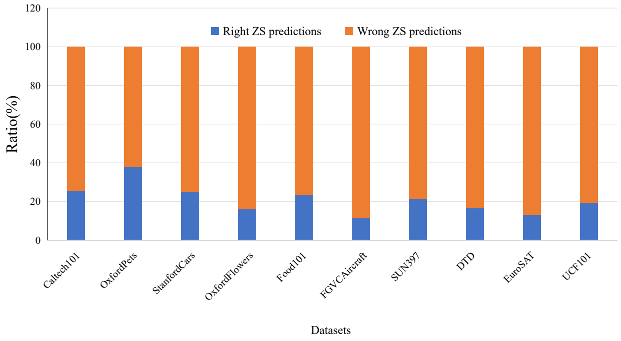

Figure 3. Failure cases analysis. We evaluate the distribution of samples that are mis-classified by KgCoOp but correctly classified by CoOp models

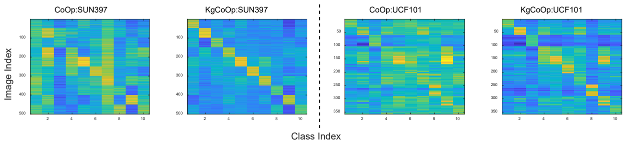

Figure 4. Confusion matrix of the prediction.ls

6. Failure cases

Similar to ProGrad, we analyze the failure cases where KgCoOp predict incorrectly but CoOp gives right predictions. Specifically, we count the percentage of the failure cases that zero-shot CLIP models also fails in Figure 4. We observe that a high proportion of the faiure cases are misclassified by CoOp model.

7. Disscussion about the generalization on new class

As show in Table 2, the proposed methods obtains the lower performance on the new class. The reason is that the domain discrepancy between seen and new classes affects the hardness of generalization to new classes. Specially, from the Table 3 in the paper, CoOp obtains more than 10% *New* performance drop on DTD, EuroSAT, and UCF101 datasets. The reason is that the new classes have a serious domain gap with the seen classes, making the learned prompt biased to the new classes (CoOp in Fig. 4). KgCoOp constrains the learnable prompts to contain the general knowledge in CLIP and discriminative to a new class(Fig. 4). Therefore, KgCoOp significantly improves CoOp for the new classes on those three datasets.

Table 2. Comparison in the base-to-new setting with different K-shot samples in terms of the average performance among all 11 datasets and backbones(ViT-B/16 and ResNet-50).

Backbones	Methods	K=4			K=8			K=16		
		Base	New	H	Base	New	H	Base	New	H
ViT-B/16	CoOp	78.43	68.03	72.44	80.73	68.39	73.5	82.63	67.99	74.60
	CoCoOp	76.72	**73.34**	74.85	78.56	72.0	74.9	80.47	71.69	75.83
	ProGrad	79.18	71.14	74.62	**80.62**	71.02	75.2	**82.48**	70.75	76.16
	KgCoOp	**79.92**	73.11	**75.90**	78.36	**73.89**	**76.06**	80.73	**73.6**	**77.0**
ResNet-50	CoOp	72.06	59.69	65.29	74.72	58.05	65.34	77.24	57.4	65.86
	CoCoOp	71.39	65.74	68.45	73.4	66.42	69.29	75.2	63.64	68.9
	ProGrad	**73.88**	64.95	69.13	**76.25**	64.74	70.03	**77.98**	64.41	69.94
	KgCoOp	72.42	**68.00**	**70.14**	74.08	**67.86**	**70.84**	75.51	**67.53**	**71.30**

Table 3. Effect of hand-crafted prompts.

Methods	CoOp	CoCoOp	ProGrad	T1	T2
H	74.60	75.83	76.16	76.02	76.85
Methods	T3	T4	T5	T6	
H	76.23	76.71	76.12	77.0	

Table 4. Comparison of different measurement methods on the average performance of all 11 datasets in the base-to-new setting.

Methods	Base	New	H
Baseline(CoOp)	82.63	67.99	74.60
CoOp+$\mathcal{L}_{pt}$	78.84	70.67	74.53
CoOp+$\mathcal{L}_{kl}$	80.42	72.43	76.22
CoOp+$\mathcal{L}_{kg}$	80.73	73.6	77.0

8. Detailed Results

To verify the effectiveness of the proposed KgCoOp, we compare KgCoOp with existing CoOp-based methods, *i.e,* CoOp [4], CoCoOp [3], and ProGrad [5], based on different backbones and different K-shot samples. Specifily, the CNN-based model ResNet-50 [2] and the transformer-based model ViT-B/16 [1] are applied as the visual encoder to extract the image's description. Furthermore, three types of few-shot settings, *i.e.,* 4-shot, 8-shot, and 16-shot, are conducted for comparison. The summarized averaged results are shown in Table 2. The detailed results of the backbone of ViT-B/16 are shown in Table 5 and Table 6 for 4-shot and 8-shot settings. For ResNet-50, the results of 4-shot, 8-shot, and 16-shot settings are shown in Table 7, Table 8, and Table 8, respectively.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 3

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 3

[3] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16795–16804. IEEE, 2022. 1, 3

[4] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *Int. J. Comput. Vis.*, 130(9):2337–2348, 2022. 1, 3

[5] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. *CoRR*, abs/2205.14865, 2022. 1, 3

Table 5. Comparison with existing methods in the base-to-new generalization based on the **ViT-B/16** and **4-shot** settings. The context length M is 4 for prompot-based methods. H: Harmonic mean.

	CoOp			CoCoOp			ProGrad			KgCoOp		
Datasets	Base	New	H	Base	New	H	Base	New	H	Base	New	H
ImageNet	73.60	63.29	68.06	75.46	69.58	72.40	74.24	65.47	69.58	74.87	69.09	71.86
Caltech101	97.27	93.01	95.09	97.25	94.90	96.06	97.37	93.92	95.61	97.53	94.43	95.95
OxfordPets	93.33	95.69	94.50	94.59	96.75	95.66	94.08	97.63	95.82	94.68	97.58	96.11
StandfordCars	70.92	69.38	70.14	67.71	75.37	71.33	72.69	69.88	71.26	69.25	74.98	72.00
Flowers	92.50	70.12	79.77	84.75	73.85	78.93	92.46	72.69	81.39	91.30	75.34	82.56
Food101	86.79	89.06	87.91	89.79	90.99	90.39	88.91	90.18	89.54	90.30	91.39	90.84
FGVCAircraft	33.21	28.57	30.72	32.07	33.93	32.97	33.73	30.09	31.81	34.21	32.81	33.50
SUN397	76.49	64.56	70.02	77.57	76.96	77.26	77.72	71.93	74.71	78.87	75.64	77.22
DTD	71.26	50.93	59.40	67.44	56.00	61.19	71.06	52.58	60.44	73.65	57.21	64.40
EuroSAT	82.56	53.04	64.59	79.27	65.44	71.69	82.48	56.43	67.01	82.63	59.98	69.51
UCF101	79.97	65.98	72.30	78.01	73.07	75.46	81.30	76.02	78.57	80.80	75.77	78.20
Avg.	78.43	68.03	72.44	76.72	73.35	74.85	79.18	71.14	74.62	78.92	73.11	75.90

Table 6. Comparison with existing methods in the base-to-new generalization based on the **ViT-B/16** and **8-shot** settings. The context length M is 4 for prompot-based methods. H: Harmonic mean.

	CoOp			CoCoOp			ProGrad			KgCoOp		
Datasets	Base	New	H	Base	New	H	Base	New	H	Base	New	H
ImageNet	75.22	65.91	70.26	75.52	70.28	72.81	75.72	66.76	70.96	75.84	69.33	72.44
Caltech101	97.81	92.58	95.12	97.76	93.63	95.65	98.00	93.38	95.63	97.68	94.10	95.86
OxfordPets	94.19	96.11	95.14	95.50	97.69	96.58	94.47	97.03	95.73	94.81	97.58	96.18
StandfordCars	73.20	67.44	70.20	69.70	74.13	71.85	75.08	70.63	72.79	69.66	75.40	72.42
Flowers	96.17	69.41	80.63	92.24	72.77	81.36	93.80	72.20	81.59	87.72	74.75	80.72
Food101	87.27	86.96	87.11	89.60	90.79	90.19	89.48	89.90	89.69	90.46	91.63	91.04
FGVCAircraft	37.01	38.45	37.72	33.71	32.15	32.91	36.89	31.67	34.08	34.53	34.95	34.74
SUN397	78.61	66.25	71.90	78.05	76.29	77.16	79.21	70.77	74.75	79.37	76.85	78.09
DTD	76.97	51.81	61.93	73.03	57.24	64.18	74.42	52.38	61.48	69.72	56.44	62.38
EuroSAT	83.27	50.59	62.94	78.68	56.03	65.45	82.27	58.52	68.39	81.07	63.13	70.98
UCF101	82.85	64.32	72.42	80.40	71.68	75.79	82.61	73.75	77.93	81.16	78.65	79.89
Avg.	80.74	68.39	73.51	78.56	72.06	74.90	80.62	71.02	75.21	78.37	73.89	76.06

Table 7. Comparison with existing methods in the base-to-new generalization based on the **ResNet-50** and **4-shot** settings. The context length M is 4 for prompot-based methods. H: Harmonic mean.

	CoOp			CoCoOp			ProGrad			KgCoOp		
Datasets	Base	New	H	Base	New	H	Base	New	H	Base	New	H
ImageNet	64.53	54.47	59.07	67.80	62.45	65.02	65.23	55.96	60.24	67.13	61.96	64.44
Caltech101	94.06	87.01	90.40	95.03	90.47	92.69	94.47	89.26	91.79	94.43	91.56	92.97
OxfordPets	87.36	93.49	90.32	91.62	94.99	93.27	91.25	94.93	93.05	92.29	94.13	93.20
StandfordCars	61.84	57.25	59.46	60.58	64.78	62.61	64.98	61.92	63.41	60.53	67.42	63.79
Flowers	89.71	57.68	70.21	81.86	71.44	76.30	90.12	68.82	78.04	78.12	72.77	75.35
Food101	77.20	76.85	77.02	83.19	84.53	83.85	81.48	82.54	82.01	83.56	84.86	84.20
FGVCAircraft	22.19	18.36	20.09	22.55	25.03	23.73	23.47	18.44	20.65	22.53	26.83	24.49
SUN397	70.68	60.87	65.41	72.03	71.76	71.89	73.53	67.04	70.14	73.68	71.92	72.79
DTD	64.74	47.18	54.58	61.77	53.34	57.25	67.90	52.94	59.49	66.24	53.54	59.22
EuroSAT	86.39	46.91	60.80	75.60	37.68	50.29	84.74	60.46	70.57	84.87	52.55	64.91
UCF101	73.96	56.53	64.08	73.27	66.70	69.83	75.56	62.13	68.19	73.20	70.43	71.79
Avg.	72.06	59.69	65.29	71.39	65.74	68.45	73.88	64.95	69.13	72.42	68.00	70.14

Table 8. Comparison with existing methods in the base-to-new generalization based on the **ResNet-50** and **8-shot** settings. The context length M is 4 for prompot-based methods. H: Harmonic mean.

	CoOp			CoCoOp			ProGrad			KgCoOp		
Datasets	Base	New	H	Base	New	H	Base	New	H	Base	New	H
ImageNet	66.69	57.36	61.67	68.06	62.71	65.28	67.25	57.83	62.19	67.62	62.27	64.83
Caltech101	94.40	83.88	88.83	95.31	91.05	93.13	95.12	88.97	91.94	94.92	91.88	93.38
OxfordPets	90.02	93.36	91.66	92.45	95.73	94.06	91.90	94.59	93.23	92.36	94.37	93.35
StandfordCars	65.49	55.89	60.31	61.61	65.98	63.72	68.33	60.10	63.95	60.91	66.55	63.61
Flowers	93.07	57.59	71.15	85.25	68.56	76.00	92.46	67.59	78.09	87.18	72.67	79.27
Food101	78.55	78.03	78.29	84.09	85.37	84.73	82.50	83.36	82.93	83.74	85.21	84.47
FGVCAircraft	25.01	18.04	20.96	23.17	23.60	23.38	27.71	20.58	23.62	24.15	26.83	25.42
SUN397	73.58	60.95	66.67	73.53	72.52	73.02	75.13	67.03	70.85	74.63	72.21	73.40
DTD	71.53	40.34	51.59	68.29	49.76	57.57	71.61	47.58	57.17	69.25	51.57	59.12
EuroSAT	85.88	42.46	56.83	80.43	48.75	60.71	87.45	59.75	70.99	83.87	52.80	64.80
UCF101	77.69	50.64	61.31	75.23	66.54	70.62	79.30	64.81	71.33	76.28	70.18	73.10
Avg.	74.72	58.05	65.34	73.40	66.42	69.29	76.25	64.74	70.03	74.08	67.87	70.84

Table 9. Comparison with existing methods in the base-to-new generalization based on the **ResNet-50** and **16-shot** settings. The context length M is 4 for prompot-based methods. H: Harmonic mean.

	CoOp			CoCoOp			ProGrad			KgCoOp		
Datasets	Base	New	H	Base	New	H	Base	New	H	Base	New	H
ImageNet	68.57	58.76	63.29	68.21	62.28	65.11	69.13	57.39	62.72	67.67	62.45	64.96
Caltech101	95.20	87.55	91.21	95.40	90.28	92.77	95.72	89.92	92.73	95.35	91.92	93.60
OxfordPets	90.15	90.70	90.42	92.10	95.81	93.92	92.36	94.48	93.41	92.57	94.61	93.58
StandfordCars	68.89	57.13	62.46	63.53	64.46	63.99	71.79	59.36	64.99	63.28	66.92	65.05
Flowers	95.22	59.53	73.26	90.66	67.19	77.18	94.71	68.86	79.74	91.45	71.75	80.41
Food101	81.70	78.13	79.88	84.44	85.80	85.11	83.77	83.74	83.75	83.90	85.23	84.56
FGVCAircraft	28.39	20.02	23.48	23.98	21.05	22.42	30.17	19.70	23.84	24.91	25.69	25.29
SUN397	76.33	62.89	68.96	74.64	72.78	73.70	76.90	68.09	72.23	75.33	72.25	73.76
DTD	75.12	37.08	49.65	71.18	47.42	56.92	73.80	46.38	56.96	74.73	48.39	58.74
EuroSAT	90.25	31.30	46.48	86.13	31.65	46.29	88.44	49.49	63.47	84.28	53.53	65.47
UCF101	79.78	48.31	60.18	76.92	61.38	68.28	81.04	60.07	69.00	77.16	70.13	73.48
Avg.	77.24	57.40	65.86	75.20	63.65	68.94	77.98	63.41	69.94	75.51	67.53	71.30