Supplementary material for LASP: Text-to-Text Optimization for Language-Aware Soft Prompting of Vision-Language Models

Adrian Bulat[1,2], Georgios Tzimiropoulos[1,3]
[1]Samsung AI Cambridge [2]Technical University of Iasi [3]Queen Mary University of London

A. Additional results

A.1. Cross-Dataset Transfer

Following [7], we measure how well the soft prompts learned on ImageNet perform when evaluated on different datasets. In this setting, the training is performed on images from all 1,000 classes, using 16 images for each class. As the results from Table 1 show, our approach surpasses CoOp by 2.5% while outperforming the more computationally demanding CoCoOp (0.8% better on average).

A.2. Domain generalization

Following the encouraging results reported in [7, 8] on domain generalization, herein we attempt to evaluate whether our approach can improve the quality of the leaned prompts under domain shift too. To this end, we trained LASP on all classes from ImageNet (16-shot setting) and evaluate the learned prompts on 5 datasets with class names compatible with those of ImageNet, but different data distribution. Following [8], we used ImageNet [1] as the source dataset, and ImageNetV2 [5], ImageNet-Sketech [6], ImageNet-A [2] and ImageNet-R [3] as the test datasets.

As the results from Table 2 show, with the exception of ImageNet-V2, our approach outperforms all prior work, showing strong domain generalization capabilities.

A.3. Effect of LN fine-tuning on CoOp

Herein, we analyze the effect of fine-tuning the LN layers of the vision encoder directly on top of our baseline, *i.e.* CoOp. As the results from Table 3 show, the improvements, especially on the new classes, are small. This shows that LN fine-tuning alone is not enough for obtaining high accuracy.

A.4. Combining CLIP with CoOp

To further show the effectiveness of our approach, we compare it with an ensemble formed by combining CLIP and CoOp. Following CLIP [4], the ensemble is formed by taking the average over the logits produced by CoOp using the learned prompts and, respectively, by CLIP using the hand-crafted templates. Perhaps unsurprisingly, the ensemble outperforms CoOp on the new classes and is outperformed on the base ones. LASP largely outperforms all these variants showcasing the importance of the proposed formulation.

A.5. Training and inference speed considerations.

Once trained, LASP is as fast as CoOp. For training, LASP, CoOp and CoCoOp differ only in the text encoder whose cost is $G \cdot M \cdot C_T$, $M \cdot C_T$ and $B \cdot M \cdot C_T$, respectively, where B is the batch size, M is the number of classes and C_T is the text encoder's cost for 1 sample. In practice, for B = 32, LASP is, on average, 2.3x slower than CoOp and up to 10x faster than CoCoOp. Note that these numbers are subject to the implementation optimizations made for each method. For G=1, LASP's training cost is the same as CoOp's while losing only 0.5% on average.

A.6. Generalized zero-shot results

As mentioned in the main paper, the current evaluation protocol used in [7, 8] computes the accuracy considering the base and new classes in isolation. A more realistic evaluation protocol should consider the classes across both subsets (i.e. base and novel) jointly. We report results using this setting in Table 5. To ground the results, as no pretrained models where available, we retrain CoCoOp using the official code released by the authors. As it can be observed, the same conclusions, previously made using the protocol proposed in [8] hold true.

B. Implementation details

Hand-engineered prompts set ζ: Unless otherwise specified, we used the following set of hand-engineered templates (borrowed from CLIP and CoOp):

```
"a photo of a {}, a type of flower.",
"a photo of a person doing {}.",
"a centered satellite photo of {}.",
"a photo of a {}, a type of aircraft.",
"{} texture.",
"itap of a {}.",
"a bad photo of the {}.",
"a origami {}.",
```

Table 1. **Comparison with state-of-the-art for the cross-dataset transfer setting.**

	Source	Target										
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	*Average*
CoOp	**71.51**	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
CoCoOp	71.02	94.43	**90.14**	65.32	**71.88**	86.06	22.94	**67.36**	**45.73**	45.37	68.21	65.74
LASP	71.30	**94.50**	89.36	**66.20**	71.74	**86.40**	**23.03**	67.0	45.54	**48.50**	**68.24**	**66.52**

Table 2. **Comparison with state-of-the-art for the domain generalization setting.**

		Source	Target			
	Learnable?	ImageNet	ImageNetV2	ImageNet-Sketch	ImageNet-A	ImageNet-R
CLIP		66.73	60.83	46.15	47.77	73.96
CoOp	✓	**71.51**	**64.20**	47.99	49.71	75.21
CoCoOp	✓	71.02	64.07	48.75	50.63	76.18
LASP	✓	71.10	63.96	**49.01**	**50.70**	**77.07**

Table 3. **Effect of LN fine-tuning on CoOp.** The results reported represent the average accuracy over the 11 datasets.

Method	Base	New	H
Baseline (CoOp)	82.69	63.22	71.66
CoOp + LN	82.80	65.17	72.94

Table 4. **Comparison between LASP and an ensemble formed by CLIP and CoOp.** The results reported represent the average accuracy over the 11 datasets.

Method	Base	New	H
Baseline (CoOp)	82.69	63.22	71.66
CoOp + CLIP	78.50	70.1	74.06
LASP	82.7	74.9	78.61
LASP-V	**83.18**	**76.11**	**79.48**

```
"a photo of the large {}.",
"a {} in a video game.",
"art of the {}.",
"a photo of the small {}.",
"a photo of a {}.",
"a photo of many {}.",
"a photo of the hard to see {}.",
"a low resolution photo of the {}.",
"a rendering of a {}.",
"a bad photo of the {}.",
"a cropped photo of the {}.",
"a pixelated photo of the {}.",
"a bright photo of the {}.",
"a cropped photo of a {}.",
"a photo of the {}.",
"a good photo of the {}.",
"a rendering of the {}.",
"a close-up photo of the {}.",
"a low resolution photo of a {}.",
"a rendition of the {}.",
"a photo of the clean {}.",
"a photo of a large {}.",
"a blurry photo of a {}.",
"a pixelated photo of a {}.",
"itap of the {}.",
"a jpeg corrupted photo of the {}.",
"a good photo of a {}."
```

Note that {} represent the placeholder for the location of the class name w.

Random prompts: For the experiments involving random prompts, we list bellow a few such examples:

```
"Ports, waterways, the subfield that
{}.",
"In TCP, prepared mind, but some
others, Milatiai, appear to have {}.",
"Iron Age, The Eastern Shore of
Virginia residents age 5 and {}.",
"Cat mostly all with {}.",
"Wind erosion. go unnoticed|it was
{}.",
"River Delta, on six different {}.",
"12 hours. few times every million
```

Table 5. **Comparison with the state-of-the-art for the generalized zero-shot setting**. We have re-trained CoCoOp using the officially released code.

(a) **Average over 11 datasets**.

	Base	New	H
CoCoOp	72.46	64.77	68.39
LASP	76.59	67.55	71.78
LASP-V	**77.23**	**68.52**	**72.61**

(b) ImageNet.

	Base	New	H
CoCoOp	71.9	67.5	69.63
LASP	**72.0**	67.33	69.51
LASP-V	71.9	**68.0**	**69.78**

(c) Caltech101.

	Base	New	H
CoCoOp	95.20	90.67	92.87
LASP	94.87	92.20	93.51
LASP-V	95.54	**92.78**	**94.13**

(d) OxfordPets.

	Base	New	H
CoCoOp	91.01	93.10	92.04
LASP	91.53	92.87	92.19
LASP-V	**92.23**	**93.17**	**92.69**

(e) StanfordCars.

	Base	New	H
CoCoOp	67.26	**69.43**	68.33
LASP	**72.27**	68.73	**70.45**
LASP-V	71.0	68.50	69.27

(f) Flowers102.

	Base	New	H
CoCoOp	86.73	64.63	74.06
LASP	90.97	68.80	78.34
LASP-V	**92.20**	**69.93**	**79.53**

(g) Food101.

	Base	New	H
CoCoOp	85.73	85.50	85.61
LASP	87.53	**87.17**	87.34
LASP-V	**87.73**	**87.17**	**87.45**

(h) FGVCAircraft.

	Base	New	H
CoCoOp	24.50	25.93	25.19
LASP	24.33	27.03	25.61
LASP-V	**28.77**	**27.80**	**28.27**

(i) SUN397.

	Base	New	H
CoCoOp	71.13	67.76	69.40
LASP	**72.60**	67.21	69.80
LASP-V	72.55	**69.11**	**70.79**

(j) DTD.

	Base	New	H
CoCoOp	59.33	42.70	49.65
LASP	**67.53**	46.93	55.37
LASP-V	65.67	**49.90**	**56.71**

(k) EuroSAT.

	Base	New	H
CoCoOp	69.20	39.23	50.14
LASP	89.38	54.87	67.99
LASP-V	**90.80**	**56.80**	**69.88**

(l) UCF101.

	Base	New	H
CoCoOp	75.16	66.10	70.34
LASP	79.57	70.0	74.47
LASP-V	**81.20**	**70.60**	**75.52**

```
{}.",
etc.
```

Additional class names for in-domain ablation: Below, we list the manually defined in-domain class name distractors used to produce the results for with in-domain distractors. For Food-101, we added the following classes:

['aroma', 'bagel', 'batter', 'beans', 'biscuit', 'broth', 'burger', 'burrito', 'butter', 'candy', 'caramel', 'caviar', 'cheese', 'chili', 'chimichanga', 'cider', 'cocoa', 'coffee', 'cobbler', 'empanada', 'fish', 'flour', 'ketchup', 'margarine', 'mousse', 'muffin', 'mushrooms', 'noodle', 'nuts', 'oil', 'olives', 'pudding', 'raclette', 'rice', 'salad', 'salsa', 'sandwitch', 'soda', 'tea', 'stew', 'toast', 'waffles', 'yogurt', 'wine', 'sopapillas', 'chilli con carne', 'banana bread', 'yorkshire pudding', 'spaghetti carbonara', 'roast potatoes', 'sausage ragu', 'avocado panzanella', 'lamb biryani']

Respectively, for Flowers102 dataset:

['Agapanthus', 'Allium', 'Alstroemerias', 'Amaranthus', 'Astilbe', 'Begonia', 'brunia', 'California poppy', 'Calla lily', 'Campanula', 'Carnations', 'Celosia', 'Chrysanthemum', 'Cornflower', 'Delphinium', 'Dianthus', 'Dusty Miller', 'Eryngium', 'Freesia', 'Gardenias', 'Gerbera daisies', 'Gladiolus', 'Gypsophila', 'Hydrangea', 'Hypericum', 'Kale', 'Larkspur', 'Liatris', 'Lilies', 'Lisianthus', 'Orchids', 'Peony', 'Periwinkle', 'Ranunculus', 'Scabiosa', 'Sunflowers', 'Yarrow', 'Zinnia', 'Bellflower', 'Bleeding Heart', 'Browallia', 'Bugleweed', 'Butterfly Weed', 'Calendula', 'Cardinal Flower', 'Celosia', 'Clary Sage', 'Coreopsis', 'Forget-Me-Not', 'Freesias', 'Gaillardia', 'Glory of the Snow', 'Heather', 'Hollyhock', 'Hyssop', 'Impatiens', 'Jack-in-the-Pulpit', 'Lilac', 'Lilies', 'Lobelia', 'Periwinkle', 'Rue', 'Thunbergia', 'Verbena', 'Wisteria']

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[2] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 1

[3] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceed-*

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15262–15271, 2021. 1

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[5] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 1

[6] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[7] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 1

[8] 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. 1