Supplementary Material: *Prompt Generation Networks for Input-Space Adaptation of Frozen Vision Transformers*

Jochem Loedeman[1]
j.m.loedeman@gmail.com
Maarten C. Stol[2]
maarten.stol@braincreators.com
Tengda Han[3]
htd@robots.ox.ac.uk
Yuki M. Asano[1*]
y.m.asano@uva.nl

[1] University of Amsterdam
[2] BrainCreators
[3] University of Oxford

1 Details of the datasets

Table 1 gives an overview of the downstream datasets used for the evaluation of our method, including the text prompt templates used to generate classifiers for CLIP.

Dataset	# Train	# Val.	# Test	Classes	Text Prompt
CIFAR100	50K	-	10K	100	"This is a photo of a { }"
CIFAR10	50K	-	10K	10	"This is a photo of a { }"
Flowers102	4K	1.6K	2.5K	102	"This is a photo of a { }"
Food101	50K	20K	30.3K	101	"This is a photo of a { }"
EuroSAT	13.5K	5.4K	8.1K	10	"This is a photo of a { }"
SUN397	15.9K	4K	19.9K	397	"This is a photo of a { }"
UCF101	7.6K	1.9K	3.7K	101	"This is a photo of a { }"
SVHN	73.3K	-	26K	10	"This is a photo of a { }"
OxfordPets	2.9K	736	3.6K	37	"This is a photo of a { }"
DTD	2.8K	1.1K	1.6K	47	"This is a photo of a { }"
Resisc45	18.9K	6.3K	6.3K	45	"This is a photo of a { }"

Table 1: Description of the datasets and the corresponding text prompt used for CLIP. The data is adapted from Bahng *et al.* [1].

*Now at Technical University of Nuremberg

Table/Figure	Epochs	Number of prompts	TL Size	PGN resolution
Table 1	1000	16	256	224×224
Table 2	500	8	64	64×64
Table 3*	1000	16	256	224×224
Table 4	500	16	256	224×224
Figure 2	1000	16	256	224×224
Figure 3	500	8	64	64×64
Figure 4	1000	16	256	224×224
Figure 5	1000	16	256	224×224

Table 2: Experimental settings for each of our tables and figures. *In Table 3 of the main paper, the reported numbers for VP do not come from our own experimentation, hence our settings do not apply.

2 Additional experimental settings

Training Details. In Table 2, we show the training details of the experiments in the main paper. We train the PGN with a learning rate of 0.1 and apply a cosine decay learning schedule ending at zero learning rate with a linear warmup for the first 50 epochs. We use an SGD optimizer with 0.9 momentum. Except when specified, we use a batch size of 128 images on one Nvidia-1080TI GPU. Compared to the 1,000 epochs of concurrent work [1], we train our network for 500 epochs by default in the motivation and ablation sections and for 1,000 in the large-scale comparisons (Table 3 of the main paper).

Architectures. In Table 3, we show the details of ResNet10 architectures.

stage	specification	output sizes $H \times W \times C$
input data	-	$224^2 \times 3$
conv_1	$7 \times 7, 16$ stride $2, 2$	$112^2 \times 16$
pool_1	$3 \times 3, 16$ stride $2, 2$	$56^2 \times 16$
res_2	$\begin{bmatrix} 3 \times 3, 16 \\ 3 \times 3, 16 \end{bmatrix} \times 1$	$56^2 \times 16$
res_3	$\begin{bmatrix} 3 \times 3, 32 \\ 3 \times 3, 32 \end{bmatrix} \times 1$	$28^2 \times 32$
res_4	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 1$	$14^2 \times 64$
res_5	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 1$	$7^2 \times 128$
pool_2	$7 \times 7, 128$ stride $1, 1$	$1^2 \times 128$

Table 3: The structure of ResNet10, which is modified from ResNet18 to be more lightweight. Modifications are marked in red. Note that the final classification layer is omitted.

Feature similarities computation. For Figure 3 in the main paper, we embed the validation set of CIFAR-100 using the three visual encoders of PGN (only), CLIP, and PGN+CLIP. For this we cluster the features into 100 clusters using k-means. After this, the representations can be easily compared with each other using the normalised mutual information score.

3 Qualitative analysis

From Table 1 in the paper, we observed that the CLIP zero-shot and the PGN backbone model's performance on their own are low with 63-64%. However, when combined, we reach performance increases of +15% yielding up to 79% on CIFAR100. In this section, we analyse how the simple mechanism behind PGN is allowing the combined model to achieve superior performances.

What do the individual Token Library items stand for? To answer this question, we pass the validation sets through the trained PGN model and pick individual tokens that we wish to visualize. We then pick the top four input samples that have the highest softmax values for the selected item. The result is shown in Figure 1 for CIFAR100. We find that while some tokens are fairly category specific, such as those for a tree or an apple, some cover much broader categories such as lighting conditions or even geometric structures. Note however that the PGN is not doing the heavy-lifting in terms of classifying the images by itself, as its output is not well-aligned with the ground-truth, as demonstrated in Figure 3 of the main paper. It rather supplies the frozen transformer model with orthogonal information that helps the task. More examples are provided at the end of this document.

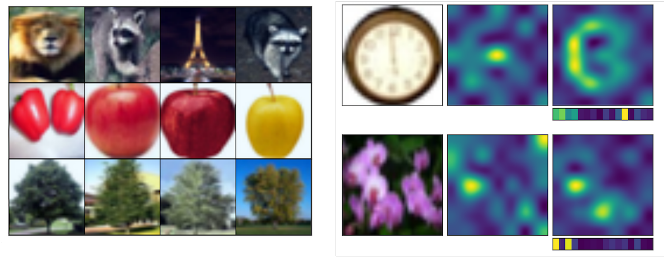

Figure 1: **Token Library items and attention values.** On the left, we show 4 CIFAR100 samples that maximally activate one of three selected items in the token library. Each row in the grid corresponds to one token library item. On the right, we show the individual attention values from the `CLS` token to the supplied prompts of PGN and IIP. We find that while PGN has an overall lower average attention, the input-dependency successfully yields a wider distribution in adapting the original model.

How is the computation changed by PGN prompts? Next, we analyse the effect of the PGN prompts to the internal computations of the frozen vision transformer. In Figure 1, we visualize the `CLS` token's attention map at the final layer of the transformer with or without our PGN. Despite showing the effect of the prompts on the *last* layer's attention map, we still find a strong effect of the PGN's additionally supplied prompts. While the effect is not interpretable for low-resolution datasets such as CIFAR, for Pets and Resisc we observe an increased focus on the foreground. We also show the attention values of the `CLS` to the 16 supplied prompts below the PGN-CLIP attention maps. A strong variance between images is seen, demonstrating that the method learns and leverages the input-dependency of the prompts that are supplied to the frozen model. More examples are provided at the end of this document.

4 Multi-dataset PGN

We retain the same setting as in our large-scale experiments and train with batches that contain samples from the four datasets in Table 4. The model is thus forced to allocate the token library items in a manner that best supports this task, reducing the overall number of additionally adapted parameters by 75%. From Table 4, we find that despite this reduction in parameters, the overall performance only decreases by a small amount of 3.7%, despite the fact that the classification problem is now 193-way and thus much more difficult.

Method	EurSAT	UCF	Pets	RESISC	Avg.	Δ	Σ params
CLIP+TP (I)	40.0	59.9	85.9	42.4	57.1	-7.7%	-
CLIP+TP (J)	4.4	59.7	85.8	47.7	49.4		-
+ PGN (I)	98.0	77.6	91.5	92.1	89.8	-3.7%	5M
+ PGN (J)	96.9	72.7	89.0	85.7	86.1		1M

Table 4: **Training multi-dataset PGN.** Adapting and inferring jointly (J) over multiple datasets compared to individual (I), per-dataset training and evaluation. Giving more text prompts (TP) for joint inference leads to a strong decrease in accuracy, yet, joint training of the PGN retains a strong performance while reducing the number parameters by 75%.

5 Details of the feature similarity analysis

In the NMI analysis in Figure 3 and Sec. 4.1 of the main paper, we measure the pairwise alignment between the outputs of the visual encoders we use and the ground truth. These are: the frozen CLIP model's visual encoder that outputs `CLS` embedding, the trained PGN model that outputs prompts (the $\hat{\mathbf{h}}_V$ in Eqn. 4), and the combined CLIP+PGN model which uses PGN prompts to modify CLIP's visual encoder's outputs (that outputs `CLS` embedding after CLIP). For this, we apply k-means clustering to the set of embeddings generated by each encoder individually, setting k equal to the number of ground-truth classes. For our experiment, we use the full CIFAR100 test split. This yields a set of 3 pseudo labelings of the dataset. After combination with the ground-truth labels, we can make 6 pairwise comparisons and calculate the normalised mutual information, which measures a generalized correlation between labelings. The results are shown in Table 5.

6 Large-scale comparisons

In Table 3 in the main paper, the results for linear finetuning are adopted from the original CLIP paper [5], whereas the results for full finetuning are taken from VP [1].

NMI	GT	PGN	CLIP	PGN+CLIP
GT	100	29.5	58.1	**70.1**
PGN		100	27.5	33.8
CLIP			100	61.2
PGN+CLIP				100

Table 5: Normalized Mutual Information (NMI) score in %.

7 Additional experiments

Comparison between linear and non-linear layer. In Table 6 we evaluate replacing the final linear layer of g_θ with a MLP with 1 hidden layer, which allows for a nonlinear mapping between image features and the logits that give rise to the combination coefficients in Eqn. 3. No significant performance gain is observed.

Type	CIFAR100	SUN397
Linear	77.9	**70.5**
MLP	**78.2**	70.4

Table 6: Feature projection layer type.

Unfreezing the classification layer. So far, we have utilized CLIP's text prompts (TP) to generate the fixed weights of a linear classifier. In Table 7, we compare this approach to a trainable classifier, which takes the TP weights as a starting point.

Cls.	CIFAR100	SUN397
TP	79.3	70.9
+ SGD	79.3	70.3

Table 7: Training a linear layer in addition to PGN.

Method	ImageNet	A	R	V2	Sketch
PGN	66.0	22.8	62.5	56.7	36.5
LP	67.0	10.6	38.1	1.0	36.1

Table 8: Evaluation accuracies on 4 robustness benchmarks. We compare adaptation with a PGN to linear finetuning (LP). We observe that PGN retains much higher scores on these robustness evaluations.

Experiments on robustness. We evaluate the robustness of PGNs by training for 100 epochs onf ImageNet [2] and evaluating on four ImageNet variations (ImageNet-A [3], ImageNet-R [4], Imagenet-V2 [6] and Imagenet-Sketch [7]). For these experiments, we

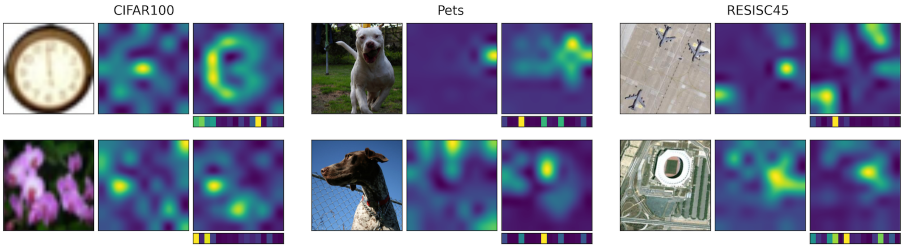

Figure 2: **Modification of `CLS` attention maps.** We show the attention map of the `CLS` token for various inputs (left) with the spatial patches for both the original CLIP model's (middle) and the PGN-modified CLIP's final layer (right). Below the PGN attention map, we show the attention to PGN's additional prompts. We observe a clear modification of the attention map as well as the diverse activation patterns to the supplied tokens.

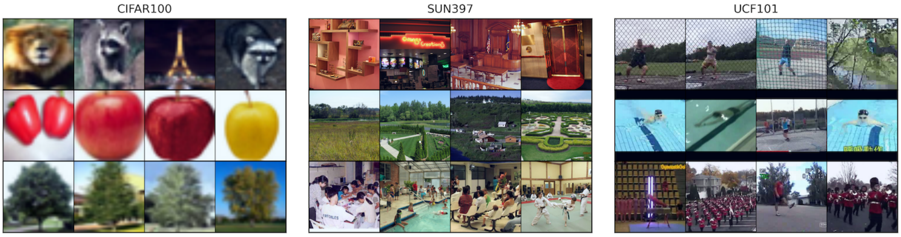

Figure 3: **Token Library example items.** We show 4 samples that maximally activate one of three selected items in the token library for three datasets. Each row in the grid corresponds to one token library item. We find that the items can stand for whole objects such as apples and trees for CIFAR100, and also for lower level features such as light warmth or net structures as in UCF101.

use identical PGN settings as in Table 3 in the paper. The results are shown in Table 8 and compared to the case of linear finetuning on the same, frozen CLIP backbone (ViT-B/32). We see that the PGN outperforms linear finetuning on all robustness benchmark, despite being comparable in terms of its performance on the upstream dataset.

References

[1] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv:2203.17274*, 2022.

[2] 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, 2009. doi: 10.1109/CVPR.2009.5206848.

[3] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *arXiv preprint arXiv:1907.07174*, 2019.

[4] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan

Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *ICML*, 2021.

[6] 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.

[7] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019.