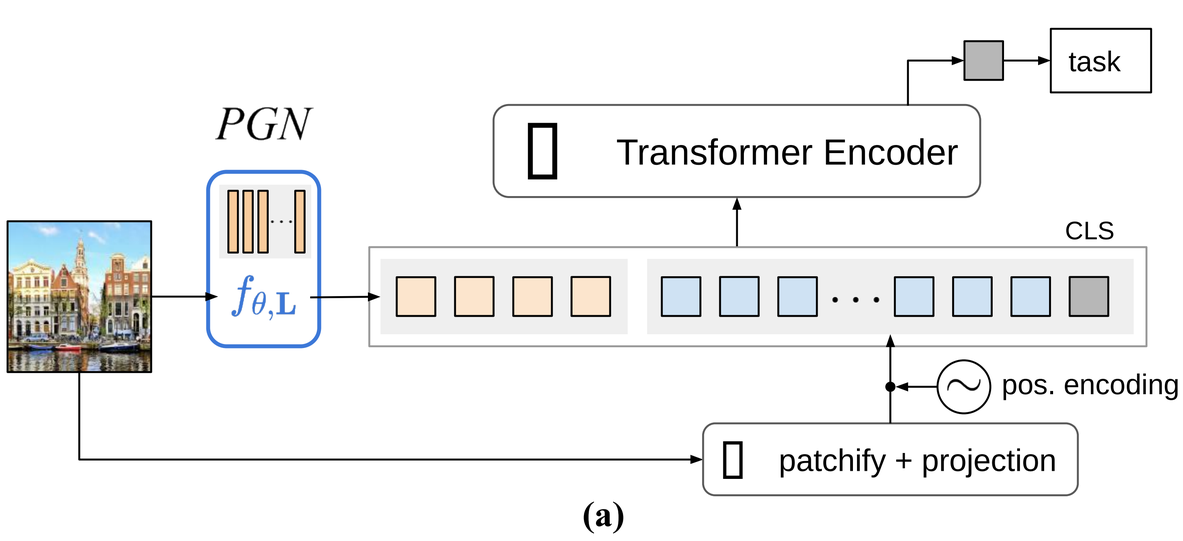
Prompt Generation Networks for Input-Space Adaptation of Frozen Vision Transformers

Jochem Loedeman Maarten C. Stol Tengda Han Yuki M. Asano

Motivation

- Input prompt learning has limited capacity because **learned prompts do not depend on the input**
- Most prompt-based adaptation techniques change the forward pass, making them **infeasible to operate when scaling up to many downstream tasks**

Method

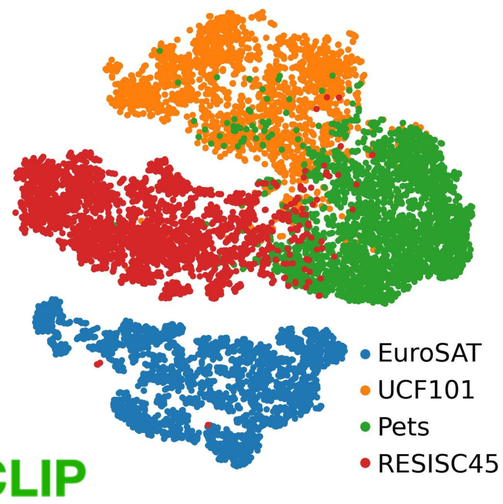


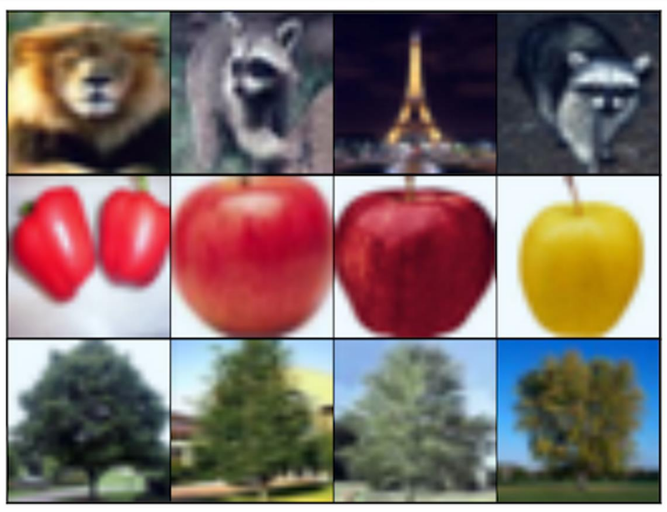
- A Prompt Generation Network adds a set of input-dependent prompts, by using a lightweight encoder that outputs coefficients with respect to a library of tokens
- Using Prompt Inversion, these prompts can be mapped to the visual input space to decouple adaptation from model computation at inference time

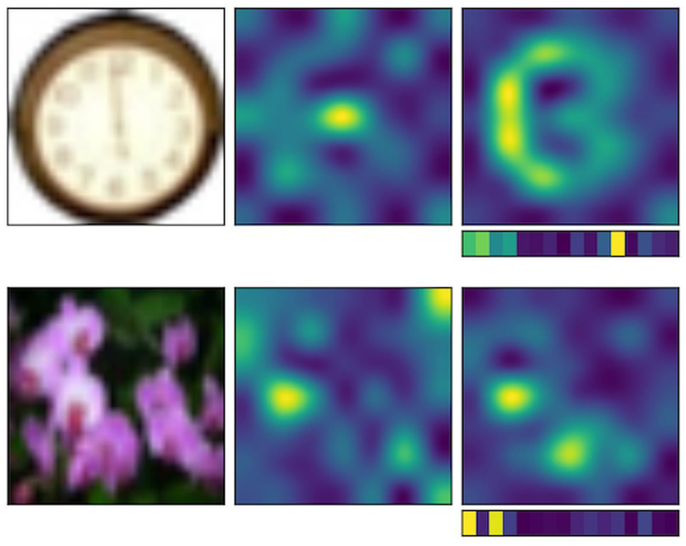
Results

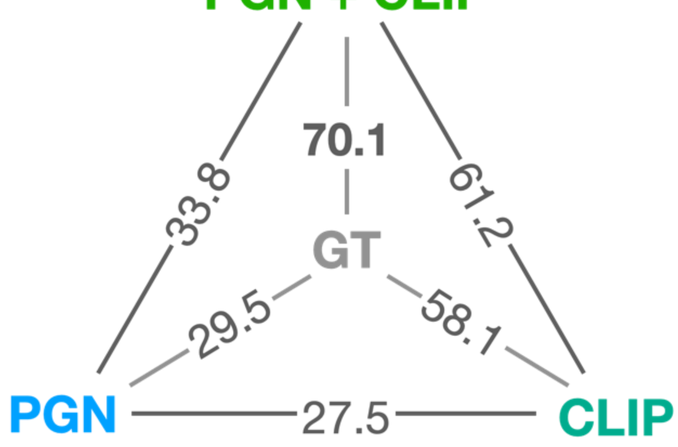
Method	C100	C10	Flwrs	Food	EuSAT	SUN	UCF	SVHN	Pets	DTD	RESISC	CLEVR	Avg.	Σ params
CLIP+TP	63.1	89.0	61.9	79.8	40.0	60.0	59.9	5.1	85.9	43.0	42.4	20.2	54.2	-
+ VP	75.3	94.2	70.3	78.9	96.4	60.6	66.1	88.4	85.0	57.1	84.5	81.4	78.2	0.94M
+ PGN (ours)	79.3	**96.1**	94.0	82.5	**98.0**	70.9	77.6	94.2	**91.5**	71.5	92.1	**95.1**	**86.9**	12.4M
Linear ft.	80.0	95.0	96.9	**84.6**	95.3	**75.0**	**83.3**	65.4	89.2	**74.6**	92.3	66.0	83.1	0.75M
full-ft.	**82.1**	95.8	**97.4**	80.5	97.9	64.0	80.9	**95.7**	88.5	72.3	**93.3**	94.4	**86.9**	1032M

[1]









Conclusions

- PGN matches full fine-tuning for vision transformers, while being significantly more parameter efficient
- Compared to other modern techniques such as LoRA, PGN is efficient during inference, because it does not change the forward pass of the model

References

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

[2] Hu, Edward J., et al. Lora: Low-rank adaptation of large language models.

Codebase