BlackVIP: Black-Box Visual Prompting for Robust Transfer Learning

Paper Review by Ravialdy

Main source : [Oh'23;CVPR] Oh *et al.*, "BlackVIP: Black-Box Visual Prompting for Robust Transfer Learning." IEEE Computer Vision and Pattern Recognition(CVPR)

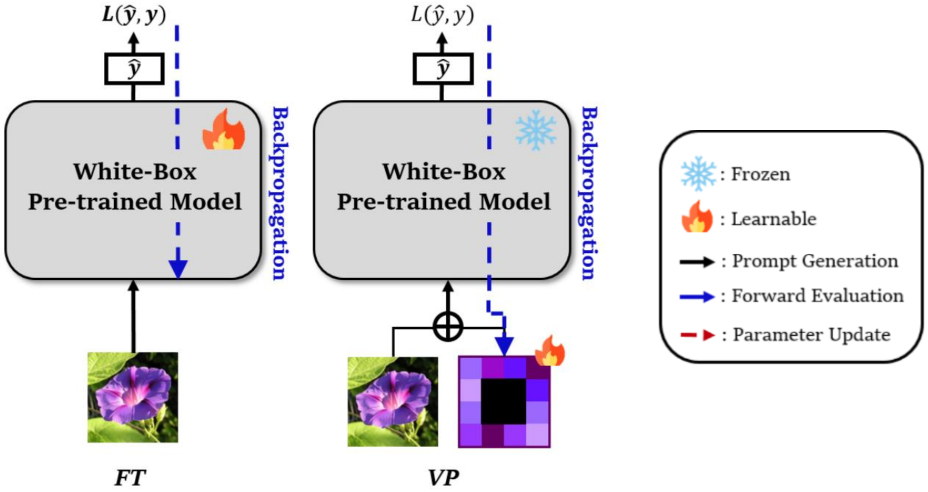
Motivation :

Figure 1. Illustration of using black-box Pre-trained Model (PTM) in image classification task.

- Currently, many high-performing AI models are in form of APIs (e.g., DeepLobe, Kony, etc).
- However, existing approaches always consider white-box setting where we can do backpropagation which means we have access to model parameters for transfer-learning.
- Thus, the condition where we don't have access to model parameters (black-box) is still an unexplored problem.

Main source : [Oh'23;CVPR] Oh *et al.*, "BlackVIP: Black-Box Visual Prompting for Robust Transfer Learning." IEEE Computer Vision and Pattern Recognition(CVPR)

Motivation :



Notation meanings :

$\hat{y}$ is the predicted label.

y is the ground-truth label.

L is an objective (loss) function.

*Note : FT = Fine-tuning.

Figure 2. Different transfer-learning methods in image classification task.

- Previous work (Visual Prompting or VP [Bahng'22]) only update parameters in the input space where Pre-trained Model (PTM) is full frozen.
- However, that approach still does backpropagation (white-box setting) -> requires large memory capacity.

Main source : [Oh'23;CVPR] Oh *et al.*, "BlackVIP: Black-Box Visual Prompting for Robust Transfer Learning." IEEE Computer Vision and Pattern Recognition(CVPR)

#1 Key Idea : Coordinator

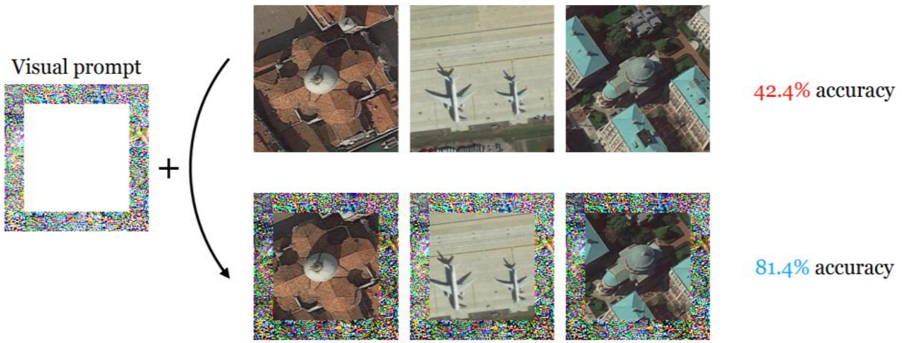


Figure 3. Illustration of visual prompt (learnable padding).

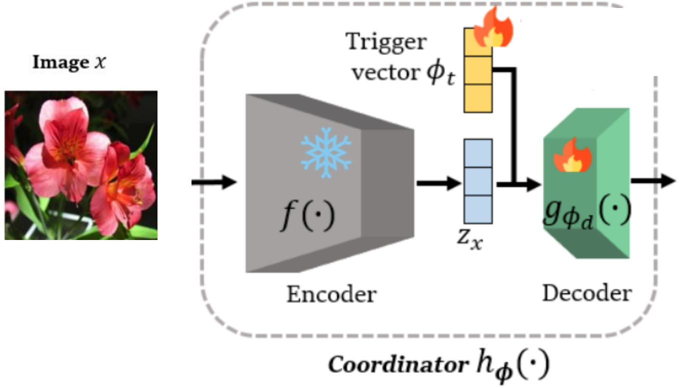


Figure 4. Diagram of coordinator part of BlackVIP model.

*Note : z_x = Image features.

- Previous work use the same learnable visual prompt (padding) for all images in downstream dataset -> limit its flexibility to change visual semantics when necessary.
- Thus, BlackVIP design “Coordinator” to automatically design visual prompt conditioned on each image.

Main source : [Oh’23;CVPR] Oh *et al.*, “BlackVIP: Black-Box Visual Prompting for Robust Transfer Learning.” IEEE Computer Vision and Pattern Recognition(CVPR)

#2 Key Idea : SPSA with Gradient Correction (SPSA-GC)

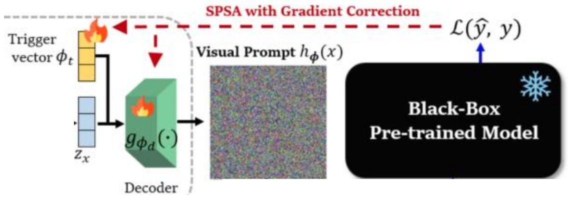


Figure 5. Diagram of SPSA-GC part of BlackVIP model.

Common SPSA parameter update rule :

$$\hat{g}_i(\phi_i) = \frac{L(\phi_i + c_i\Delta_i) - L(\phi_i - c_i\Delta_i)}{2c_i}\Delta_i^{-1} \quad (1)$$

$$\phi_{i+1} = \phi_i - a_i\hat{g}_i(\phi_i) \quad (2)$$

SPSA-GC parameter update rule :

$$\phi_{i+1} = \phi_i + m_{i+1} \quad (3)$$
$$m_{i+1} = \beta m_i - a_i\hat{g}_i(\phi_i + \beta m_i)$$

Notation meanings :

$a_i > 0$ is the positive decaying sequences.

$c_i \in [0, 1]$ is a constant between 0 and 1.

$\hat{g}$ is the gradient approximation.

L is an objective function.

$\phi_i \in \mathbb{R}^d$ is d-dimensional learnable parameters.

$\Delta_i \in \mathbb{R}^d$ is a i^{th}-step random perturbation vector.

$\beta \in [0, 1]$ is smoothing parameter.

m_i is a momentum at step i.

*Note : SPSA = Simultaneous Perturbation Stochastic Approximation.

- SPSA is well-known black-box optimization method due to its theoretically guarantees convergence [Spall'92].
- However, that method still requires many iterations because of data noise within visual prompt.
- Thus, inspired by Nesterov's Accelerated Gradient (NAG), the parameter update will become eq. (3) in order to speed up the process by using "look-ahead" gradient.

Main source : [Oh'23;CVPR] Oh *et al.*, "BlackVIP: Black-Box Visual Prompting for Robust Transfer Learning." IEEE Computer Vision and Pattern Recognition(CVPR)

BlackVIP Overall Architecture :

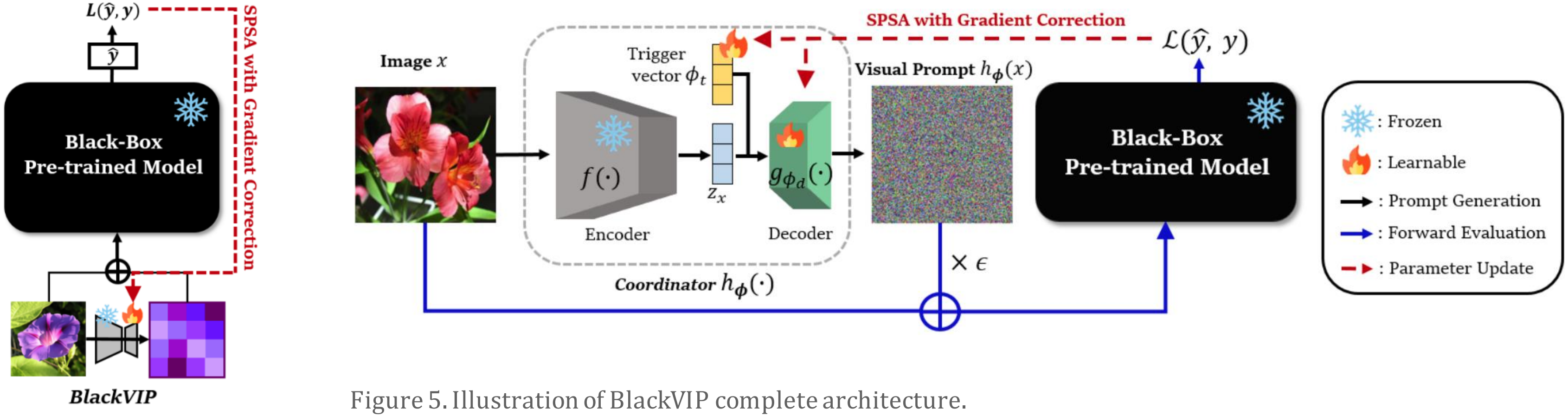


Figure 5. Illustration of BlackVIP complete architecture.

- Consists of two key parts, Coordinator and SPSA with Gradient Correction (SPSA-GC).
- Coordinator has a purpose to automatically create visual prompt given an image (input-dependent).
- SPSA-GC is used for parameters update without the need to access PTM's model parameters.

Main source : [Oh'23;CVPR] Oh *et al.*, "BlackVIP: Black-Box Visual Prompting for Robust Transfer Learning." IEEE Computer Vision and Pattern Recognition(CVPR)

Performance :

Method	Caltech	Pets	Cars	Flowers	Food	Aircraft	SUN	DTD	SVHN	EuroSAT	RESISC	CLEVR	UCF	IN	*Avg.*	*Win*
VP (white-box)	94.2	90.2	66.9	86.9	81.8	31.8	67.1	61.9	60.4	90.8	81.4	40.8	74.2	67.4	71.1	13
ZS	92.9	89.1	65.2	**71.3**	86.1	24.8	62.6	44.7	18.1	47.9	57.8	14.5	66.8	66.7	57.6	-
BAR	**93.8**	88.6	63.0	71.2	84.5	24.5	62.4	**47.0**	34.9	**77.2**	**65.3**	18.7	64.2	64.6	61.4	6
VP w/ SPSA-GC	89.4	87.1	56.6	67.0	80.4	23.8	61.2	44.5	29.3	70.9	61.3	25.8	64.6	62.3	58.8	4
BlackVIP	93.7	**89.7**	**65.6**	70.6	**86.6**	**25.0**	**64.7**	45.2	**44.3**	73.1	64.5	**36.8**	**69.1**	**67.1**	**64.0**	**13**

Figure 6. Performance of BlackVIP across downstream tasks (only compared with black-box model).

Method	Peak Memory (MB)		Params	
	ViT-B	ViT-L	ViT-B	ViT-L
FT (white-box)	21,655	76,635	86M	304M
LP (white-box)	**1,587**	3,294	513K	769K
VP (white-box)	11,937	44,560	69K	69K
BAR	1,649	3,352	37K	37K
VP w/ SPSA-GC	1,665	3,369	69K	69K
BlackVIP	2,428	**3,260**	**9K**	**9K**

Figure 7. Peak memory allocation and number of learnable parameters on ImageNet dataset.

Main source : [Oh’23;CVPR] Oh *et al.*, “BlackVIP: Black-Box Visual Prompting for Robust Transfer Learning.” IEEE Computer Vision and Pattern Recognition(CVPR)

Thank You