JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

THU-PM-346

BlackVIP: Black-Box Visual Prompting for Robust Transfer Learning

Towards realistic adaptation of pre-trained vision models

Changdae Oh[1], Hyeji Hwang[1], Hee-young Lee[2], YongTaek Lim[1], Geunyoung Jung[1], Jiyoung Jung[1], Hosik Choi[1], and Kyungwoo Song[3]

[1] University of Seoul

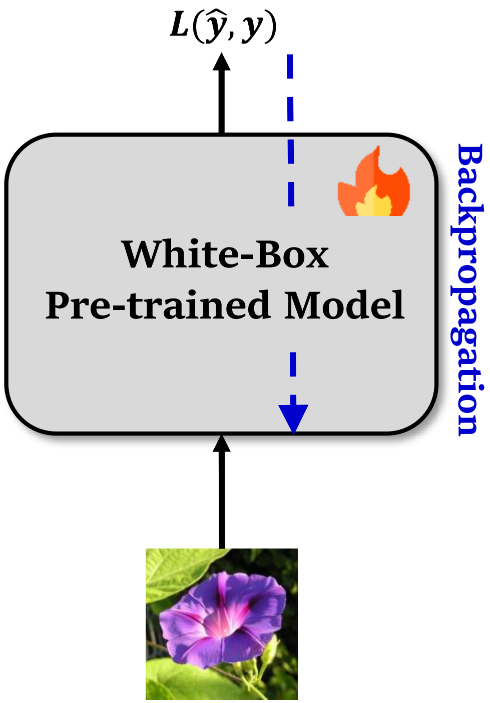
[2] Sungkyunkwan University

[3] Yonsei University

2023.06.01

Problem define

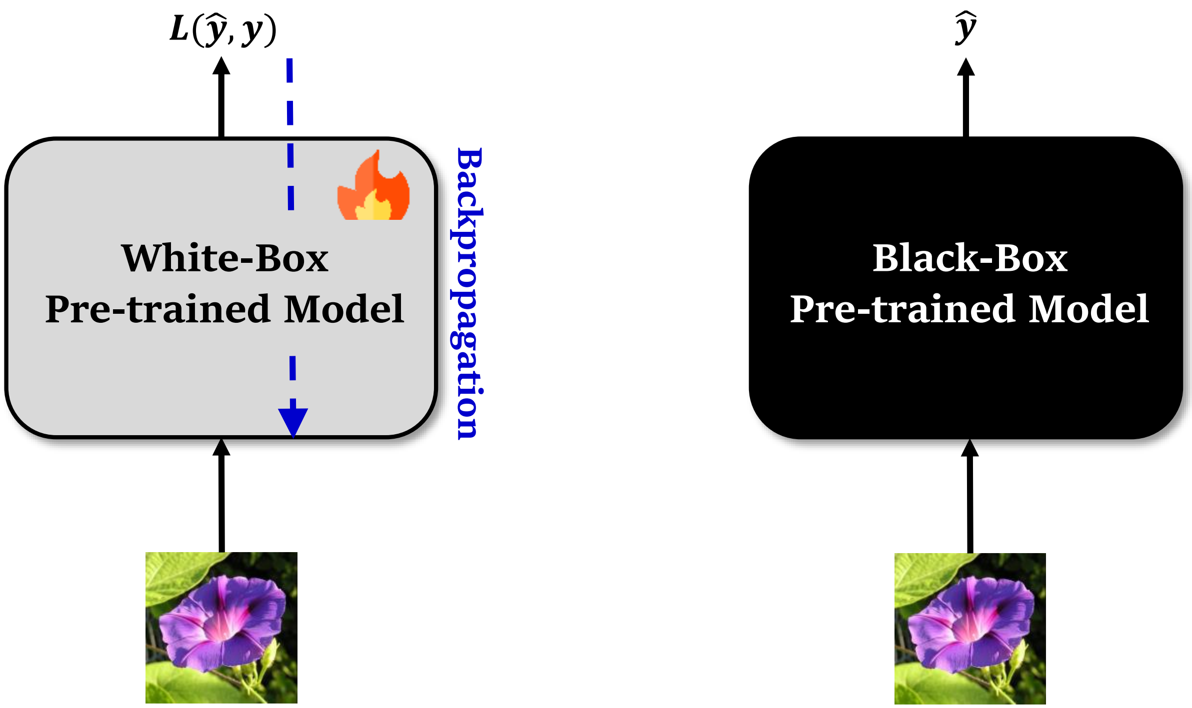
- Adapting a large-scale pre-trained model (PTM) to diverse downstream tasks
- Existing works assume the parameter accessability and large-memory capacity



Brief Overview

Problem define

- Adapting a large-scale pre-trained model (PTM) to diverse downstream tasks
- Existing works assume the parameter accessability and large-memory capacity

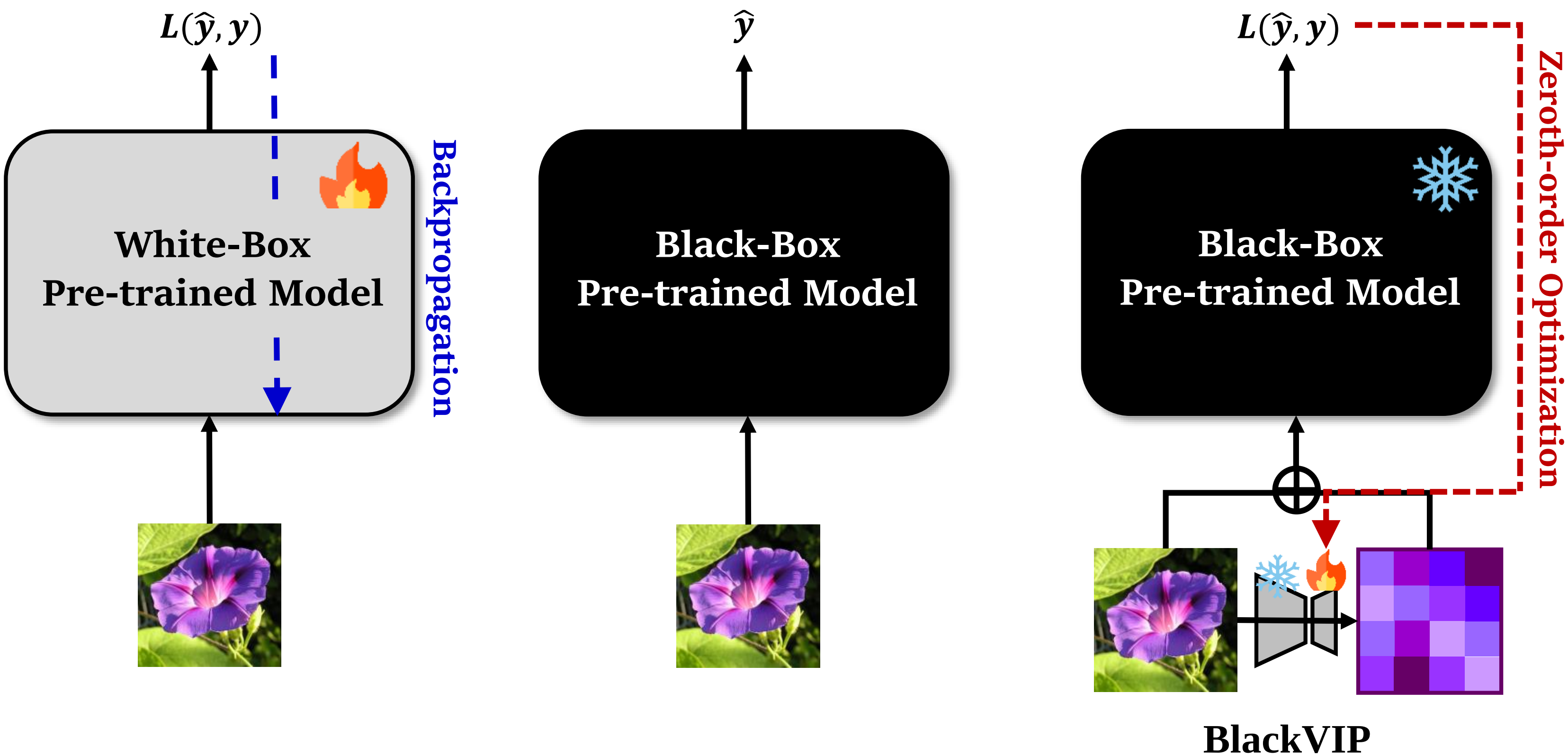


However…
PTMs are provided as inference-only black-box API service in many real-world applications!!

Brief Overview

Problem define

- Adapting a large-scale pre-trained model (PTM) to diverse downstream tasks
- Existing works assume the parameter accessability and large-memory capacity



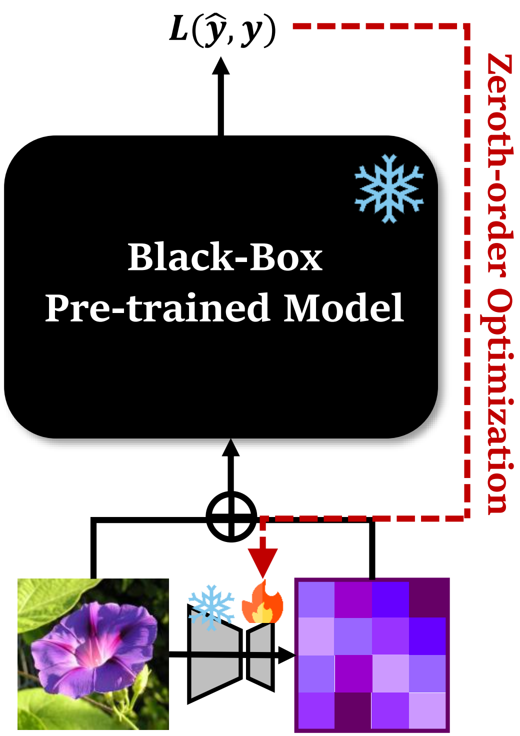
Our approach: *black-box visual prompting (BlackVIP)*

- Let's tune the input!! rather than the model components.
- Learning is progressed by just forward evaluations without backpropagation

Brief Overview

BlackVIP in a nutshell

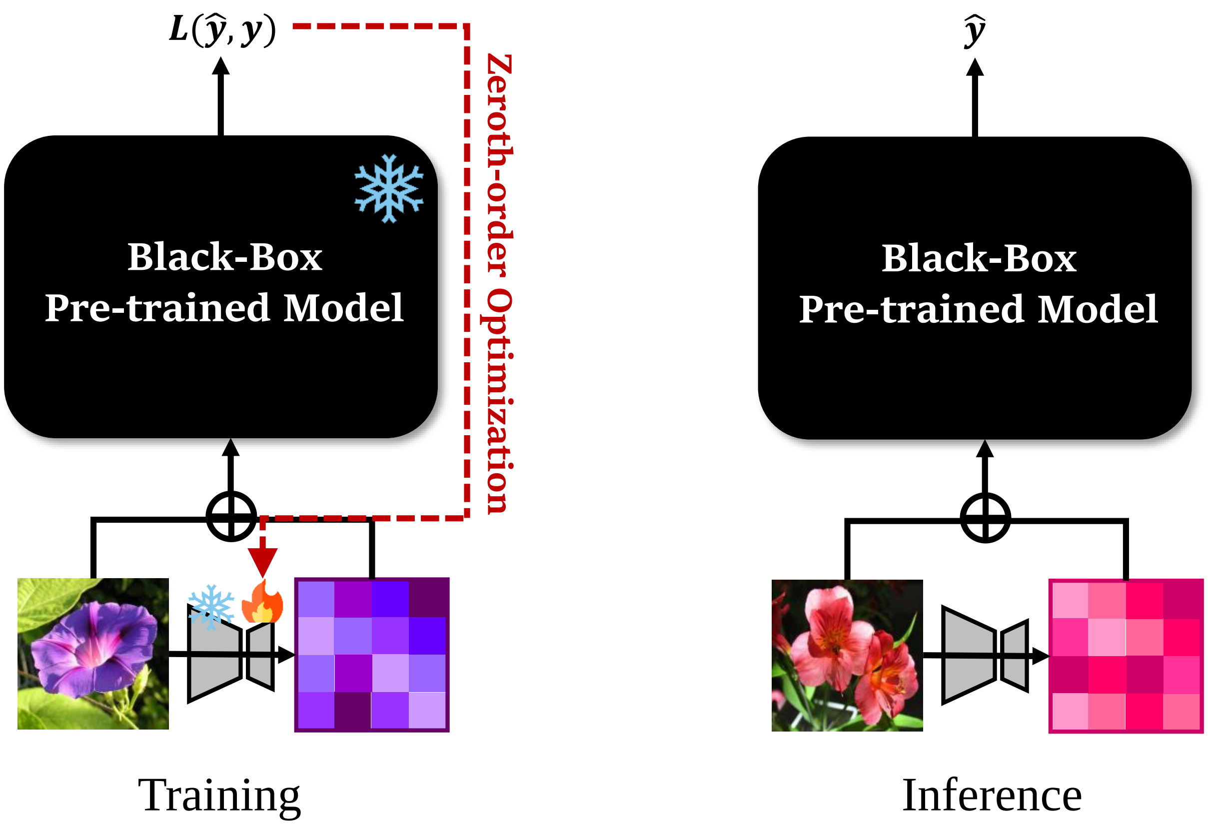
- Learns **visual prompts** (perturbations) to steer the model to produce desired output
- Prompts are obtained by training a **prompt generator** via **zeroth-order optimization**



Training

BlackVIP in a nutshell

- Learns **visual prompts** (perturbations) to steer the model to produce desirable output
- Prompts are obtained by training a **prompt generator** via **zeroth-order optimization**
- After training, BlackVIP automatically generates an **input-dependent visual prompt** for a query image to be better recognized by the black-box API model



Brief Overview

Validation

- On two synthetic datasets and 14 transfer learning benchmarks that cover diverse tasks
- From the perspective of few-shot adaptability, robustness, and practical usefulness

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**

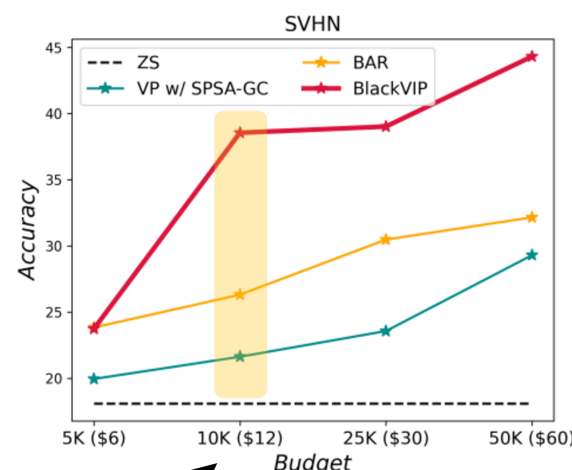
Train set

Test set

Method	Biased MNIST 16-Shot $\rho = 0.8$	16-Shot $\rho = 0.9$	32-Shot $\rho = 0.8$	32-Shot $\rho = 0.9$
VP (white-box)	57.92	43.55	69.65	42.91
ZS	37.56	37.25	37.56	37.25
BAR	53.25	53.07	53.93	53.30
VP w/ SPSA-GC	60.34	53.86	59.58	51.88
BlackVIP	**66.21**	**62.47**	**65.19**	**64.47**

Robustness

When spurious correlation exists, BlackVIP holds strong robustness under the distribution shift



Query efficiency

$12 (USD) with 10K query makes about ×2 performance compared to no-adaptation baseline

Motivation

Pre-trained General-purpose Models

Motivation

- The era of surging large-scale pre-trained models (PTMs)
  - Build a strong generalist model [1], then adapt it to a wide range of downstream tasks!
- Problem
  - *How can we (pre-) train the general-purpose models?*
  - *How can we **adapt** the PTMs to our specific problems?*

Our focus

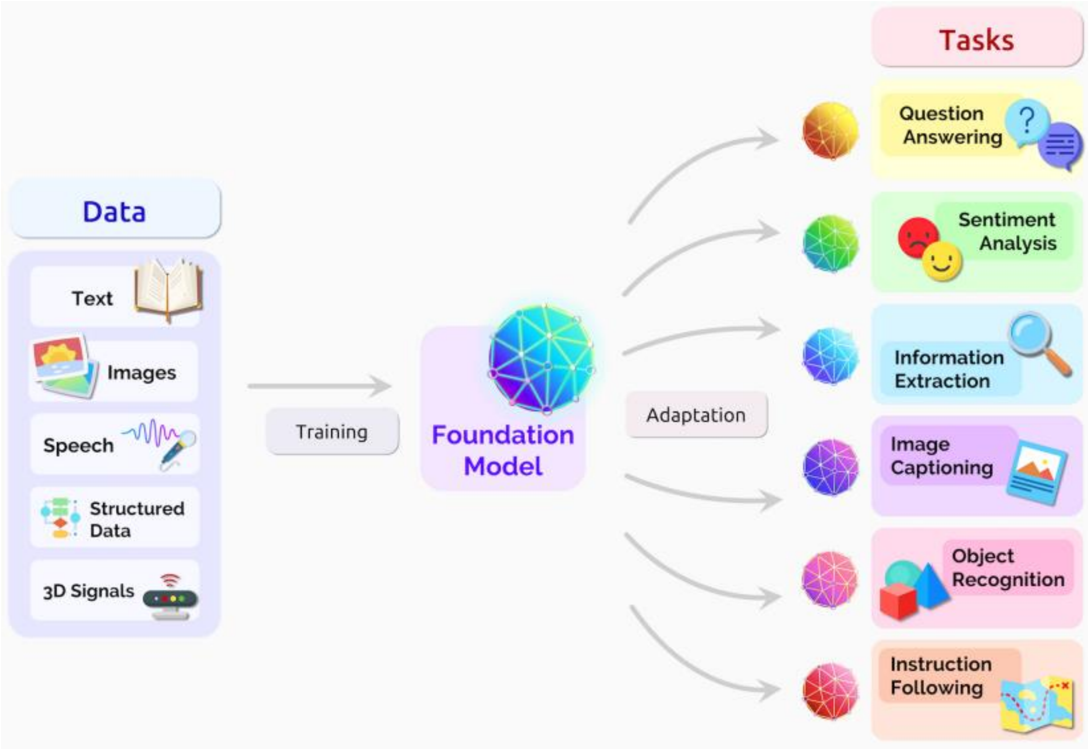


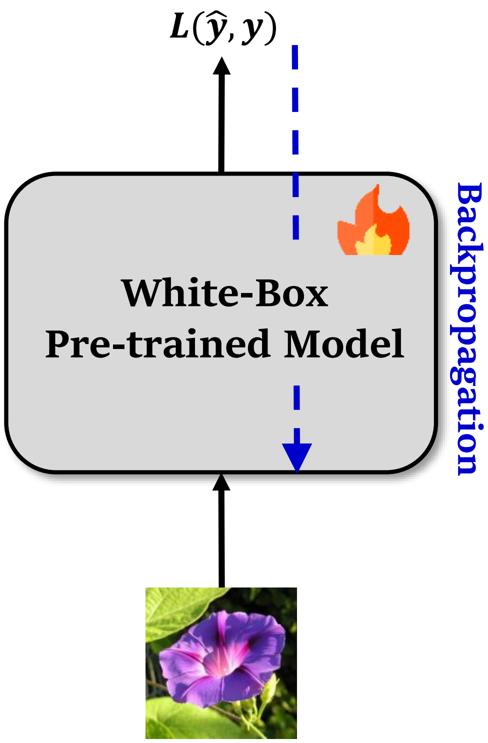
Figure from [1]

[1] On the Opportunities and Risks of Foundation Models, Bommasani et al. 2021

Adapting PTMs

Motivation

- Existing approaches
  - FT: Update entire model parameters



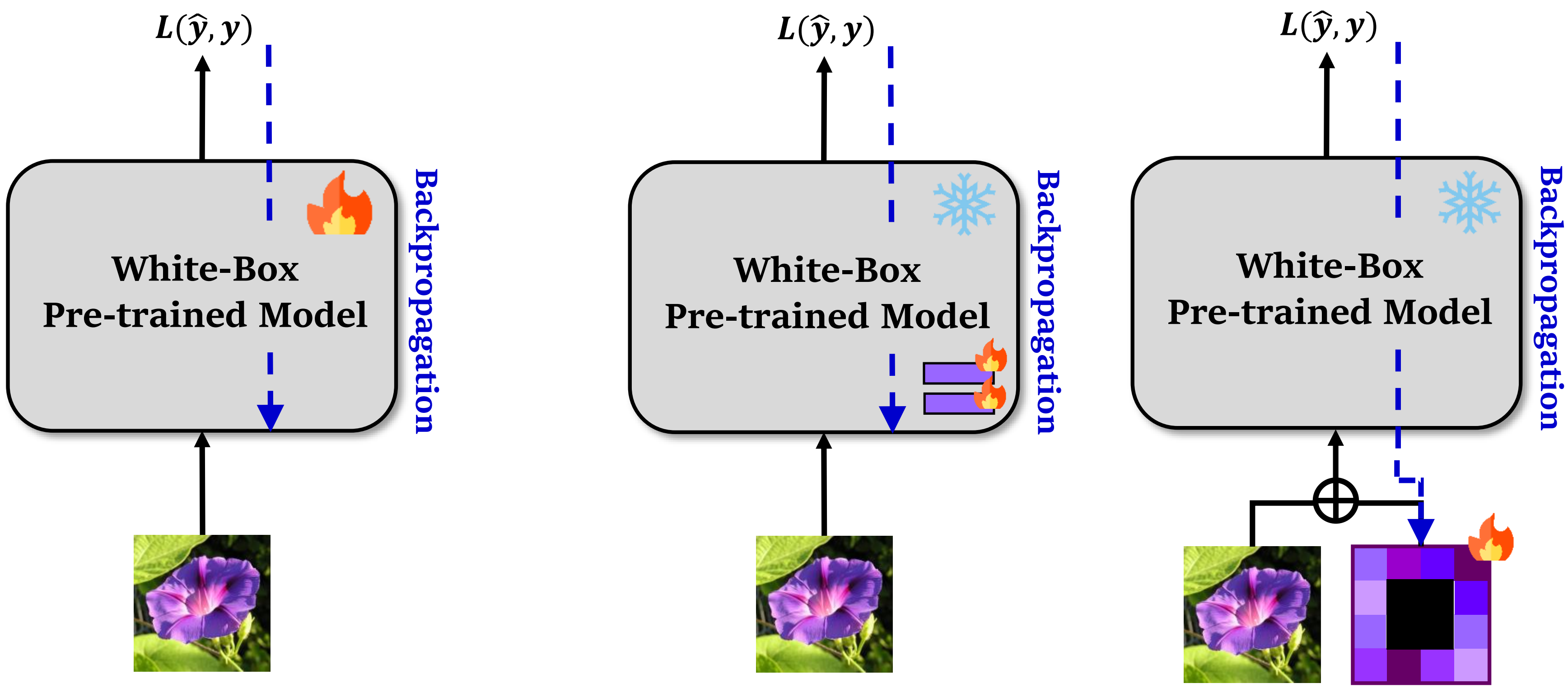
Full Fine-tuning (FT)

Adapting PTMs

Motivation

- Existing approaches
  - FT: Update entire model parameters
  - VPT: Update only a small amount of parameters inside the model
  - VP: Update a single image perturbation (visual prompt)

Parameter Efficient Transfer Learning (PETL)



Full Fine-tuning (FT)

Visual Prompt Tuning (VPT)
(Jia et al. 2022 [2])

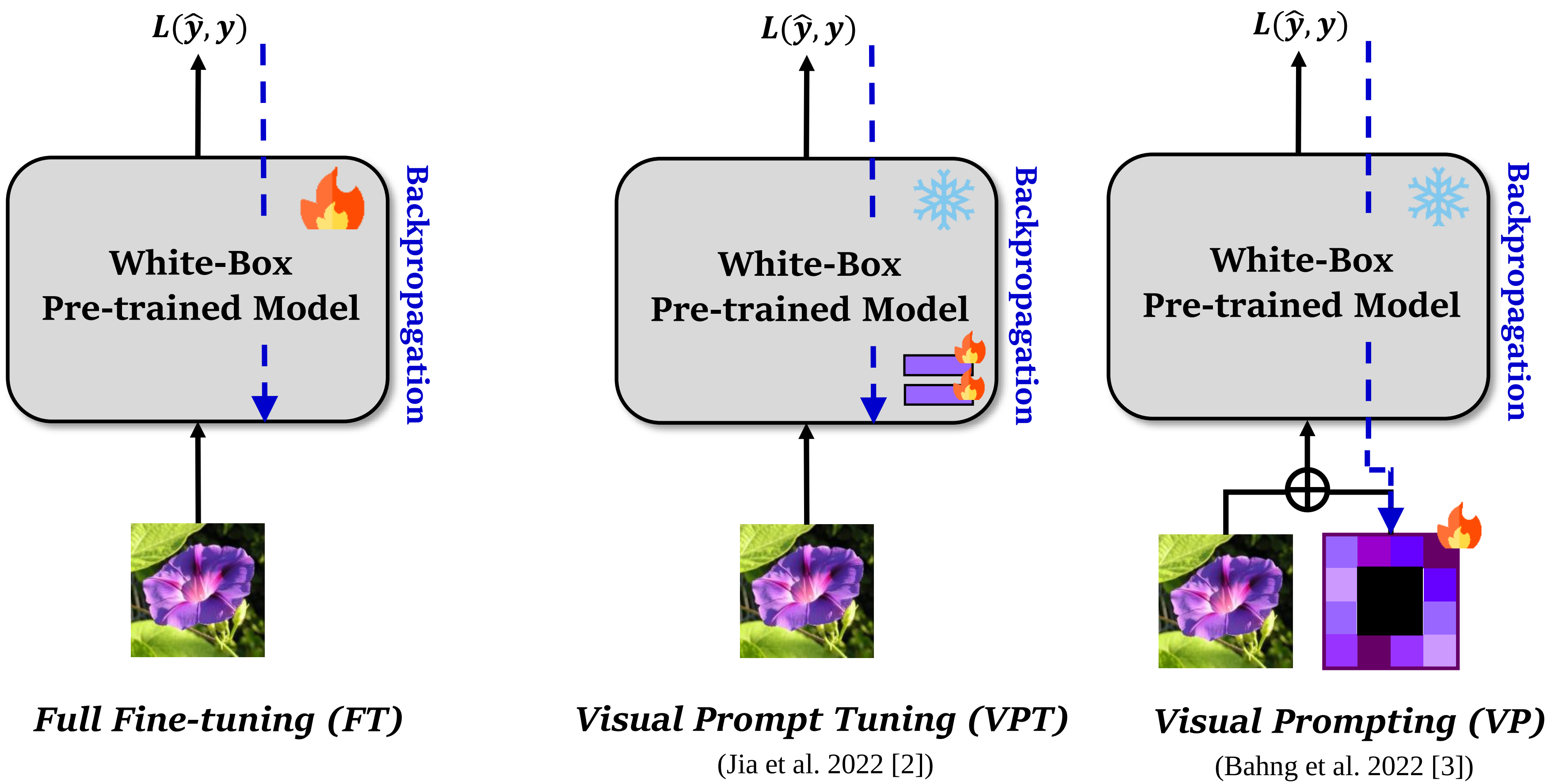
Visual Prompting (VP)
(Bahng et al. 2022 [3])

[2] Visual Prompt Tuning, Jia et al. 2022
[3] Exploring Visual Prompts for Adapting Large-Scale Models, Bahng et al. 2022

Adapting PTMs

Motivation

- Existing methods rely on two optimistic assumptions:
  1. The **model parameters are fully accessible**.
  2. A **sufficiently large memory capacity** is equipped to fine-tune the model.

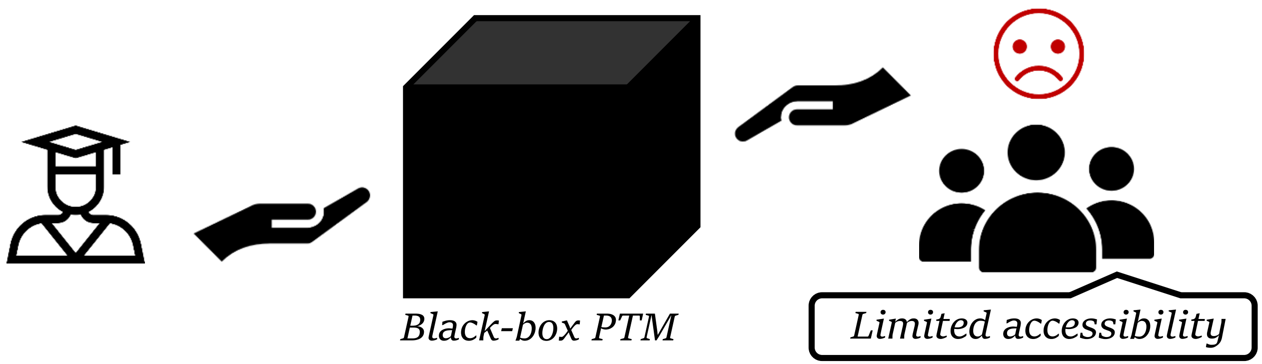


[2] *Visual Prompt Tuning, Jia et al. 2022*
[3] *Exploring Visual Prompts for Adapting Large-Scale Models, Bahng et al. 2022*

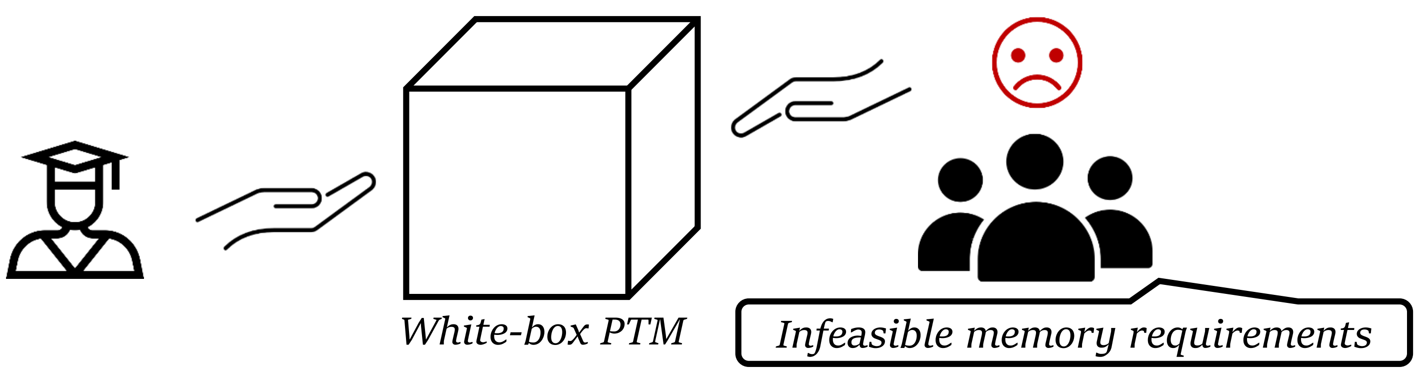
Challenges of adapting PTMs in real-world

Motivation

- However, due to commercial issues,
  PTMs in many real-world applications are provided in the form of **black-box API**



- Even though the full model is leased with parameters,
  end-users (usually low-resourced) are hard to meet the **large memory requirements**



Figures are inspired by [4] Offsite-tuning: Transfer learning without full model, Xiao et al. 2023

Challenges of adapting PTMs in real-world

- However, due to commercial issues,
  PTMs in many real-world applications are provided in the form of **black-box API**
- Even though the full model is leased with parameters,
  end-users (usually low-resourced) are hard to meet the **large memory requirements**

Therefore, the desirable adaptation method **should be**:

- Free from dependence on parameter accessibility
- Efficient (and/or cheap) enough to be affordable for end-users

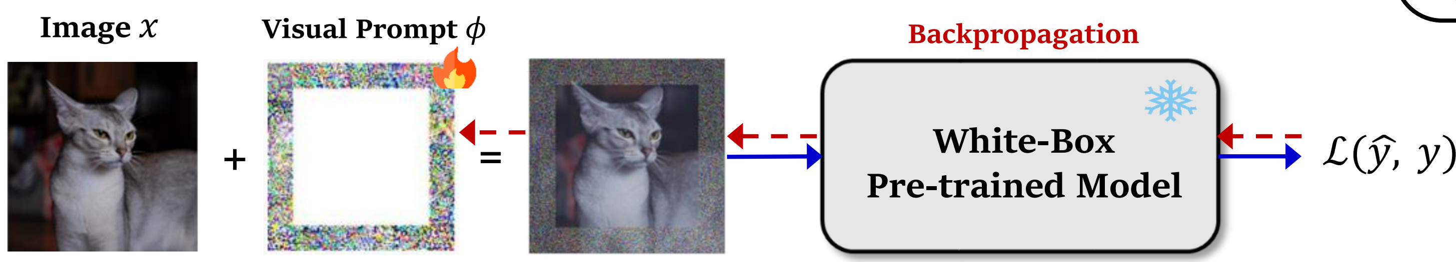
Our Approach

Black-box Visual Prompting

BlackVIP: Black-box Visual Prompting

- : Frozen
- : Learnable
- : Prompt Generation
- : Forward Evaluation
- : Parameter Update

- Previous approach: (white-box) visual prompting, VP [3]



1. Require parameter accessibility and large memory capacity
2. Manually designed input-independent prompt

Learning Objective

$$\arg\min_{\phi} -\log P_{\theta;\phi}(y|x+\phi)$$

$P_\theta(\cdot \mid \cdot)$: PTM

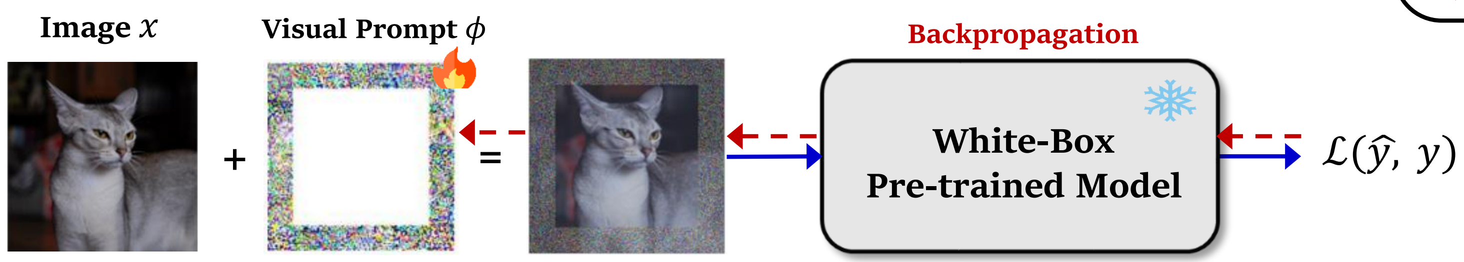
[3] Exploring Visual Prompts for Adapting Large-Scale Models, Bahng et al. 2022

BlackVIP: Black-box Visual Prompting

Methodology

: Frozen
: Learnable
: Prompt Generation
: Forward Evaluation
: Parameter Update

- Previous approach: (white-box) visual prompting, VP [3]



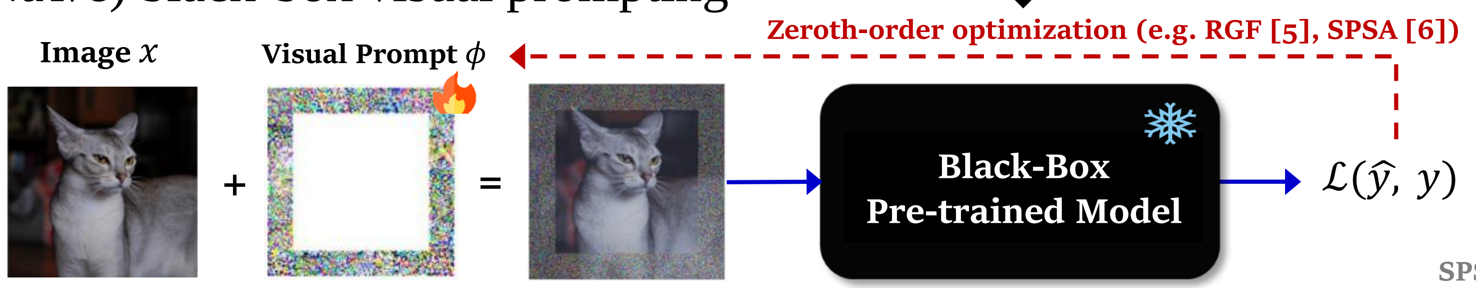
1. Require parameter accessibility and large memory capacity
2. Manually designed input-independent prompt

Learning Objective

$$\arg\min_{\phi} -\log P_{\theta;\phi}(y|x+\phi)$$

$P_\theta(\cdot\,|\,\cdot)$: PTM

- (Naïve) black-box visual prompting



1. Free from parameter accessibility and memory capacity
2. Manually designed input-independent prompt
3. Slow convergence of standard zeroth-order optimizers (RGF, SPSA)

SPSA algorithm

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

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

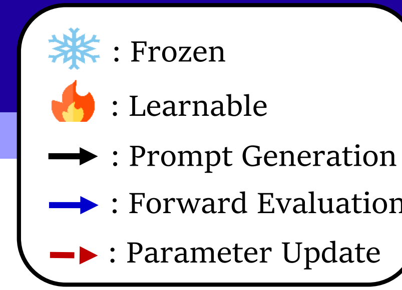
Δ_i: random perturbation

a_i, c_i: positive decaying sequences

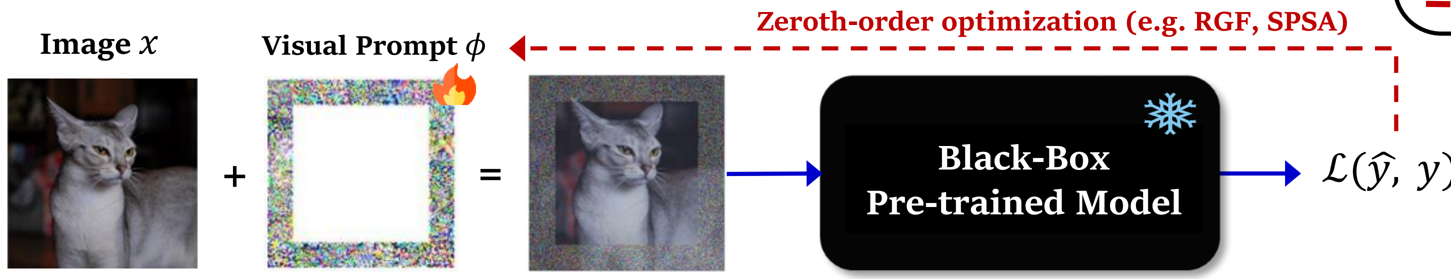
[3] Exploring Visual Prompts for Adapting Large-Scale Models, Bahng et al. 2022
[5] Random Gradient-Free Minimization of Convex Functions, Nesterov et al. 2015
[6] Multivariate stochastic approximation using a simultaneous perturbation gradient approximation, Spall et al. 1992

BlackVIP: Black-box Visual Prompting

Methodology



- (Naïve) black-box visual prompting

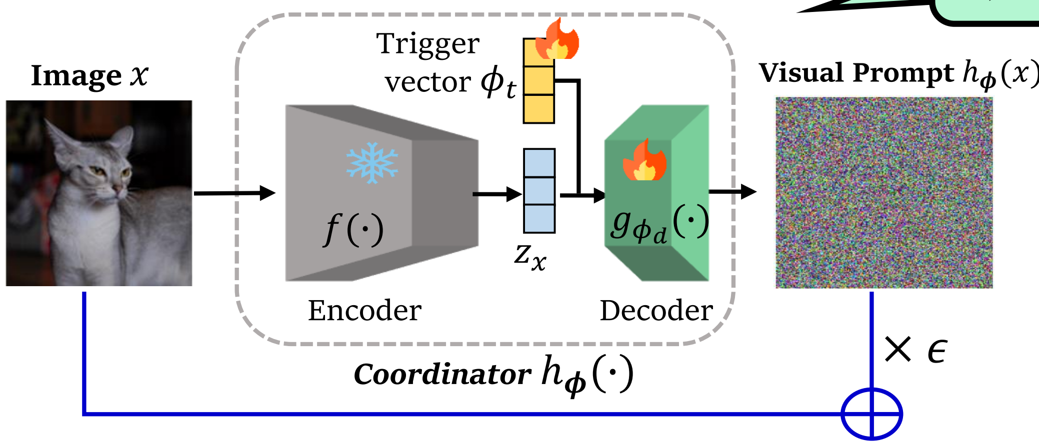


Manually designed input-independent prompt

- **Our improvement 1.**

Coordinator, prompt reparameterization

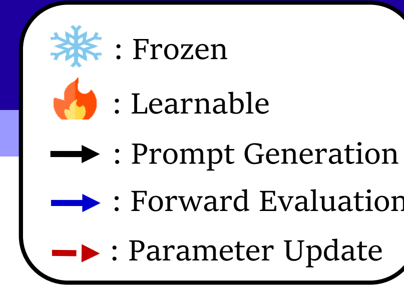
1) Reduce the number of parameters (69K→9K)
2) Input-dependent automatic prompt design



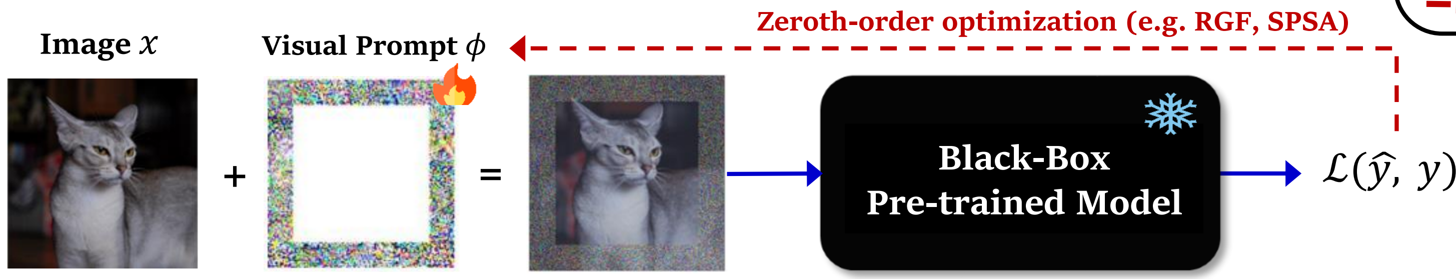
$$\arg\min_{\phi} -\log P_{\theta;\phi}(y|\tilde{x})$$
$$\tilde{x} = \text{clip}(x + \epsilon h_\phi(x))$$
$$h_\phi(x) = g_{\phi_d}(z_x, \phi_t)$$

BlackVIP: Black-box Visual Prompting

Methodology



- (Naïve) black-box visual prompting

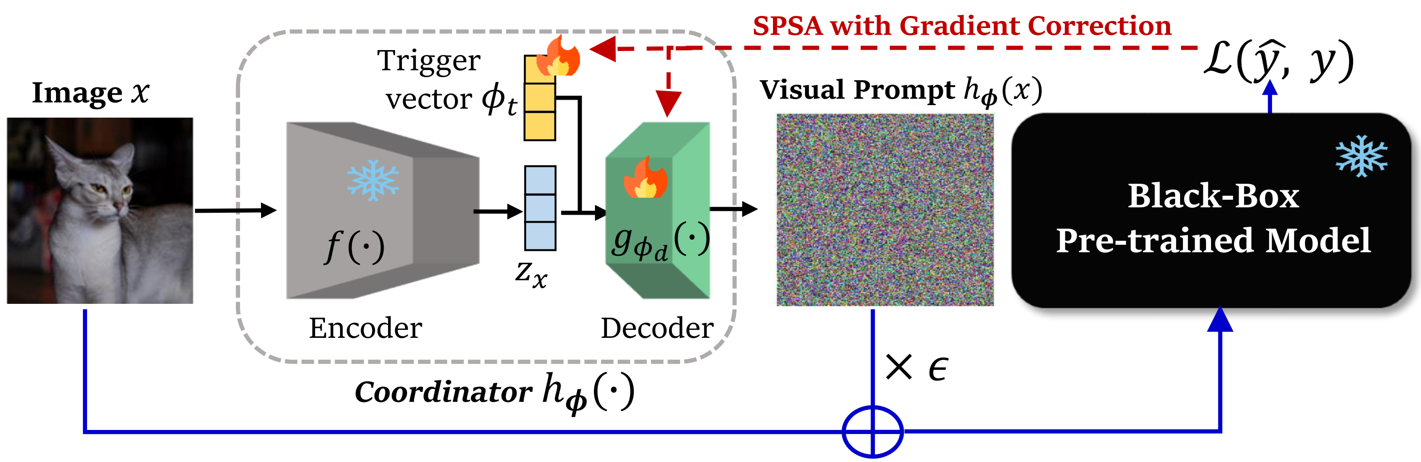


Slow convergence of standard zeroth-order optimizers (RGF, SPSA)

- **Our improvement 2.**

SPSA-GC, enhanced zeroth-order optimizer

Faster convergence by gradient correction effect (Sutskever et al. 2013 [7])



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

$$\downarrow$$

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

[7] *On the importance of initialization and momentum in deep learning, Sutskever et al. 2013*

Empirical Results

- Robustness on distribution shift
- Robustness on object-location shift
- Few-shot adaptation
- Practical Usefulness

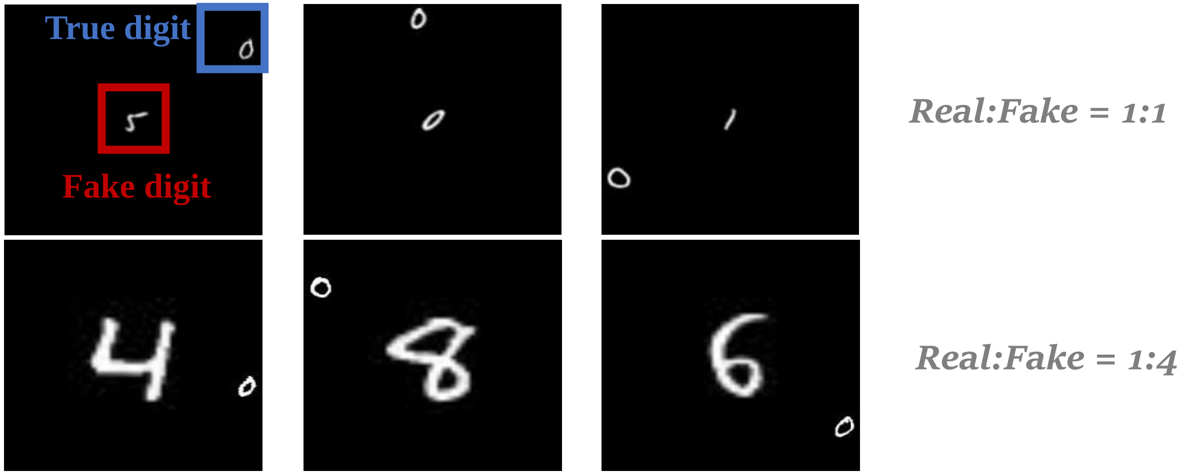
All experiments are done on a few-shot evaluation setting with CLIP pre-trained ViT/B-16 backbone model.

Robustness Analysis

Empirical Results

- Biased MNIST [8]
  - There is a **spurious correlation** between the background colors and target digits.

- Loc-MNIST
  - Unlike most benchmarks where the objects are centered, Loc-MNIST **arbitrarily distributes the target object to the edge side** of the image.
  - We additionally put the random fake digit in the center to increase task difficulty.



[8] Learning De-biased Representations with Biased Representations, Bahng et al. 2020

- Existing methods struggle to deal with spurious correlation.
- Our input-dependent image-shaped visual prompt can be beneficial in domain generalization setting!

Method	Biased MNIST				Loc-MNIST			
	16-Shot		32-Shot		16-Shot		32-Shot	
	$\rho = 0.8$	$\rho = 0.9$	$\rho = 0.8$	$\rho = 0.9$	1:1	1:4	1:1	1:4
VP (white-box)	57.92	43.55	69.65	42.91	86.79	86.54	90 .18	92.09
ZS	37.56	37.25	37.56	37.25	29.70	22.70	29.70	22.70
BAR	53.25	53.07	53.93	53.30	33.98	26.05	34.73	27.72
VP w/ SPSA-GC	60.34	53.86	59.58	51.88	16.21	25.68	18.43	30.13
BlackVIP	**66.21**	**62.47**	**65.19**	**64.47**	**69.08**	**60.86**	**76.97**	**67.97**

- BlackVIP shows significantly better performance than baseline methods under object-non-centered and adversarially-disturbanced setting.

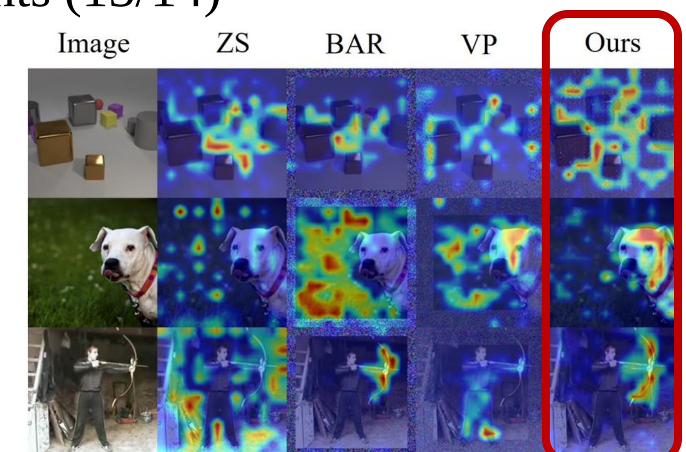
Few-shot Adaptation on Diverse Domains

Empirical Results

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**

BlackVIP achieves consistent performance improvements (13/14) across diverse image domains.

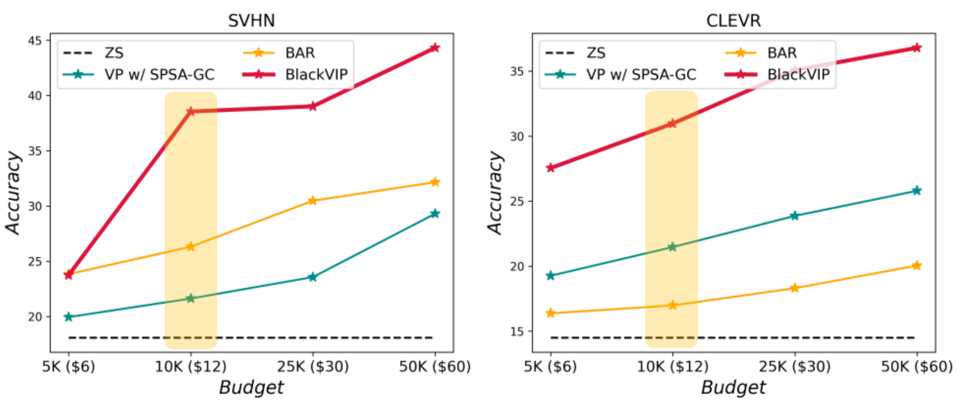
by controlling the attention of PTM
to focus on the proper region of targeted semantics.



Practical Usefulness

Empirical Results

- Moreover, BlackVIP shows outstanding **query efficiency**, which is crucial for adaptation in real-world applications.
- And greatly **reduces** the trainable **parameters** and required **pick memory allocation**.



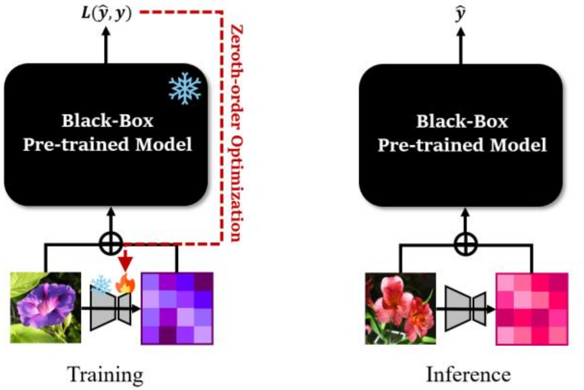
(Costs are based on Clarifai Vision API)

Table 4. Train-time peak memory allocation (Peak Memory) and the number of learnable parameters (Params) on ImageNet.

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**

$12 (USD) with 10k query makes about ×2 performance
e.g., (left) 18% to 38%, (right) 14% to 32%

Conclusion



- We pioneered the ***black-box visual prompting*** for realistic and robust adaptation of pre-trained models.
- For this, we devised **Coordinator**, which reparameterizes the prompt as an autoencoder to handle the input-dependent visual prompt with tiny parameters.
- Besides, we provided the new zeroth-order optimizer **SPSA-GC**, which gives look-ahead corrections to the SPSA's estimated gradient for fast convergence.
- We extensively validated BlackVIP on 16 datasets and demonstrate its effectiveness regarding few-shot adaptability, robustness on distribution/object-location shift, and practical usefulness.

Thank you for watching!

Contact:

changdae730@gmail.com

@Changdae_Oh

Paper

https://arxiv.org/pdf/2303.14773.pdf

Code

https://github.com/changdaeoh/BlackVIP