Black-Box Forgetting

Yusuke Kuwana[1] **Yuta Goto**[1] **Takashi Shibata**[2] **Go Irie**[1]

[1]Tokyo University of Science [2]NEC Corporation

Motivation

Large-scale pre-trained models (PTMs) have strong capabilities of zero-shot classification for everyday objects.

ex. CLIP [Radford+, ICML21]

Dataset Examples

	ImageNet ResNet101	Zero-Shot CLIP	Δ Score
ImageNet	**76.2**	**76.2**	0%
ObjectNet	32.6	**72.3**	+39.7%
ImageNet Sketch	25.2	**60.2**	+35.0%

Motivation

Practical applications do not always require the classification of all kinds of objects.

ex. Autonomous driving system

necessary

unnecessary

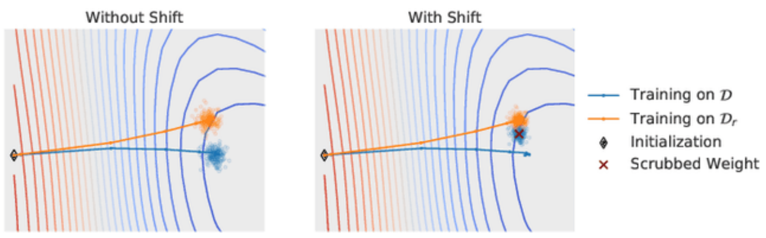
Retaining unnecessary classes may have disadvantages.

- <u>Decrease overall accuracy</u>
- <u>Information leakage</u>

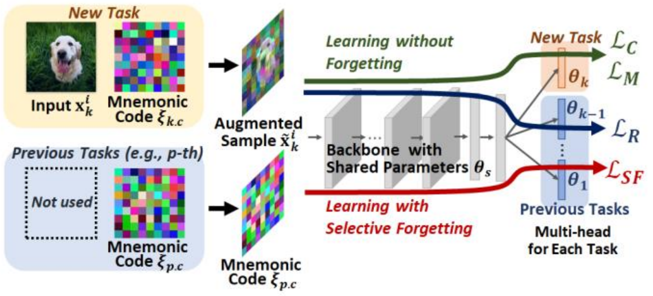
We address the problem of selective forgetting.

Motivation

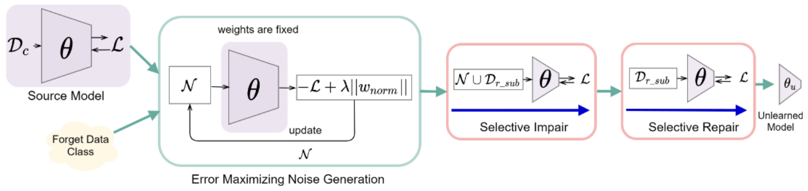
Existing selective forgetting methods



[Golatker+, ECCV20]



[Shibata+, IJCAI21]



[Tarun+, IEEE TNNLS23]

All the existing methods assume “white-box” settings, where model information is available for training.

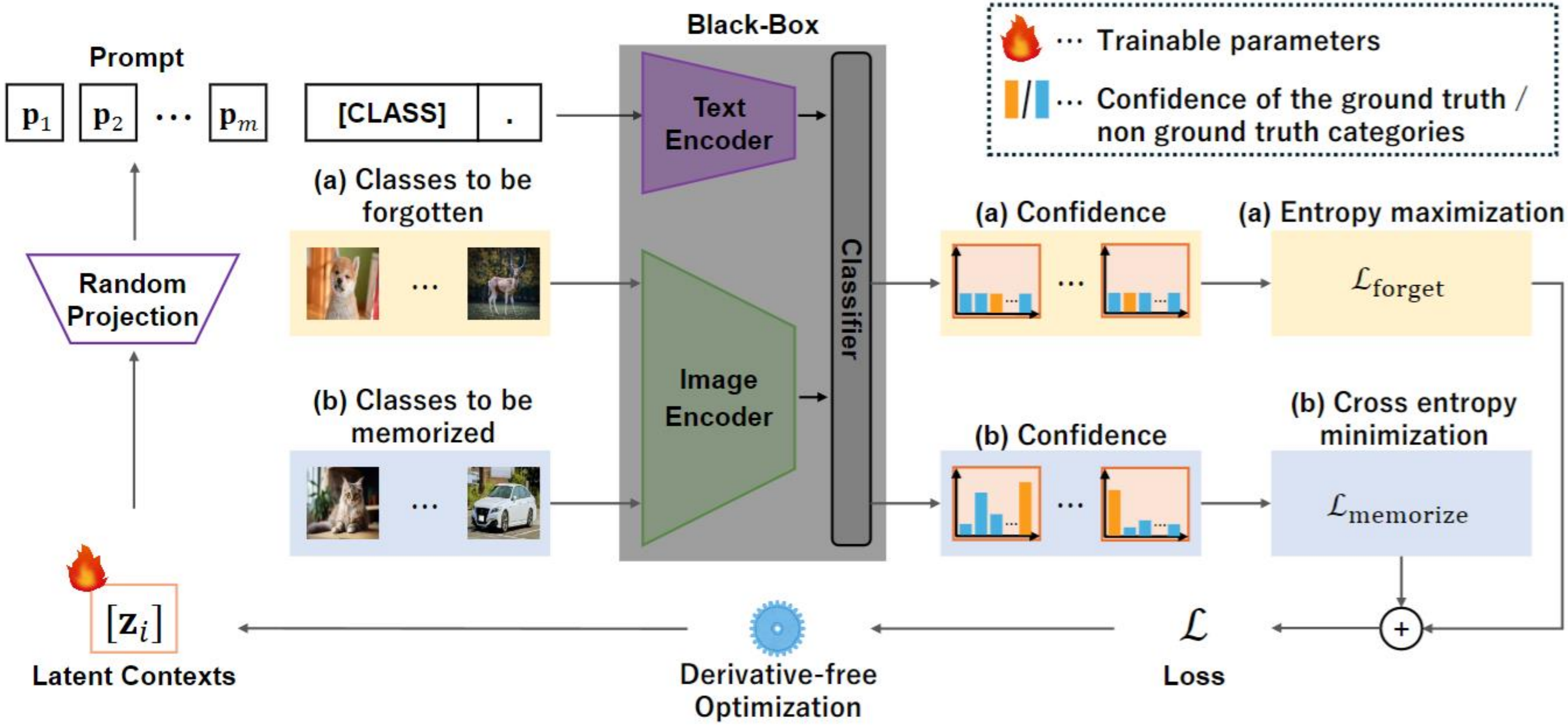
However, PTMs are often “black-box,” where model information is unavailable for commercial reasons or social responsibilities.

→ **Inapplicable for black-box models.**

We address a novel problem of selective forgetting for black-box models.

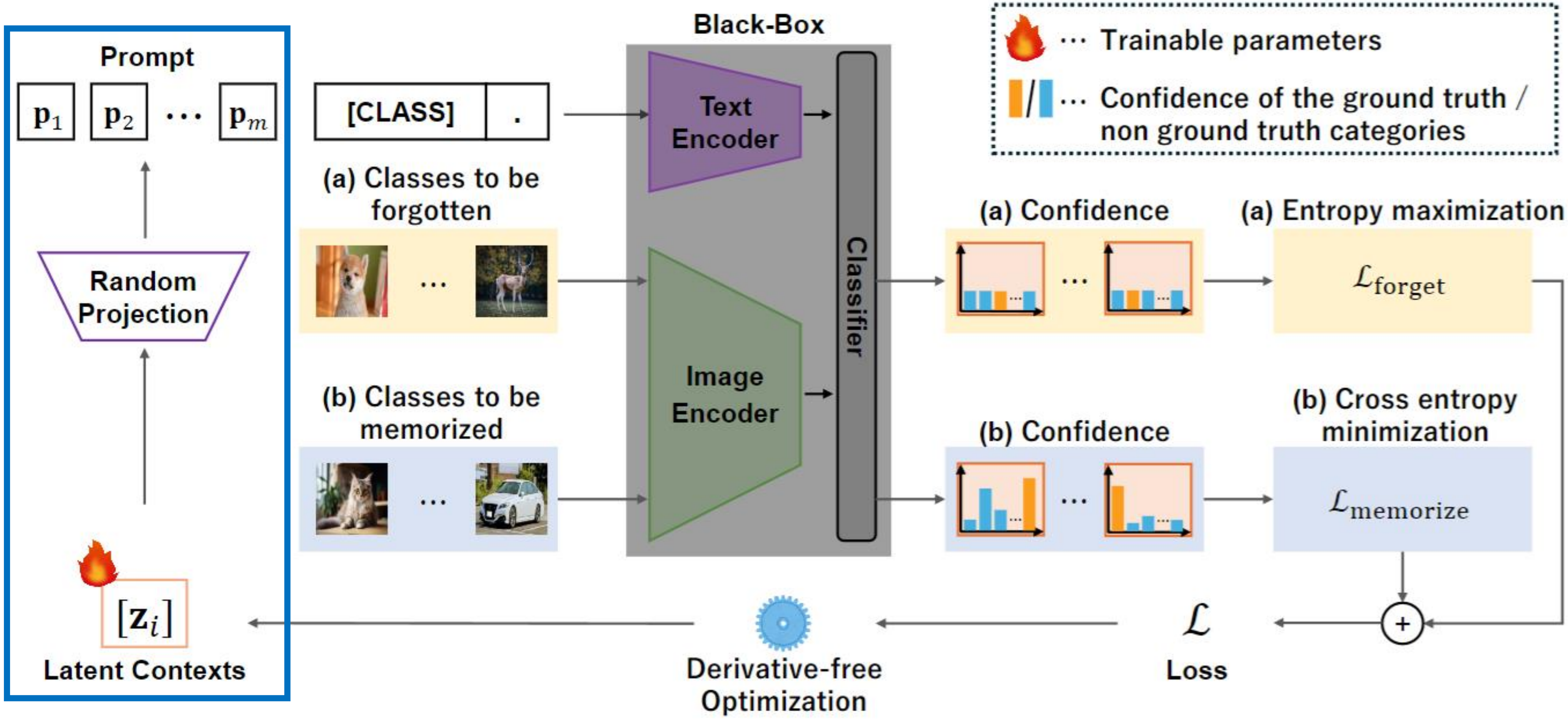
Our Method

We optimize the textual prompt to decrease the accuracy of specified classes through derivative-free optimization, because the gradients of the loss are unavailable in black-box models.



Our Method

We optimize lower-dimensional latent contexts instead of optimizing contexts for the textual prompt directly to mitigate high-dimensional optimization.

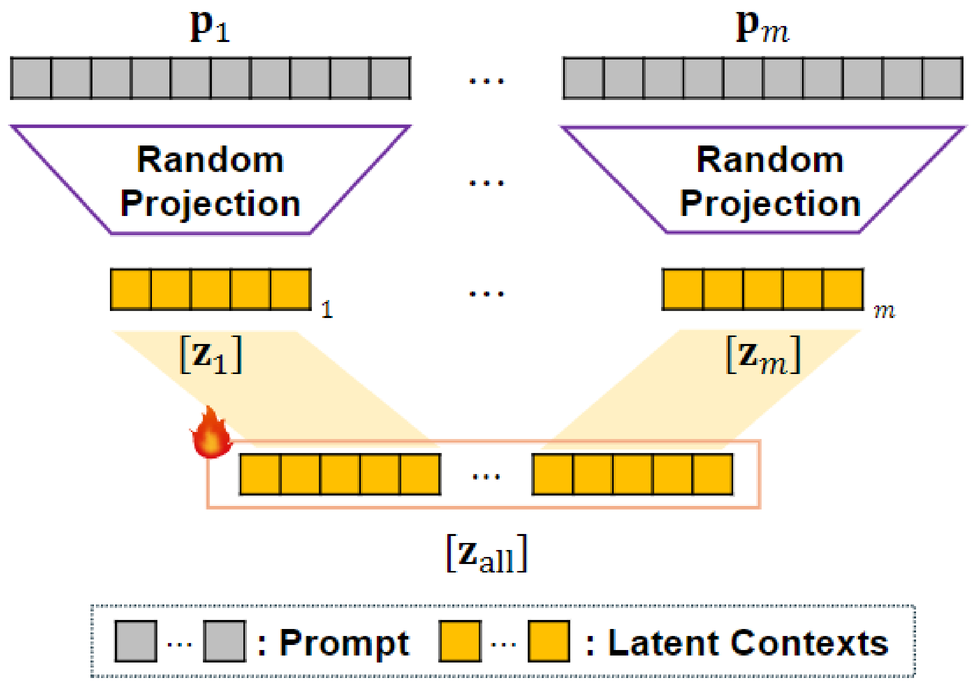


Our Method

BBT [Sun et al., ICML22] optimizes a lower-dimensional latent context instead of directly optimizing textual prompt to mitigate high dimensionality.

We found that the effectiveness of context parametrization through BBT is limited for Black-Box Forgetting.

BBT [Sun et al., ICML22]



Method	CIFAR-10		
	$H\uparrow$	$Err_{\text{for}}\uparrow$	$Acc_{\text{mem}}\uparrow$
Zero-Shot CLIP	15.30	8.37	89.05
BBT [Sun et al., ICML22]	85.69	79.31	93.19
CoOp [Zhou et al., IJCV22] (White-Box)	96.49	96.95	96.04

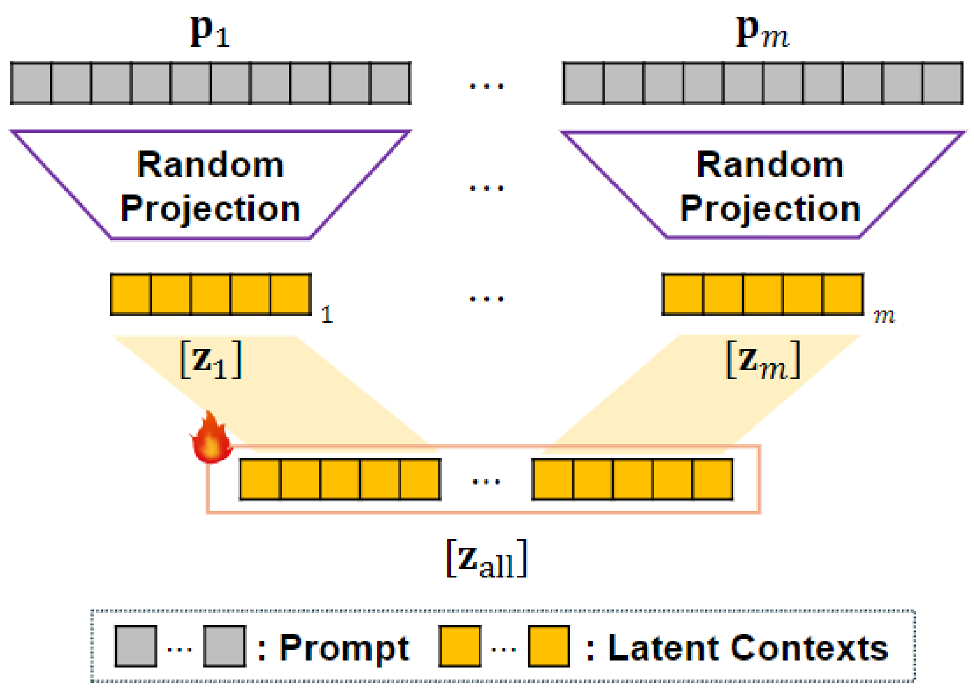
Err_{for} ⋯ the error of the classes to be forgotten
Acc_{mem} ⋯ the accuracy of the classes to be memorized
H ⋯ the harmonic mean of Err_{for} and Acc_{mem}

Our Method

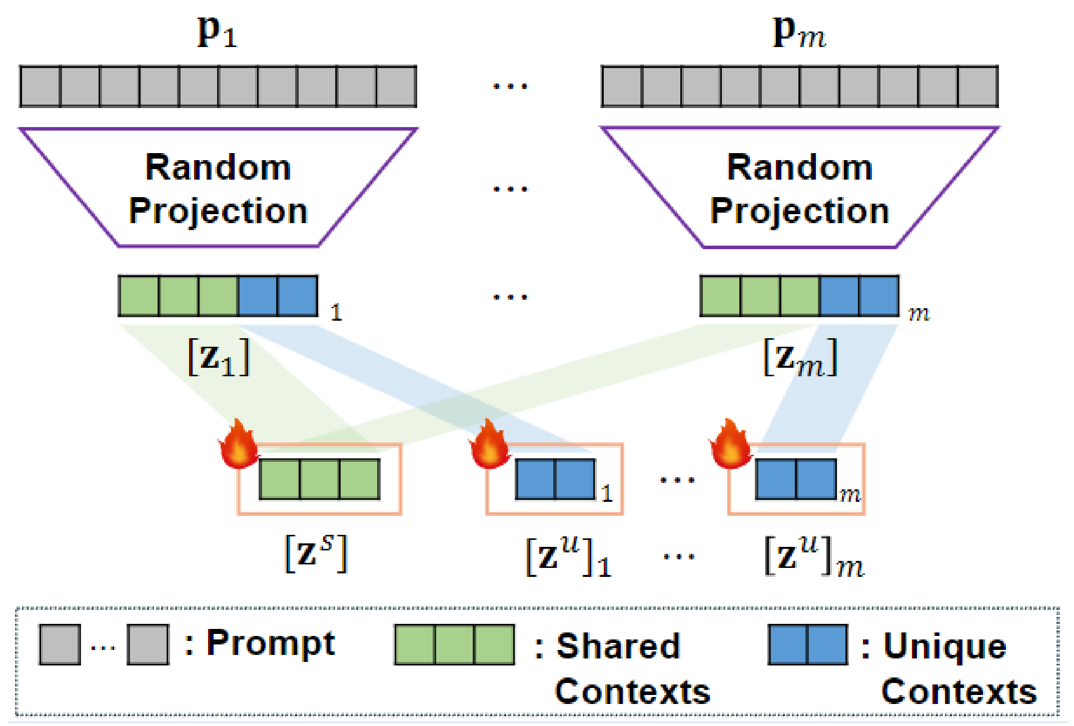
We propose **Latent Context Sharing (LCS)** for more effective context parametrization.

In **LCS**, a latent context is composed of **unique** components and **common** components among multiple latent contexts, and each component is optimized independently.

BBT [Sun et al., ICML22]



Ours: Latent Context Sharing (LCS)



Experiments

Our method outperforms the reasonable baselines on all the datasets.

Table 1: Comparisons with the baselines. The best value is shown in **bold**. BBT [Sun et al., 2022b] and CBBT (w/o adapter) [Guo et al., 2023] are the reasonable baselines as these are for black-box prompt tuning. CoOp [Zhou et al., 2022b] is a white-box method and is included for a reference. Performance is evaluated using the three metrics: the error Err_{for} for the classes to be forgotten, the accuracy Acc_{mem} for the classes to be memorized, the harmonic mean H of Err_{for} and Acc_{mem}. Higher values mean better performance.

Method	CIFAR-10			CIFAR-100		
	$H \uparrow$	$Err_{\mathrm{for}} \uparrow$	$Acc_{\mathrm{mem}} \uparrow$	$H \uparrow$	$Err_{\mathrm{for}} \uparrow$	$Acc_{\mathrm{mem}} \uparrow$
Zero-Shot CLIP	15.30	8.37	89.05	42.14	31.17	65.03
BBT	$85.69_{\pm 0.02}$	$79.31_{\pm 0.03}$	$93.19_{\pm 0.01}$	$78.36_{\pm 0.01}$	$87.30_{\pm 0.01}$	$71.09_{\pm 0.00}$
CBBT	$93.48_{\pm 0.02}$	$90.99_{\pm 0.04}$	$\mathbf{96.11}_{\pm 0.00}$	$73.20_{\pm 0.00}$	$72.69_{\pm 0.01}$	$\mathbf{73.72}_{\pm 0.00}$
Ours (w/o LCS)	$72.37_{\pm 0.13}$	$58.57_{\pm 0.17}$	$94.68_{\pm 0.01}$	$79.38_{\pm 0.02}$	$89.17_{\pm 0.03}$	$71.52_{\pm 0.01}$
Ours	$\mathbf{95.07}_{\pm 0.01}$	$\mathbf{96.10}_{\pm 0.02}$	$94.06_{\pm 0.01}$	$\mathbf{80.99}_{\pm 0.01}$	$\mathbf{93.37}_{\pm 0.02}$	$71.52_{\pm 0.01}$
CoOp (White-Box)	$96.49_{\pm 0.00}$	$96.95_{\pm 0.01}$	$96.04_{\pm 0.00}$	$82.22_{\pm 0.00}$	$99.81_{\pm 0.00}$	$69.90_{\pm 0.01}$
Method	**CUB-200-2011**			**ImageNet30**		
	$H \uparrow$	$Err_{\mathrm{for}} \uparrow$	$Acc_{\mathrm{mem}} \uparrow$	$H \uparrow$	$Err_{\mathrm{for}} \uparrow$	$Acc_{\mathrm{mem}} \uparrow$
Zero-Shot CLIP	46.30	46.20	**46.41**	2.31	1.17	98.00
BBT	$58.75_{\pm 0.01}$	$88.98_{\pm 0.04}$	$43.85_{\pm 0.01}$	$94.22_{\pm 0.05}$	$90.17_{\pm 0.08}$	$99.06_{\pm 0.01}$
CBBT	$56.84_{\pm 0.01}$	$73.52_{\pm 0.02}$	$46.33_{\pm 0.01}$	$87.88_{\pm 0.08}$	$79.69_{\pm 0.12}$	$\mathbf{99.32}_{\pm 0.02}$
Ours (w/o LCS)	$58.78_{\pm 0.01}$	$85.85_{\pm 0.01}$	$44.69_{\pm 0.01}$	$95.26_{\pm 0.02}$	$92.19_{\pm 0.03}$	$98.59_{\pm 0.01}$
Ours	$\mathbf{59.67}_{\pm 0.01}$	$\mathbf{89.29}_{\pm 0.01}$	$44.81_{\pm 0.01}$	$\mathbf{97.28}_{\pm 0.01}$	$\mathbf{95.94}_{\pm 0.01}$	$98.67_{\pm 0.01}$
CoOp (White-Box)	$63.20_{\pm 0.02}$	$98.09_{\pm 0.02}$	$46.62_{\pm 0.02}$	$99.30_{\pm 0.01}$	$99.72_{\pm 0.00}$	$98.89_{\pm 0.01}$

Conclusion

- We proposed **Black-Box Forgetting**, a novel problem of selective forgetting for black-box models.
- We introduced **Latent Context Sharing (LCS),** an efficient and effective parametrization method of prompt, which is suitable for derivative-free optimization.
- Experimental results demonstrated that our method outperforms the reasonable baselines.

arXiv

Thank you!

code