Bayesian Optimization Meets Self-Distillation (Supplementary Materials)

HyunJae Lee* Heon Song* Hyeonsoo Lee* Gi-hyeon Lee Suyeong Park Donggeun Yoo
Lunit Inc.
{hjlee, hslee, dgyoo}@lunit.io, {songheony, lghsigma597, suyeong.park0}@gmail.com

Method	CIFAR-10	CIFAR-100	Tiny-ImgNet
Baseline	93.75	74.43	53.33
Grid	93.87	74.51	53.94
SD	94.21	76.08	56.46
BO	94.52	76.48	55.39
SD+BO	94.57	76.53	56.89
BOHB	94.66	76.64	56.13
BOSS (Ours)	**94.98**	**77.69**	**58.55**

Table A.1. Extended comparison of Top-1 accuracy (%) on CIFAR-10/100 and Tiny-ImageNet with VGG-16, incorporating additional methods like Grid, SD+BO, and BOHB.

A. More Comparisons with Various Methods

Expanding upon the experiments conducted in Section 4.1, our comparative analysis is extended by incorporating additional methods on CIFAR-10/100 and Tiny ImageNet datasets, namely Grid, SD+BO, and BOHB. The Grid method conducts hyper-parameter searches within the same search space and budget as the other methods. The SD+BO employs the standard SD process but integrates BO directly into the training of the student model. Additionally, for benchmarking BOSS against a state-of-the-art approach, we select BOHB [1], recognized for its exceptional performance across various benchmarks through the fusion of BO and Hyperband. The results of the comparative analysis are consolidated in Table A.1. While both SD+BO and BOHB exhibit performance enhancements over individual SD and BO strategies, BOSS demonstrates superior performance by effectively leveraging prior knowledge.

B. Extra Computational Cost

BOSS involves multiple SD trials throughout the BO process. Despite the increased computational resource requirements, BO is widely adopted in numerous applications [6,24,25] due to its ability to enhance task performance while minimizing manual hyperparameter search. Following a similar principle, BOSS proposes an effective design choice that combines BO and SD for a substantial boost in model performance. While BOSS introduces a negligible amount of additional training time compared to the naive combination of SD and BO (as shown in Table B.1), it shows meaningful performance improvement over SD+BO (see Table A.1). More importantly, BOSS introduces no further computational burden during inference. Therefore, when the deployed model is used in practical applications, there is no extra computational overhead. Given its performance benefits, analogous to the established value of BO in diverse machine learning applications, BOSS stands as a worthwhile extension despite the slightly higher computational demand.

	CIFAR10	CIFAR100	Tiny-ImgNet
SD+BO	1h 11m 2s	1h 12m 38s	3h 34m 40s
BOSS	1h 11m 4s (+2s)	1h 12m 41s (+3s)	3h 34m 42s (+2s)

Table B.1. Training duration of VGG-16 for a single trial on one T4 GPU, 4 core Intel Skylake CPU and 15GB RAM.

References

[1] Stefan Falkner, Aaron Klein, and Frank Hutter. Bohb: Robust and efficient hyperparameter optimization at scale. In *ICML*, 2018. 1

[*] Authors contributed equally.