Supplementary material for ST-CoNAL: Consistency-Based Acquisition Criterion Using Temporal Self-Ensemble for Active Learning

Jae Soon Baik[0000−0002−1508−0619], In Young Yoon[0000−0003−3173−3947], and Jun Won Choi[0000−0002−3733−0148]

Hanyang University, Seoul, Korea
{jsbaik, inyoungyoon}@spa.hanyang.ac.kr, junwchoi@hanyang.ac.kr

A Experimental Setup

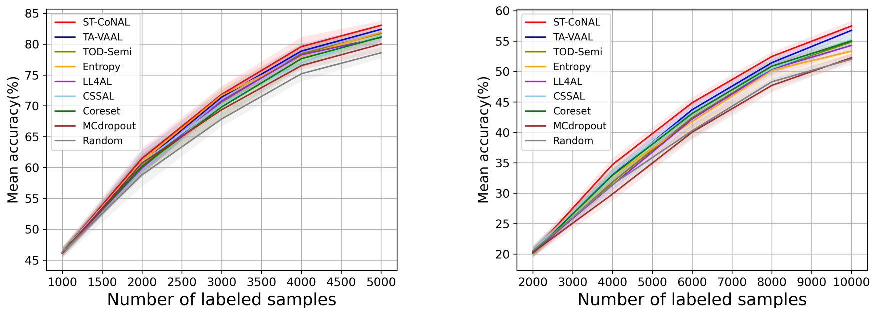
In this section, we provide the detailed experimental setups. The configuration parameters used for ST-CoNAL are provided in Table 1. As for SSL setup, we followed the configurations used in [8].

Dataset	CIFAR-10	CIFAR-100	Caltech-256	Tiny ImageNet
Initial learning rate (l_0)	0.1	0.1	0.01	0.005
Nesterov momentum	0.9	0.9	0.9	0.9
Weight decay	0.0004	0.0004	0.001	-
Batch size	128	128	128	128
Labeled batch size	32	32	32	32
Total epochs	200	200	200	100
Size of unlabeled subset ($\|\mathcal{S}\|$)	10k	20k	10k	20k
Budget (b)	1k	2k	1k	2k
Size of initially labeled set	1k	2k	1k	2k
Storing interval (c)	10	10	10	10
Learning rate decay (γ)	1.0	0.5	0.3	0.5
Learning rate decay point (T_0)	160	160	160	60

Table 1: The configuration parameters used for ST-CoNAL.

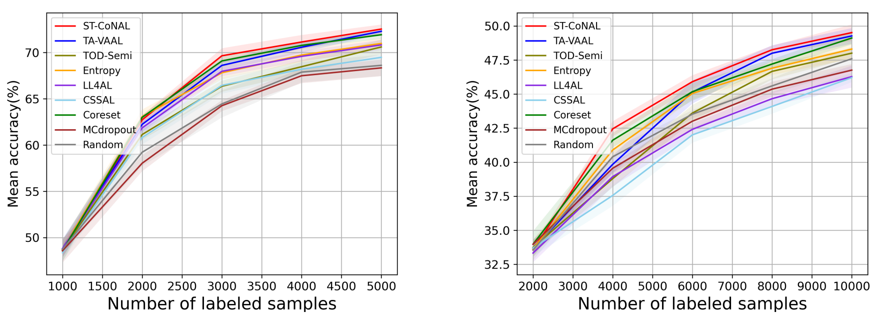
B Experimental Results

In the main manuscript, we used the performance gap from the random sampling baseline as a performance measure. To supplement this, we provide the absolute values of the average accuracy achieved by the AL methods. Fig. 1 (a) and (b) show the performance results evaluated on CIFAR-10 and CIFAR-100 datasets and Fig. 2 (a) and (b) represent those on Caltech-256 and Tiny ImageNet datasets.



(a) (b)

Fig. 1: Average test accuracy versus the number of labeled samples on (a) CIFAR-10 and (b) CIFAR-100 dataset.

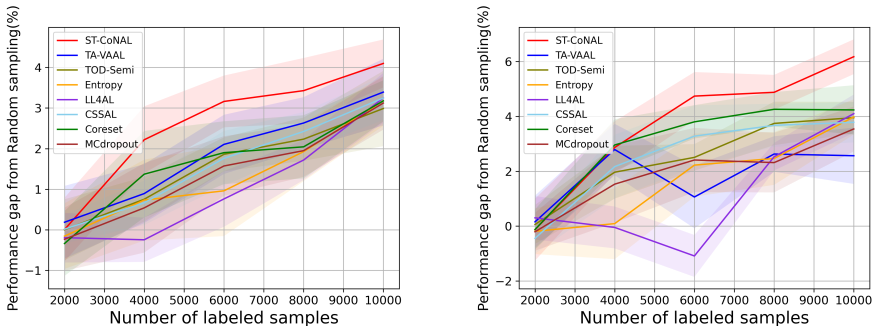


(a) (b)

Fig. 2: Average test accuracy versus the number of labeled samples on (a) Caltech-256 and (b) Tiny ImageNet dataset.

C Performance Comparison of AL Methods on Differently Imbalanced Datasets

In the main manuscript, we reported the performance of ST-CoNAL obtained on two imbalanced versions of CIFAR-10. In this section, we evaluate the algorithms on two differently imbalanced CIFAR-100 datasets, the step-imbalanced CIFAR-100 [2] and long-tailed CIFAR-100 [1, 2]. The imbalance ratio was set to 100 for all cases. Fig. 3 (a) and (b) present the performance of ST-CoNAL on the step imbalanced CIFAR-100 [5] and the long-tailed CIFAR-100 [1], respectively. Even when different imbalance setups are used, ST-CoNAL consistently outperforms other AL methods. After the last acquisition step, ST-CoNAL achieves a performance improvement of 4.10% and 6.18% over the random sampling on the step-imbalanced CIFAR-100 and long-tailed CIFAR-100, respectively.



(a) (b)

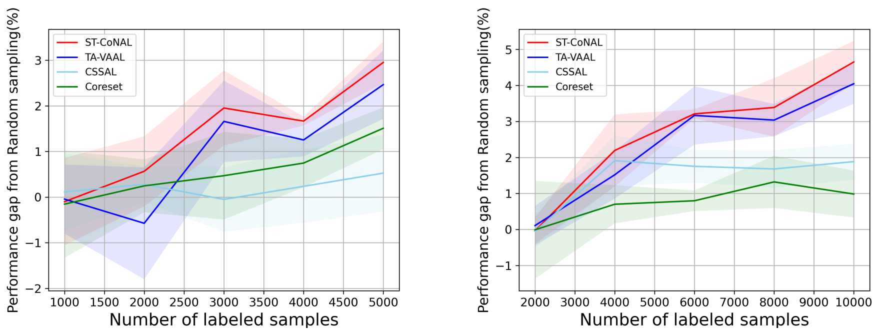
Fig. 3: Average accuracy improvement from random sampling versus the number of labeled samples on (a) the step imbalanced CIFAR-100 and (b) the long-tailed CIFAR-100. The imbalance ratio was set to 100 for both cases.

D Performance Comparison of AL Methods With Different Backbones

In this section, we evaluate the performance of ST-CoNAL when VGG16 [7] and ResNet-50 [4] are used as a backbone network. Fig. 4 (a) and (b) show the performance of ST-CoNAL on CIFAR-10 and CIFAR-100 when VGG16 backbone is used. After the last acquisition step, ST-CoNAL achieves 2.95% and 4.65% performance gains over random sampling on CIFAR-10 and CIFAR-100, respectively. Fig. 5 presents the performance of ST-CoNAL when ResNet-50 backbone is used. ST-CoNAL achieves the performance improvement of 4.99% and 8.91% over random sampling on CIFAR-10 and CIFAR-100, respectively.

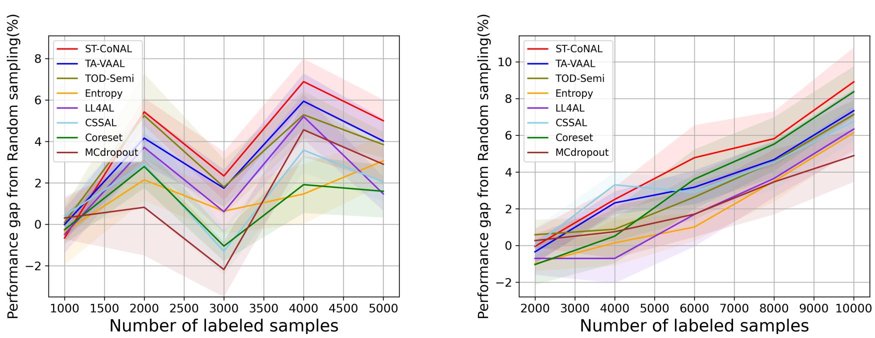
E Performance Versus Other Parameters

We evaluate the performance of ST-CoNAL as a function of the budget sizes b, the number of student models Q and temperature parameter T on CIFAR-10. We try the different values of $b \in \{500, 1k, 2k\}$, $Q \in \{2, 4, 8, 10\}$, and $T \in \{0.3, 0.5, 0.7, 1.0\}$. For the budget sizes b, we compare our ST-CoNAL with three competitive methods, Entropy [6], and TA-VAAL [3]. Table 2 provides the classification accuracy as a function of the acquisition step for different values of labeling budget b. The proposed ST-CoNAL maintains the performance gain over Entropy [6] and TA-VAAL [3] with different budget sizes. Additionally, we provide the performance of ST-CoNAL as a function of acquisition step versus the number of student models Q and temperature parameter T. Table 3 shows that ST-CoNAL achieves good performance for different Q and T values. For CIFAR-10 dataset, we set $Q = 4$ and $T = 0.7$ to provide decent performance.



(a) (b)

Fig. 4: Average accuracy improvement from random sampling versus the number of labeled samples evaluated on CIFAR-10 and (b) CIFAR-100. We evaluate the performance of AL methods using VGG16 backbone [7].



(a) (b)

Fig. 5: Average accuracy improvement from random sampling versus the number of labeled samples evaluated on (a) CIFAR-10 and (b) CIFAR-100. We evaluate the performance of AL methods using ResNet-50 backbone [4].

$b = 500$	$1k$ samples labeled	$2k$ samples labeled	$3k$ samples labeled	$4k$ samples labeled	$5k$ samples labeled
Entropy	46.83	59.29	68.33	76.10	81.16
TA-VAAL	47.26	60.71	69.71	76.35	82.27
ST-CoNAL	46.93	**60.91**	**70.39**	**77.03**	**83.40**
$b = 1k$	$1k$ samples labeled	$2k$ samples labeled	$3k$ samples labeled	$4k$ samples labeled	$5k$ samples labeled
Entropy	46.96	61.00	71.69	78.09	81.92
TA-VAAL	46.94	60.79	71.40	78.87	82.40
ST-CoNAL	47.13	**61.48**	**71.85**	**79.61**	**83.05**
$b = 2k$	$1k$ samples labeled		$3k$ samples labeled		$5k$ samples labeled
Entropy	47.19		71.00		81.29
TA-VAAL	47.18		71.22		81.40
ST-CoNAL	46.99		**72.48**		**82.53**

Table 2: Mean accuracy versus the number of labeled data samples as the function of budget size b.

Q	$1k$ samples labeled	$2k$ samples labeled	$3k$ samples labeled	$4k$ samples labeled	$5k$ samples labeled
2	46.67	60.99	70.95	79.18	82.63
4	47.13	61.48	**71.85**	**79.61**	83.05
8	46.86	**61.49**	71.77	79.23	82.83
10	46.70	61.37	71.64	79.40	**83.14**
T	$1k$ samples labeled	$2k$ samples labeled	$3k$ samples labeled	$4k$ samples labeled	$5k$ samples labeled
0.3	47.00	60.66	70.62	78.03	82.57
0.5	47.23	60.91	70.82	78.51	82.35
0.7	47.13	**61.48**	**71.85**	79.61	**83.05**
1.0	46.87	60.91	71.41	**79.78**	82.96

Table 3: Mean accuracy versus the number of labeled data samples as the function of the number of student models Q and temperature parameter T.

References

1. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 32 (2019)
2. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9268–9277 (2019)
3. Gao, M., Zhang, Z., Yu, G., Arık, S.Ö., Davis, L.S., Pfister, T.: Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In: European Conference on Computer Vision (ECCV). pp. 510–526 (2020)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
5. Kim, K., Park, D., Kim, K.I., Chun, S.Y.: Task-aware variational adversarial active learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8166–8175 (2021)

6. Shannon, C.E.: A mathematical theory of communication. Bell system technical journal **27**(3), 379–423 (1948)
7. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR) (2015)
8. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 1195–1204 (2017)