Supplementary Material:
Improving Open-Set Semi-Supervised Learning with Self-Supervision

Erik Wallin[1,2], Lennart Svensson[2], Fredrik Kahl[2], Lars Hammarstrand[2]
[1]Saab AB, [2]Chalmers University of Technology
{walline,lennart.svensson,fredrik.kahl,lars.hammarstrand}@chalmers.se

1. Choices of hyperparameters

While SeFOSS may appear to have an extensive set of hyperparameters, most values are gathered from existing works and have been used successfully without further tuning. For example, the architectures for the CIFAR experiments follow FixMatch [3] and the architecture for ImageNet-30 follows OpenMatch [2]. Parameters such as the labeled and unlabeled batch sizes, weight regularization, SGD momentum, EMA momentum, learning rate, and the learning rate schedule all follow FixMatch and have not been further tuned for SeFOSS. However, the initial constant learning rate for the pre-training phase is a new addition for SeFOSS.

The decay rates for the learning rates for CIFAR-10 and CIFAR-100 follow DoubleMatch [4] without further tuning for SeFOSS. The temperature scaling of the energy score is set to $\beta = 1$ as suggested by the original work [1]. The weight for the self-supervised feature loss, w_s, was largely determined based on results from DoubleMatch, which found that a smaller w_s works better for CIFAR-10 and a larger w_s works better for CIFAR-100. The number of training steps is empirically determined and is not a very sensitive hyperparameter; no large performance gains were found by increasing the number of training steps further. The weight for the energy regularization, w_e, is new for SeFOSS and was determined by grid searches using CIFAR-10 as ID and CIFAR-100 as OOD. The hyperparameters determining the thresholds for pseudo inliers and outliers, and the margin for the hinge loss (ξ_{id}, ξ_{ood} and ζ_{ood}), were initially given rough estimates by observing the empirical distributions of energies for ID and OOD test data in relation the the distribution of energies for labeled ID training data. These hyperparameters were then more carefully tuned by grid search using CIFAR-10 as ID and CIFAR-100 as OOD.

2. Motivating the free-energy score

The original work that proposes the free-energy score for OOD detection [1] finds that the free-energy score generally outperforms the baseline of using the maximum softmax confidence, $\max_y p(y|x)$. To further motivate the use of the free-energy score in SeFOSS, we conduct experiments where we evaluate AUROC for the confidence score and for the free-energy score at the end of the pre-training phase of SeFOSS. In the pre-training phase, the only losses used are the supervised cross-entropy, l_l, the self-supervised feature consistency, l_s, and the weight regularization, l_w. We perform these evaluations at the end of the pre-training phase, because until then, the energy regularization that amplifies the performance of the energy-score is not applied. The results are shown in Tab. 1. We see that the energy score performs better or equal to the confidence score in all scenarios, giving strong support to the design choice of using the free-energy score in SeFOSS.

3. OSR for previously unseen OOD

The results in the main paper evaluate open-set recognition at testing for the OOD classes that appear in the unlabeled set. While we argue that this scenario is closest to real-world applications, it can also be of interest to evaluate the performance for open-set recognition on classes that are completely unseen during training. For example, the model may be trained to classify CIFAR-10, encounters OOD data from CIFAR-100 in the unlabeled training data, but is then evaluated for open-set recognition with CIFAR-10 as ID and SVHN as OOD. Results from these experiments are shown in Tabs. 2 to 7. We see that SeFOSS generally achieves high AUROCs also for unseen OOD data. The scenarios where we observe poor performance are when we evaluate using CIFAR-10 as unseen OOD when CIFAR-100 is ID. This seems reasonable because many of the classes in CIFAR-10 are very similar to classes that appear in CIFAR-100, and the model has had no chance to learn to distinguish between these in training.

	CIFAR-10: 1,000 labels			CIFAR-10: 4,000 labels			CIFAR-100: 2,500 labels			CIFAR-100: 10,000 labels		
	CIFAR-100	SVHN	Noise	CIFAR-100	SVHN	Noise	CIFAR-10	SVHN	Noise	CIFAR-10	SVHN	Noise
Confidence	0.83	0.99	1.0	0.84	0.98	1.0	0.64	0.98	1.00	0.72	0.99	1.00
Energy	0.84	1.00	1.0	0.86	1.00	1.0	0.66	1.00	1.00	0.74	1.00	1.00

Table 1. AUROC at the end of the pretraining phase for SeFOSS using either the softmax confidence or the free-energy score.

Test OOD	SVHN	Noise
AUROC	0.99	0.98

Table 2. ID: CIFAR-10 4,000 labels, Train OOD: CIFAR-100

Test OOD	CIFAR-100	Noise
AUROC	0.90	0.99

Table 3. ID: CIFAR-10 4,000 labels, Train OOD: SVHN

Test OOD	CIFAR-100	SVHN
AUROC	0.91	0.95

Table 4. ID: CIFAR-10 4,000 labels, Train OOD: Noise

Test OOD	SVHN	Noise
AUROC	0.91	0.94

Table 5. ID: CIFAR-100 10,000 labels, Train OOD: CIFAR-10

Test OOD	CIFAR-10	Noise
AUROC	0.76	0.85

Table 6. ID: CIFAR-100 10,000 labels, Train OOD: SVHN

Test OOD	CIFAR-10	SVHN
AUROC	0.77	0.92

Table 7. ID: CIFAR-100 10,000 labels, Train OOD: Noise

References

[1] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2020. 1

[2] Kuniaki Saito, Donghyun Kim, and Kate Saenko. Openmatch: Open-set semi-supervised learning with open-set consistency regularization. In *Advances in Neural Information Processing Systems*, 2021. 1

[3] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, 2020. 1

[4] Erik Wallin, Lennart Svensson, Fredrik Kahl, and Lars Hammarstrand. DoubleMatch: Improving semi-supervised learning with self-supervision. In *International Conference on Pattern Recognition*, 2022. 1