# S4L: Self-Supervised Semi-Supervised Learning (2019)

Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, Lucas Beyer

**Summary**

## Contributions

The authors propose to take advantage of self-supervised techniques for semi-supervised learning, the proposed framework works well in both settings, a semi-supervised settings using the additional knowledge extracted from the pretext task, and transfer learning tasks that can greatly benefit from the additional labels. For labeled images, the model outputs predictions for both the pretext task and the classification task, and only the pretext task for the unlabeled examples.

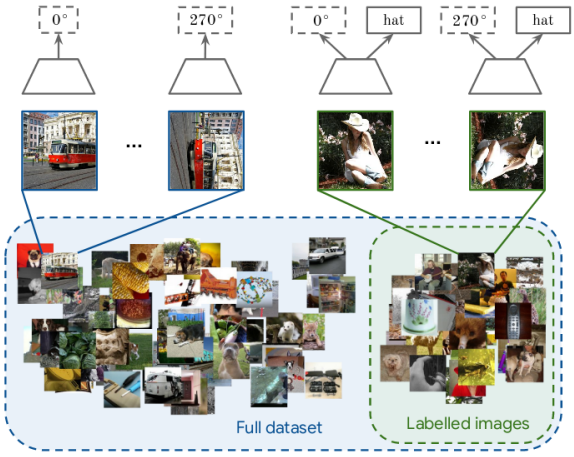


Figure 1. A schematic illustration of one of the proposed self-supervised semi-supervised techniques: $S^4L$-Rotation. Our model makes use of both labeled and unlabled images. The first step is to create four input images for any image by rotating it by 0°, 90°, 180° and 270° (inspired by [10]). Then, we train a single network that predicts which rotation was applied to all these images and, additionally, predicts semantic labels of annotated images. This conceptually simple technique is competitive with existing semi-supervised learning methods.

## Method

In semi supervised learning the training objective is to minimize the supervised loss where the labeled images with used in a standard cross-entropy loss and the unsupervised loss, which will be defined using a pretext task in this case.

$$\min_{\theta} \mathcal{L}_l\left(D_l, \theta\right) + w\mathcal{L}_u\left(D_u, \theta\right)$$

**Self-supervised learning** The authors propose two pretext tasks to train on both the labeled and unlabeled examples (in fact $\mathcal{L}_u is computed over both the labeled and unbelaled sets$.

- **S4L-Rotation.** For a set of possible rotations $\{0°, 90°, 180°, 270°\}$. The model must predict the correct rotation value that was applied to the input. The loss in this case is.

$$\mathcal{L}_{rot} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \sum_{x \in D_u} \mathcal{L}\left(f_\theta\left(x^r\right), r\right)$$

- **Exemplar.** In exemplar, various augmentation are applied to a single given image, all of the resulting augmented images (8 instances in this case, with random H flip and random HSV space randomization) with the original ones will construct a new class. The model needs to predict the correct class for a given image.

# Results

The authors compare the proposed methods: Rotation and Exemplar, when applied on in a self-supervised setting, with fine tuning and only using a linear layer over the representations. And then when used with a supervised loss over the labeled examples. ANd compare then to the supervised baseline when the model is only trained using the labeled examples, VAT and VAT with entropy minimization, and pseudo labeling where the baseline model is used to labels the whole dataset and then used to re-train the network. The authors also propose a new model called MOAM, mix of all models, where its trained on three steps, first the model is trained using rotation in a semi-supervised setting with VAT and Entropy minimization. The trained model is then used to generate pseudo labels and its then retrained with the same losses as the first step, and the last step is a fine tuning step on the labeled set.

Table 1. Top-5 accuracy [%] obtained by individual methods when training them on ILSVRC-2012 with a subset of labels. All methods use the same standard width ResNet50v2 architecture.

| ILSVRC-2012 labels: (i.e. images per class) | 10 % (128) | 1 % (13) |
|---|---|---|
| Supervised Baseline (Section 4.1) | 80.43 | 48.43 |
| Pseudolabels [20] | 82.41 | 51.56 |
| VAT [24] | 82.78 | 44.05 |
| VAT + Entropy Minimization [11] | 83.39 | 46.96 |
| Self-sup. Rotation [17] + Linear | 39.75 | 25.98 |
| Self-sup. Exemplar [17] + Linear | 32.32 | 21.33 |
| Self-sup. Rotation [17] + Fine-tune | 78.53 | 45.11 |
| Self-sup. Exemplar [17] + Fine-tune | 81.01 | 44.90 |
| $S^4L$-Rotation | 83.82 | 53.37 |
| $S^4L$-Exemplar | 83.72 | 47.02 |

Table 2. Comparing our MOAM to previous methods in the literature on ILSVRC-2012 with 10 % of the labels. Note that *different models use different architectures*, larger than those in Table 1.

| | labels | Top-5 | Top-1 |
|---|---|---|---|
| MOAM full (*proposed*) | 10% | **91.23** | **73.21** |
| MOAM + pseudo label (*proposed*) | 10% | 89.96 | 71.56 |
| MOAM (*proposed*) | 10% | 88.80 | 69.73 |
| ResNet50v2 (4×wider) | 10% | 81.29 | 58.15 |
| VAE + Bayesian SVM [32] | 10% | 64.76 | 48.41 |
| Mean Teacher [41] | 10% | 90.89 | - |
| †UDA [43] | 10% | 88.52† | 68.66† |
| †CPCv2 [13] | 10% | 84.88† | 64.03† |
| *Training with all labels:* | | | |
| ResNet50v2 (4×wider) | 100% | 94.10 | 78.57 |
| MOAM (*proposed*) | 100% | **94.97** | **80.17** |
| †UDA [43] | 100% | 94.45† | 79.04† |
| †CPCv2 [13] | 100% | 93.35† | - |

† *marks concurrent work.*

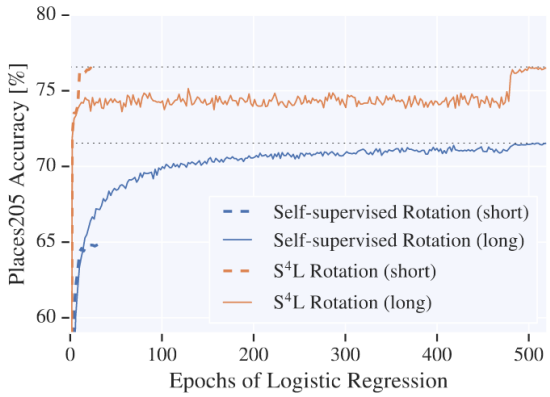


Figure 2. Places205 learning curves of logistic regression on top of the features learned by pre-training a self-supervised versus $S^4L$-Rotation model on ILSVRC-2012. The significantly faster convergence ("long" training schedule vs. "short" one) suggests that more easily separable features are learned.

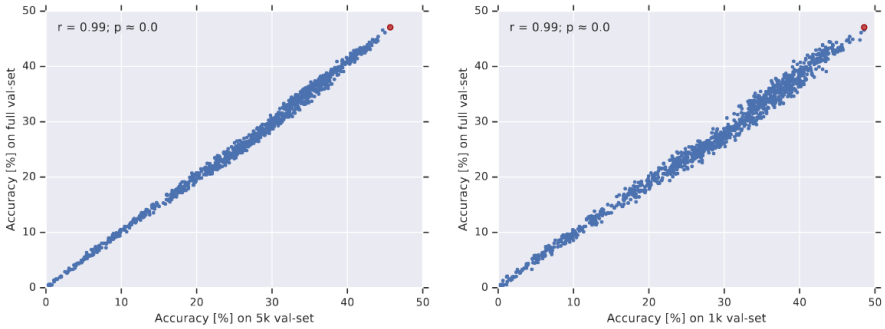


Figure 3. Correlation between validation score on a (custom) validation set of 1000, 5000, and 50 046 images on ILSVRC-2012. Each point corresponds to a *trained model* during a sweep for plain supervised baseline for the 1 % labeled case. The best model according to the validation set of 1 000 is marked in red. As can be seen, evaluating our models even with only a single validation image per class is robust, and in particular selecting an optimal model with this validation set works as well as with the full validation set.