S^{4}L: Self-Supervised Semi-Supervised Learning Supplemental Material

Xiaohua Zhai,* Avital Oliver,* Alexander Kolesnikov,* Lucas Beyer*
Google Research, Brain Team
{xzhai, avitalo, akolesnikov, lbeyer}@google.com

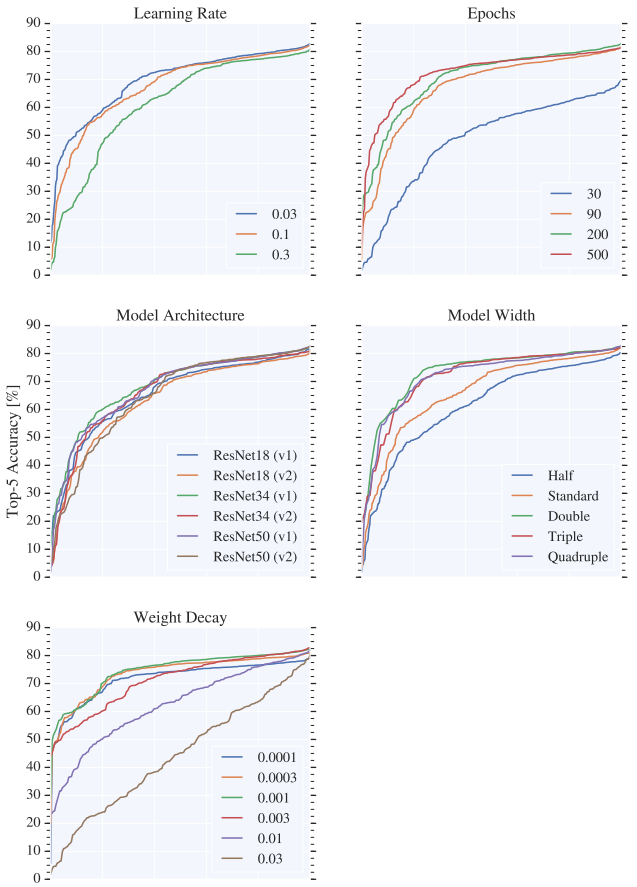

Figure 1. The "hypersweep curves" for the supervised baseline trained on 10 % of ILSVRC-2012. See text for details.

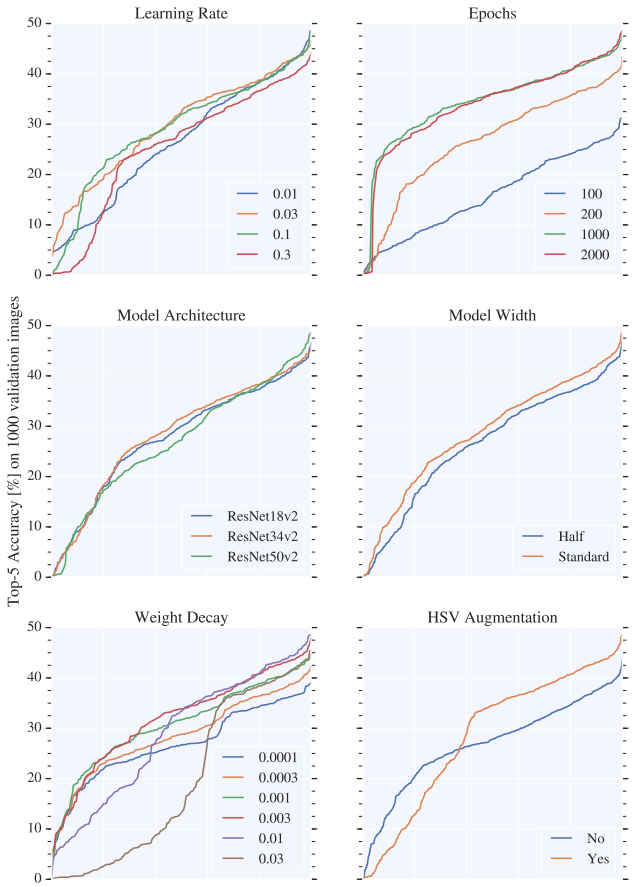

Figure 2. The "hypersweep curves" for the supervised baseline trained on 1 % of ILSVRC-2012. See text for details.

A. Detailed Results of the Supervised Baselines

Since we performed quite extensive hyperparameter search and trained many models in order to find a solid fully-supervised baseline on 10 % and 1 % of ILSVRC-2012, we believe that it is valuable to report the full results to the community, instead of just providing the final best model.

We present the results in the form of what we call "hypersweep curves" in Figures 1 and 2.

Each plot shows a large collection of models – *each point on each plot is a fully trained model*. The curves are sorted by accuracy, allowing testing sensitivity to different hyper-

*equal contribution

parameters, not only comparing the best model.

For each curve, we plot the accuracy of models where one of the hyperparameters is fixed.

Thus, by comparing curves, one can see:

1. Which value of a hyperparameter performs best by looking at which curve's rightmost point is highest.
2. How sensitive the model is to a hyperparameter *in the best case* by looking at how far apart the curves are from eachother at their rightmost point.
3. How robust a hyperparameter is *on average* by looking at how similar the curves are overall.
4. How independent a specific hyperparameter value is from all others by looking at the curve's shape, and whether curves cross-over (strong interplay) or not (strong independence).

While the results shown in Figure 1 use the full (custom) validation set, those in Figure 2 were computed using the validation set of size 1000, *i.e*. with only one image per class. As we have shown in Section 7, this is sufficient to determine the best hyperparameters, and we encourage the community to follow this more realistic protocol.

As can be seen, weight decay and number of training epochs are the two things which matter most when training using only a fraction of ILSVRC-2012.

Perhaps the most surprising finding is that, contrary to current folklore, *reducing model capacity is detrimental* to performance on the smaller dataset. Neither reducing depth, nor reducing width improve performance. In fact, the deeper and wider models still outperform their shallower and thinner counterparts, even when using only $1\,\%$ of the training data. Even more so, the wider models are more robust to other hyperparameter's values as evidenced by their curves being significantly higher on the left end. This is in line with recent findings suggesting wider models ease optimization [5, 1, 7].

Furthermore, when reducing the dataset size to $1\,\%$, we found that adding the same color augmentation as introduced by Exemplar is helpful. We thereafter tried adding it to our best few models on $10\,\%$, but it did not help there.

Finally, while in the $1\,\%$ case, learning-rate of 0.1 and 0.01 seem to perform equally well in the good cases (right hand side of curves), we manually inspected training curves and found that 0.1 is significantly less robust, typically not learning anything before the first decay, and only catching up later on.

While we trained thousands of models in order to rigorously test multiple hypotheses (such as that of reducing model capacity), almost all boost in performance could have been achieved in just a few dozen trials with intuitively important hyperparameters (weight decay and epochs), which would take about a week on a modern four-GPU machine.

Overall, we hope that this thorough baseline investigation inspires the semi-supervised learning community to be more careful with baselines, as those that we found perform almost $20\,\%$ absolute better than those previously reported in the literature.

B. Randomness of S^4L

Table 1. S^4L performance for 9 runs with random image subsets. Top-5 accuracies [%] are reported as mean±standard deviation.

Method	10% ImageNet	1% ImageNet
S^4L-Rotation	83.91 ± 0.13	53.47 ± 0.22
S^4L-Exemplar	83.76 ± 0.06	46.61 ± 0.25

There are two factors of randomness of a semi supervised model: (1) labeled subset sampling, (2) run with different seeds. In order to estimate the randomness in the performance we train 9 models with random data subsets and random seeds for our proposed S^4L method. Table 1 presents the detailed results. Overall, we observe that standard deviation is fairly small across both subsets and different runs and, therefore, our empirical evaluation provides robust comparison of various techniques.

C. More Results in the Transfer Setup

In this section we present more results from the transfer evaluation task on Places205 [8]. Table 2 shows the results for the models mentioned in our main paper. For each method, we select the best model and evaluate its transfer to Places205.

We follow the same setup as [3] to train a linear models with SGD on top of frozen representations. The only difference is the training epochs, we train for 30 epochs in total with learning rate decayed at 10 and 20 epochs respectively. The learning rate is linearly ramped up for the first epoch. Kolesnikov et.al. [3] train for 520 epochs with learning rate decays at 480 and 500 epochs. The schedule used in our paper is much shorter because of our finding that representation learned with labels are more separable and converges significantly faster. (See in Section 6 of the main paper for details.) To make fair comparison with the self-supervised models, results in Table 2 with 0% labels are trained for 520 epochs to ensure their convergence.

From the plain supervised baselines, we observe that either more labels or wider networks lead to more transferable representations. Surprisingly, we found that pseudo labels outperforms the other two semi-supervised baselines in the transfer setup. On the 1% labels evaluation setup, pseudo labels achieves the best result comparing to the other methods. With 10% labels, S^4L is comparable to the semi-supervised baselines, and our MOAM clearly outperforms

Table 2. Accuracy (in percent) obtained by various individual methods when transferring their representation to the Places205 dataset using linear models on frozen representations. All methods use the same plain ResNet50v2 base model, except for the ones marked by *, which use a $4\times$ wider network. When it was necessary, a $^+$ marks longer transfer training of 520 epochs. The "%-labels" column shows the percentage of ILSVRC-2012 labels that was used for training the model.

Method	%-labels	top-5	top-1
Supervised	1	65.4	36.2
Supervised	10	75.0	44.7
Supervised	100	81.9	52.5
Supervised*	100	83.1	53.7
SS Rotation$^+$ [3]	0	71.4	41.7
SS Exemplar$^+$ [3]	0	69.0	39.8
Pseudolabels [4]	1	71.6	41.8
VAT [6]	1	64.9	35.9
VAT + EntMin [2]	1	65.9	36.4
Pseudolabels [4]	10	78.1	48.2
VAT [6]	10	76.4	45.8
VAT + EntMin [2]	10	76.4	46.2
SS Rotation [3] + Fine-tune	1	66.1	36.3
SS Exemplar [3] + Fine-tune	1	60.0	31.1
SS Rotation [3] + Fine-tune	10	75.4	45.9
SS Exemplar [3] + Fine-tune	10	75.6	45.9
S^4L-Rotation	1	67.3	38.0
S^4L-Exemplar	1	61.2	32.2
S^4L-Rotation	10	76.4	46.6
S^4L-Exemplar	10	75.9	45.9
MOAM* full	10	83.3	54.2
MOAM* + pseudo label	10	83.3	54.2
MOAM*	10	79.2	49.5

all other models trained on 10% of labels. More interestingly, the *MOAM (full)* model on 10% is slightly better than the 100% supervised baseline with the same $4\times$ wider network. This indicates that learning a model with multiple losses may lead to representations that generalize better to unseen tasks.

References

[1] C Daniel Freeman and Joan Bruna. Topology and geometry of half-rectified network optimization. *arXiv preprint arXiv:1611.01540*, 2016. 2

[2] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 529–536. MIT Press, 2005. 3

[3] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3

[4] Dong-Hyun Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 07 2013. 3

[5] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems*, pages 6391–6401, 2018. 2

[6] Takeru Miyato, Shin-ichih Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *arXiv preprint arXiv:1704.03976*, 2017. 3

[7] Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2019. 2

[8] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014. 2