Supplementary Material: Rethinking few-shot image classification: a good embedding is all you need?

Yonglong Tian[1]*, Yue Wang[1]*, Dilip Krishnan[2], Joshua B. Tenenbaum[1], and Phillip Isola[1]

[1] MIT
[2] Google Research

A Architectures

The architectures of ResNet-12 ans SEResNet-12 is show in Figure 1.

B More training details

For SEResNet-12, we use the same training setup as ResNet-12 on all four benchmarks, as described in Sec 4.1.

For 4-layer convnet, we also the same training setup as ResNet-12 on tiered-ImageNet, CIFAR-FS, and FC100, For miniImageNet, we train for 240 epochs with learning rate decayed at epochs 150, 180, and 210 with a factor of 0.1. We found that using the logit layer as feature results in slightly better accuracy ($\leq 1\%$) on miniImageNet, so we report this number in Table 6 for miniImageNet.

C Details of unsupervised learning

We adapt the first layer of a standard ResNet-50 to take images of size 84×84 as input. We only train on the meta-train set of miniImageNet dataset (do not use meta-val set). We follow the training recipe in CMC [6] and MoCo [2] (which also follows InstDis [7]) except for two differences. The first one is that we only use 2048 negatives for each positive sample as miniImageNet contains less than 40k images in total. The second difference is that we train for 2000 epochs, with a learning rate initialized as 0.03 and decayed by consine annealing.

D Further discussion

1. What is the intuition of this paper?
A: We hope this paper will shed new light on few-shot classification. We believe representations play an important role. Shown by our empirical experiments, a linear model can generalize well as long as a good representation of the data is given.

2. Why does this simple baseline work? Is there anything that makes few-shot classification special?
A: Few-shot classification is a special case of meta-learning in terms of compositionality of tasks. Each task is an K-way classification problem, and on current benchmarks the classes, even between tasks, are all mutually exclusive. This means we can merge all N of the K-way classification tasks into a single but harder NK-way classification task. Our finding is that training an embedding model on this new NK-way task turns out to transfer well to meta-testing set. On the other hand, we also find that self-supervised embedding, which does not explicitly require this NK compositionality, achieves a similar level of performance. A concurrent work [1] studies the representations for few-shot learning from the theoretical point of view.
3. Does your work negate recent progress in meta-learning?
A: No. Meta-learning is much broader than just few-shot classification. Although we show a simple baseline outperforms other complicated meta-learning algorithms in few-shot classification, methods like MAML may still be favorable in other meta-learning domains (e.g., meta-reinforcement learning).
4. Why does distillation work? What does it suggest?
A: The soft-labels [3] from the teacher model depict the fact that some classes are closer to each other than other classes. For example, a white goat is much more similar to a brown horse than to an airplane. But the one-hot label does not capture this.

After being regularized by soft-labels, the network learns to capture the metric distance. From theoretical perspective, [5] provides analysis for linear case. Ongoing work [4] argues distillation amplifies regularization in Hilbert space.

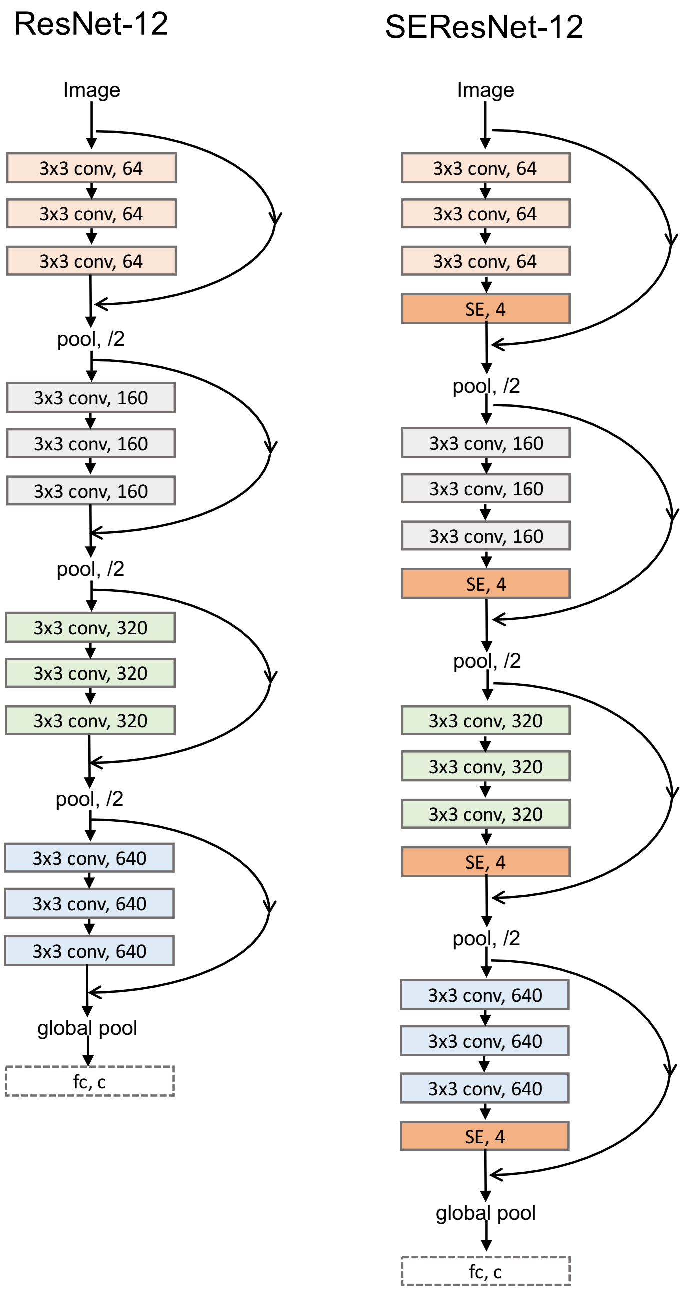

Fig. 1. Nework architectures of ResNet-12 and SEResNet-12 used in this paper. The "SE, 4" stands for a Squeeze-and-Excitation layer with reduction parameter of 4. Dotted box will be removed during meta-testing stage.

References

1. Du, S.S., Hu, W., Kakade, S.M., Lee, J.D., Lei, Q.: Few-shot learning via learning the representation, provably. ArXiv **abs/2002.09434** (2020)
2. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.B.: Momentum contrast for unsupervised visual representation learning. ArXiv **abs/1911.05722** (2019)
3. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: NIPS Deep Learning and Representation Learning Workshop (2015)
4. Mobahi, H., Farajtabar, M., Bartlett, P.L.: Self-distillation amplifies regularization in hilbert space. ArXiv **abs/2002.05715** (2020)
5. Phuong, M., Lampert, C.: Towards understanding knowledge distillation. In: ICML (2019)
6. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. arXiv preprint arXiv:1906.05849 (2019)
7. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: CVPR (2018)