Author Response for “The Unreasonable Effectiveness of Big Models for Semi-Supervised Learning”

We thank the reviewers for feedback, as well as efforts in reviewing. We respond to each comment below.

R1: *“The framework is mostly incremental . . . Overall, there is no significant contribution to unsupervised pre-training.”* **R3**: *“Most major parts in this work . . . are proposed in previous works...the novelties in individual parts are somehow limited.”* The fact that our main contribution is a detailed procedure, rather than a theorem, architecture, or other artifact, does not make the contribution less novel. Compared to previous work, our procedure achieves a $\mathbf{10}\times$ improvement in label efficiency on ImageNet. That we are able to obtain such a large improvement on a well-studied benchmark indicates that our procedure was not previously known to the research community. Beyond the procedure itself, we provide an extensive empirical evaluation of the importance of model size, and ablations for our specific design choices.

We believe our contributions are significant. Our results set a new, much higher bar for future work, and our procedure is likely to have major practical applications. Indeed, R3 recognizes that “the simple semi-supervised framework is still valuable for industry. . . I think it will inspire several future works.” The fact that our approach is simple and intuitive should count in favor of it rather than against it.

R1: *“The main claim...is supported only with empirical results on ImageNet.”* **R3**: *“My major concern is that this paper only conducts experiments on ImageNet... In small dataset, it is well known big models are hard to train.”* While we believe ImageNet is a much more difficult and convincing dataset, we have conducted further experiments on CIFAR-10 given the reviewers’ concern. More specifically, we pretrain ResNet on CIFAR-10 without labels. The ResNet variants we trained are of 6 depths, namely 18, 34, 50, 101, 152 and 200. To keep experiments tractable, by default we use Selective Kernel, and a width multiplier of $1\times$. After the models are pretrained, we then fine-tune them (using simple augmentations of random crop and horizontal flipping) on different numbers of labeled examples (250, 4000, and total of 5000 labeled examples), following MixMatch’s protocol and running on 5 seeds. The results are shown in the Figure 1, and suggest similar trends to our results on ImageNet: big pretrained models can perform well, often better, with a few labeled examples. These results can be further improved with better augmentations during fine-tuning and an extra distillation step.

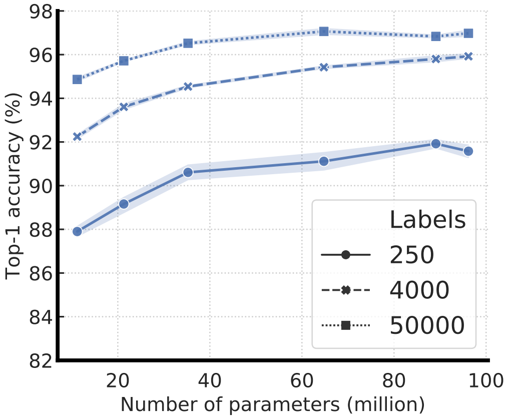

Figure 1: CIFAR-10 fine-tuning.

R2: *“If I understand correctly, the combination of using a larger MLP projection head, and finetuning from the middle layer of the MLP head, is the same as just adding one more FC layer to the base network.”* Fine-tuning from the first layer of the MLP head is the same as adding an FC layer to the base network and removing an FC layer from the head. The impact of this extra FC layer is contingent on the label fraction during fine-tuning (Figure 5b). We will clarify this.

R2: *“What exactly is named SimCLRv2? The contrastive learning part, or the entire method (including finetuning & distillation)?”* We understand the confusion regarding our naming. The name “SimCLRv2” is intended for the whole procedure (including fine-tuning and distillation with unlabeled data), and we will say this explicitly.

R3: *“The ablation in Table 2 of 1% labels shows that distillation can provide 50%+ performance gain which is quite surprising. The student model is much better than the teacher model. Are there any insights and explanations why this happen?”* The student model in Table 2 is not better than the teacher. The teacher (pretrained and fine-tuned) gets 70.6% with 1% of the data and 77.0% with 10%, as shown in in Table 1. We will add the teacher’s accuracy to Table 2. The 50%+ improvement is over pure supervised training on 1% of data (which is heavily overfitting). *“...if self-distillation loss is used...why the distillation can bring so much improvement compared to teacher network?”* The improvement from self-distillation is meaningful but modest: 74.9%→76.6% with 1% of labels and 80.1%→80.9% with 10% of labels for our biggest architecture. It may result either from the use of weaker augmentation during distillation or from regularization induced by distillation. We will add more discussion.

R3: *“It’s interesting to see the semi-supervised trained models’ transfer learning performances on downstream tasks, using fine-tuning.”* In this work, we focus primarily on the problem of semi-supervised learning. But we agree it is interesting to look at the transfer performance of these semi-supervised models, and we will do so after the rebuttal period. We will also address rest of the minor issues (e.g. additional references) brought up by the reviewer.

R4: *“All experiments are conducted on ImageNet, which is a curated dataset, with centered objects, well-balanced class/image distribution.”* We agree that ImageNet is a well curated dataset, and may not reflect all real world scenarios. However, ImageNet is also a well-studied benchmark dataset, and in that sense, it is ideal for evaluating our algorithms on. Nonetheless, we agree this is an important point, and will discuss this issue in the revision.