Big Self-Supervised Models are Strong Semi-Supervised Learners

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, Geoffrey Hinton
Google Research, Brain Team

Compiled by Hongming Shan

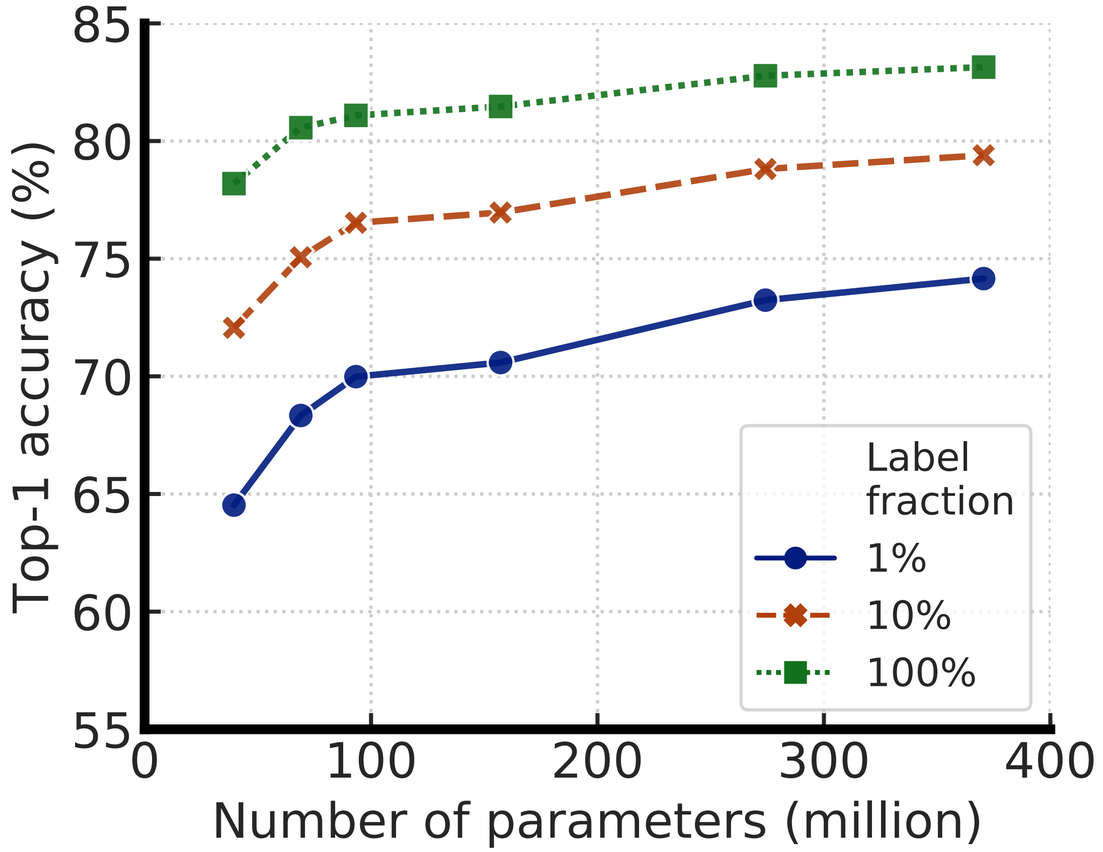
Contributions

- One paradigm for learning from few labeled examples while making best use of a large amount of unlabeled data is **unsupervised pretraining** followed by **supervised fine-tuning** .
- Although this paradigm uses unlabeled data in a **task-agnostic** way, in contrast to most previous approaches to semi-supervised learning for computer vision, we show that it is **surprisingly effective** for semi-supervised learning on ImageNet.
- A key ingredient of our approach is the use of a **big (deep and wide) network** during pretraining and fine-tuning.
- We find that, the fewer the labels, the more this approach (task-agnostic use of unlabeled data) benefits from a bigger network. After fine-tuning, the big network can be further improved and **distilled** into a much smaller one with little loss in classification accuracy by using the unlabeled examples for a second time, but in a **task-specific** way.

Workflow and Achievement

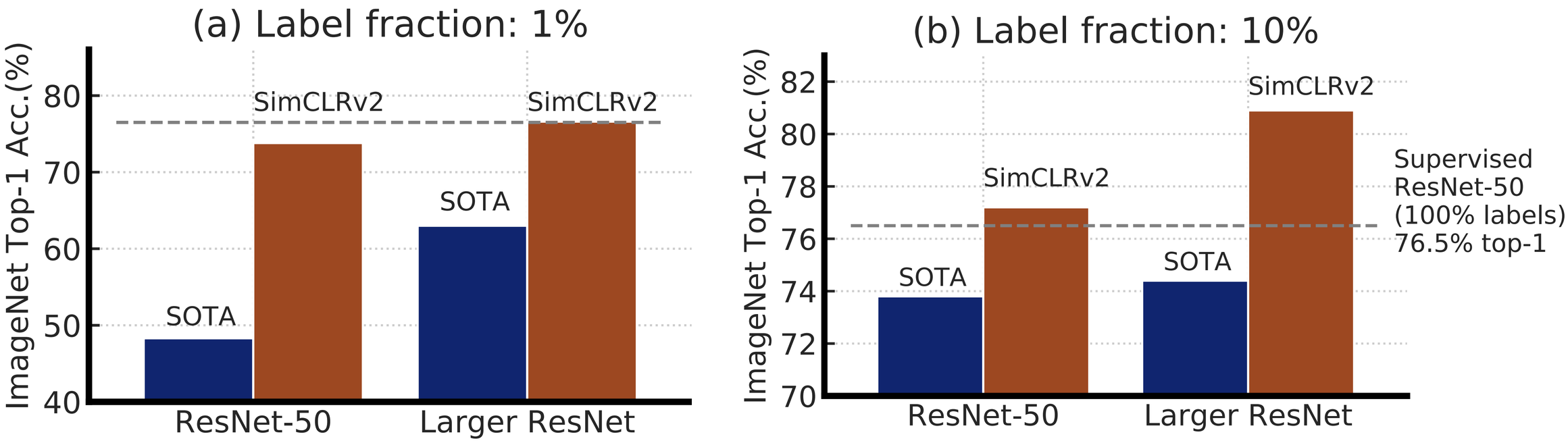
- The proposed semi-supervised learning algorithm can be summarized in three steps:
    1. **unsupervised pretraining** of a big ResNet model using SimCLRv2 (a modification of SimCLR [1]),
    2. **supervised fine-tuning** on a few labeled examples, and
    3. **distillation** with unlabeled examples for refining and transferring the task-specific knowledge.
- This procedure achieves 73.9% ImageNet top-1 accuracy with just 1% of the labels (13 labeled images per class) using ResNet-50, a **10x improvement** in label efficiency over the previous state-of-the-art. With 10% of labels, ResNet-50 trained with our method achieves 77.5% top-1 accuracy, **outperforming** standard supervised training with all of the labels

Glance at results



- Bigger models yield larger gains when fine-tuning with fewer labeled examples.

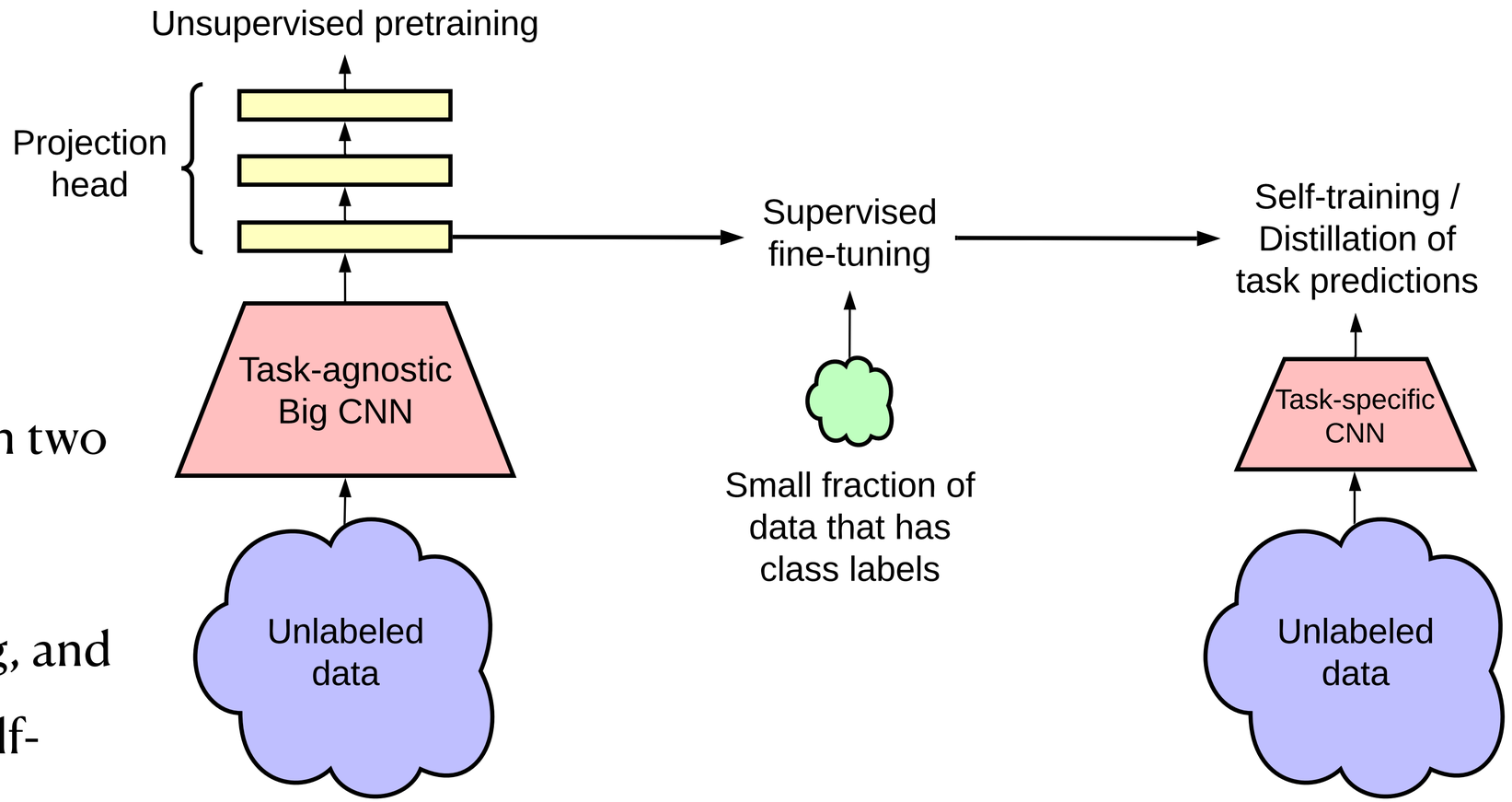
Glance at results



- Top-1 accuracy of previous state-of-the-art (SOTA) methods and our method (SimCLRv2) on ImageNet using only 1% or 10% of the labels. Dashed line denotes fully supervised ResNet-50trained with 100% of labels.

Proposed Semi-supervised Learning Framework

- Three main steps:
  - (1) pretrain;
  - (2) fine-tune; and
  - (3) distill
- Leverages unlabeled data in two ways
  - (1) task-agnostic use in unsupervised pretraining, and
  - (2) task-specific use in self-training / distillation.

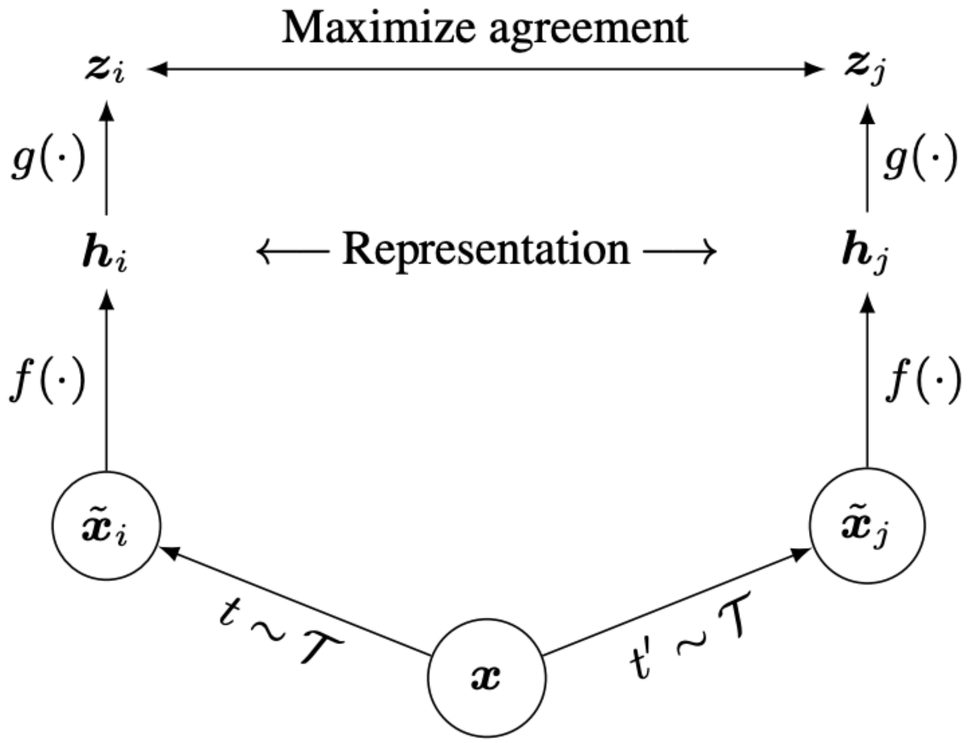


Framework

- Pretrain
  - Self-supervised pretraining with SimCLRv2
- Fine-tune
- Distill
  - Self-training / knowledge distillation via unlabeled examples

SimCLR

- A simple framework for contrastive learning of visual representation
- Two separate data augmentation operators are sampled from the same family of augmentations and applied to each data example to obtain two correlated views.
- A base encoder network f(·) and a projection head g(·) are trained to maximize agreement using a contrastive loss.
- After training is completed, we throw away the projection head g(·) and use encoder f(·) and representation h for downstream tasks.



SimCLR

- With a mini-batch of augmented examples, the contrastive loss between a pair of positive example i, j (augmented from the same image) is give as follows

$$\ell_{i,j}^{\text{NT-Xent}} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)},$$

- sim() is cosine similarity between two vectors, and \tau is a temperature scalar.

SimCLRv2

First change

- To fully leverage the power of general pretraining, we explore **larger ResNet** models.
- Unlike SimCLR [1] and other previous work [27, 20], whose largest model is ResNet-50 (4x), we train models that **are deeper but less wide**. The largest model we train is a 152-layer ResNet [25] with 3x wider channels and selective kernels (SK) [28], a channel-wise attention mechanism that improves the parameter efficiency of the network.
- By scaling up the model from ResNet-50 to ResNet-152 (3x +SK), we obtain a **29% relative improvement** in top-1 accuracy when fine-tuned on 1% of labeled examples.

SimCLRv2

Second change

- We also increase the capacity of the **non-linear network g()** (a.k.a. projection head), by making it deeper.
- Furthermore, instead of throwing away g() entirely after pretraining as in SimCLR [1], we **fine-tune from a middle layer** (detailed later). This small change yields a significant improvement for both linear evaluation and fine-tuning with only a few labeled examples.
- Compared to SimCLR with 2-layer projection head, by using a 3-layer projection head and fine-tuning from the 1st layer of projection head, it results in as much as **14% relative improvement** in top-1 accuracy when fine-tuned on 1% of labeled examples

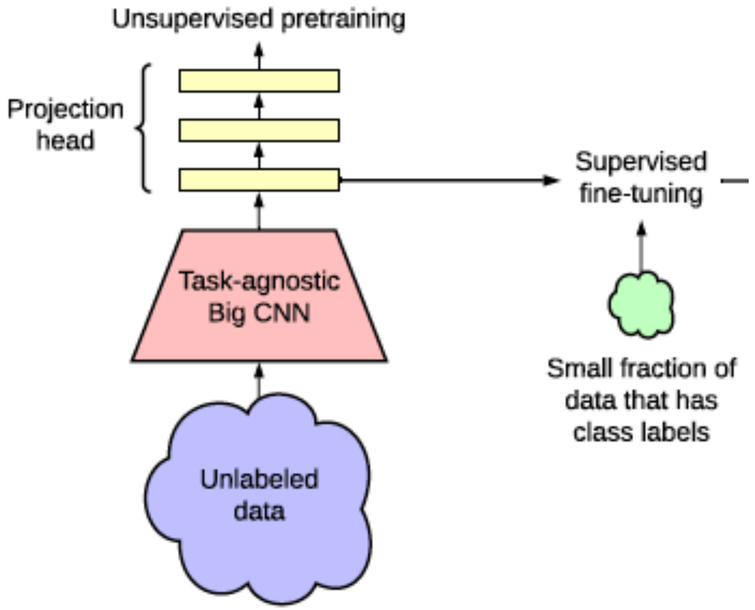
SimCLRv2

Third change

- Motivated by [29], we also incorporate the **memory mechanism** from MoCo [20], which designates a memory network (with a moving average of weights for stabilization) whose output will be buffered as negative examples.
- Since our training is based on large mini-batch which already supplies many contrasting negative examples, this change yields an improvement of **~1%** for linear evaluation as well as when fine-tuning on 1% of labeled examples

Fine-tuning

- Fine-tuning is a common way to adapt the task-agnostically pretrained network for a specific task.
- In SimCLR [1], the MLP projection head g() is discarded entirely after pretraining, while only the ResNet encoder f() is used during the fine-tuning.
- Instead of throwing it all away, we propose to incorporate part of the MLP projection head into the base encoder during the fine-tuning.
- This is equivalent to fine-tuning from a **middle layer** of the projection head, instead of the input layer of the projection head as in SimCLR.



Self-training/knowledge distillation via unlabled examples

- To further improve the network for the target task, here we leverage the unlabeled data directly for the target task. We use the fine-tuned network as a teacher to impute labels for training a student network. Specifically, we minimize the following distillation loss where no real labels are used:

$$\mathcal{L}^{\text{distill}} = -\sum_{\boldsymbol{x}_i \in \mathcal{D}} \left[\sum_{y} P^T(y|\boldsymbol{x}_i;\tau) \log P^S(y|\boldsymbol{x}_i;\tau) \right]$$

- The teacher network is fixed during the distillation. Only the student network is trained.

Self-training/knowledge distillation via unlabled examples

- While we focus on distillation using only unlabeled examples in this work, when the number of labeled examples is significant, one can combine the distillation loss with ground-truth labeled examples using a weighted combination

$$\mathcal{L} = -(1-\alpha) \sum_{(\boldsymbol{x}_i, y_i) \in \mathcal{D}^L} \Big[\log P^S(y_i|\boldsymbol{x}_i) \Big] - \alpha \sum_{\boldsymbol{x}_i \in \mathcal{D}} \Big[\sum_y P^T(y|\boldsymbol{x}_i;\tau) \log P^S(y|\boldsymbol{x}_i;\tau) \Big].$$

Experiments

Settings and Implementation details

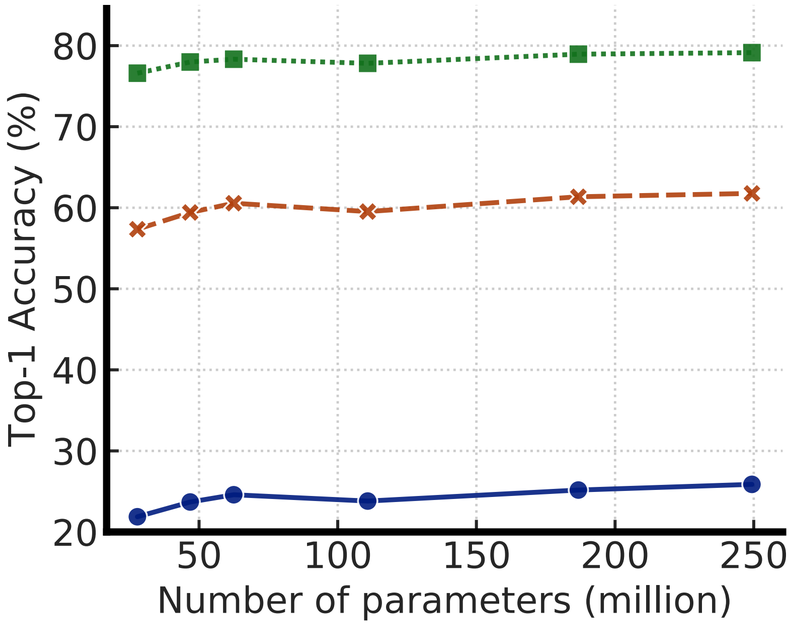
- Evaluate the method on ImageNet, 1% or 10% with associated labels
- Report performance when training a linear classifier on top of a fixed representation
- Use LARS (large batch) optimizer throughout for pretraining, fine-tuning, and distillation.
- Train this model on 128 Cloud TPUs with a batch size of 4096
- For distillation using unlabeled example only, two types:
  - Self-distillation where the student has the same model architecture as the teacher
  - Big-to-Small distillation where the student is a much smaller network

Bigger Models Are More Label-Efficient

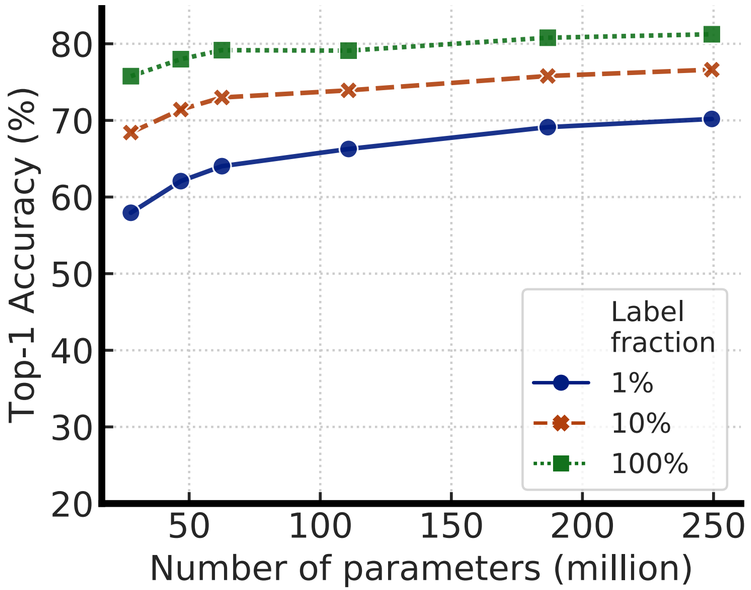
Top-1 accuracy of fine-tuning SimCLRv2

Depth	Width	SK	Param (M)	F-T (1%)	F-T (10%)	F-T (100%)	Linear eval	Supervised
50	1×	False	**24**	**57.9**	**68.4**	**76.3**	**71.7**	**76.6**
		True	35	64.5	72.1	78.7	74.6	78.5
	2×	False	94	66.3	73.9	79.1	75.6	77.8
		True	140	70.6	77.0	81.3	77.7	79.3
101	1×	False	43	62.1	71.4	78.2	73.6	78.0
		True	65	68.3	75.1	80.6	76.3	79.6
	2×	False	170	69.1	75.8	80.7	77.0	78.9
		True	257	73.2	78.8	82.4	79.0	80.1
152	1×	False	58	64.0	73.0	79.3	74.5	78.3
		True	89	70.0	76.5	81.3	77.2	79.9
	2×	False	233	70.2	76.6	81.1	77.4	79.1
		True	354	74.2	79.4	82.9	79.4	80.4
152	3×	True	**795**	**74.9**	**80.1**	**83.1**	**79.8**	**80.5**

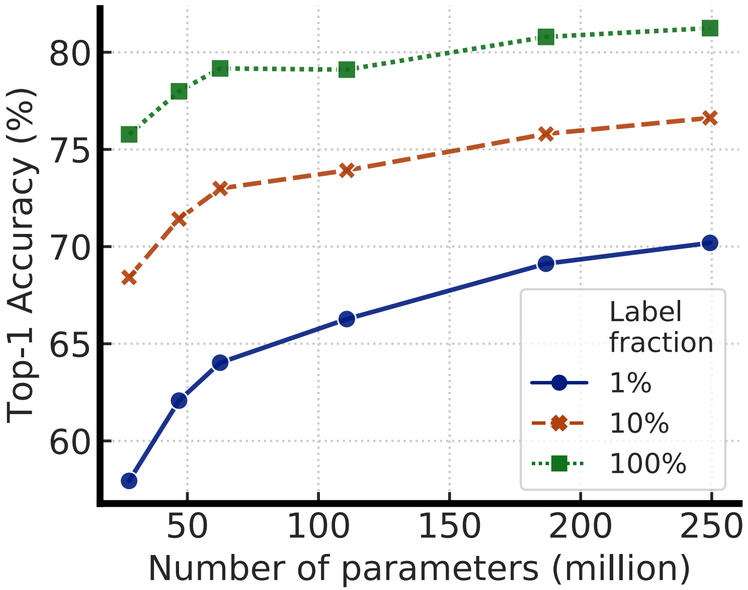
Performance as model size and label fraction vary



Supervised



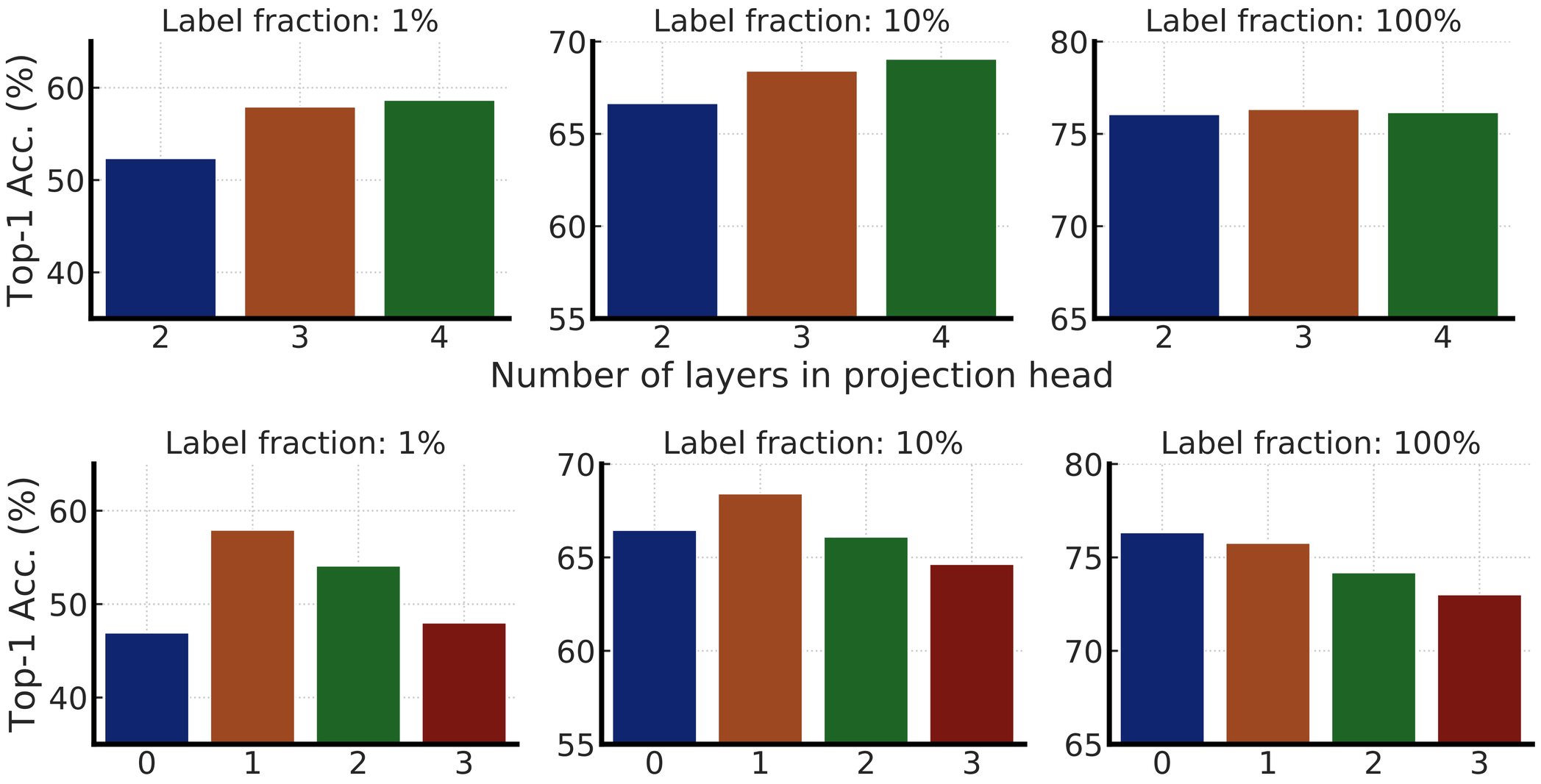
Semi-supervised



Zoomed

- These results show that bigger models are more label-efficient for both supervised and semi-supervised learning, but gains appear to be larger for semi-supervised learning

Bigger/Deeper Projection Heads Improve Representation Learning

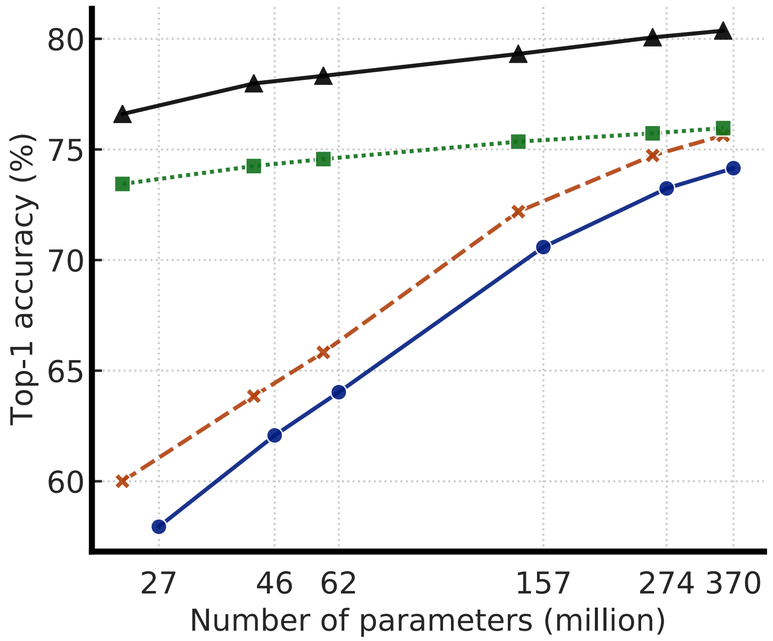


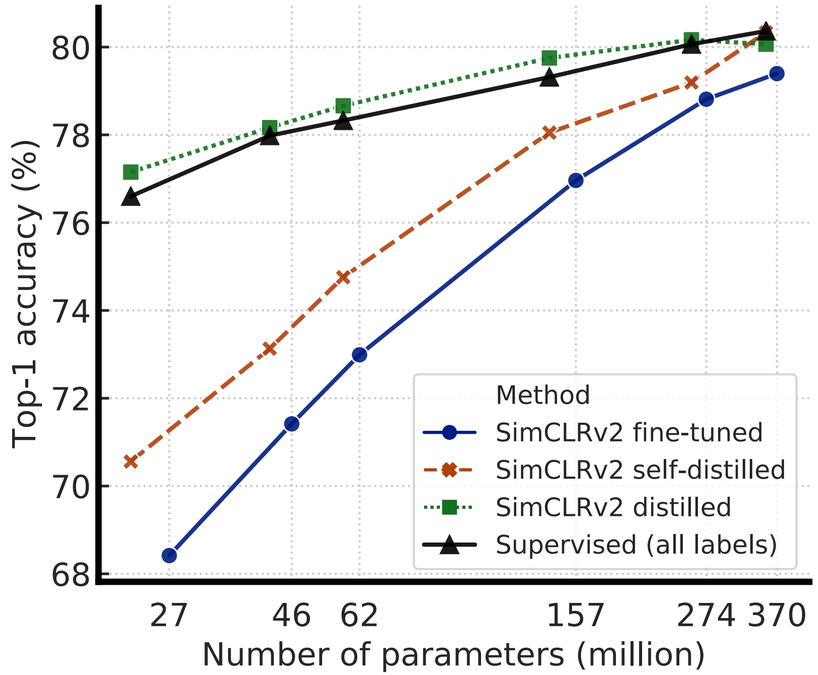
Distillation Using Unlabeled Data Improves Semi-Supervised Learning

Method	Label fraction 1%	10%
Label only	12.3	52.0
Label + distillation loss (on labeled set)	23.6	66.2
Label + distillation loss (on labeled+unlabeled sets)	69.0	75.1
Distillation loss (on labeled+unlabeled sets; our default)	68.9	74.3

- Distillation typically involves both **a distillation loss** that encourages the student to match a teacher and an **ordinary supervised cross-entropy** loss on the labels (Eq. 3).
- In Table 2, we demonstrate the **importance of using unlabeled examples** when training with the distillation loss. Furthermore, using the distillation loss alone (Eq. 2) works almost as well as balancing distillation and label losses (Eq. 3) when the labeled fraction is small.

Distillation Using Unlabeled Data Improves Semi-Supervised Learning





- (1) when the student model has a smaller architecture than the teacher model, it improves the model efficiency by transferring task-specific knowledge to a student model,
- (2) even when the student model has the same architecture as the teacher model (excluding the projection head after ResNet encoder), self-distillation can still meaningfully improve the semi-supervised learning performance.
- To obtain the best performance for smaller ResNets, the big model is self-distilled before distilling it to smaller models.

Comparison with SOTA

Method	Architecture	Top-1 Label fraction 1%	Top-1 Label fraction 10%	Top-5 Label fraction 1%	Top-5 Label fraction 10%
Supervised baseline [30]	ResNet-50	25.4	56.4	48.4	80.4
Methods using unlabeled data in a task-specific way:					
Pseudo-label [11, 30]	ResNet-50	-	-	51.6	82.4
VAT+Entropy Min. [37, 38, 30]	ResNet-50	-	-	47.0	83.4
UDA (w. RandAug) [14]	ResNet-50	-	68.8	-	88.5
FixMatch (w. RandAug) [15]	ResNet-50	-	71.5	-	89.1
S4L (Rot+VAT+Entropy Min.) [30]	ResNet-50 (4×)	-	73.2	-	91.2
MPL (w. RandAug) [2]	ResNet-50	-	73.8	-	-
Methods using unlabeled data in a task-agnostic way:					
BigBiGAN [39]	RevNet-50 (4×)	-	-	55.2	78.8
PIRL [40]	ResNet-50	-	-	57.2	83.8
CPC v2 [19]	ResNet-161(∗)	52.7	73.1	77.9	91.2
SimCLR [1]	ResNet-50	48.3	65.6	75.5	87.8
SimCLR [1]	ResNet-50 (4×)	63.0	74.4	85.8	92.6
Methods using unlabeled data in both ways:					
SimCLRv2 distilled (ours)	ResNet-50	73.9	77.5	91.5	93.4
SimCLRv2 distilled (ours)	ResNet-50 (2×+SK)	75.9	80.2	93.0	95.0
SimCLRv2 self-distilled (ours)	ResNet-152 (3×+SK)	**76.6**	**80.9**	**93.4**	**95.5**

Broader Impact

- The findings can potentially be harnessed to improve accuracy in any application of computer vision where it is more expensive or difficult to label additional data than to train larger models.
- Some such applications are clearly beneficial to society. For example, in medical applications where acquiring high-quality labels requires careful annotation by clinicians, better semi-supervised learning approaches can potentially help save lives. Applications of computer vision to agriculture can increase crop yields, which may help to improve the availability of food.
- However, we also recognize that our approach could become a component of harmful surveillance systems.
- Moreover, there is an entire industry built around human labeling services, and technology that reduces the need for these services could lead to a short-term loss of income for some of those currently employed or contracted to provide labels.