Big Self-Supervised Models are Strong Semi-Supervised Learners

Presented at MLBBQ by Minoo Jafarlou

Outline

1. Introduction
2. Proposed Method
3. Experimental Results
4. Conclusion
5. Q&A

Introduction

- **Problem Statement:** Highlight the challenge of learning from a small number of labeled examples while leveraging a large amount of unlabeled data.

- **Motivation:** Discuss the traditional paradigm of unsupervised pre training followed by supervised fine-tuning and its effectiveness in computer vision, particularly on the ImageNet dataset.

- **Key Approach:** Introduce the approach of using large (deep and wide) networks for both pretraining and fine-tuning to improve semi-supervised learning efficiency.

Proposed Method

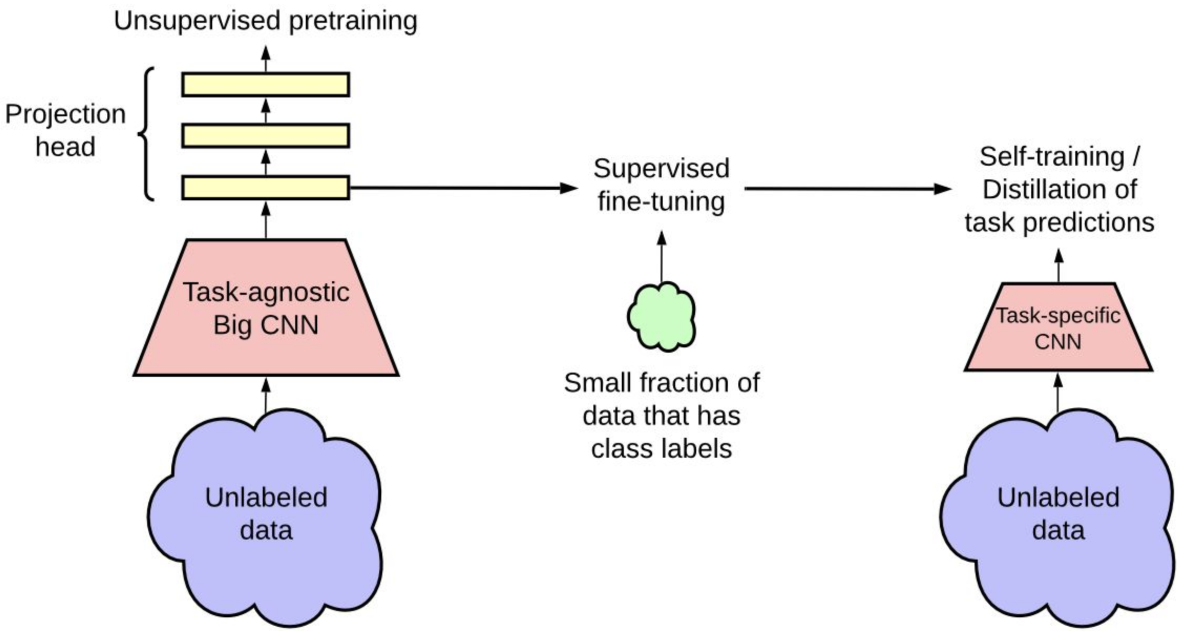


Figure 3: The proposed semi-supervised learning framework leverages unlabeled data in two ways: (1) task-agnostic use in unsupervised pretraining, and (2) task-specific use in self-training / distillation.

Contrastive Loss

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

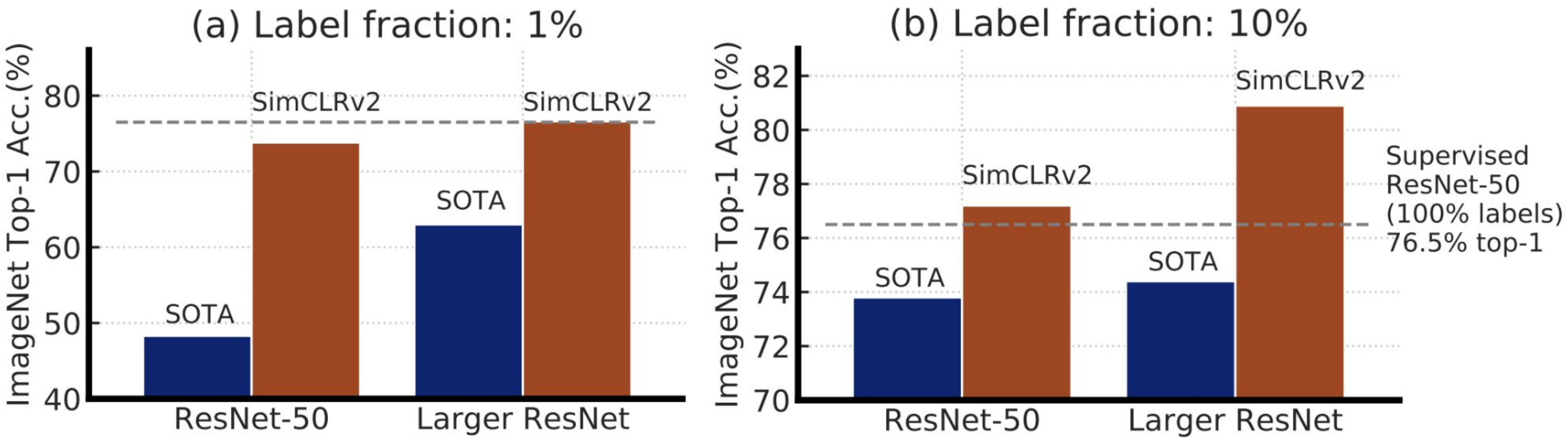
Where $\mathrm{sim}(\cdot, \cdot)$ is cosine similarity between two vectors, and τ is a temperature scalar.

Knowledge distillation

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

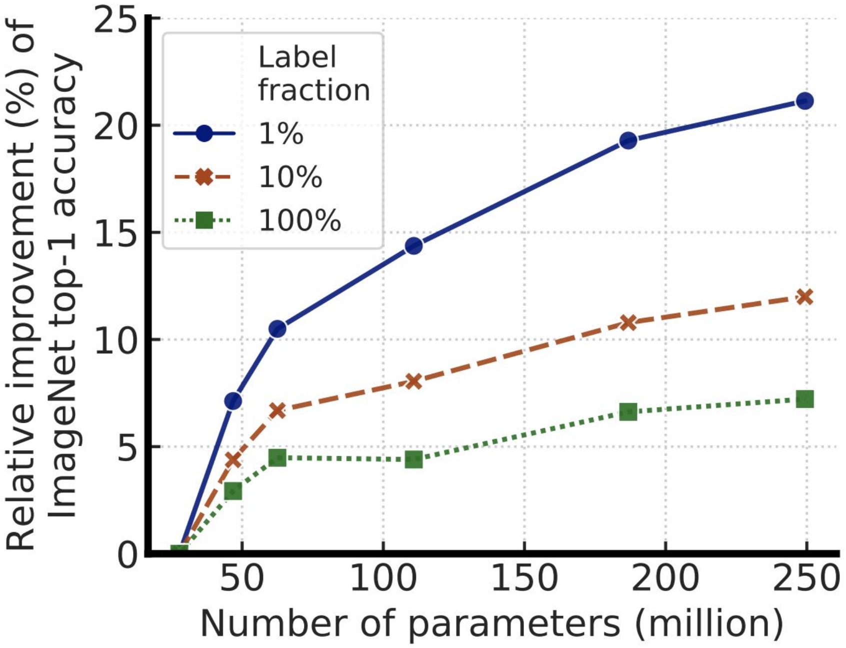
Experiment Results

Top-1 accuracy of previous state-of-the-art (SOTA) methods [1, 2] and our method (SimCLRv2) on ImageNet using only 1% or 10% of the labels. Dashed line denotes fully supervised ResNet-50 trained with 100% of labels.



Experiment Results

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

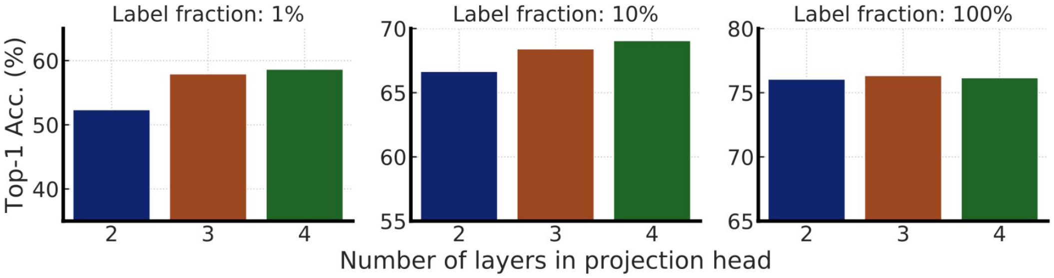


Experiment Results

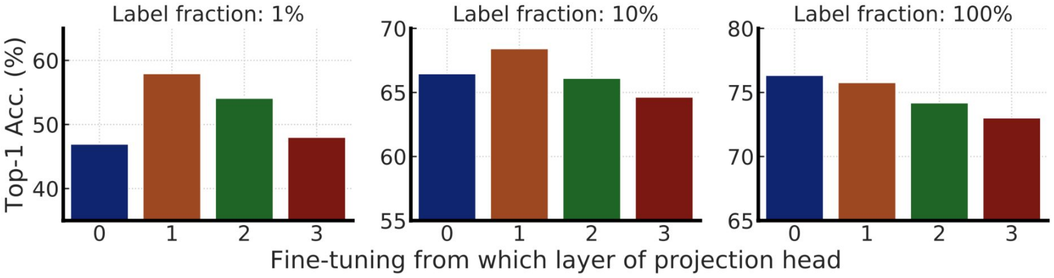
Depth	Width	Use SK [28]	Param (M)	Fine-tuned on 1%	Fine-tuned on 10%	Fine-tuned on 100%	Linear eval	Supervised
50	1×	False	**24**	**57.9**	**68.4**	**76.3**	**71.7**	**76.6**
		True	35	64.5	72.1	78.7	74.6	78.5
	2×	False	94	66.3	73.9	79.1	75.6	77.8
		True	140	70.6	77.0	81.3	77.7	79.3
101	1×	False	43	62.1	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**

Experiment Results

Top-1accuracyviafine-tuningunderdifferentprojectionheadsettingsandlabelfractions (using ResNet-50).



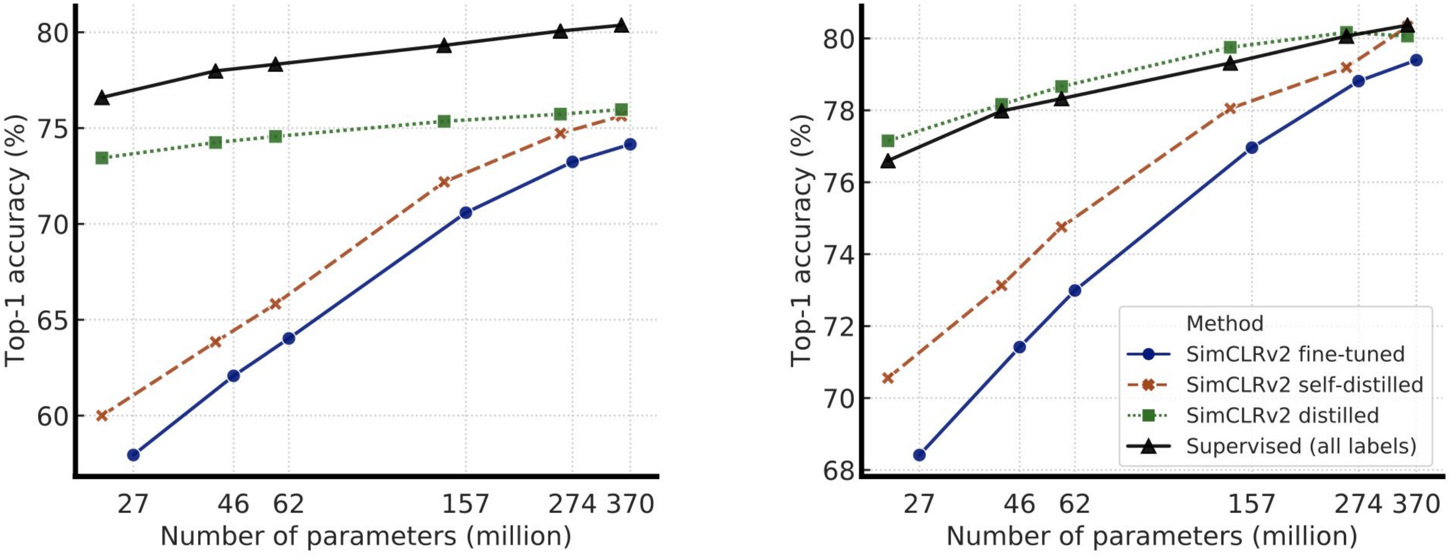
(a) Effect of projection head's depth when fine-tuning from optimal middle layer.



(b) Effect of fine-tuning from middle of a 3-layer projection head (0 is SimCLR).

Experiment Results

Method	Label fraction 1%	Label fraction 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



(a) Label fraction 1% (b) Label fraction 10%

Conclusion

- method for semi-supervised ImageNet classification
- larger models can significantly improve performance even with fewer labeled examples
- task-agnostic representations learned by large models can be distilled into smaller, task-specific networks using unlabeled data

Thank You!