Big Self-Supervised Models are Strong Semi-Supervised Learners (SimCLR v2)

SimCLR v1 review

- 각 이미지에 서로 다른 augmentation 적용

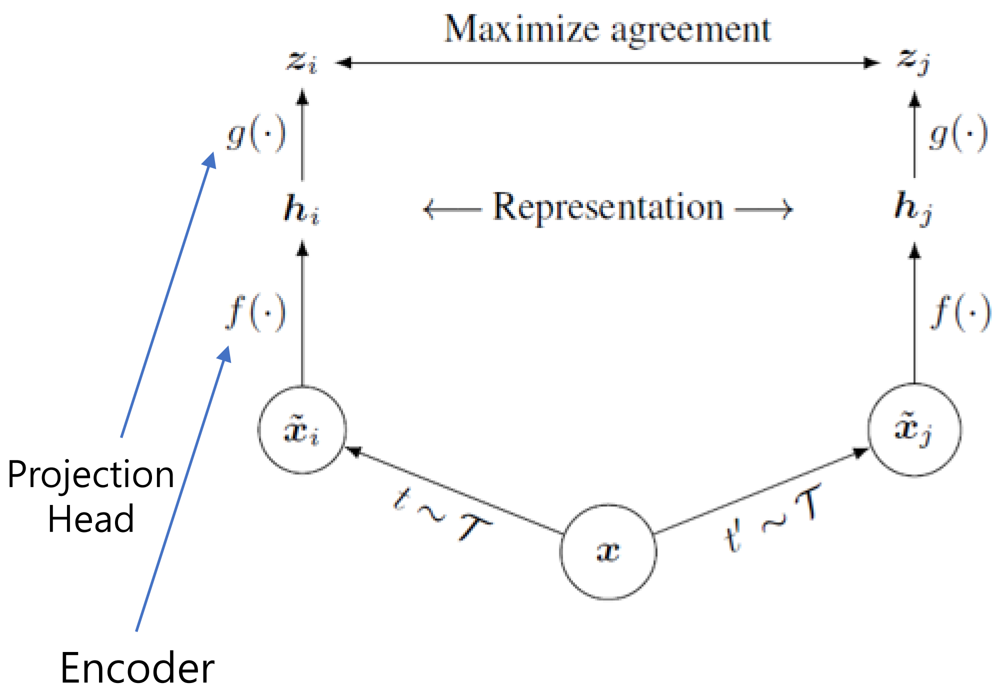
- Encoder : ResNet
- Projection Head : MLP - ReLU - MLP

- Batch size : N -> 2N 개의 이미지

  positive pair 1쌍, negative pair (N-1)쌍

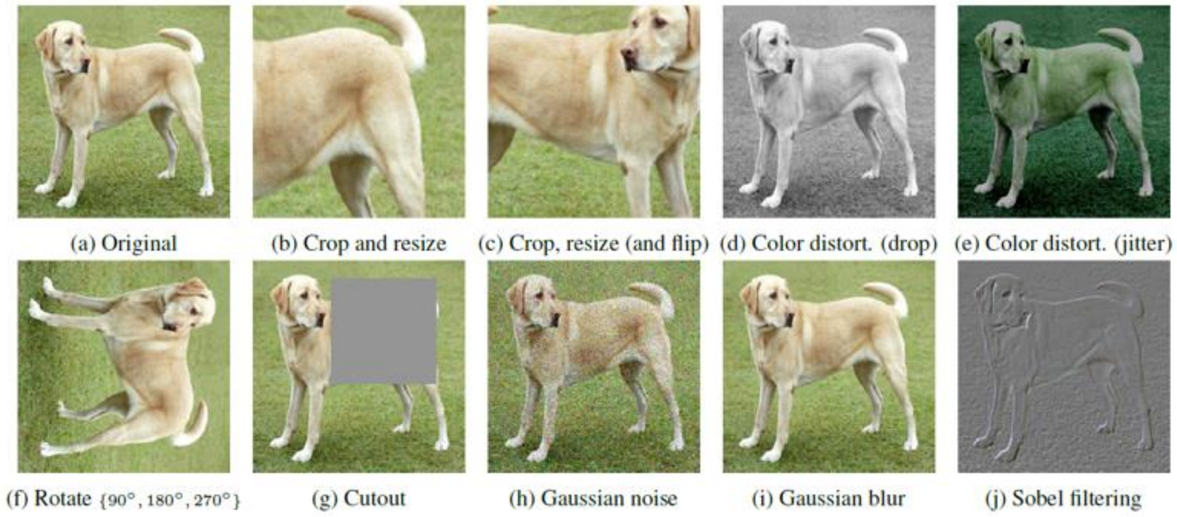
  -> NT-Xent loss로 학습

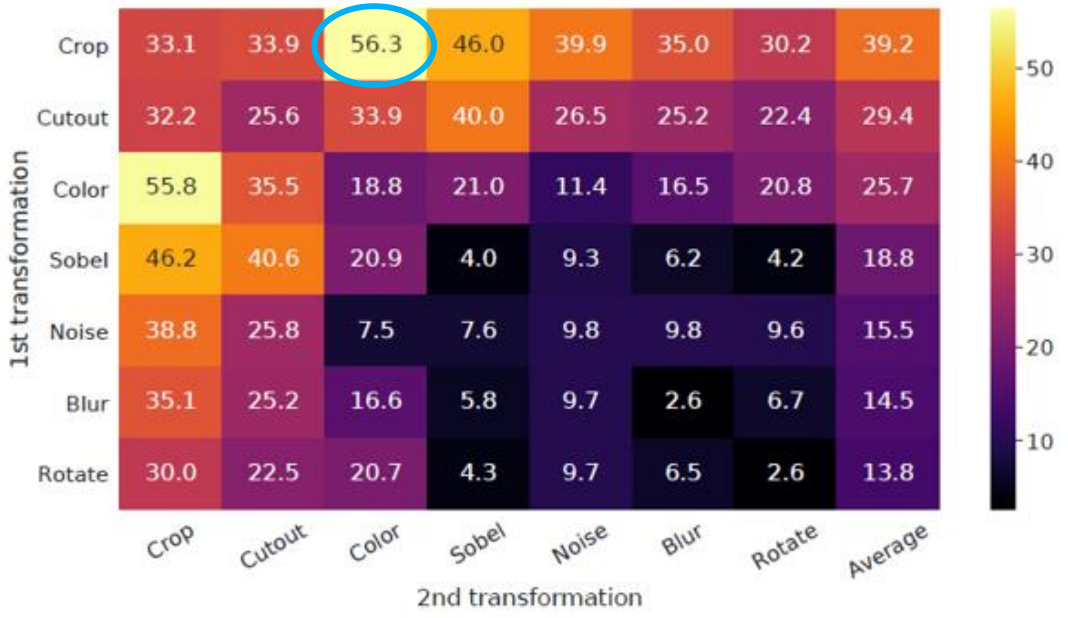
$$\ell_{i,j}^{\text{NT-Xent}} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)} ,$$



SimCLR v1 review

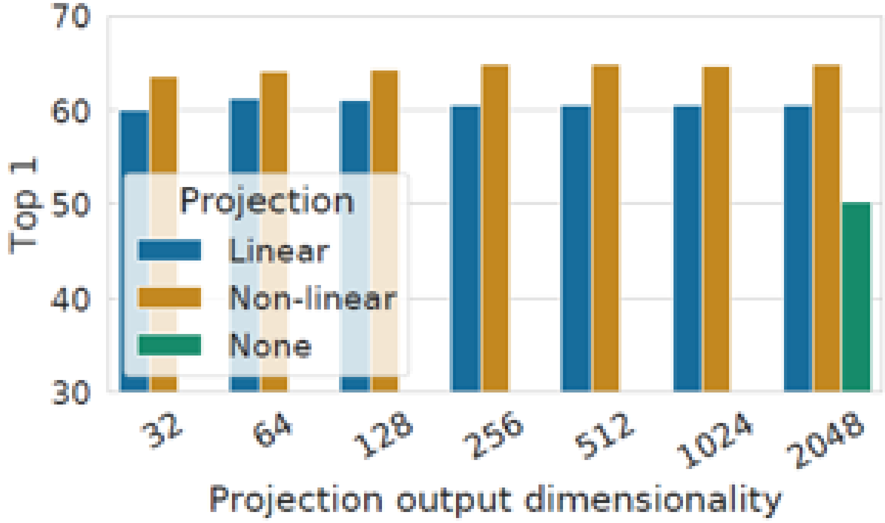
- 널리 쓰이는 crop, resize 뿐만 아니라 color distortion 등을 적용하여 성능 향상

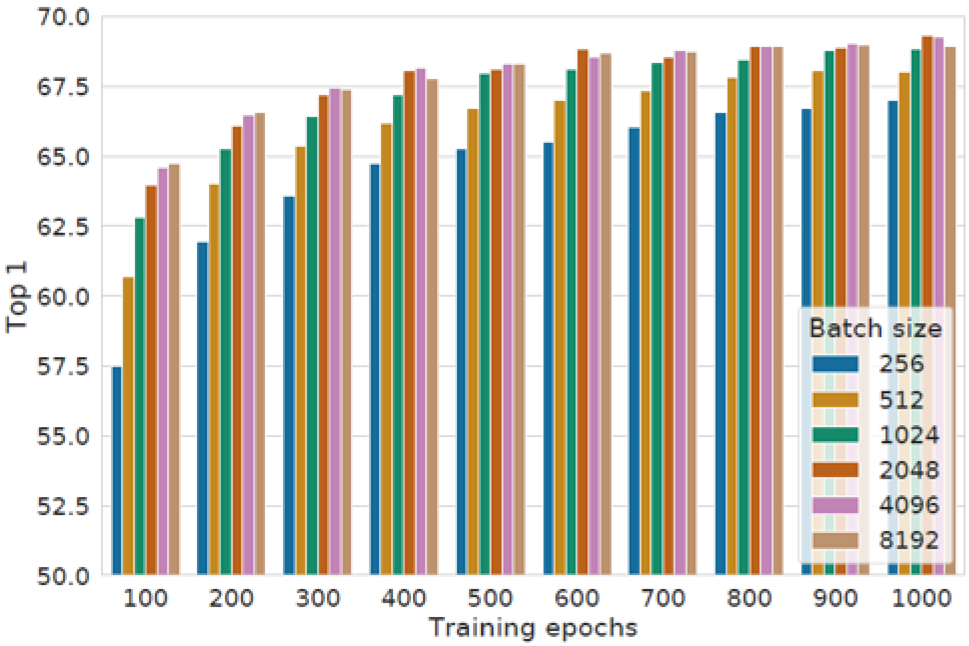




SimCLR v1 review

- Projection Head는 사용하지 않는 것보다 사용하는 것이, Linear 구조보다는 Non-linear 구조가 유리
- Batch size는 2048 이상으로 크게하는 것이 유리

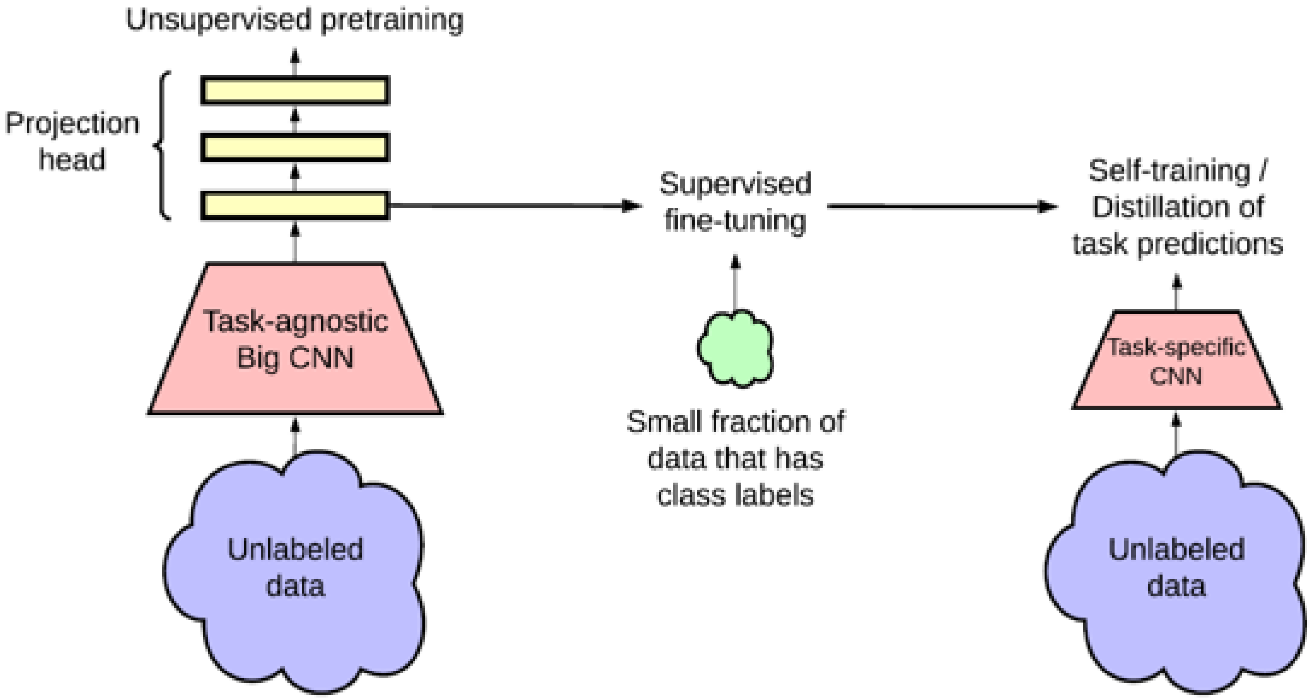




SimCLR v2

구조 및 개념

- few-labeled 데이터셋을 활용하여 fine tune
- distillation을 통한 semi-supervised learning에서도 SOTA 달성



SimCLR v2

Self-supervised pretraining with SimCLRv2

1. 기존 ResNet-50 (4x) 모델에서 ResNet-152 (3x) 모델로 모델 크기 증대 + Selective Kernel 적용
   - 모델 파라미터 수는 약 2배 증가
   - Label이 일부 존재할 때 더욱 효과적인 성능 향상을 얻을 수 있음

Method	Architecture	Param (M)	Top 1	Top 5
Methods using other architectures:				
Rotation	RevNet-50 (4×)	86	55.4	-
BigBiGAN	RevNet-50 (4×)	86	61.3	81.9
AMDIM	Custom-ResNet	626	68.1	-
CMC	ResNet-50 (2×)	188	68.4	88.2
MoCo	ResNet-50 (4×)	375	68.6	-
CPC v2	ResNet-161 (∗)	305	71.5	90.1
SimCLR (ours)	ResNet-50 (2×)	94	74.2	92.0
SimCLR (ours)	ResNet-50 (4×)	375	**76.5**	**93.2**

Depth	Width	Use SK [28]	Param (M)	Fine-tuned on			Linear eval	Supervised
				1%	10%	100%		
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**

SimCLR v2

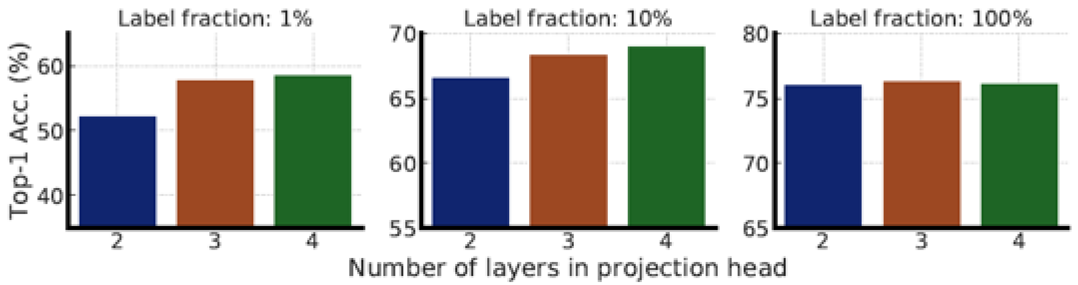
Self-supervised pretraining with SimCLRv2

2. Projection Head 고도화

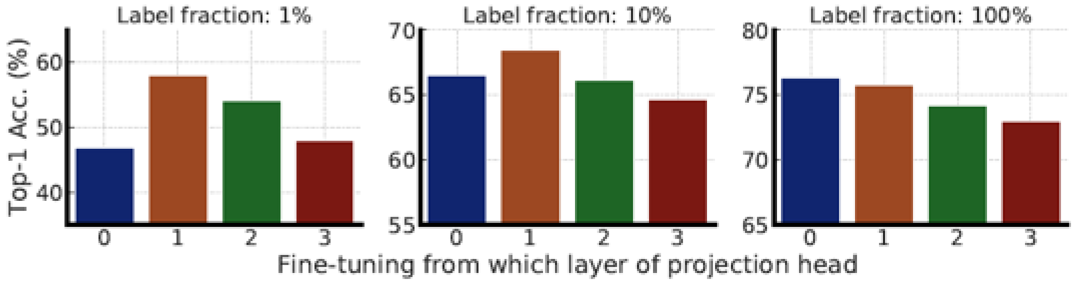
2-(a) Projection Head의 linear layer 개수 2 -> 3

2-(b) Projection Head의 middle layer를 encoder에 포함

-> 1% label sample로 fine tune 했을 때 14%의 top-1 Acc 향상



(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).

SimCLR v2

Self-supervised pretraining with SimCLRv2

3. Memory network 추가

- MoCo에서 영감 받아 negative example을 최대한 늘리기 위함
- 하지만 이미 충분히 큰 batch size(4096)로 약 1% 정도의 성능 향상만을 얻음

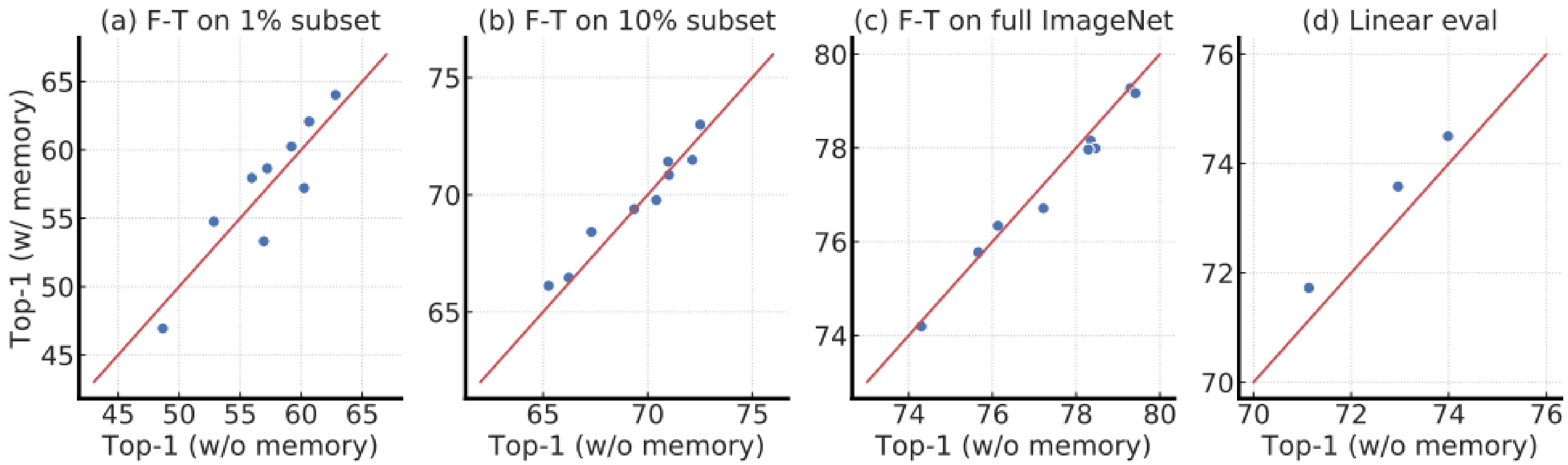


Figure D.1: Top-1 results of ResNet-50, ResNet-101, and ResNet-152 trained with or without memory.

SimCLR v2

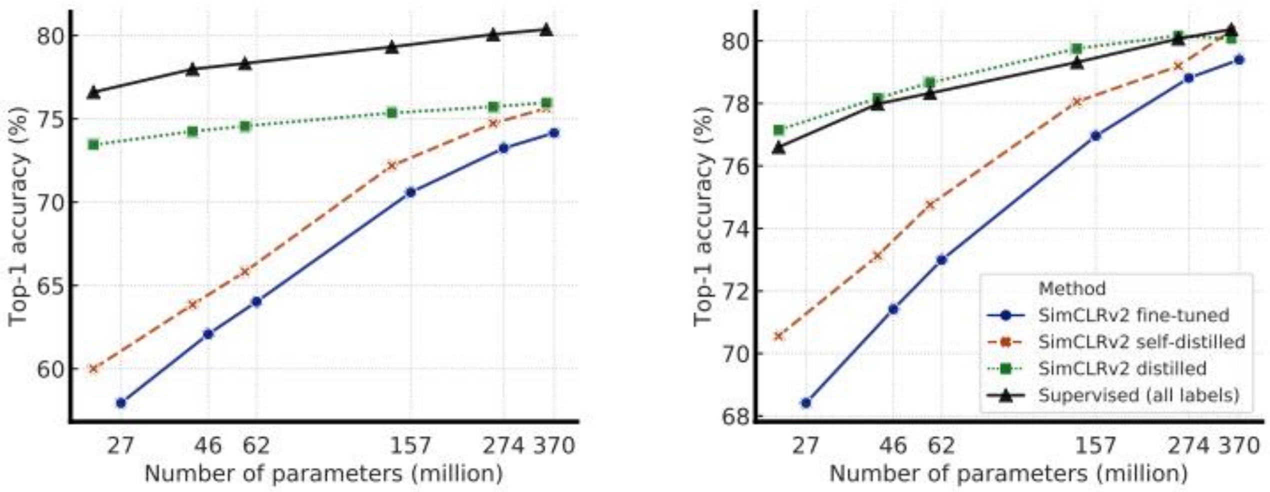
Fine-tuning

- Projection head의 middle layer를 encoder에 붙여 fine-tune 진행
- Projection head의 첫 번째 layer를 포함하는 것은 그냥 encoder에 MLP 하나 더하는 것과 다르지 않음
- 1% label : 60 epochs
- 10% label : 30 epochs

SimCLR v2

Distillation

- Teacher : SimCLRv2 fine-tuned (only Encoder)
- Student self-distilled : same as Teacher
- Student distilled : SimCLRv2 ResNet-152 (2x+SK)
- $\alpha = 0.1, \mathcal{T} = 0.1$



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

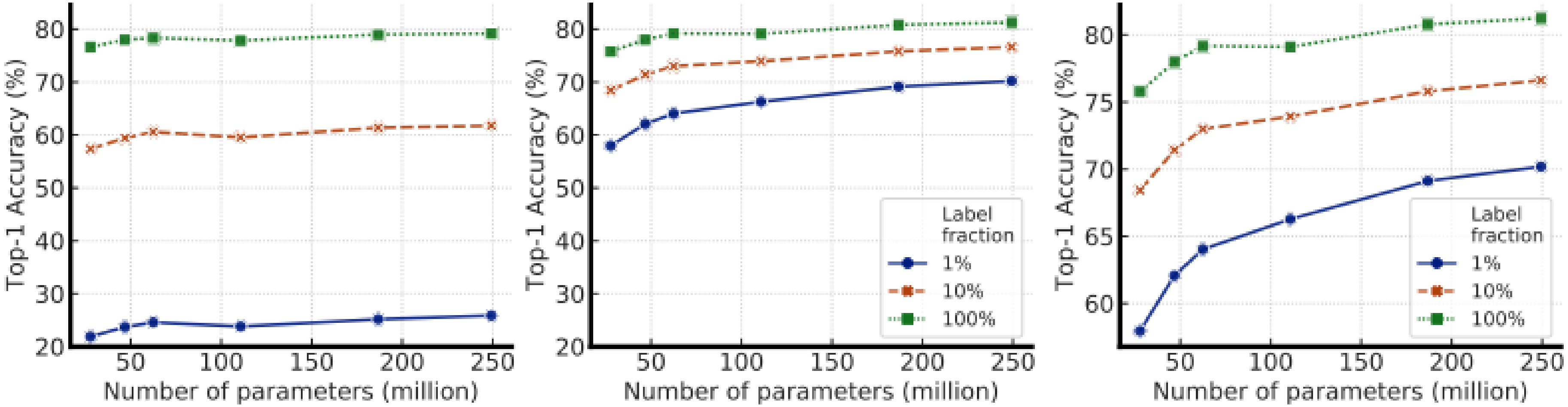
$$\mathcal{L}^{\text{distill}} = -\sum_{\boldsymbol{x}_i \in \mathcal{D}} \left[\sum_y P^T(y|\boldsymbol{x}_i;\tau) \log P^S(y|\boldsymbol{x}_i;\tau) \right]$$

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

SimCLR v2

Discussion and Result (Cont.)

- 큰 모델일수록 좋은 성능을 보이며, 그 정도는 Semi-supervised에서 더 큼
- 모델 크기가 클수록 데이터를 외워 일반화 성능이 저하될 수 있지만 본 연구에서는 그렇지 않음
- Unlabeled data를 task-agnostic하게 사용하면서 더 큰 모델이 일반적인 특징을 더 잘 배웠을 수 있다고 추측
- 더 좋은 설명을 위해 후속 연구가 필요하다



(a) Supervised (b) Semi-supervised (c) Semi-supervised (y-axis zoomed)

SimCLR v2

Discussion and Result

Table 3: ImageNet accuracy of models trained under semi-supervised settings. For our methods, we report results with distillation after fine-tuning. For our smaller models, we use self-distilled ResNet-152 (3×+SK) as the teacher.

Method	Architecture	Top-1 Label fraction		Top-5 Label fraction	
		1%	10%	1%	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
Mean teacher [39]	ResNeXt-152	-	-	-	90.9
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	-	-
CowMix [40]	ResNet-152	-	73.9	-	91.2
Methods using unlabeled data in a task-agnostic way:					
InstDisc [17]	ResNet-50	-	-	39.2	77.4
BigBiGAN [41]	RevNet-50 (4×)	-	-	55.2	78.8
PIRL [42]	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 (2×)	58.5	71.7	83.0	91.2
SimCLR [1]	ResNet-50 (4×)	63.0	74.4	85.8	92.6
BYOL [43] (concurrent work)	ResNet-50	53.2	68.8	78.4	89.0
BYOL [43] (concurrent work)	ResNet-200 (2×)	71.2	77.7	89.5	93.7
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**