A Simple Framework for Contrastive Learning of Visual Representations

Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey Hinton

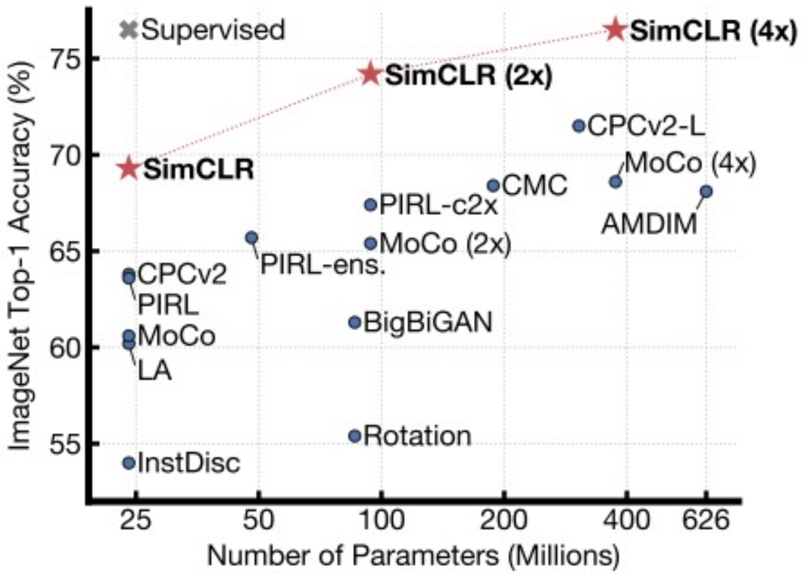
ICML 2020

https://arxiv.org/abs/2002.05709

발표자 : 박민호

Task

- Learning effective visual representations without human supervision



Contribution

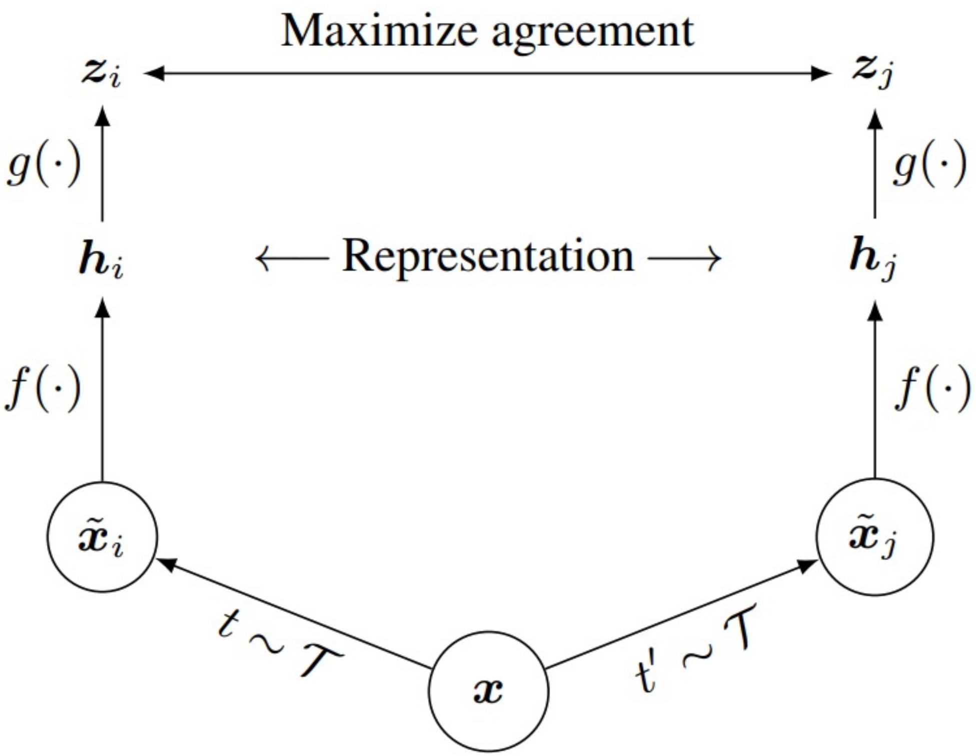
1. Composition of data augmentation plays a critical role
2. Introducing a learnable nonlinear transformation
3. Contrastive learning benefits from large batch sizes and more training steps compared to supervised learning

Method

Projection

Encoder

Augmentation



Method

Algorithm 1 SimCLR's main learning algorithm.

input: batch size N, constant τ, structure of f, g, $\mathcal{T}$.
for sampled minibatch $\{\boldsymbol{x}_k\}_{k=1}^{N}$ **do**
  for all $k \in \{1, \ldots, N\}$ **do**
    draw two augmentation functions $t \sim \mathcal{T}$, $t' \sim \mathcal{T}$
    # the first augmentation
    $\tilde{\boldsymbol{x}}_{2k-1} = t(\boldsymbol{x}_k)$
    $\boldsymbol{h}_{2k-1} = f(\tilde{\boldsymbol{x}}_{2k-1})$ # representation
    $\boldsymbol{z}_{2k-1} = g(\boldsymbol{h}_{2k-1})$ # projection
    # the second augmentation
    $\tilde{\boldsymbol{x}}_{2k} = t'(\boldsymbol{x}_k)$
    $\boldsymbol{h}_{2k} = f(\tilde{\boldsymbol{x}}_{2k})$ # representation
    $\boldsymbol{z}_{2k} = g(\boldsymbol{h}_{2k})$ # projection
  end for
  for all $i \in \{1, \ldots, 2N\}$ and $j \in \{1, \ldots, 2N\}$ **do**
    $s_{i,j} = \boldsymbol{z}_i^\top \boldsymbol{z}_j / (\|\boldsymbol{z}_i\| \|\boldsymbol{z}_j\|)$ # pairwise similarity
  end for
  define $\ell(i,j)$ **as** $\ell(i,j) = -\log \frac{\exp(s_{i,j}/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(s_{i,k}/\tau)}$
  $\mathcal{L} = \frac{1}{2N} \sum_{k=1}^{N} [\ell(2k-1, 2k) + \ell(2k, 2k-1)]$
  update networks f and g to minimize $\mathcal{L}$
end for
return encoder network $f(\cdot)$, and throw away $g(\cdot)$

Let,

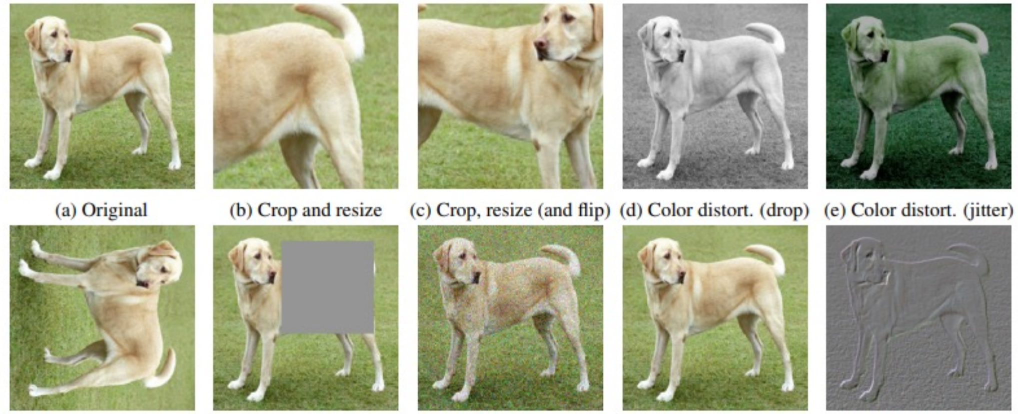
$$Z = \begin{bmatrix} \vdots & \vdots & & \vdots \\ \boldsymbol{z}_1 & \boldsymbol{z}_2 & \cdots & \boldsymbol{z}_{2k} \\ \vdots & \vdots & & \vdots \end{bmatrix}$$

we want

$$Z^T Z = \begin{bmatrix} \boldsymbol{z}_1^T \boldsymbol{z}_1 & \boldsymbol{z}_1^T \boldsymbol{z}_2 & 0 & \cdots & \cdots & 0 \\ \boldsymbol{z}_2^T \boldsymbol{z}_1 & \boldsymbol{z}_2^T \boldsymbol{z}_2 & 0 & & & 0 \\ 0 & 0 & \ddots & & & \vdots \\ \vdots & & & \ddots & 0 & 0 \\ \vdots & & & 0 & \boldsymbol{z}_{2k-1}^T \boldsymbol{z}_{2k-1} & \boldsymbol{z}_{2k-1}^T \boldsymbol{z}_{2k} \\ 0 & 0 & \cdots & 0 & \boldsymbol{z}_{2k}^T \boldsymbol{z}_{2k-1} & \boldsymbol{z}_{2k}^T \boldsymbol{z}_{2k} \end{bmatrix}$$

Data Augmentation

- No single transformation suffices to learn good representation



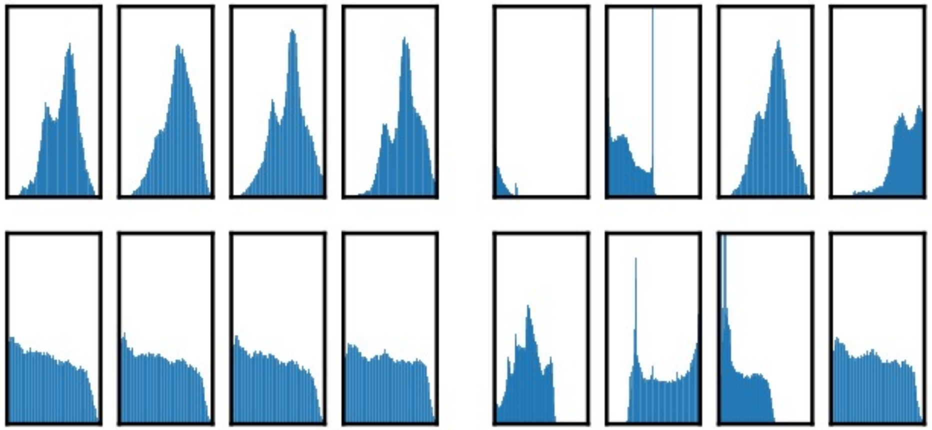
(a) Original (b) Crop and resize (c) Crop, resize (and flip) (d) Color distort. (drop) (e) Color distort. (jitter)

(f) Rotate {90°, 180°, 270°} (g) Cutout (h) Gaussian noise (i) Gaussian blur (j) Sobel filtering

Fig 5. Linear evaluation (ImageNet top-1 accuracy)

Data Augmentation

- Crop and Color distortion are good
- Conjecture: Most patches from an image share a similar color distribution. Neural nets may exploit this shortcut.



(a) Without color distortion. (b) With color distortion.

Stronger Data Augmentation

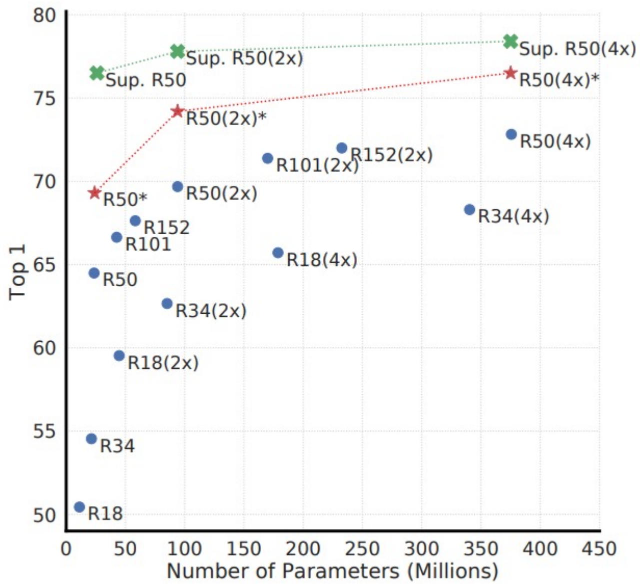
- Contrastive learning needs stronger data augmentation than supervised learning
- Data augmentation that does not yield accuracy benefits for supervised learning can still help considerably with contrastive learning

Methods	Color distortion strength 1/8	1/4	1/2	1	1 (+Blur)	AutoAug
SimCLR	59.6	61.0	62.6	63.2	64.5	61.1
Supervised	77.0	76.7	76.5	75.7	75.4	77.1

AutoAugment (Cubuk et al., 2019) : A sophisticated augmentation policy found using supervised learning.

Architecture (Parameters)

- Unsupervised contrastive learning benefits (more) from bigger models

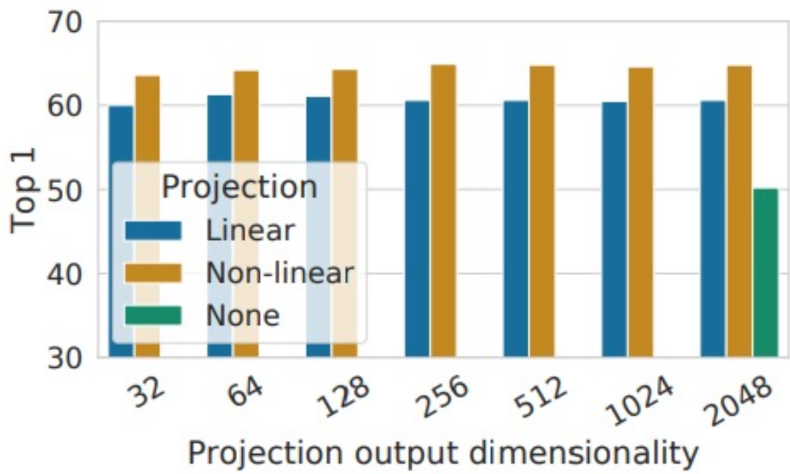


Green: Trained for 90 epochs
Red: Trained for 1000 epochs
Blue: Trained for 100 epochs

x2, x4: Width
R18, R50: Depth

Nonlinear Projection

- A nonlinear projection head improves the representation quality of the layer before it.
- Better than 3%, 10+%



Nonlinear Projection

- Loss of information induced by the contrastive loss.
- In particular, $\boldsymbol{z} = g(\boldsymbol{h})$ is trained to be invariant to data transformation. ⇒ g can remove information that may be useful for the downstream task, such as the color or orientation of objects.

What to predict?	Random guess	Representation $\boldsymbol{h}$	Representation $g(\boldsymbol{h})$
Color vs grayscale	80	99.3	97.4
Rotation	25	67.6	25.6
Orig. vs corrupted	50	99.5	59.6
Orig. vs Sobel filtered	50	96.6	56.3

Loss Function (omit)

- Ours: Normalized cross entropy loss with adjustable temperature
- Simpler, and better than NT-Xent loss, Logistic loss, and margin loss.

Batch size

- When the number of training epochs is small (e.g. 100 epochs), larger batch sizes have a significant advantage over t he smaller ones.

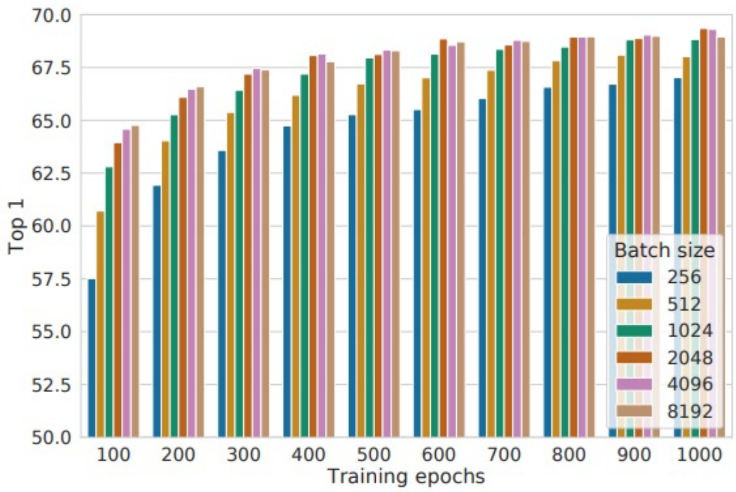


Fig 9. Linear evaluation models (ResNet-50)

Batch size

- Larger batch size provides more negative examples.
- Training longer also provides more negative examples.
- Why do more negative examples improve the result?

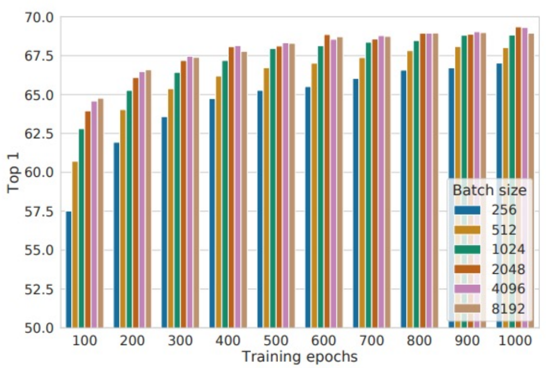


Figure 9. Linear evaluation models (ResNet-50) trained with different batch size and epochs. Each bar is a single run from scratch.[10]

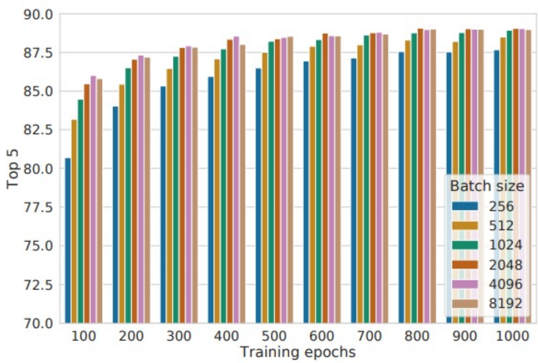


Figure B.1. Linear evaluation (top-5) of ResNet-50 trained with different batch sizes and epochs. Each bar is a single run from scratch. See Figure 9 for top-1 accuracy.

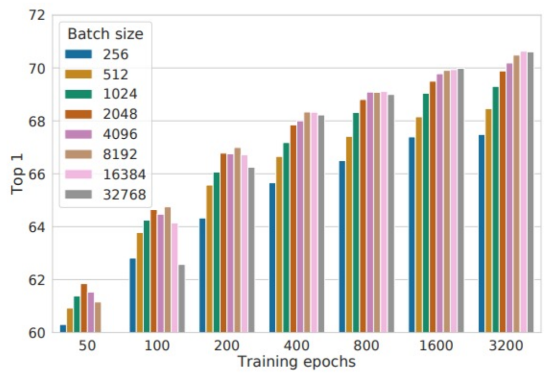


Figure B.2. Linear evaluation (top-1) of ResNet-50 trained with different batch sizes and *longer* epochs. Here a *square root* learning rate, instead of a linear one, is utilized.

Comparison with State-of-the-art

- Linear evaluation

Method	Architecture	Param (M)	Top 1	Top 5
Methods using ResNet-50:				
Local Agg.	ResNet-50	24	60.2	-
MoCo	ResNet-50	24	60.6	-
PIRL	ResNet-50	24	63.6	-
CPC v2	ResNet-50	24	63.8	85.3
SimCLR (ours)	ResNet-50	24	**69.3**	**89.0**
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**

Table 6. ImageNet accuracies of linear classifiers trained on representations learned with different self-supervised methods.

Comparison with State-of-the-art

- Semi-supervised learning

Method	Architecture	Label fraction 1%	10%
		Top 5	
Supervised baseline	ResNet-50	48.4	80.4
Methods using other label-propagation:			
Pseudo-label	ResNet-50	51.6	82.4
VAT+Entropy Min.	ResNet-50	47.0	83.4
UDA (w. RandAug)	ResNet-50	-	88.5
FixMatch (w. RandAug)	ResNet-50	-	89.1
S4L (Rot+VAT+En. M.)	ResNet-50 (4×)	-	91.2
Methods using representation learning only:			
InstDisc	ResNet-50	39.2	77.4
BigBiGAN	RevNet-50 (4×)	55.2	78.8
PIRL	ResNet-50	57.2	83.8
CPC v2	ResNet-161(∗)	77.9	91.2
SimCLR (ours)	ResNet-50	75.5	87.8
SimCLR (ours)	ResNet-50 (2×)	83.0	91.2
SimCLR (ours)	ResNet-50 (4×)	**85.8**	**92.6**

Table 7. ImageNet accuracy of models trained with few labels.

Comparison with State-of-the-art

- Semi-supervised learning

Architecture	Label fraction 1%		10%		100%	
	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5
ResNet-50	49.4	76.6	66.1	88.1	76.0	93.1
ResNet-50 (2×)	59.4	83.7	71.8	91.2	79.1	94.8
ResNet-50 (4×)	64.1	86.6	74.8	92.8	80.4	95.4

Table B.2. Classification accuracy obtained by fine-tuning the SimCLR (which is pretrained with broader data augmentations) on 1%, 10% and full of ImageNet. As a reference, our ResNet-50 (4×) trained from scratch on 100% labels achieves 78.4% top-1 / 94.2% top-5.

Comparison with State-of-the-art

- Transfer learning

	Food	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech-101	Flowers
Linear evaluation:												
SimCLR (ours)	**76.9**	**95.3**	80.2	48.4	**65.9**	60.0	61.2	**84.2**	**78.9**	89.2	**93.9**	**95.0**
Supervised	75.2	**95.7**	**81.2**	**56.4**	64.9	**68.8**	**63.8**	83.8	**78.7**	**92.3**	**94.1**	94.2
Fine-tuned:												
SimCLR (ours)	**89.4**	**98.6**	**89.0**	**78.2**	**68.1**	**92.1**	**87.0**	**86.6**	**77.8**	92.1	**94.1**	97.6
Supervised	88.7	98.3	**88.7**	**77.8**	67.0	91.4	**88.0**	86.5	**78.8**	**93.2**	**94.2**	**98.0**
Random init	88.3	96.0	81.9	**77.0**	53.7	91.3	84.8	69.4	64.1	82.7	72.5	92.5

Table 8. Comparison of transfer learning performance of our self-supervised approach with supervised baselines across 12 natural image classification datasets, for ResNet-50 $(4\times)$ models pretrained on ImageNet. Results not significantly worse than the best ($p > 0.05$, permutation test) are shown in bold. See Appendix B.8 for experimental details and results with standard ResNet-50.

Thank you