A Simple Framework for Contrastive Letive Learning of Visual Representations

Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey Hinton

1 Contributions

In this work, the authors introduced a contrastive learning framework, that learns the representations of different augmented views of images. The agreement is maximized when the representations come from same image. Major components of the framework are summarized as follows:

- Composition of data augmentation plays an important role in defining the contrastive predicting tasks that yield effective representations. Stronger data augmentation benefits more in the case of unsupervised learning than in supervised learning.
- A non-linear transformation between the representation and the contrastive loss improves the quality of representations.
- Representation learning with Contrastive CE loss benefits from normalized embeddings.
- Contrastive Learning benefits from larger batch sizes and longer training compared to its supervised counterpart. Like SL, deeper and wider networks benefit SSL too.

2 Method

Stage 1: Data Augmentation The first stage, transforms any data to two correlated views, $\tilde{x_i}$ and $\tilde{x_j}$. Sequentially, 2 data augmentations have been applied, random cropping, random color distortions, and random Gaussian Blur.
Stage 2: Encoder n/w ResNet as base encoder, $h_i = f(\tilde{x_i}) = ResNet(\tilde{x_i})$
Stage3: Projection head Maps representation to space where contrastive loss is applied. They used an MLP with 1 hidden layer, $z_i = g(h_i)$. Then, the contrastive loss function between the positive pair is defined as

$$l_{i,j} = -log\frac{exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} 1_{[k\neq i]]} exp(sim(z_i, z_k)/\tau)} \tag{1}$$

with sim(...) is cosine similarity between the embeddings of two views. The authors used some modifications at the time of training, (1) Training with larger batch size, varies from

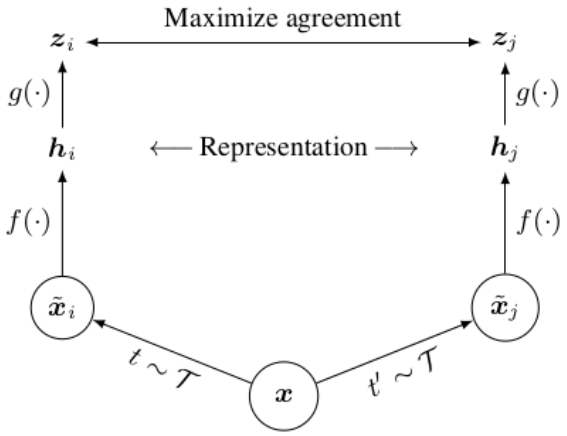

Figure 2. A simple framework for contrastive learning of visual representations. Two separate data augmentation operators are sampled from the same family of augmentations ($t \sim \mathcal{T}$ and $t' \sim \mathcal{T}$) and applied to each data example to obtain two correlated views. A base encoder network $f(\cdot)$ and a projection head $g(\cdot)$ are trained to maximize agreement using a contrastive loss. After training is completed, we throw away the projection head $g(\cdot)$ and use encoder $f(\cdot)$ and representation $\boldsymbol{h}$ for downstream tasks.

256 to 8192. LARS optimizer for stabilization, and, (2) Global Batch Norm to exploit local information leakage.

3 Results & Discussion

Based on the ablation studies, some points that are discussed in the paper:

- Data Augmentation defines predictive tasks. No single transformation is sufficient. Strong er color aug is required.
- Unsupervised Learning benefits more from bigger models than its Supervised Learning counterpart.
- Normalized cross-entropy loss with adjustable temperature works better than other contrastive loss functions.

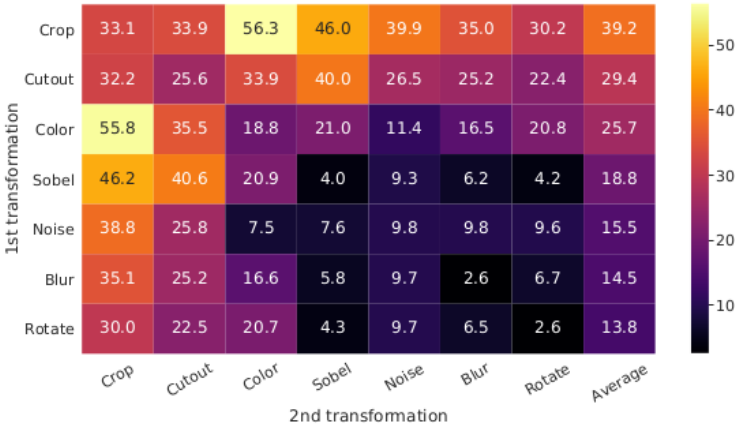

Figure 5. Linear evaluation (ImageNet top-1 accuracy) under individual or composition of data augmentations, applied only to one branch. For all columns but the last, diagonal entries correspond to single transformation, and off-diagonals correspond to composition of two transformations (applied sequentially). The last column reflects the average over the row.

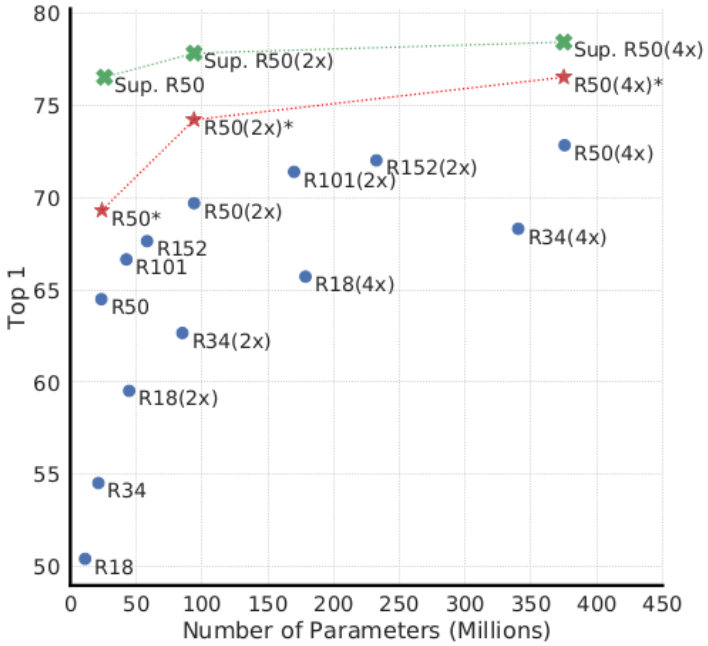

Figure 7. Linear evaluation of models with varied depth and width. Models in blue dots are ours trained for 100 epochs, models in red stars are ours trained for 1000 epochs, and models in green crosses are supervised ResNets trained for 90 epochs[7] (He et al., 2016).

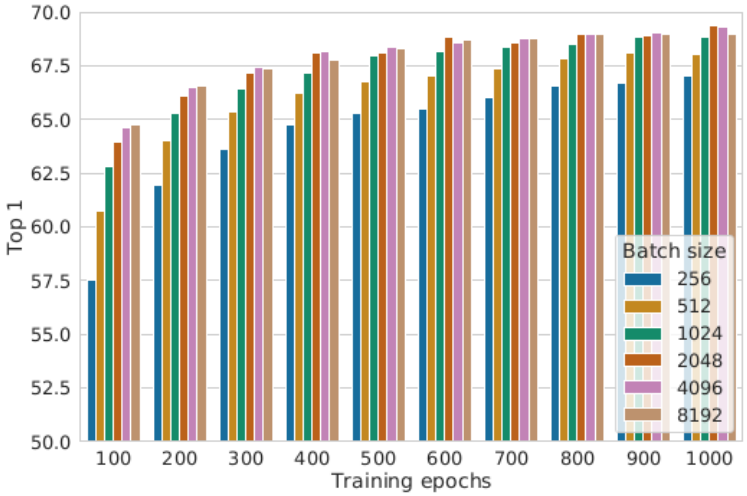

Figure 9. Linear evaluation models (ResNet-50) trained with different batch size and epochs. Each bar is a single run from scratch.[10]

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 ($4\times$)	86	55.4	-
BigBiGAN	RevNet-50 ($4\times$)	86	61.3	81.9
AMDIM	Custom-ResNet	626	68.1	-
CMC	ResNet-50 ($2\times$)	188	68.4	88.2
MoCo	ResNet-50 ($4\times$)	375	68.6	-
CPC v2	ResNet-161 ($*$)	305	71.5	90.1
SimCLR (ours)	ResNet-50 ($2\times$)	94	74.2	92.0
SimCLR (ours)	ResNet-50 ($4\times$)	375	**76.5**	**93.2**

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