Understanding Collapse in Non-Contrastive Siamese Representation Learning

Alexander C. Li, Alexei A. Efros, and Deepak Pathak
ECCV 2022

September 8, 2022

Reporter: Zhang Yajie

Outline

Outline

Self-Supervised Learning

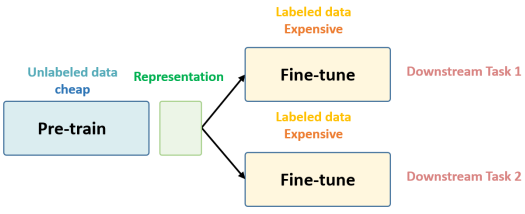


Figure 1: Self-Supervised Learning (SSL)

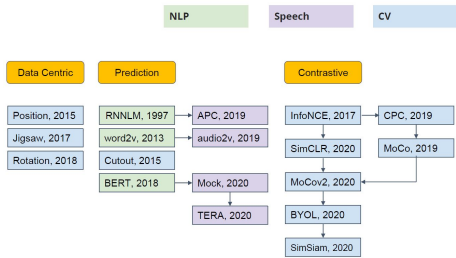


Figure 2: History of SSL in Different fields

Self-Supervised Learning

Contemporary self-supervised learning methods can roughly be broken down into two classes of methods:

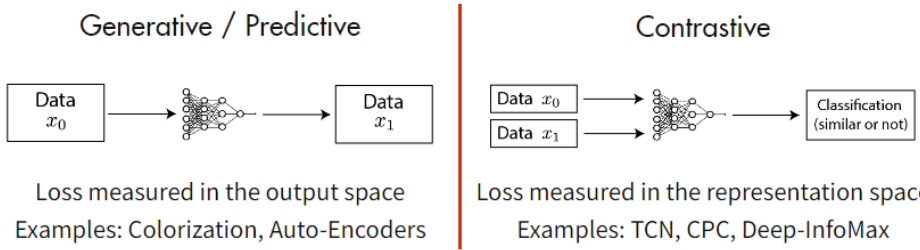


Figure 3: Generative vs Contrastive Methods

Contrastive SSL

- Pull together positive pairs (from same images)
- Push apart negative pairs (from different images)

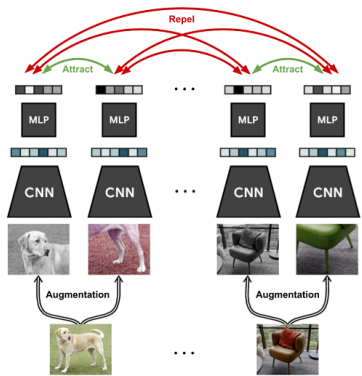


Figure 4: A Simple Framework for Contrastive Learning of Visual Representations(SimCLR)

Non-Contrastive SSL

BYOL, SimSiam

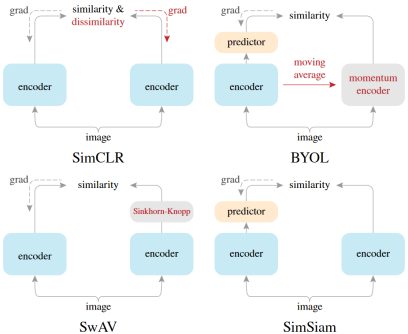


Figure 5: Comparison on Siamese Architectures

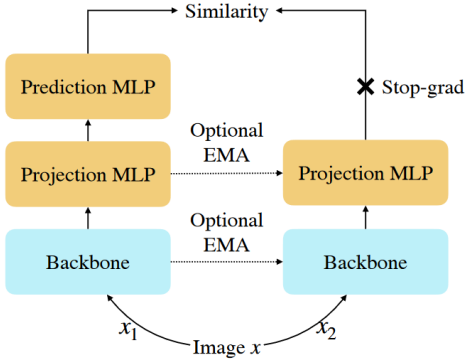


Figure 6: Non-contrastive Siamese Architecture

What causes collapse?

An undesired trivial solution to Siamese networks is all outputs "collapsing" to a constant.

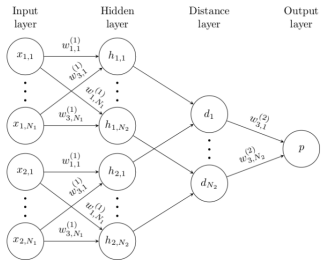


Figure 7: A simple Siamese Network

Q: Why these methods used Siamese network do not collapse?

What prevents collapse?

There have been several general strategies for preventing Siamese networks from collapsing.

- Contrastive learning: SimCLR
- Clustering: SwAV
- Momentum encoder: BYOL

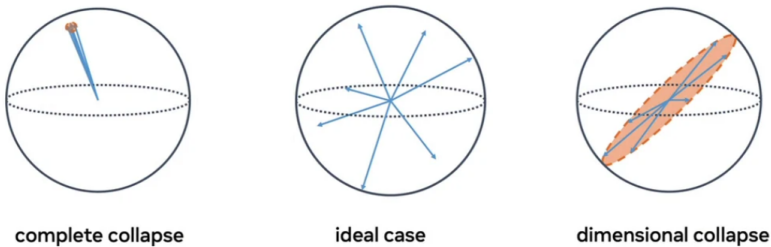


Figure 8: Collapse type

Related Work

Understanding Self-Supervised Learning Dynamics without Contrastive Pairs (ICLR 2021)

- Analyze the empirical effects of multiple hyperparameters of non-contrastive SSL (BYOL), including
  (1) Exponential Moving Average (EMA) or momentum encoder,
  (2) Higher relative learning rate α_p of the predictor,
  (3) Weight decay η

Understanding dimensional collapse in contrastive self-supervised learning (ICLR 2022)

- Contrastive SSL suffers from dimensional collapse.
- Case1: dimensional collapse caused by strong augmentation.
  Case2: implicit regularization driving models toward low-rank solutions.

Outline

Relative Underparameterization Causes Collapse

Unexpectedly, SimSiam performance drops significantly when using ResNet18 instead of ResNet50, while BYOL, which uses the same loss function, still have enough capacity to fit the SimSiam objective.

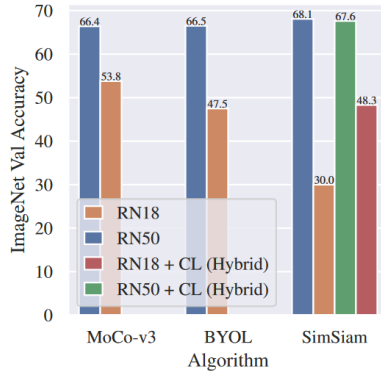


Figure 9: Linear Probe Evaluation

Performance Impact of Model Size Relative to Dataset Size

Hypothesis: The ratio of model capacity relative to the dataset complexity determines the SimSiam performance.

Experiment: Train ResNet18 and ResNet34 SimSiam models for the same number of gradient steps but on different amounts of data from ImageNet1k, ranging in $\{1\%, 5\%, 10\%, 20\%, 50\%, 100\%\}$

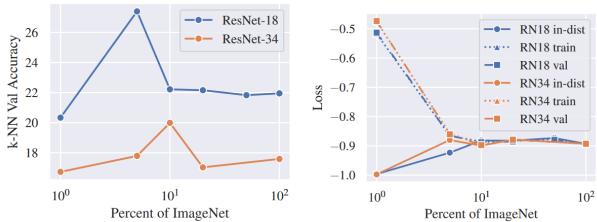


Figure 10: SimSiam performance as a function of dataset size.

Performance Impact of Model Architecture

Hypothesis: Increased depth may make it easier for SimSiam to lose information at every layer and compute collapsed representations, since the size of the vector passed between layers is limited.

Bottleneck residual block : 1×1 conv to decrease the number of channels, then 3×3 conv, then 1×1 to increase the number of channels.

Table 1: Network width matters more than depth or number of parameters.

Block type	Layers	Width Multiplier	Repr. dim.	Params	Lin. Acc.
Basic	18	1x	512	11.7M	30.0%
Basic	34	1x	512	21.8M	16.8%
Bottleneck	50	1x	2048	25.6M	68.1%
Bottleneck	26	1x	2048	16.0M	61.7%
Bottleneck	26	2x	2048	39.6M	62.6%
Basic	50	1x	512	31.9M	17.5%

Performance Drops due to Partial Dimensional Collapse

Partial dimensional collapse: Some parts of the representations either are constant across the dataset or covary with other parts of the representation.

Degree of collapse: Perform PCA on the resulting $N \times d$ representation matrix to obtain d singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$, then we get

$$(\text{Cumulative explained variance})_j = \frac{\sum_{i=1}^{j} \sigma_i}{\sum_{k=1}^{d} \sigma_k} \tag{1}$$

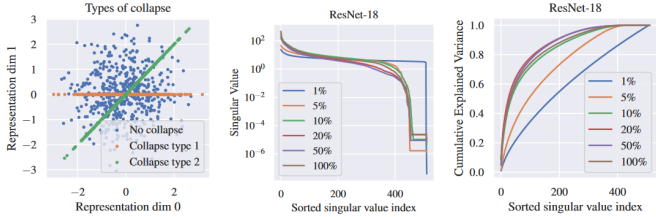


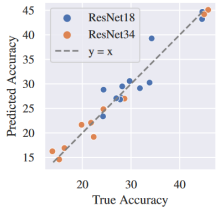
Figure 11: Partial dimensional collapse for large subsets

Predicting Performance from Collapse Metric and SimSiam Loss

Collapse metric: the area under the cumulative explained variance of the singular values:

$$AUC = \frac{\frac{1}{d}\sum_{i=1}^{d}\sum_{j=1}^{i}\sigma_j}{\sum_{j=1}^{i}\sigma_j} \in (0.5, 1) \tag{2}$$

A linear model to predict the validation accuracy from the loss and AUC:



	AUC Only	Loss Only	Use Both
R^2	0.21	0.06	0.95
Pearson's r	0.46	0.24	0.98
Spearman's ρ	0.48	0.09	0.97
AUC coeff.	-34.7	-	-79.5
Train loss coeff.	-	-14.9	-106.0

Figure 12: Accuracy is predictable from loss and collapse.

Outline

Continual Leaning prevents collapse

Figure 13: Illustration of each data ordering method.

Results

Continual training (“single pass”) improves validation accuracy by 14.5% (ResNet18) and 32.5% (ResNet34) over the “multiple pass” baseline, but leads to a 12.2% drop for ResNet50.

Table 2: ImageNet top-1 linear probing validation accuracy for different SimSiam training methods. We show the mean and standard deviation over 3 random seeds.

Training method	ResNet-18	ResNet-34	ResNet-50
Mutiple pass	30.0 ± 1.8	16.8 ± 3.2	**68.1**
Cumulative	33.0 ± 1.9	22.2 ± 2.3	67.7
Single pass	44.5 ± 0.8	45.0 ± 1.1	55.9
Hybrid (switch at 40)	**48.3** ± 0.7	**50.3** ± 0.6	67.6

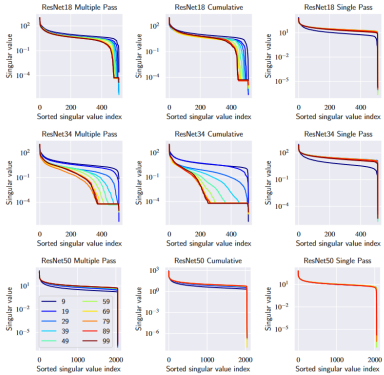
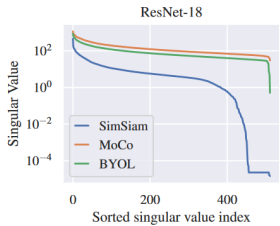


Figure 14: Evolution of dimensional collapse across training.

Results

However, continual training does not help with other SSL algorithms, such as MoCo-v3 or BYOL.



Training method	MoCo-v3	BYOL
Multiple pass	**53.8**	**47.6**
Single pass	48.9	44.2

Figure 15: MoCo and BYOL do not collapse in small models.

Exponential moving average (EMA) in BYOL prevents collapse, but it's expensive. Continual methods are faster to train, especially with limited resources.

Hybrid of Continual and Multi-epoch Training

New method: continual training first, then switching to multi-epoch training for the rest of training.

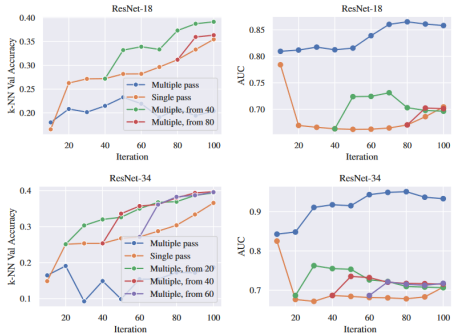


Figure 16: Hybrid of Continual and Multi-epoch Training Improves Performance

Outline

Summary

- SimSiam performance drops significantly when the model capacity is too small relative to the data complexity.
- Define a rank-based metric to measure the degree of partial dimensional collapse.
- Linear regression, using our collapse metric and the SimSiam loss, can accurately predict the linear probing accuracy.
- Model width is more important for downstream performance than depth.
- Switching to a continual learning setting eliminates collapse and restores SimSiam accuracy.