Self-distillation Family

Lifan Lin

Department of Statistics and Data Science, SUSTech

July 13, 2023

Overview

Bootstrap Your Own Latent (BYOL)

Objective

Learn representation without labels.

Contrastive Learning

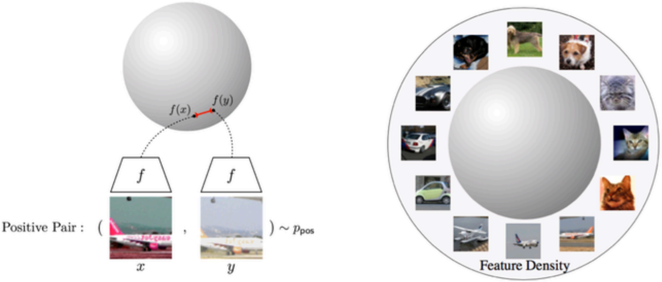
Learn a representation invariant of augmentation.

Need to Avoid Collapse

- Negative pairs
- BYOL's structure

Bootstrap Your Own Latent (BYOL)

Alignment & Uniformity



Alignment：相似实例有相近的特征

Uniformity：保留尽可能多的信息

From:Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere

Illustrations

Bootstrap Your Own Latent (BYOL)

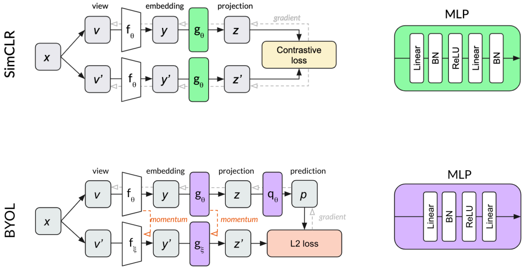
BYOL somehow prevent Collapse (and with small batch)

How BYOL prevent Collapse?

- Asymmetry (mdl&sg)
- Predictor in Student(online)
- Momentum Teacher(Target)

Implicit Negative Comparison?

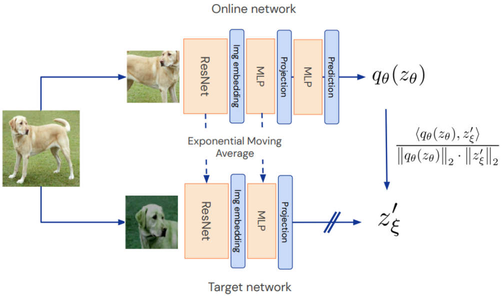
- Accumulated Negative Pairs
- Addition of Predictor



Structure of SimCLR&BYOL

Question in BYOL

- Projector
  Works better. From SimCLR.
- EMA Exponential Moving Average
  - update slowly
  - Learn from student
  - Chase Between Fast&Slow
- Predictor
  - Flexibility
  - **Learn expectation**



BYOL Arch. Summary

Why BYOL

- Which component is essential to prevent collapse (trivial sol.)?
- Semi-supervised Learning: Momentum encoder works without Predictor.
  Mean Teacher (2018)
- Stable target matters.
  Update instantly, collapse(BYOL).
- Without Target (Teacher): Predictor is important.
  A model without target network can work with a near-optimal predictor (BYOL).
  Predictor parameters should have a larger lr and carefully schedule.

Remove Momentum Teacher - SimSIAM

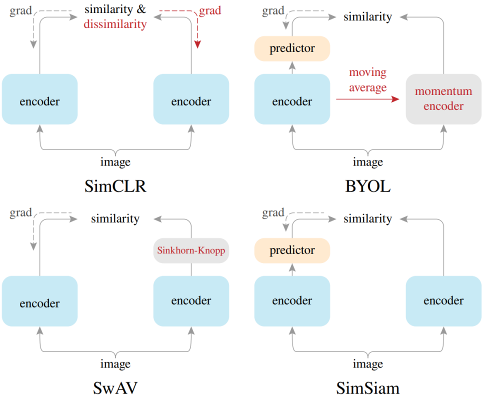
Where are the implicit Negative pairs:

- Large Batch size
- Momentum Encoder?
- Batch Normalization?
- Cosine Loss?
- Asymmetric structure

Simpler model: Shared weights

Prevent collapse without the use of:

- Negative Pair
- Large batch size
- **Momentum encoder**

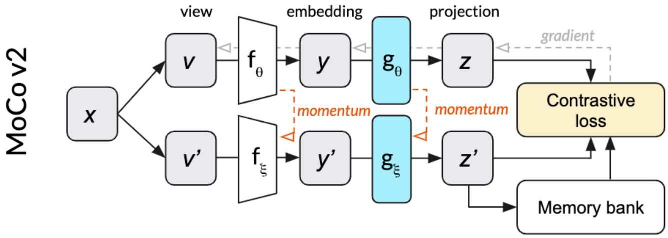


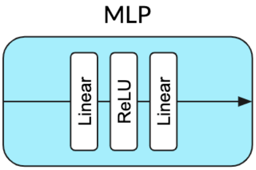
SimSIAM (SIAM means siamese.

SimSIAM - Experience

Remove Components in Model, see whether it collapses. (ablations)

Momentum encoder





May also act as "Memory bank".

SimSIAM - Experience

Remove Components in Model, see whether it collapses. (ablations)

Momentum encoder

Spacial case $\tau = 0$ in:

$$\xi \leftarrow \tau\xi + (1-\tau)\theta$$

Conclusion: Effective, but not necessary for preventing collapse.

method	batch size	negative pairs	momentum encoder	100 ep	200 ep	400 ep	800 ep
SimCLR (repro.+)	4096	✓		66.5	68.3	69.8	70.4
MoCo v2 (repro.+)	**256**	✓	✓	67.4	69.9	71.0	72.2
BYOL (repro.)	4096		✓	66.5	**70.6**	**73.2**	**74.3**
SwAV (repro.+)	4096			66.5	69.1	70.7	71.8
SimSiam	**256**			**68.1**	70.0	70.8	71.3

Table 4. **Comparisons on ImageNet linear classification**. All are based on **ResNet-50** pre-trained with **two 224×224 views**. Evaluation is on a single crop. All competitors are from our reproduction, and "+" denotes *improved* reproduction *vs*. original papers (see supplement).

Exponential Moving Average is still an effective update strategy.

SimSIAM - Ablation Experience

Remove Components in Model, see whether it collapses.

Batch Normalization

$\Rightarrow$ A mapping to hyper-sphere.

	case	proj. MLP's BN hidden	proj. MLP's BN output	pred. MLP's BN hidden	pred. MLP's BN output	acc. (%)
(a)	none	-	-	-	-	34.6
(b)	hidden-only	✓	-	✓	-	67.4
(c)	default	✓	✓	✓	-	68.1
(d)	all	✓	✓	✓	✓	unstable

Table 3. **Effect of batch normalization on MLP heads** (ImageNet linear evaluation accuracy with 100-epoch pre-training).

Exponential Moving Average is still an effective update strategy.

SimSIAM - Ablation Experience

Neither Loss function nor Batch size

method	batch size	negative pairs	momentum encoder	100 ep	200 ep	400 ep	800 ep
SimCLR (repro.+)	4096	✓		66.5	68.3	69.8	70.4
MoCo v2 (repro.+)	**256**	✓	✓	67.4	69.9	71.0	72.2
BYOL (repro.)	4096		✓	66.5	**70.6**	**73.2**	**74.3**
SwAV (repro.+)	4096			66.5	69.1	70.7	71.8
SimSiam	**256**			**68.1**	70.0	70.8	71.3

Table 4. **Comparisons on ImageNet linear classification**. All are based on **ResNet-50** pre-trained with **two 224×224 views**. Evaluation is on a single crop. All competitors are from our reproduction, and "+" denotes *improved* reproduction *vs*. original papers (see supplement).

Exponential Moving Average is still an effective update strategy.

SimSIAM - Ablation Experience

Indispensable Predictor

- Maintain asymmetric
  Without predictor h the stop gradient is equivalent to removing stop gradient and scaled by $\frac{1}{2}$.
  The gradient of $\frac{1}{2}D(z_1, \text{stopgrad}(z_2) + \frac{1}{2}D(z_2, \text{stopgrad}(z_1)$ has the same direction as th gradient of $D(z_1, z_2)$.

Maintain asymmetry.

SimSIAM - Ablation Experience

Indispensable Predictor

- Relate to representation

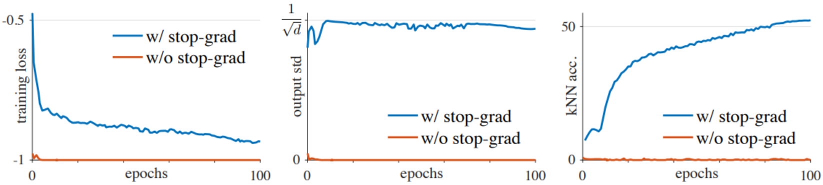
	pred. MLP h	acc. (%)
baseline	lr with cosine decay	67.7
(a)	no pred. MLP	0.1
(b)	fixed random init.	1.5
(c)	lr not decayed	68.1

Table 1. **Effect of prediction MLP** (ImageNet linear evaluation accuracy with 100-epoch pre-training). In all these variants, we use the same schedule for the encoder f (lr with cosine decay).

Closely follow new representation yield better performance.

SimSIAM - Ablation Experience

Stop Gradient is needed.



	acc. (%)
w/ stop-grad	67.7±0.1
w/o stop-grad	0.1

Figure 2. **SimSiam with *vs.* without stop-gradient**. **Left plot**: training loss. Without stop-gradient it degenerates immediately. **Middle plot**: the per-channel std of the ℓ_2-normalized output, plotted as the averaged std over all channels. **Right plot**: validation accuracy of a kNN classifier [36] as a monitor of progress. **Table**: ImageNet linear evaluation ("w/ stop-grad" is mean±std over 5 trials).

Insufficient to prevent collapsing solely by the architecture
Indicate underlying another optimization problem.

Hypothesis: Underlying Optimization Problem

The implement of SimSIAM is an EM-like algorithm, following form:

$$\mathcal{L}(\theta, \eta) = \mathbb{E}_{x,\mathcal{T}} \left[||\mathcal{F}_\theta(\mathcal{T}(x)) - \eta_x||_2^2 \right]$$

where x is the input, $\mathcal{T}$ is the augmentation, $\mathcal{F}_\theta$ is the encoder(without predictor yet).
Ideally, let η_x be the representation of image x.
The procedure of SimSIAM is thought to be alternatively optimize $\mathcal{L}$ w.r.t. θ and η.

Hypothesis: Underlying Optimization Problem

Short review of EM algorithm (k-means clustering)

1. E-Step
   Estimate the cluster centers. It is the learnable parameter of **encoder**.
2. M-Step
   Assign the label vector of x.
3. repeat until convergence.

Hypothesis: Underlying Optimization Problem

For SimSIAM

1. E-Step
   Estimate θ. It is the learnable parameter of **encoder**.
2. M-Step
   η_x is analogous to the assignment vector of the sample x (a one-hot vector in k-means): **it is the representation of** x.
3. repeat until convergence.

Hypothesis: Underlying Optimization Problem

For sample x in B:

$$\begin{aligned}
\theta^t &\leftarrow \arg\min_{\theta} \mathcal{L}(\theta, \eta^{t-1}) \\
\eta^t &\leftarrow \arg\min_{\eta} \mathcal{L}(\theta^t, \eta)
\end{aligned}$$

- Stop-gradient is a natural result of EM-algorithm.
- By basic statistic, the second optimization gives:

$$\eta \leftarrow \mathbb{E}_{\mathcal{T}}[\mathcal{F}(\mathcal{T}(x))]$$

under L_2 norm.

Hypothesis: Underlying Optimization Problem

If we sample $\mathcal{T}(x)$ exactly once, this is what SimSIAM do (ignore predictor for now):

$$\mathcal{L}(\theta, \eta) = \mathbb{E}_{x,\mathcal{T}}\left[||\mathcal{F}_\theta(\mathcal{T}(x)) - \mathcal{F}_\theta(\mathcal{T}'(x))||_2^2\right]$$

Now θ^t is a constant in this sub-problem, and $\mathcal{T}'$ implies another view due to its random nature.

If we reducing the loss above with one SGD step, then SimSIAM algorithm is approached: A Siamese network with stop-gradient.

Hypothesis: Underlying Optimization Problem

Predictor: fill the gap

We have not introduced predictor yet. The predictor h aims to minimize:

$$\mathbb{E}_{z_1}[||h(z_1) - z_2||_2^2]$$

where z is the projection(s). The optimal solution of h should be able to minimize the loss for ANY image x. Similarly, the ideal form of h is

$$h(z_1) = \mathbb{E}_z[z_2] = \mathbb{E}_{\mathcal{T}}[f(\mathcal{T}(x))]$$

Hypothesis: Underlying Optimization Problem

Predictor: fill the gap

- The approximation of η_x is rough, only sample once(the Expectation is ignored.
- It could be impractical to compute the Expectation. A neural network may approximate it.
$$h(z) \sim \mathbb{E}[z] = \mathbb{E}_{\mathcal{T}}[f(\mathcal{T}(x))]$$
- The predictor fill this gap. The task of calculate $\mathbb{E}$ is switch to the other side, but keeping the loss like $\mathcal{D}$(sample, average)

Proof of Hypothesis

Multi-step alternation
Optimize θ^t for k SGD step instead of only once.

	1-step	10-step	100-step	1-epoch
acc. (%)	68.1	68.7	68.9	67.0

The model work slightly better. SimSIAM

Proof of Hypothesis

Expectation over augmentation

- Remove predictor h
- Maintain a Expectation by EMA

$$\eta^t \leftarrow m * \eta_x^{t-1} + (1-m) * \mathcal{F}_{\theta^t}(\mathcal{T}'(x))$$

The model achieve an accuracy of 55% , does not collapse.

Short Summary

Simple is more!

- Find key Component to prevent collapse by **Ablation**.
- Propose an underlying **EM-like optimization**.

Loss in the form $\mathcal{D}$(sample, average)
Is BYOL achieve this by momentum encoder?

1. Continue with BYOL
2. Why Collapse is Prevent? -SimSIAM
3. Transformer trained by Self-Distillation
4. Conclusion

ViT - Vision Transformer

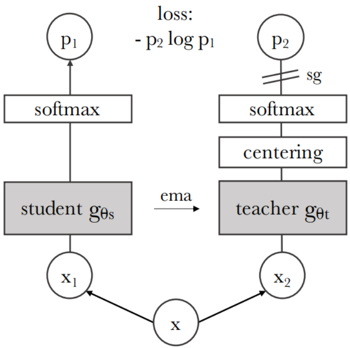
Difficulty:
Unlike language model, tokenizer of which can learn deep semantic info via statistic approach(e.g. coexistence).
Images(videos) are hard to tokenized.

- high dimension
- full of noise
- information is scarce

DINO

DINO used transformer to learn deep semantic info without ground true.
Its structure is similar to a extent of BYOL and SimSIAM

- no predictor
- Teacher model updated by EMA
- Entropy loss

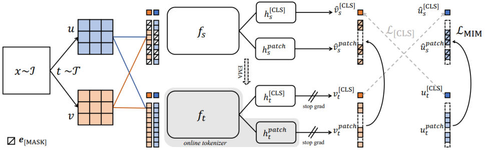


DINO (2022)

iBot

iBot used cropped image to learn deep semantic representation.
Its loss consist of two parts, difference of cropped image and difference of rest image

- no predictor
- Teacher model updated by EMA
- Entropy loss



iBot (2022)

They have no much innovation in the framework of self-distillation, but introduce transformer to it(replace covnet).

Conclusion

- What is Contrastive learning
- Learn without negative pairs
- Expectation prevents collapse in the absence of negative pair.
- Self-distillation and Transformer