Bootstrap Your Own Latent A New Approach to Self-Supervised Learning[1]

Nuriev Ainur, Sergey Petrovich, Vadim Pavlov, Sasha Latyshev

HSE University, 2021

We introduce Bootstrap Your Own Latent (BYOL), a new approach to self-supervised image representation learning. BYOL relies on two neural networks, referred to as online and target networks, that interact and learn from each other. From an augmented view of an image, we train the online network to predict the target network representation of the same image under a different augmented view. At the same time, we update the target network with a slow-moving average of the online network. While state-of-the art methods rely on negative pairs, BYOL achieves a new state of the art without them. BYOL reaches 74.3% top-1 classification accuracy on ImageNet using a linear evaluation with a ResNet-50 architecture and 79.6% with a larger ResNet. We show that BYOL performs on par or better than the current state of the art on both transfer and semi-supervised benchmarks. Our implementation and pretrained models are given on GitHub

Introduction

Bootstrap Your Own Latent
A New Approach to Self-Supervised Learning

Jean-Bastien Grill*[1] Florian Strub*[1] Florent Altché*[1] Corentin Tallec*[1] Pierre H. Richemond*[1,2]
Elena Buchatskaya[1] Carl Doersch[1] Bernardo Avila Pires[1] Zhaohan Daniel Guo[1]
Mohammad Gheshlaghi Azar[1] Bilal Piot[1] Koray Kavukcuoglu[1] Rémi Munos[1] Michal Valko[1]

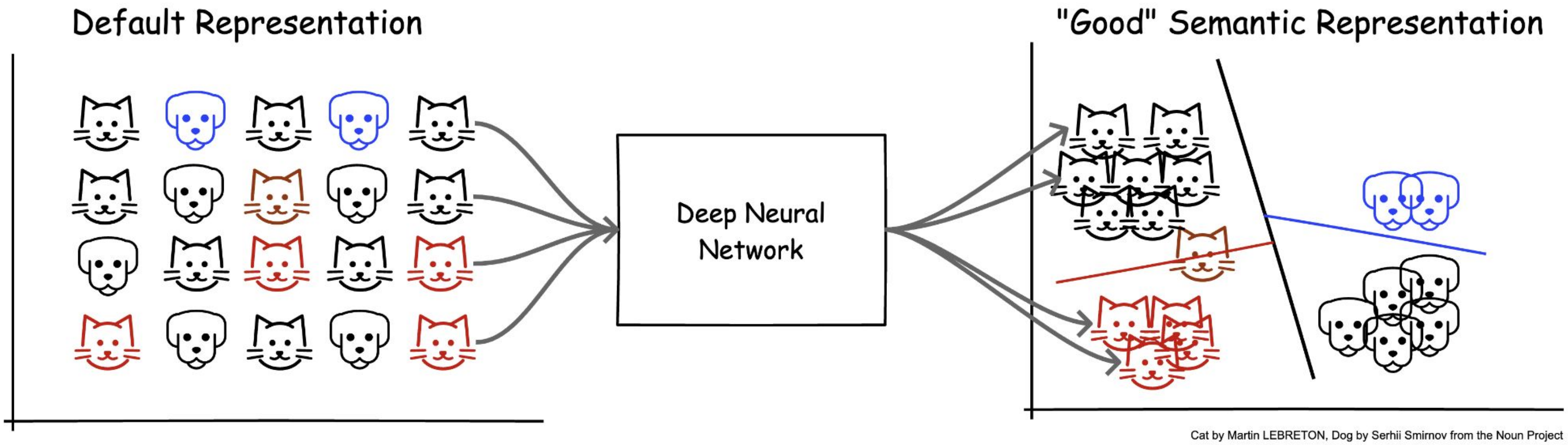
[1]DeepMind [2]Imperial College

[jbgrill,fstrub,altche,corentint,richemond]@google.com

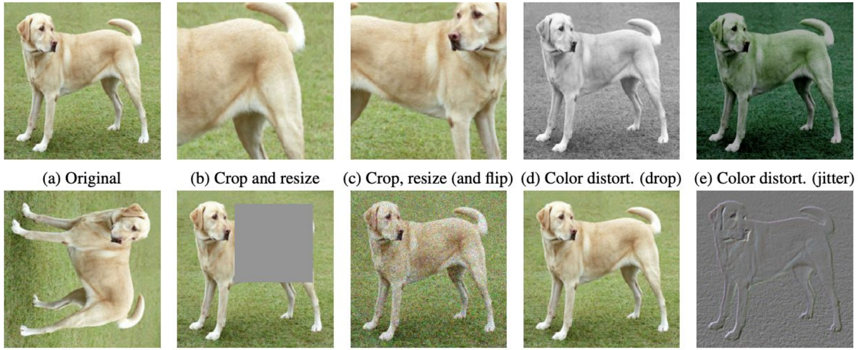
Abstract

We introduce **B**ootstrap **Y**our **O**wn **L**atent (BYOL), a new approach to self-supervised image representation learning. BYOL relies on two neural networks, referred to as *online* and *target* networks, that interact and learn from each other. From an augmented view of an image, we train the online network to predict the target network representation of the same image under a different augmented view. At the same time, we update the target network with a slow-moving average of the online network. While state-of-the art methods rely on negative pairs, BYOL achieves a new state of the art *without them*. BYOL reaches 74.3% top-1 classification accuracy on ImageNet using a linear evaluation with a ResNet-50 architecture and 79.6% with a larger ResNet. We show that BYOL performs on par or better than the current state of the art on both transfer and semi-supervised benchmarks. Our implementation and pretrained models are given on GitHub.[3]

Image Representation Learning

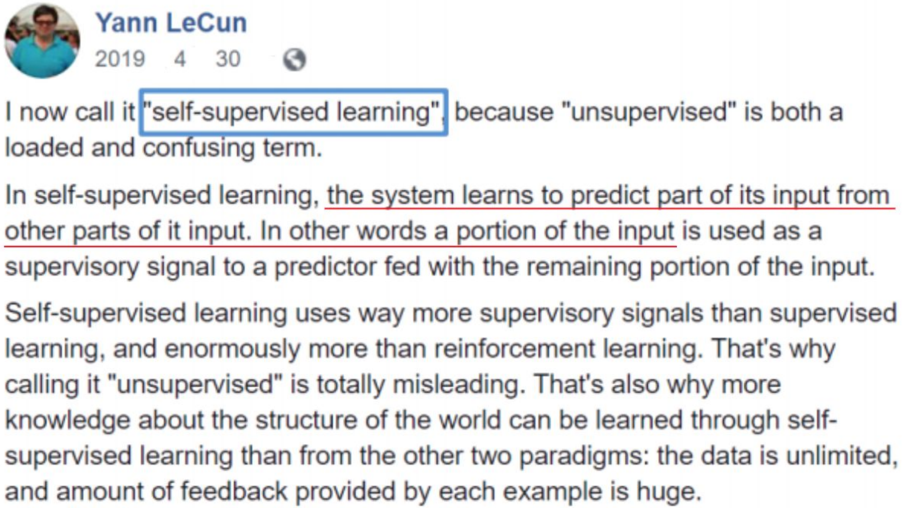


Self-Supervised Learning



(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



Yann LeCun

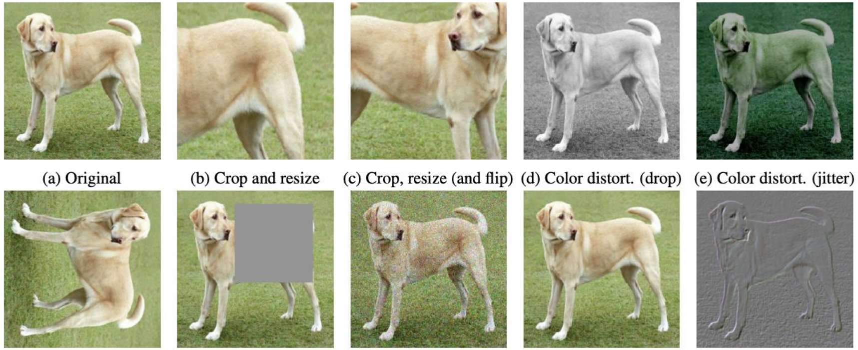
2019 4 30

I now call it "self-supervised learning", because "unsupervised" is both a loaded and confusing term.

In self-supervised learning, the system learns to predict part of its input from other parts of it input. In other words a portion of the input is used as a supervisory signal to a predictor fed with the remaining portion of the input.

Self-supervised learning uses way more supervisory signals than supervised learning, and enormously more than reinforcement learning. That's why calling it "unsupervised" is totally misleading. That's also why more knowledge about the structure of the world can be learned through self-supervised learning than from the other two paradigms: the data is unlimited, and amount of feedback provided by each example is huge.

Self-Supervised Learning



(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

Negative Samples

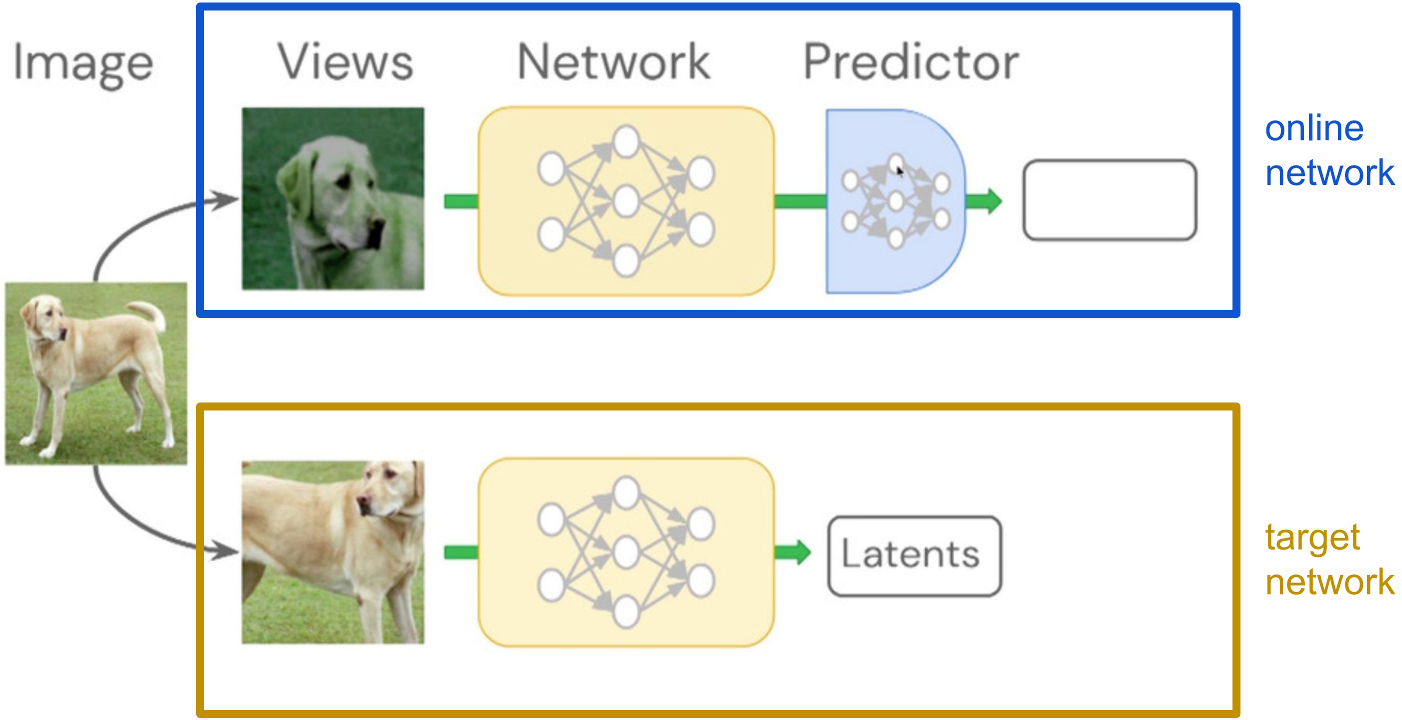
Yann LeCun
2019 4 30

I now call it "self-supervised learning", because "unsupervised" is both a loaded and confusing term.

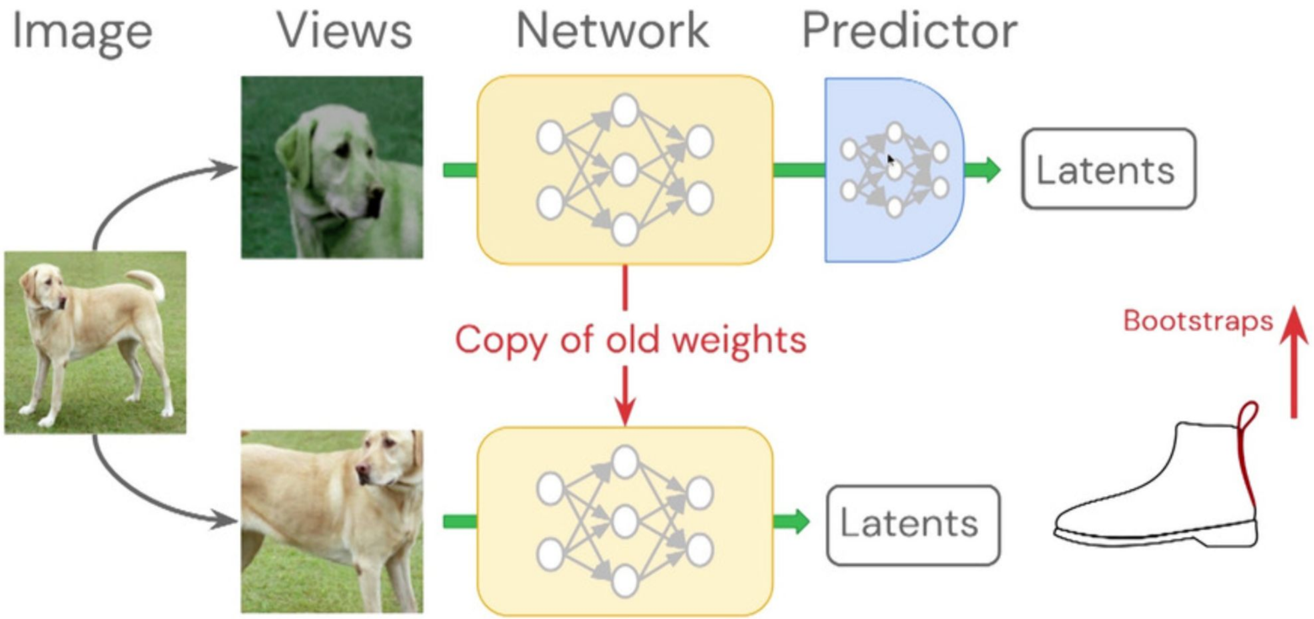
In self-supervised learning, the system learns to predict part of its input from other parts of it input. In other words a portion of the input is used as a supervisory signal to a predictor fed with the remaining portion of the input.

Self-supervised learning uses way more supervisory signals than supervised learning, and enormously more than reinforcement learning. That's why calling it "unsupervised" is totally misleading. That's also why more knowledge about the structure of the world can be learned through self-supervised learning than from the other two paradigms: the data is unlimited, and amount of feedback provided by each example is huge.

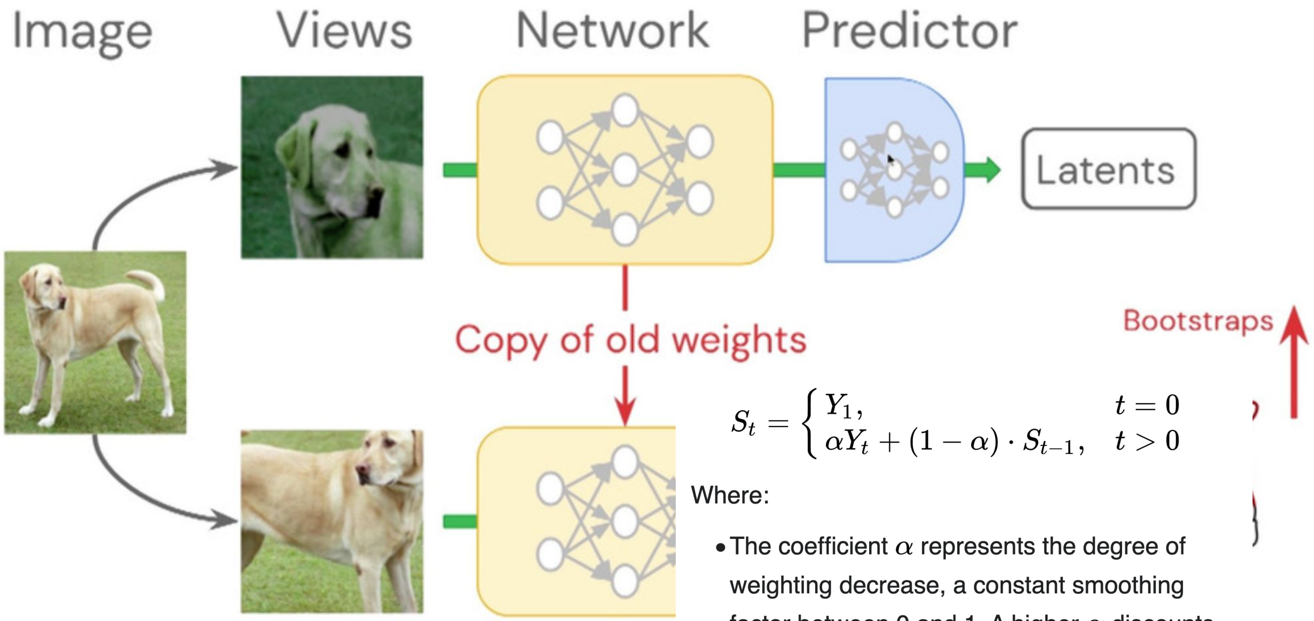
Architecture



Architecture. Copy of Old Weights



Architecture. Exponential Moving Average

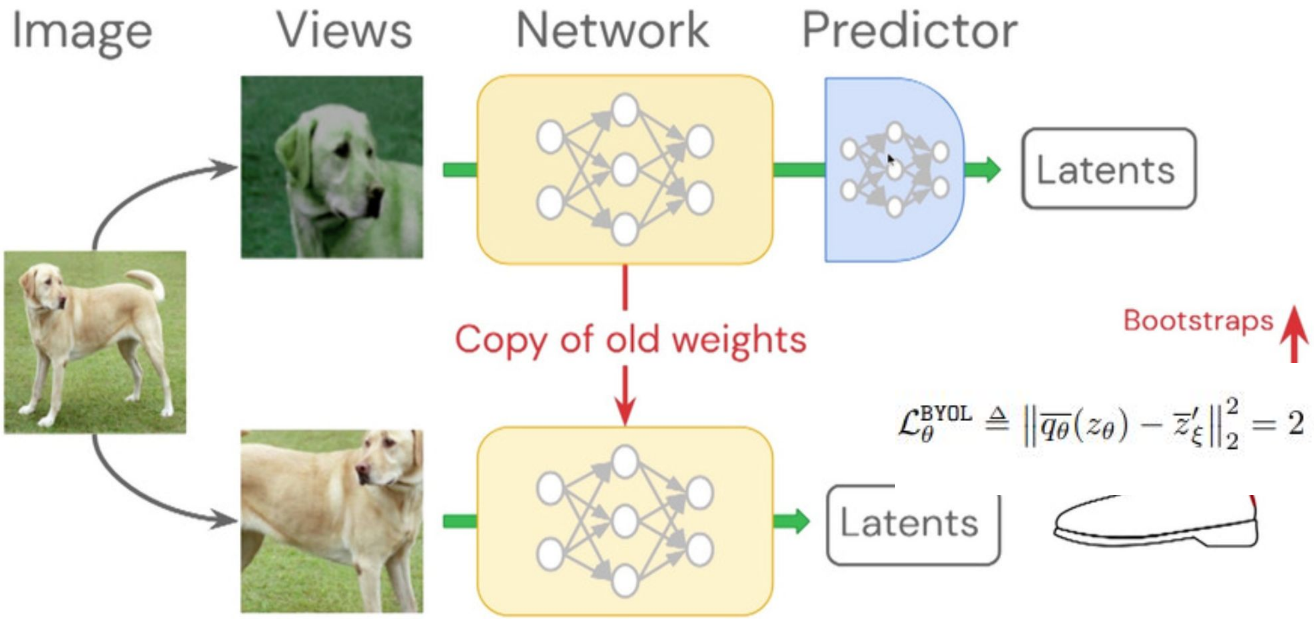


$$S_t = \begin{cases} Y_1, & t = 0 \\ \alpha Y_t + (1-\alpha)\cdot S_{t-1}, & t > 0 \end{cases}$$

Where:

- The coefficient α represents the degree of weighting decrease, a constant smoothing factor between 0 and 1. A higher α discounts older observations faster.
- Y_t is the value at a time period t.
- S_t is the value of the EMA at any time period t.

Architecture. Loss Function



$$\mathcal{L}_\theta^{\mathrm{BYOL}} \triangleq \left\|\overline{q_\theta}(z_\theta) - \bar{z}'_\xi\right\|_2^2 = 2 - 2\cdot\frac{\langle q_\theta(z_\theta), z'_\xi\rangle}{\left\|q_\theta(z_\theta)\right\|_2 \cdot \left\|z'_\xi\right\|_2}.$$

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

Experiments

Method	Top-1	Top-5
Local Agg.	60.2	-
PIRL [35]	63.6	-
CPC v2 [32]	63.8	85.3
CMC [11]	66.2	87.0
SimCLR [8]	69.3	89.0
MoCo v2 [37]	71.1	-
InfoMin Aug. [12]	73.0	91.1
BYOL (ours)	**74.3**	**91.6**

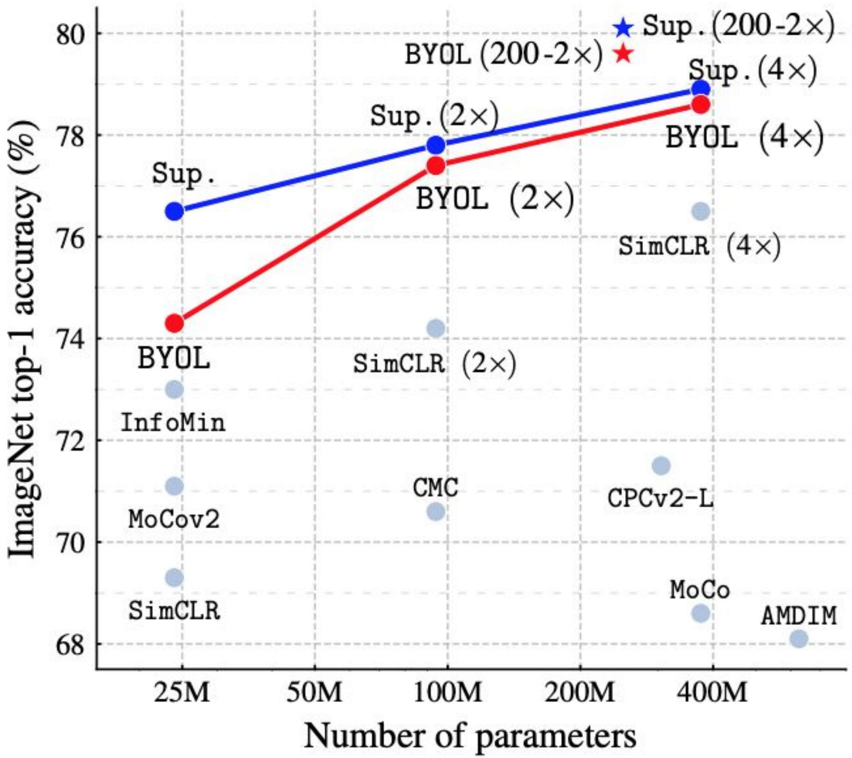
(a) ResNet-50 encoder.

Method	Architecture	Param.	Top-1	Top-5
SimCLR [8]	ResNet-50 (2×)	94M	74.2	92.0
CMC [11]	ResNet-50 (2×)	94M	70.6	89.7
BYOL (ours)	ResNet-50 (2×)	94M	77.4	93.6
CPC v2 [32]	ResNet-161	305M	71.5	90.1
MoCo [9]	ResNet-50 (4×)	375M	68.6	-
SimCLR [8]	ResNet-50 (4×)	375M	76.5	93.2
BYOL (ours)	ResNet-50 (4×)	375M	78.6	94.2
BYOL (ours)	ResNet-200 (2×)	250M	**79.6**	**94.8**

(b) Other ResNet encoder architectures.

a batch size of 4096 split over 512 Cloud TPU v3 cores. With this setup, training takes approximately 8 hours for a ResNet-50(×1)

Experiments

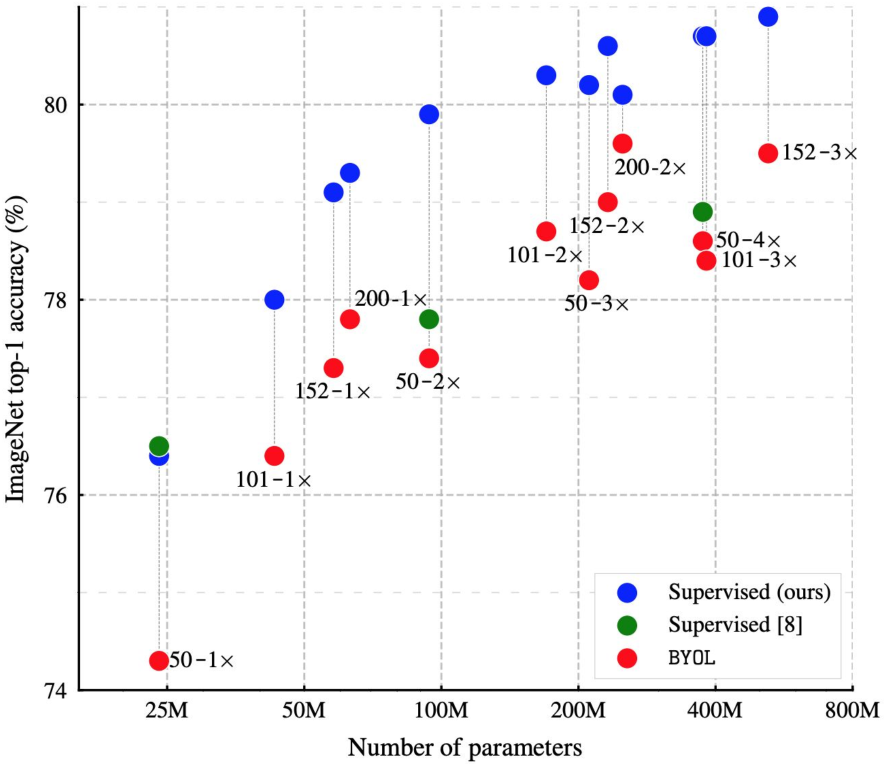


Experiments

Method	Food101	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech-101	Flowers
Linear evaluation:												
BYOL (ours)	**75.3**	91.3	**78.4**	**57.2**	**62.2**	**67.8**	60.6	82.5	75.5	90.4	94.2	**96.1**
SimCLR (repro)	72.8	90.5	74.4	42.4	60.6	49.3	49.8	81.4	**75.7**	84.6	89.3	92.6
SimCLR [8]	68.4	90.6	71.6	37.4	58.8	50.3	50.3	80.5	74.5	83.6	90.3	91.2
Supervised-IN [8]	72.3	**93.6**	78.3	53.7	61.9	66.7	**61.0**	**82.8**	74.9	**91.5**	**94.5**	94.7
Fine-tuned:												
BYOL (ours)	**88.5**	**97.8**	86.1	**76.3**	63.7	91.6	**88.1**	**85.4**	**76.2**	91.7	**93.8**	97.0
SimCLR (repro)	87.5	97.4	85.3	75.0	63.9	91.4	87.6	84.5	75.4	89.4	91.7	96.6
SimCLR [8]	88.2	97.7	85.9	75.9	63.5	91.3	88.1	84.1	73.2	89.2	92.1	97.0
Supervised-IN [8]	88.3	97.5	**86.4**	75.8	**64.3**	**92.1**	86.0	85.0	74.6	**92.1**	93.3	**97.6**
Random init [8]	86.9	95.9	80.2	76.1	53.6	91.4	85.9	67.3	64.8	81.5	72.6	92.0

Table 3: Transfer learning results from ImageNet (IN) with the standard ResNet-50 architecture.

Experiments



Рецензия. Положительные стороны статьи

- BYOL позволяет обучать представления картинок в режиме self-supervised без использования негативных примеров, и при этом не сходиться к тривиальному решению. Эта особенность позволяет избавиться от некоторых проблем предыдущих подходов - необходимость аккуратного выбора негативных примеров и большого размера батча.
- Авторы проводят много экспериментов, сравнивая с лучшими существующими на момент выхода статьи решениями, а также показывают влияние различных гиперпараметров и частей модели на итоговое качество.
- Актуальность задачи обучения векторных представлений картинок без разметки, ровно как и данной работы, не вызывает вопросов.

Рецензия. Отрицательные стороны статьи

- Отсутствие теоретической обоснованности. В первой версии статьи авторы вообще никак не объясняли, почему их метод не сходится к тривиальному решению, потом добавили некое интуитивно-эвристическое рассуждение
- Вопросы к приведенным в работе числам. Непонятно, насколько значимо улучшение в качестве, если репродукция SimCLR приносит больше процентных пунктов, чем сам BYOL
- Не все эксперименты согласуются с последующими работами:

> branch is a momentum encoder.[2] It is hypothesized in [15] that the momentum encoder is important for BYOL to avoid collapsing, and it reports failure results if removing the momentum encoder (0.3% accuracy, Table 5 in [15]).[3] Our empirical study challenges the *necessity* of the momentum encoder for preventing collapsing. We discover that the

- Воспроизводимость - полностью представлен псевдокод, но нет полной официальной реализации, а запустить сторонние оказалось очень тяжело

Рецензия. Рецензии с конференции (NeurIPS 2020)

Вот о чём пишут чаще всего

Этот отзыв написал наш умный алгоритм — он всё прочитал и выделил главное

Достоинства:

1. Актуальная область
2. Круто, что удалось избавится от негативных примеров
3. Статья написана понятно

Недостатки:

1. Непонятно почему метод не сходится к тривиальному решению
2. Недостаточно уделено внимания сравнению с MoCo и MeanTeacher, как будто авторы специально пытаются приуменьшить связь с этими работами
3. Нет открытого исходного кода

Исследование. Общие сведения

Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, Michal Valko

Oral presentation: Orals & Spotlights Track 27: Unsupervised/Probabilistic
on Thu, Dec 10th, 2020 @ 17:15 – 17:30 MSK

Poster Session 6 (more posters)
on Thu, Dec 10th, 2020 @ 20:00 – 22:00 MSK

Toggle Abstract   Paper (in Proceedings / .pdf)

Исследование. Авторы статьи

Основные области интересов авторов:
Reinforcement Learning, Representation Learning и Computer Vision.

Семь авторов несколькими месяцами ранее опубликовали статью по Reinforcement Learning, на которую впоследствии ссылаются в данной работе.

Bootstrap Your Own Latent
A New Approach to Self-Supervised Learning

Jean-Bastien Grill[*,1] **Florian Strub**[*,1] **Florent Altché**[*,1] **Corentin Tallec**[*,1] **Pierre H. Richemond**[*,1,2]
Elena Buchatskaya[1] **Carl Doersch**[1] **Bernardo Avila Pires**[1] **Zhaohan Daniel Guo**[1]
Mohammad Gheshlaghi Azar[1] **Bilal Piot**[1] **Koray Kavukcuoglu**[1] **Rémi Munos**[1] **Michal Valko**[1]

[1]DeepMind [2]Imperial College

[jbgrill,fstrub,altche,corentint,richemond]@google.com

Исследование. Источники вдохновения

Bootstrap Latent-Predictive Representations for Multitask Reinforcement Learning

Daniel Guo [*1] **Bernardo Avila Pires** [*1] **Bilal Piot** [1] **Jean Bastien Grill** [2] **Florent Altché** [2] **Rémi Munos** [2] **Mohammad Gheshlaghi Azar** [1]

BYOL - продолжение работы на обучением представлений, но в более общем случае

Исследование. Цитирования

BYOL works even without batch statistics — [PDF] arxiv.org

PH Richemond, JB Grill, F Altché, C Tallec… - arXiv preprint arXiv …, 2020 - arxiv.org

Bootstrap Your Own Latent (**BYOL**) is a self-supervised learning approach for image representation. From an augmented view of an image, **BYOL** trains an online network to predict a target network representation of a different augmented view of the same image …

Cite Cited by 14 Related articles All 3 versions

BYOL for Audio: Self-Supervised Learning for General-Purpose Audio Representation — [PDF] arxiv.org

D Niizumi, D Takeuchi, Y Ohishi, N Harada… - arXiv preprint arXiv …, 2021 - arxiv.org

Inspired by the recent progress in self-supervised learning for computer vision that generates supervision using data augmentations, we explore a new general-purpose audio representation learning approach. We propose learning general-purpose audio …

Cite Cited by 6 Related articles All 3 versions

Bootstrap your own latent: A new approach to self-supervised learning — [PDF] arxiv.org

JB Grill, F Strub, F Altché, C Tallec… - arXiv preprint arXiv …, 2020 - arxiv.org

… We show that **BYOL** performs on par or better than the current state of the art on both transfer and semi-supervised benchmarks. Our … (iii) We show that **BYOL** is more resilient to changes in the batch size and in the set of image augmentations compared to its contrastive …

Cite Cited by 675 Related articles All 11 versions

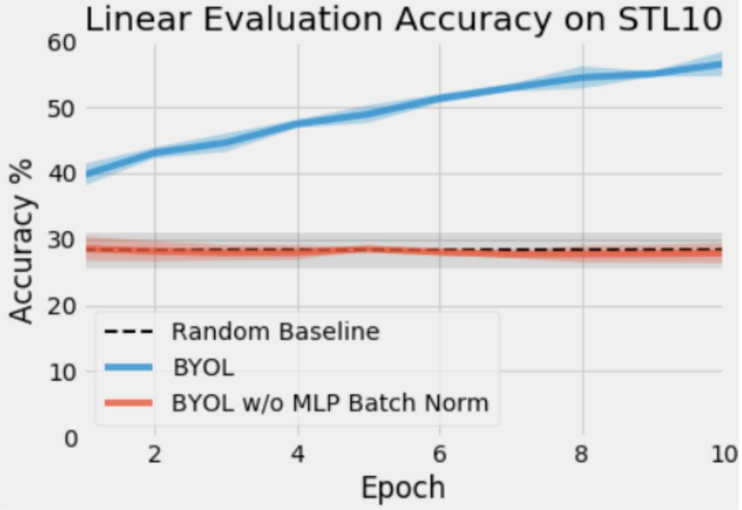
Run away from your teacher: Understanding **byol** by a novel self-supervised approach — [PDF] arxiv.org

H Shi, D Luo, S Tang, J Wang, Y Zhuang - arXiv preprint arXiv:2011.10944, 2020 - arxiv.org

Recently, a newly proposed self-supervised framework Bootstrap Your Own Latent (**BYOL**) seriously challenges the necessity of negative samples in contrastive learning frameworks. **BYOL** works like a charm despite the fact that it discards the negative samples completely …

Cite Cited by 2 Related articles All 2 versions

Исследование. Продолжение статьи



BYOL works *even* without batch statistics

Pierre H. Richemond[*,1,2] **Jean-Bastien Grill**[*,1] **Florent Altché**[*,1] **Corentin Tallec**[*,1] **Florian Strub**[*,1]

Andrew Brock[1] **Samuel Smith**[1] **Soham De**[1] **Razvan Pascanu**[1]

Bilal Piot[1] **Michal Valko**[1]

[1]DeepMind [2]Imperial College

phr17@ic.ac.uk [jbgrill,fstrub,altche,corentint]@google.com

Оказалось, что базовая версия BYOL не работает без batch-normalization. Была выдвинута гипотеза, что batch-norm неявно моделирует contrastive learning. Авторы показали, модель можно запустить и без batch-normalization.

Воспроизведение результатов

Было найдено 3 реализации BYOL, планировалось запустить одну, чтобы убедиться, что работает, а потом заменить внутри ResNet на простую сверточную сеть, чтобы проверить, сможет ли алгоритм работать с простой сетью.

Но ни одна из версией не запустилась по различным ошибкам.

- реализация в lightly
- сторонняя реализация
- официальная версия от deepmind

References

[1] Grill, Jean-Bastien, et al. "Bootstrap your own latent: A new approach to self-supervised learning." arXiv preprint arXiv:2006.07733 (2020).

[2] Richemond, Pierre H., et al. “BYOL works even without batch statistics”. arXiv preprint arXiv: 2010:10241 (2020).