DeepMind

Bootstrap Your Own Latent:
A new approach to self-supervised learning

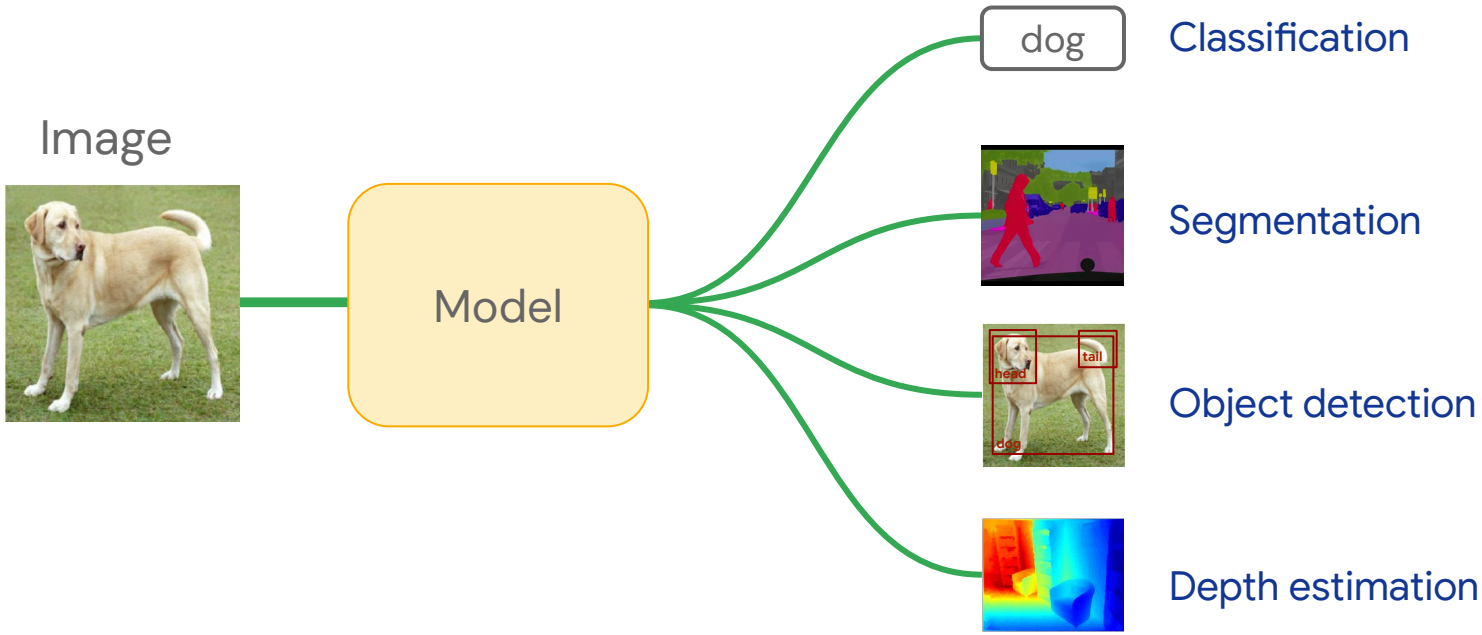
Jean-Bastien Grill*, Florian Strub*, Florent Altché*, Corentin Tallec*, Pierre H. Richemond*
Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, Michal Valko

Deep Learning: Classics and Trends

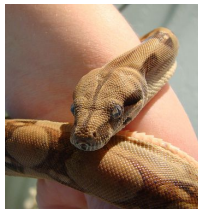
DeepMind

1 Self-Supervised Learning

Computer Vision Goal

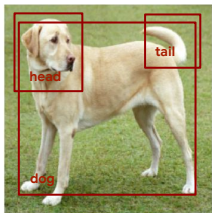


Motivation

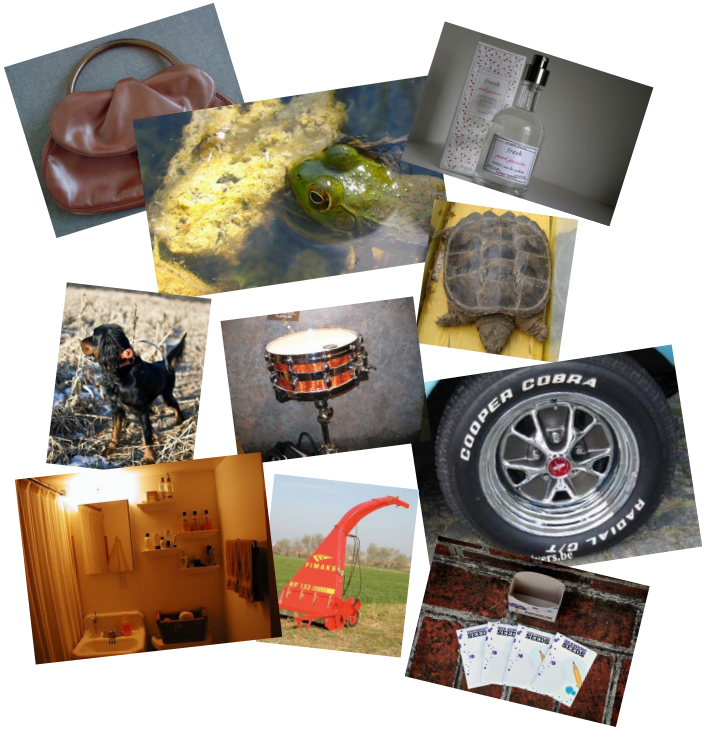


Dog

Snake

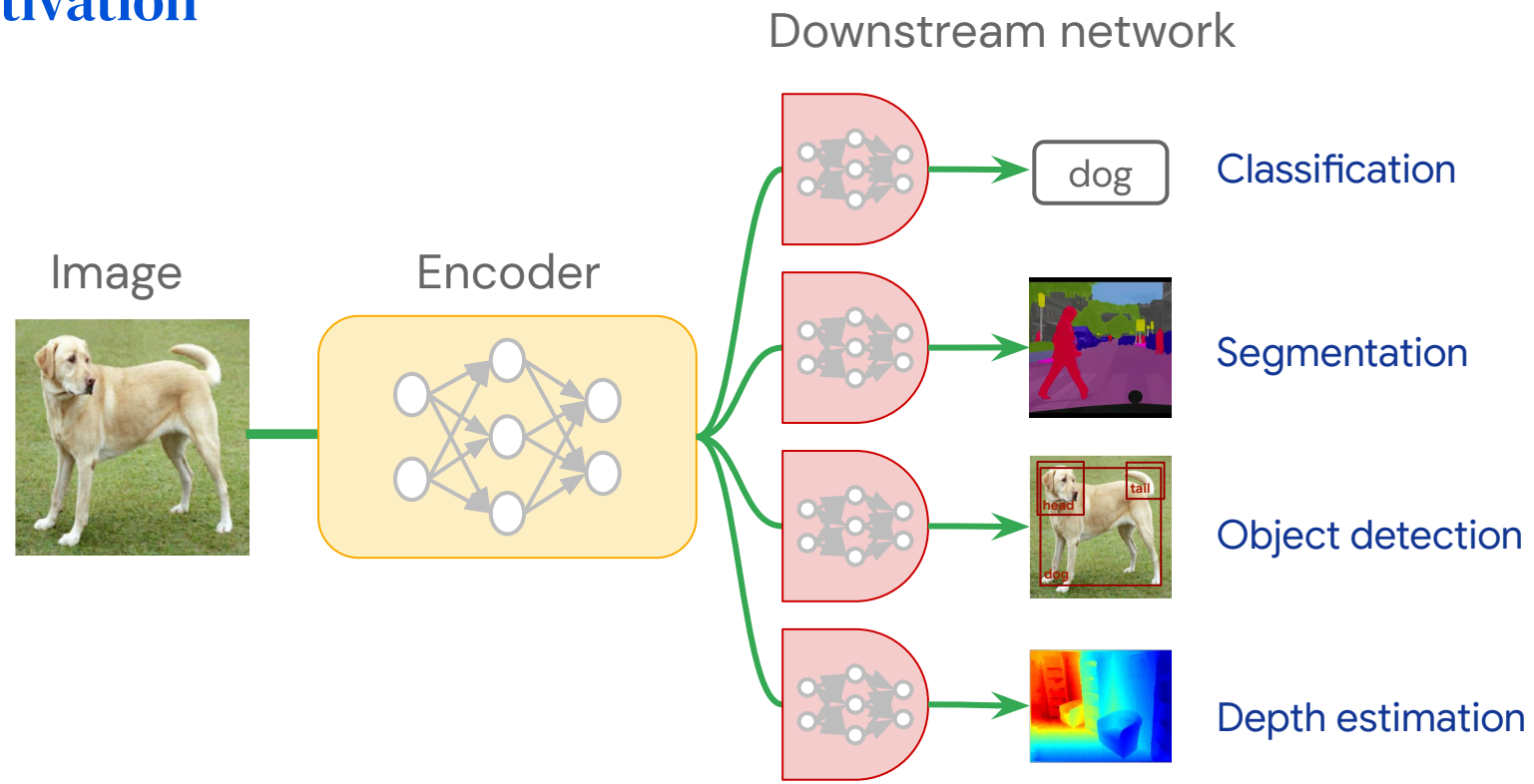


Labelled, but costly/few data



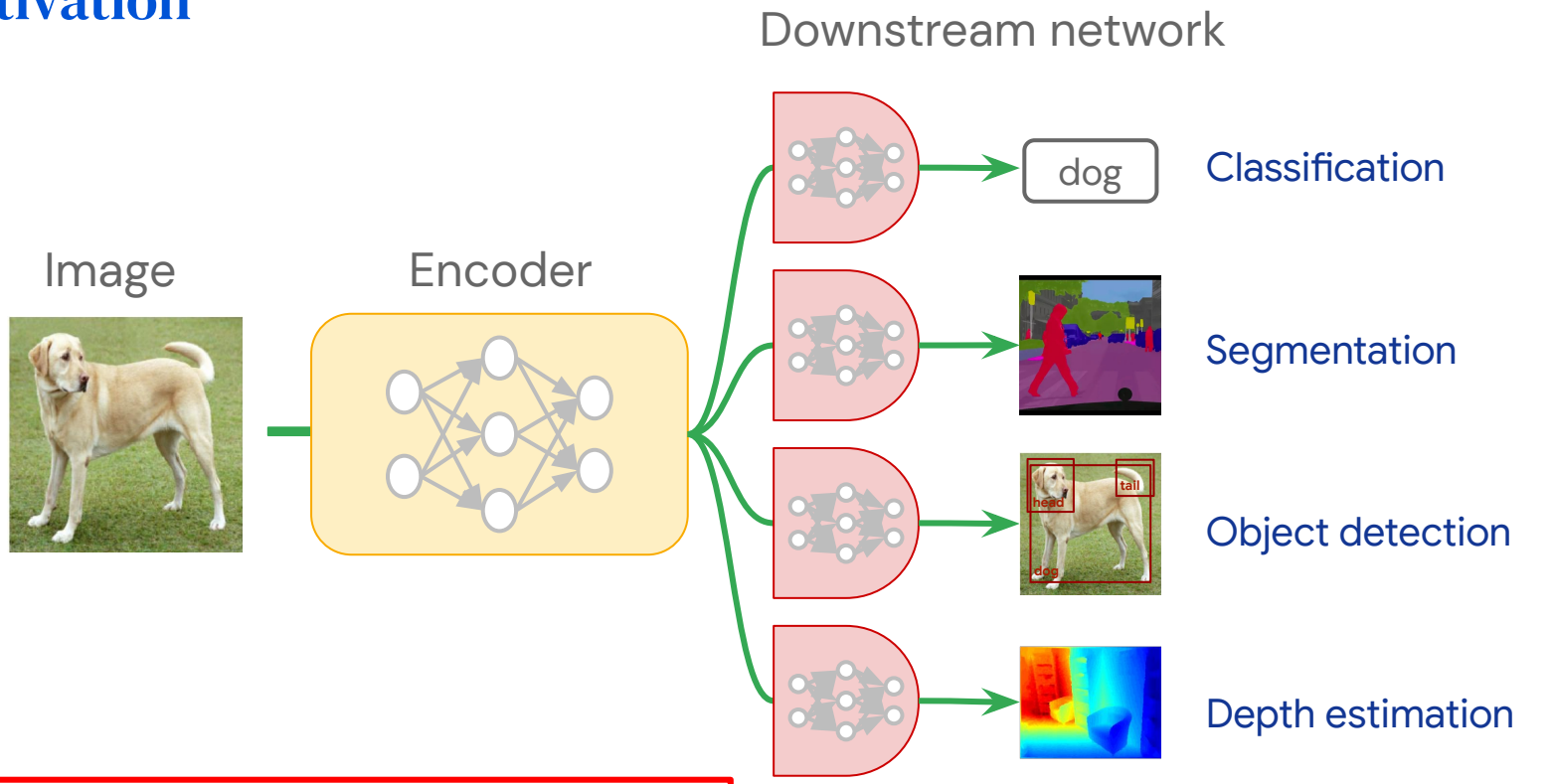
Unlabelled, **free** data!

Motivation



How to train the encoder?

Motivation



BYOL

Self-supervised
Free unlabeled data

Supervised
Few labeled data

DeepMind

2 Method

Intuition: Two different views (augmentations) of the same picture should be predictive of each other.

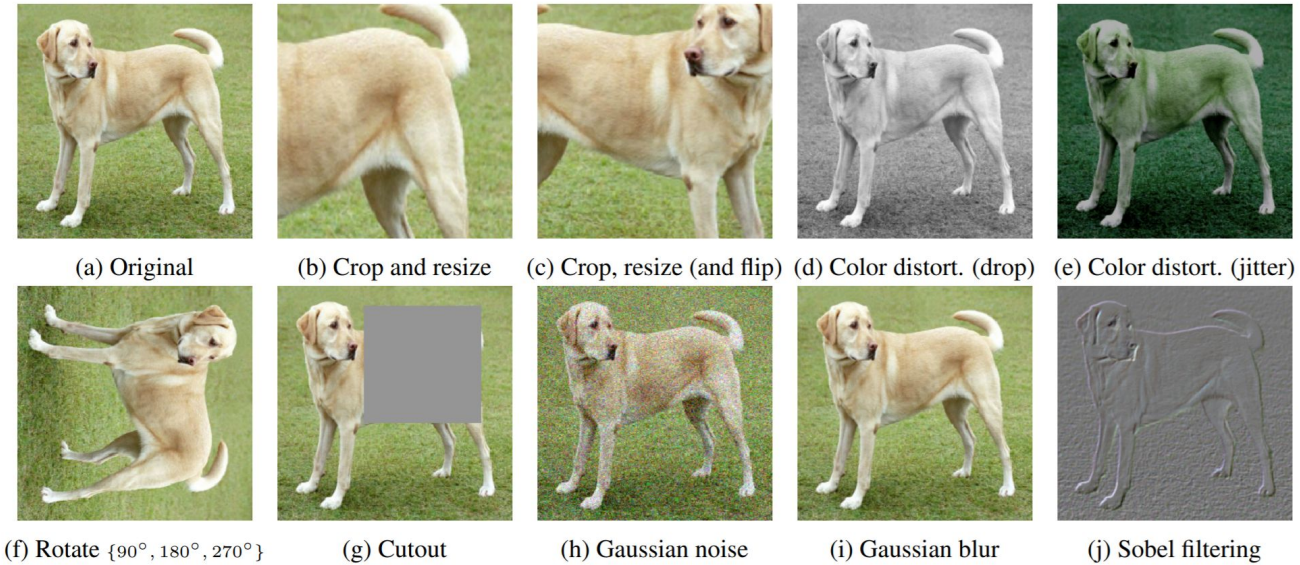
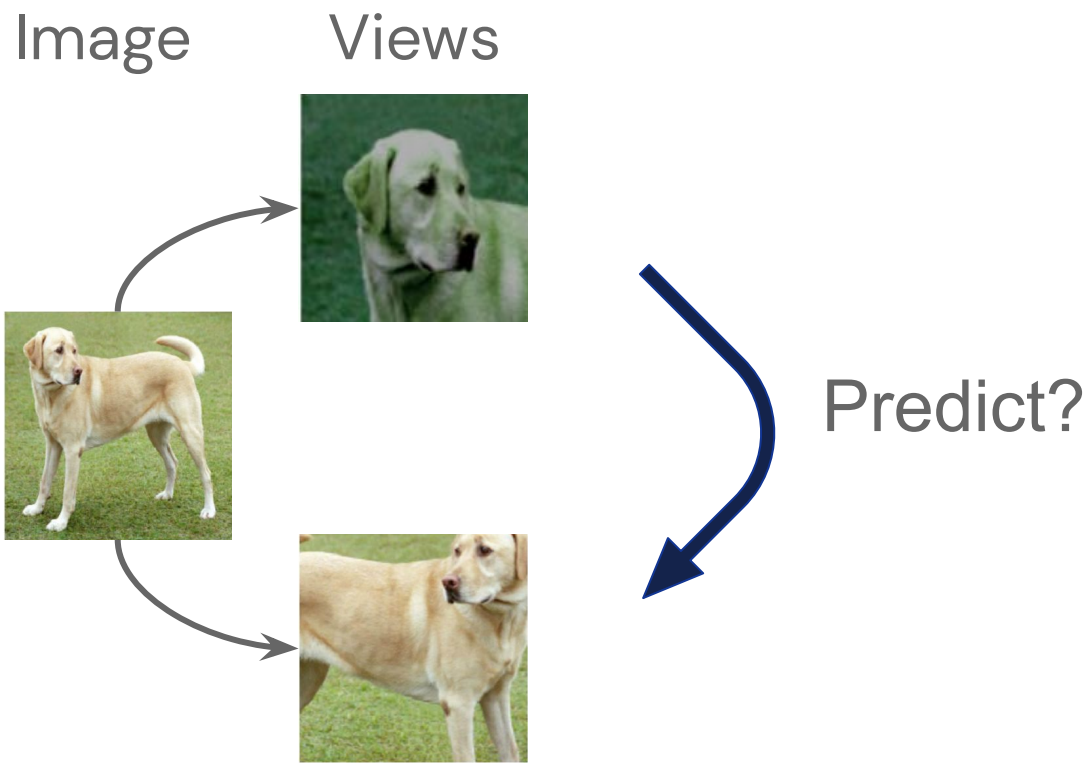


Figure from SimCLR[1]

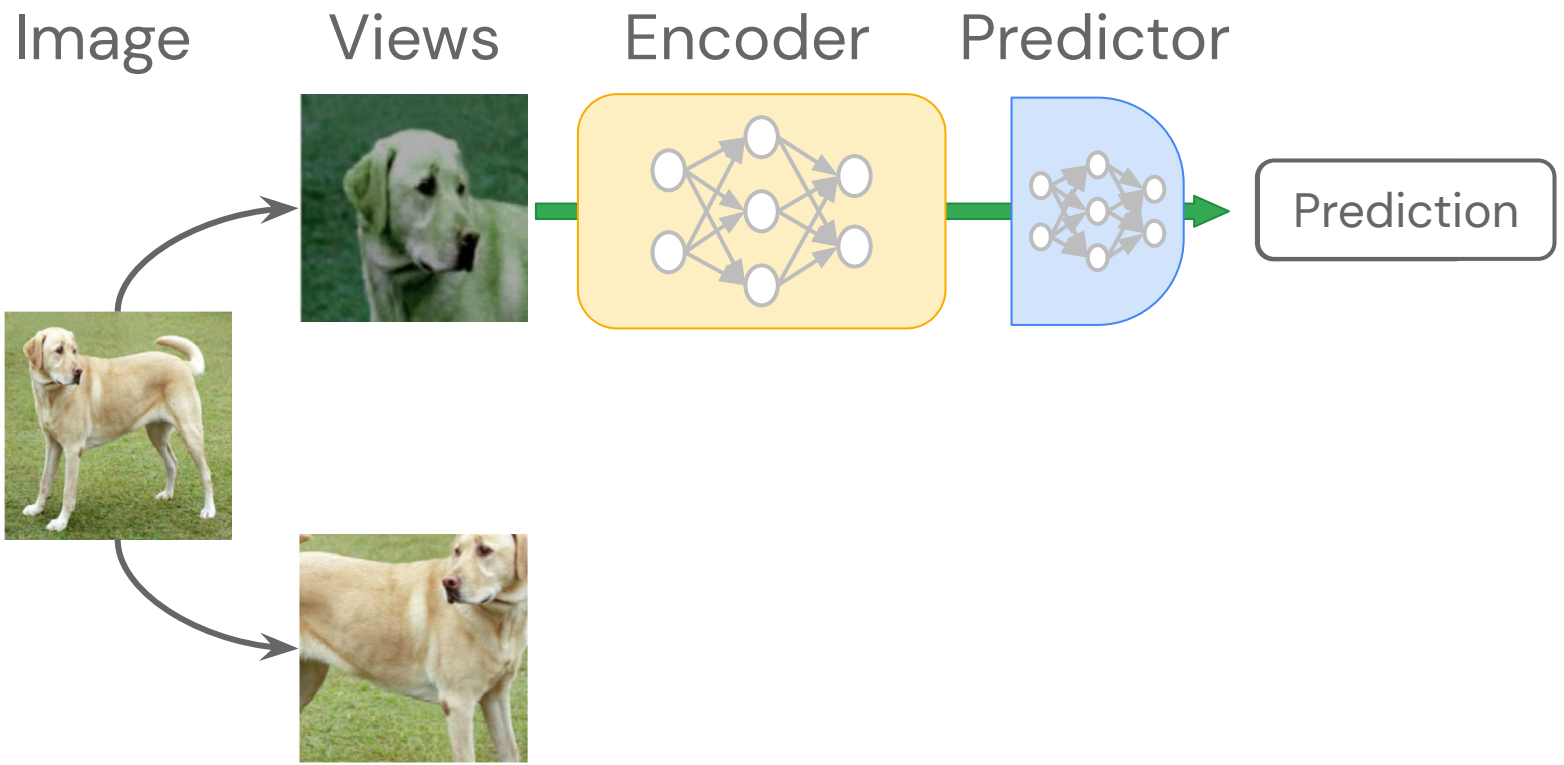
A view of a dog is still a dog, i.e. semantic information is **invariant** to transformations.

[1] SimCLR: Chen et al., A simple framework for contrastive learning of visual representations. ICML. 2020

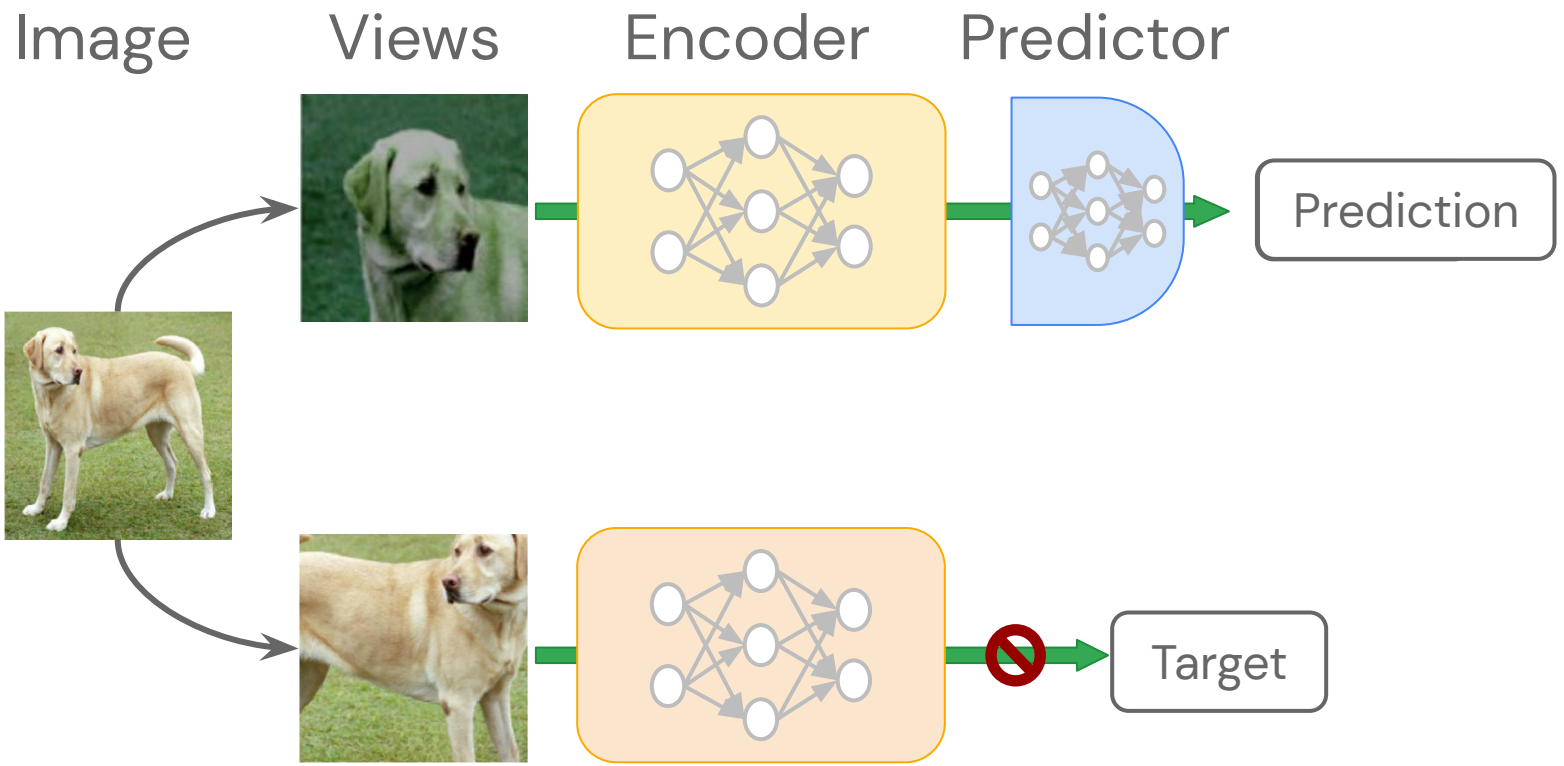
BYOL main intuition



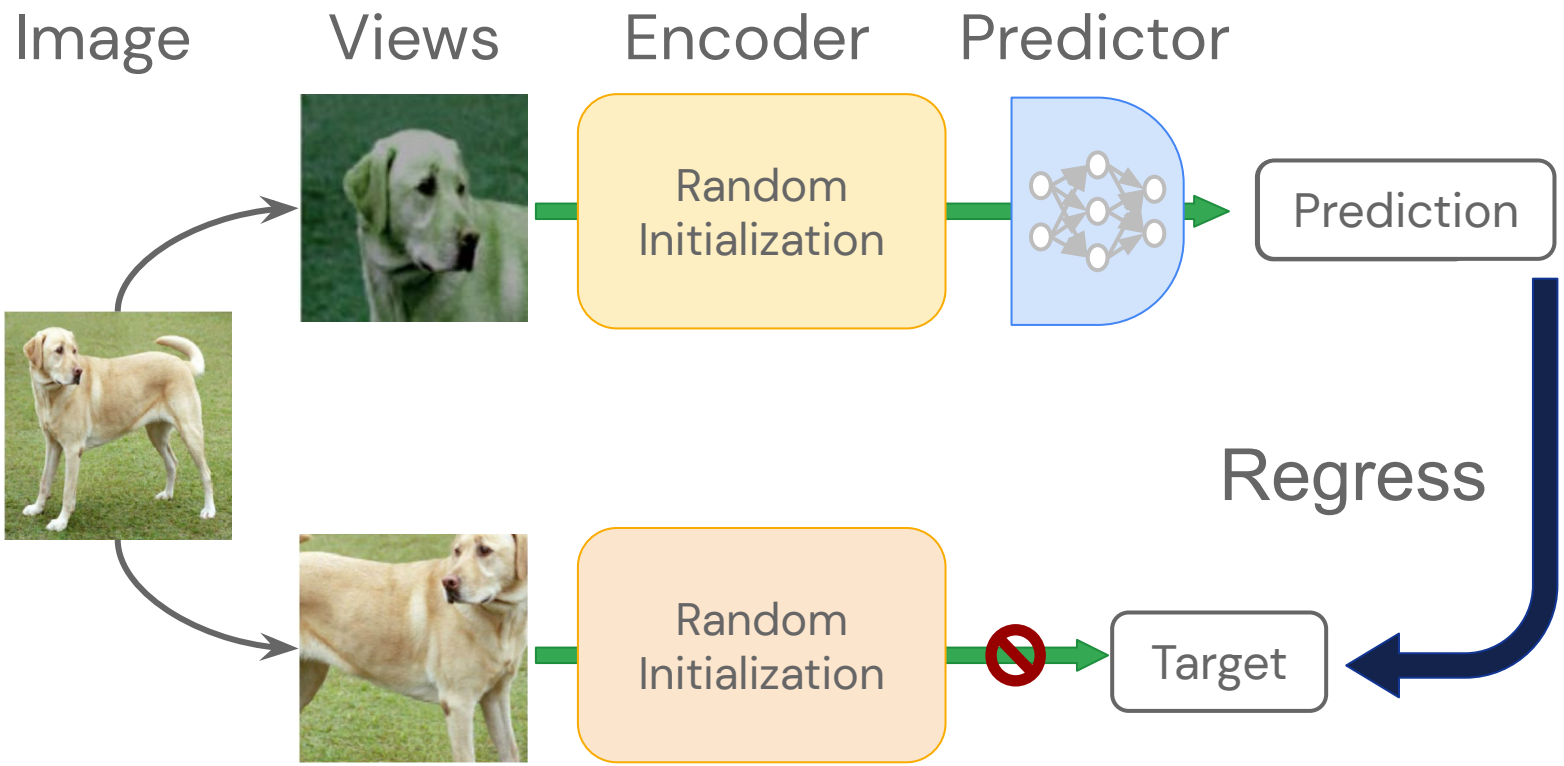
BYOL main intuition



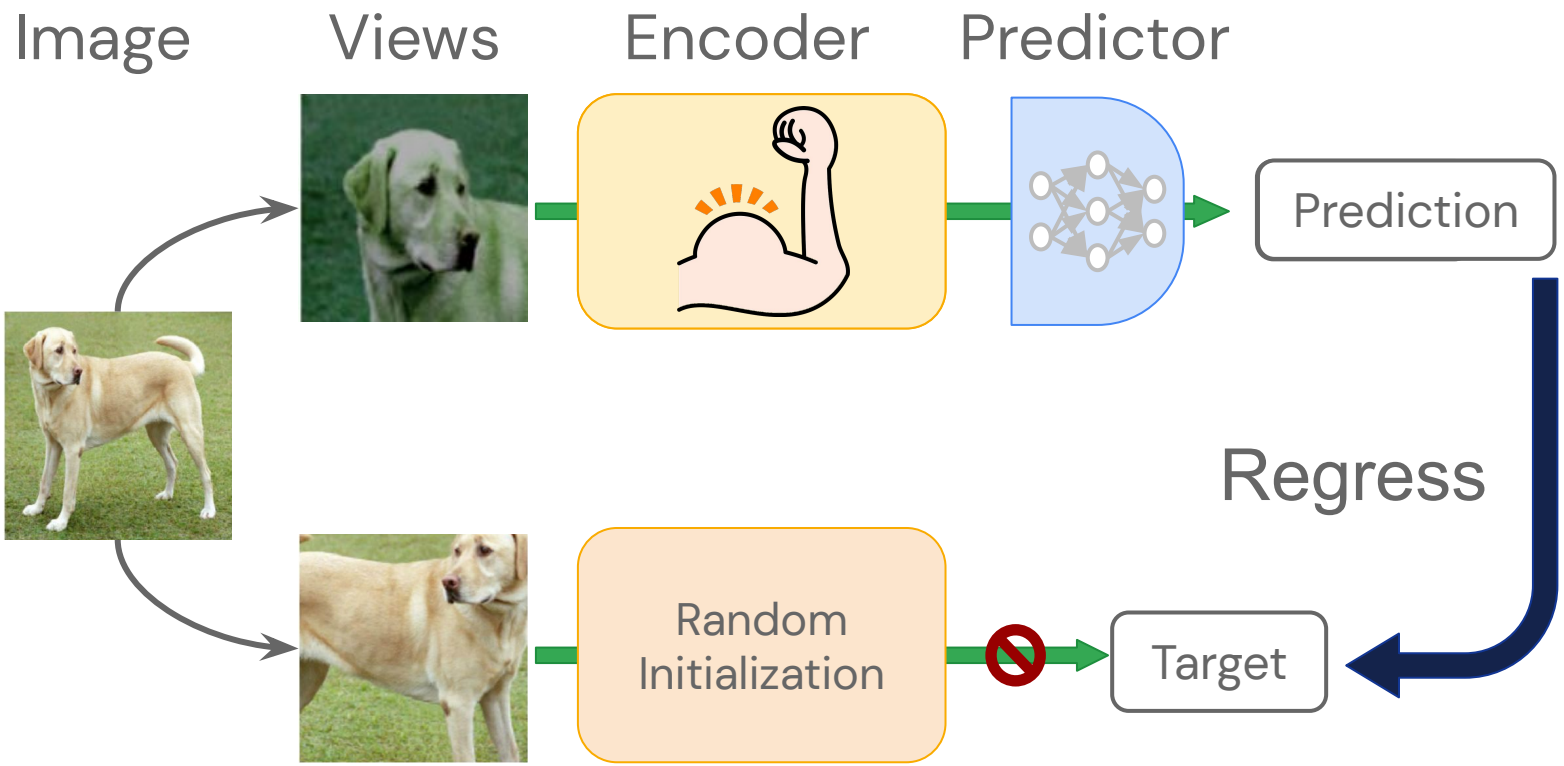
BYOL main intuition



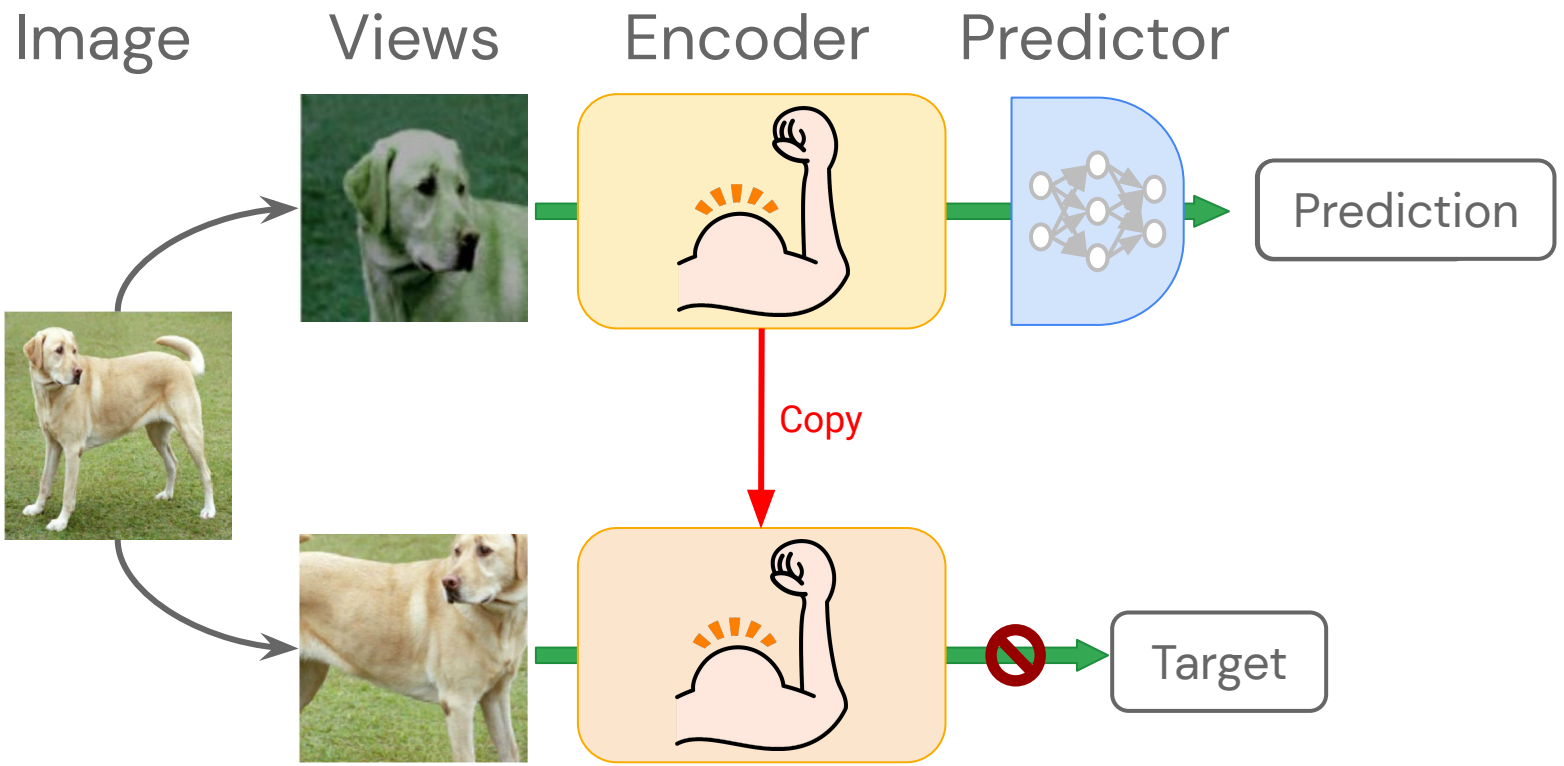
BYOL main intuition



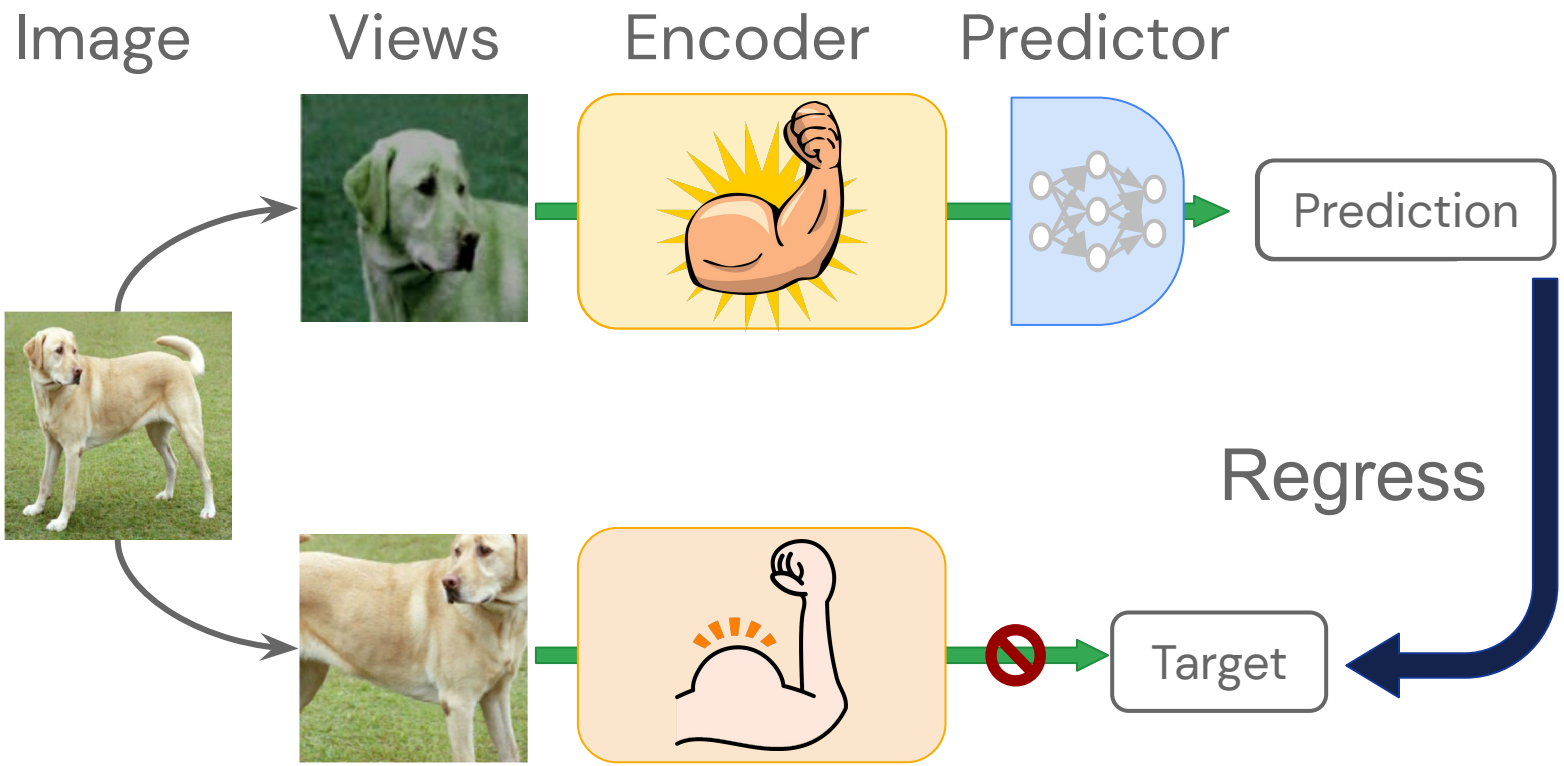
BYOL main intuition



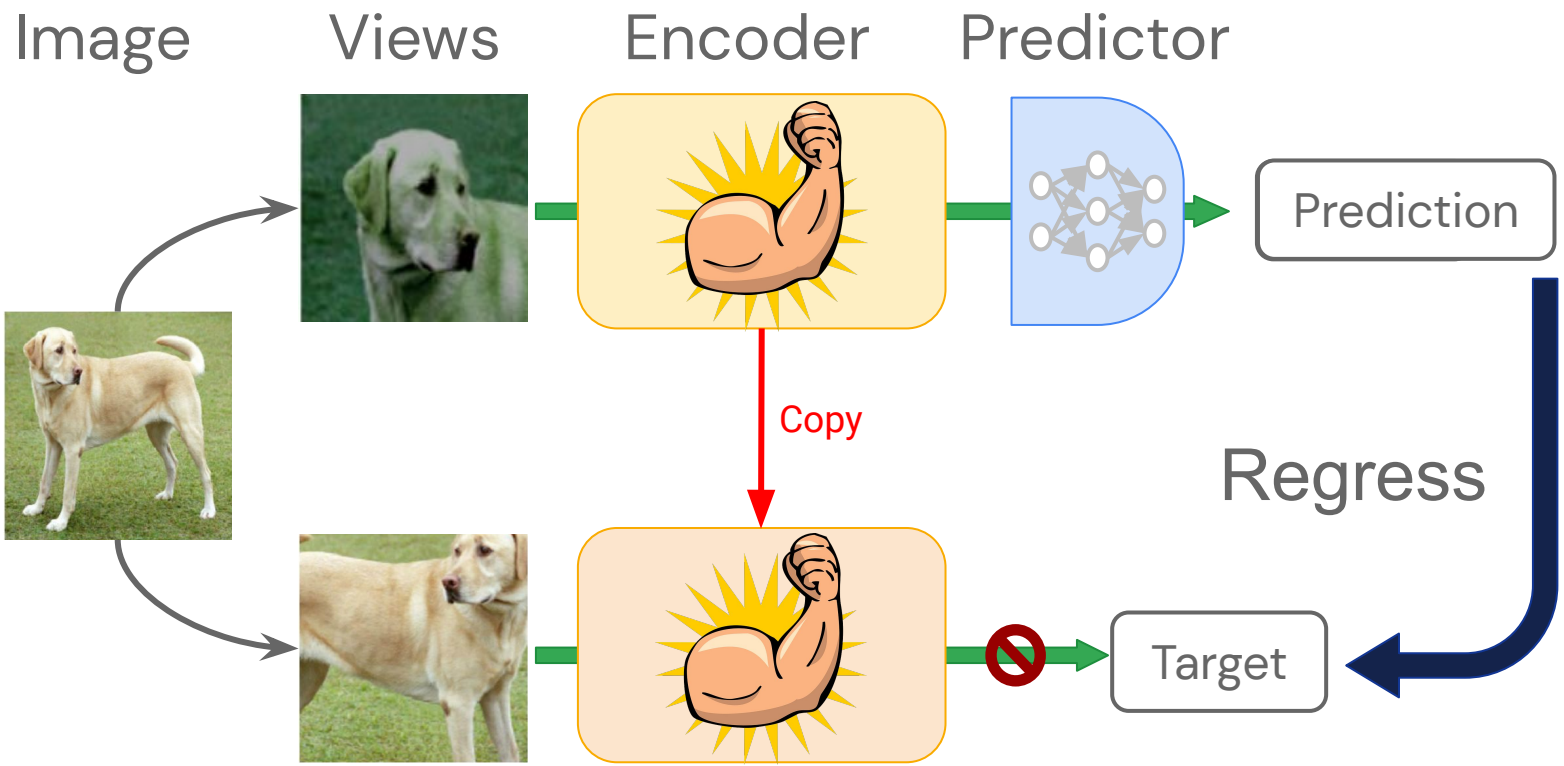
BYOL main intuition



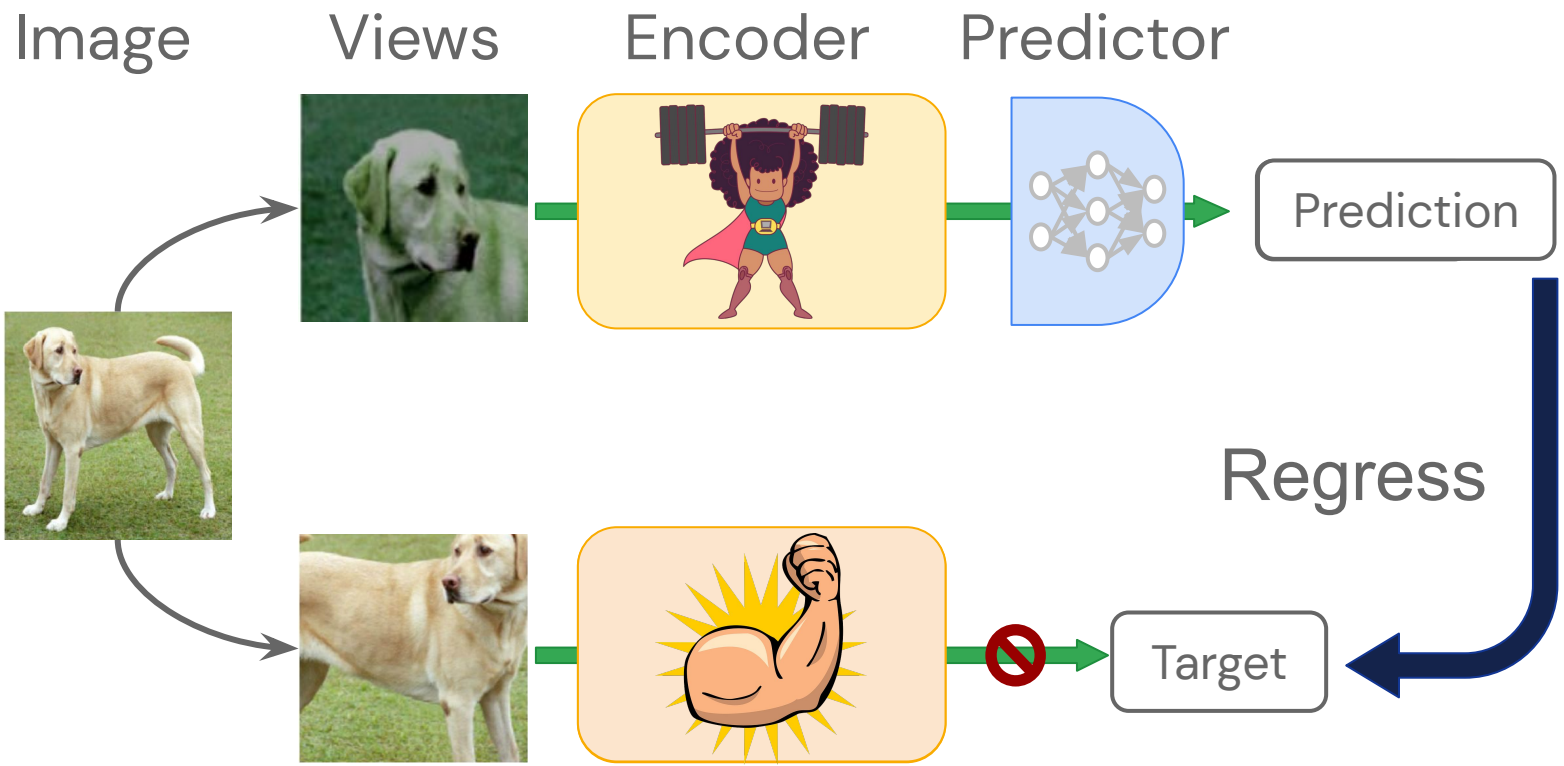
BYOL main intuition



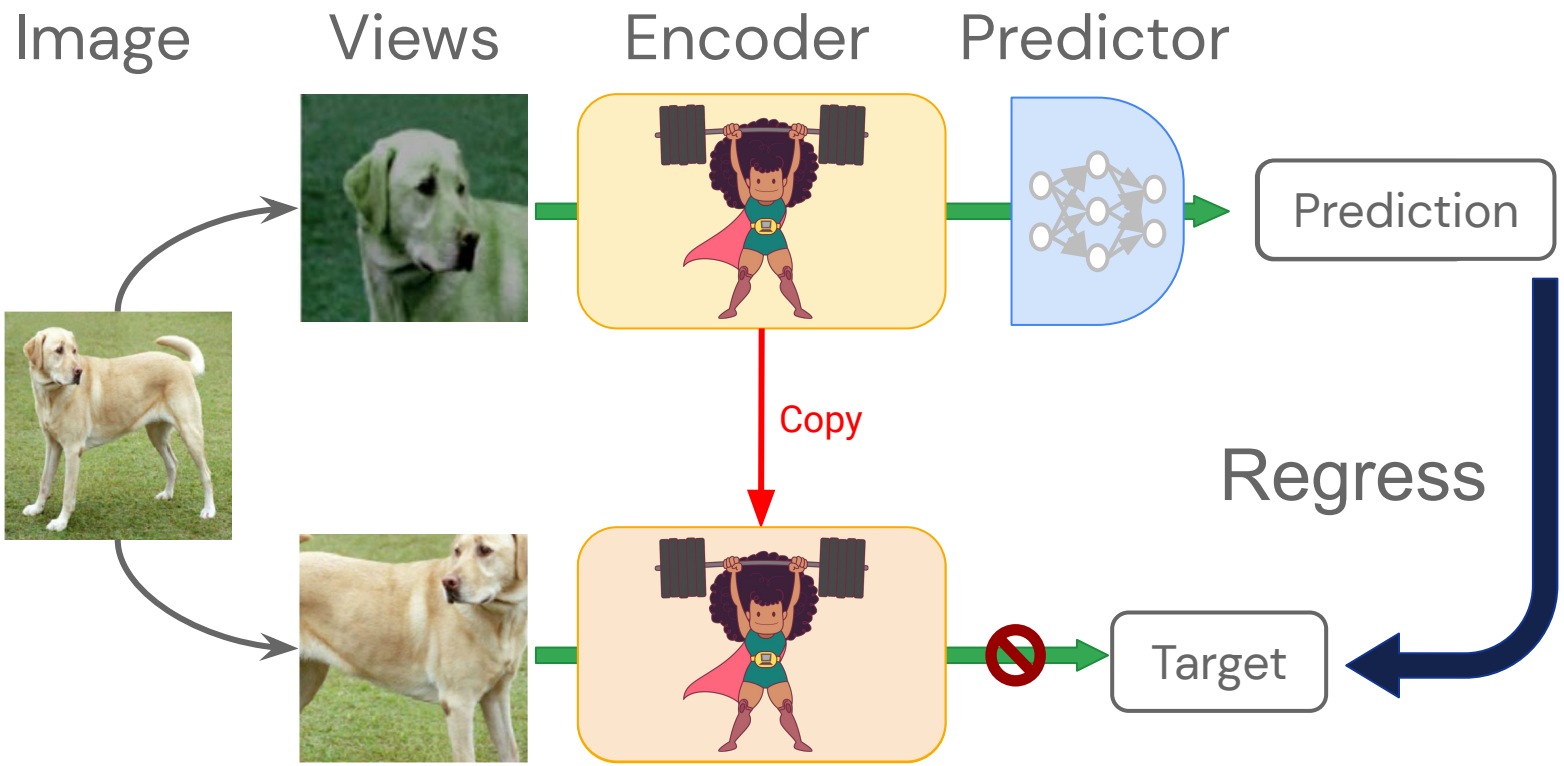
BYOL main intuition



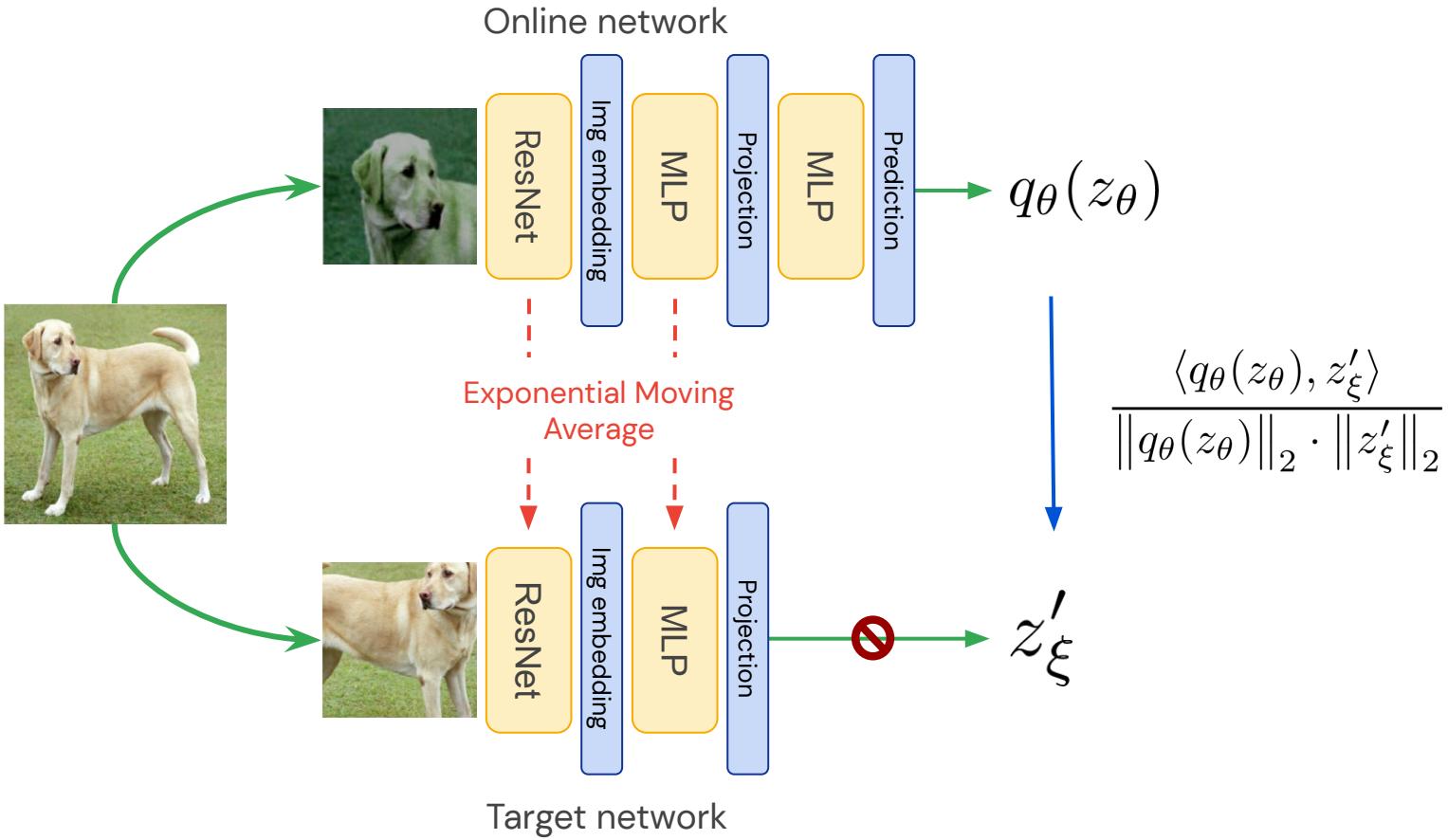
BYOL main intuition



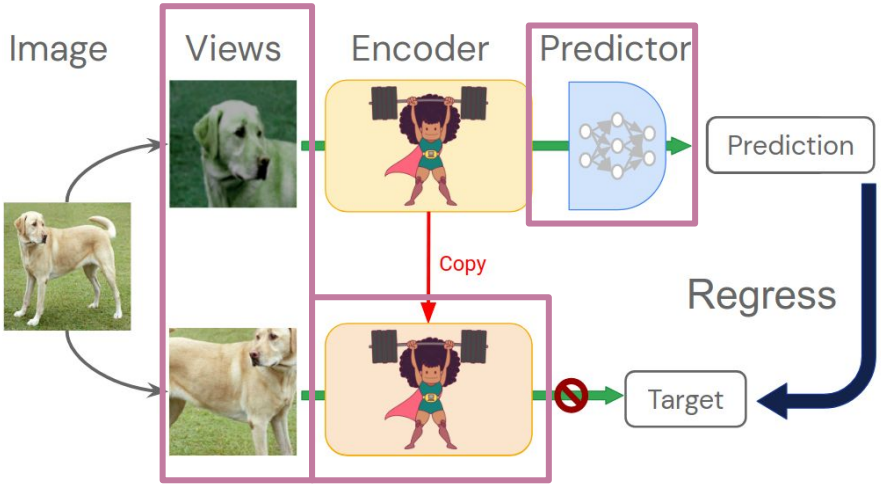
BYOL main intuition



BYOL Architecture



BYOL's highlight



Key ingredients:

- Image transformations.
- Target network.
- Additional predictor on top of online network.

Interest of the method:

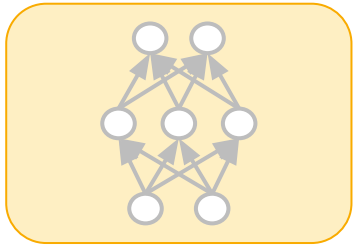
- Simple training procedure.
- No negative examples.
- Work at the embedding level, e.g. no-pseudo labels.

DeepMind

3 Performance

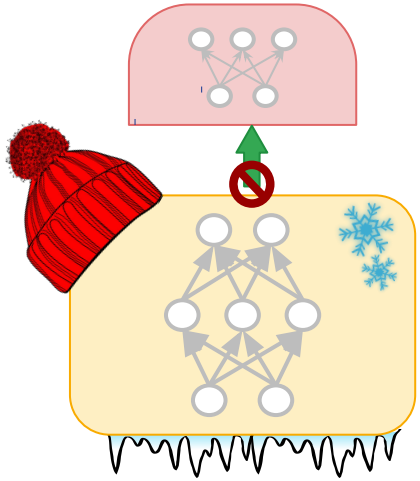
Linear Evaluation Protocol on ImageNet

Step 1: Train a “representation” on ImageNet without any labels.



ResNet

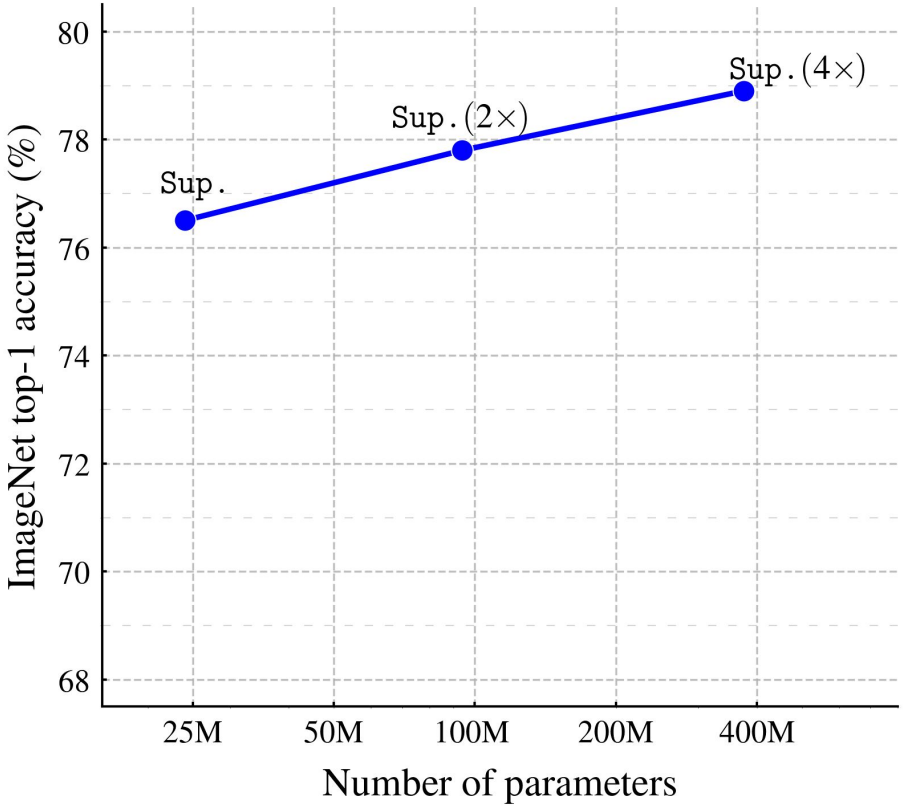
Step 2: On top of the frozen representation, train a linear classifier on ImageNet with label information.



Linear Classifier

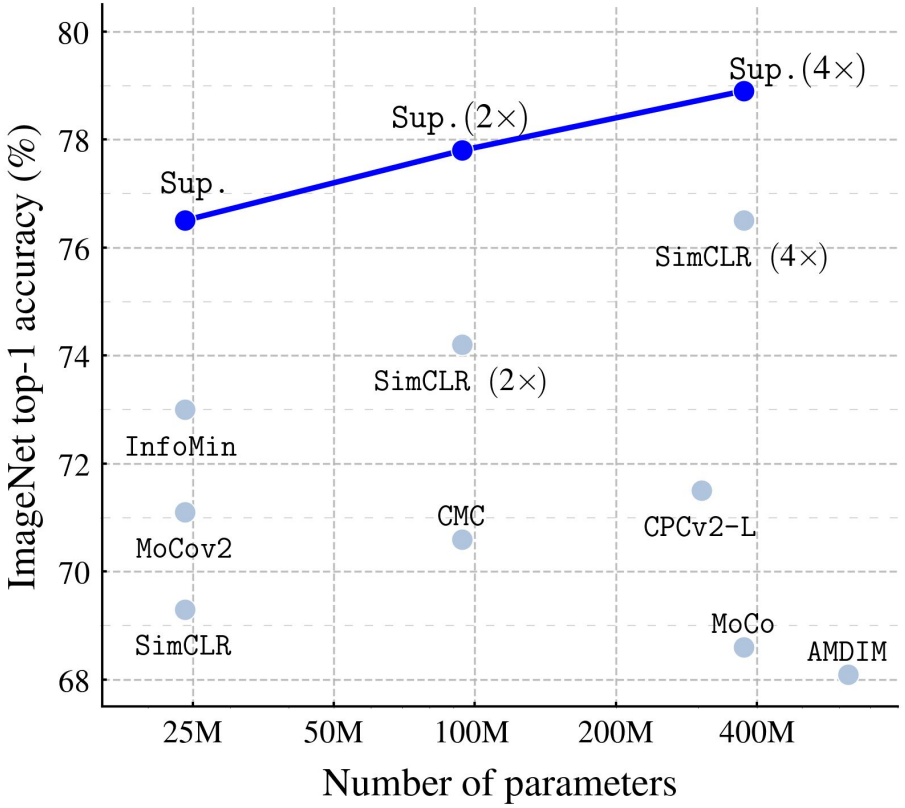
ResNet

Linear Evaluation Performance on ImageNet



Note: these supervised baselines are from SimCLR (Chen & Hinton, ICML 2020)

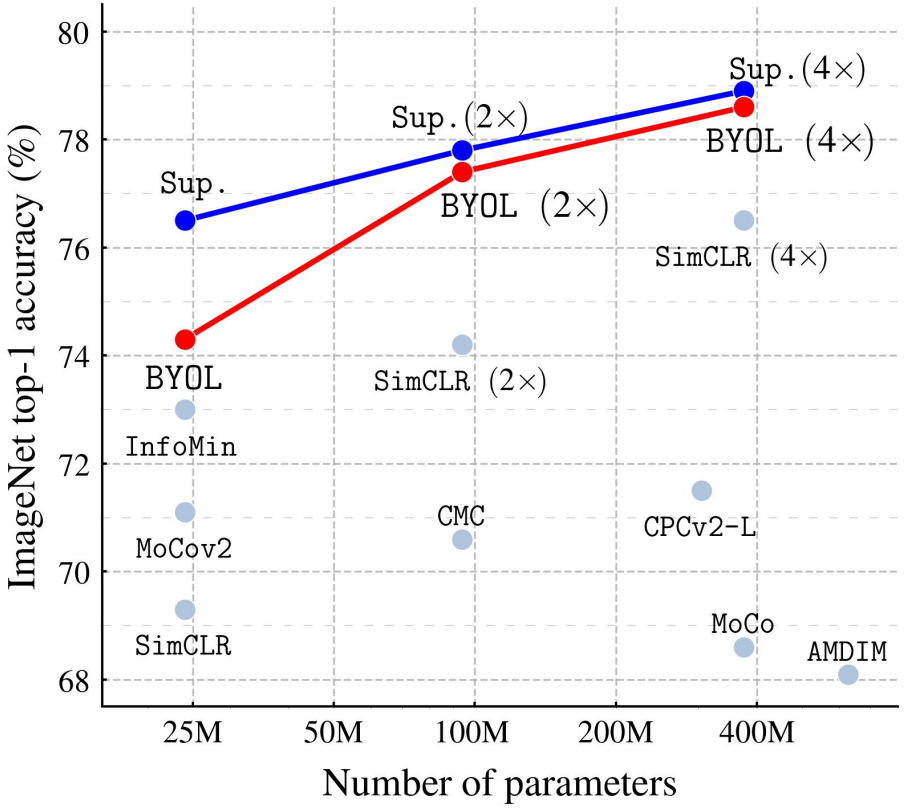
Linear Evaluation Performance on ImageNet



Note: these supervised baselines are from SimCLR (Chen & Hinton, ICML 2020)

CPCv2: van den Oord et al., *Representation learning with contrastive predictive coding*. 2018
AMDIM: Bachman et al., *Learning representations by maximizing mutual information across views*. 2019
CMC: Tian et al.,*Contrastive multiview coding*. 2019.
MoCo: He et al., *Momentum contrast for unsupervised visual representation learning*. 2019
InfoMin: Tian et al., *What makes for good views for contrastive learning*. 2020
MoCov2: Jain et al., *Improved baselines with momentum contrastive learning*. 2020
SimCLR: Chen et al., *A simple framework for contrastive learning of visual representations*. 2020

Linear Evaluation Performance on ImageNet



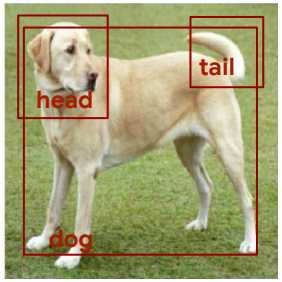
Note: these supervised baselines are from SimCLR (Chen & Hinton, ICML 2020)

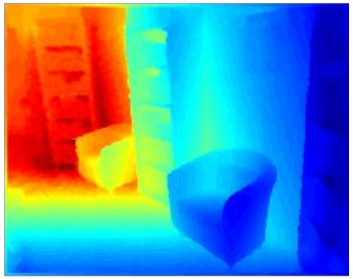
CPCv2: van den Oord et al., *Representation learning with contrastive predictive coding*. 2018
AMDIM: Bachman et al., *Learning representations by maximizing mutual information across views*. 2019
CMC: Tian et al.,*Contrastive multiview coding*. 2019.
MoCo: He et al., *Momentum contrast for unsupervised visual representation learning*. 2019
InfoMin: Tian et al., *What makes for good views for contrastive learning*. 2020
MoCov2: Jain et al., *Improved baselines with momentum contrastive learning*. 2020
SimCLR: Chen et al., *A simple framework for contrastive learning of visual representations*. 2020

Transfer Results

Semantic segmentation and object detection:

Method	AP_{50}	mIoU
Supervised-IN[1]	74.4	74.4
MoCo[1]	74.9	72.5
SimCLR (repro)	75.2	75.2
BYOL (ours)	**77.5**	**76.3**





Depth estimation:

Method	Higher better pct.<1.25	pct.$<1.25^2$	pct.$<1.25^3$	Lower better rms	rel
Supervised-IN[1]	81.1	95.3	98.8	0.573	**0.127**
SimCLR (repro)	83.3	96.5	99.1	0.557	0.134
BYOL (ours)	**84.6**	**96.7**	**99.1**	**0.541**	0.129

[1] He, Kaiming, et al. "Momentum contrast for unsupervised visual representation learning." *CVPR*. 2020.

Further comparison with SimCLR

BYOL outperforms other self-supervised learning methods on the following benchmarks:

- Semi-supervised learning on ImageNet
- Fine-tuning on small classification datasets (such as CIFAR or Flowers)
- Transfer tasks when pretraining on Places365 instead of ImageNet

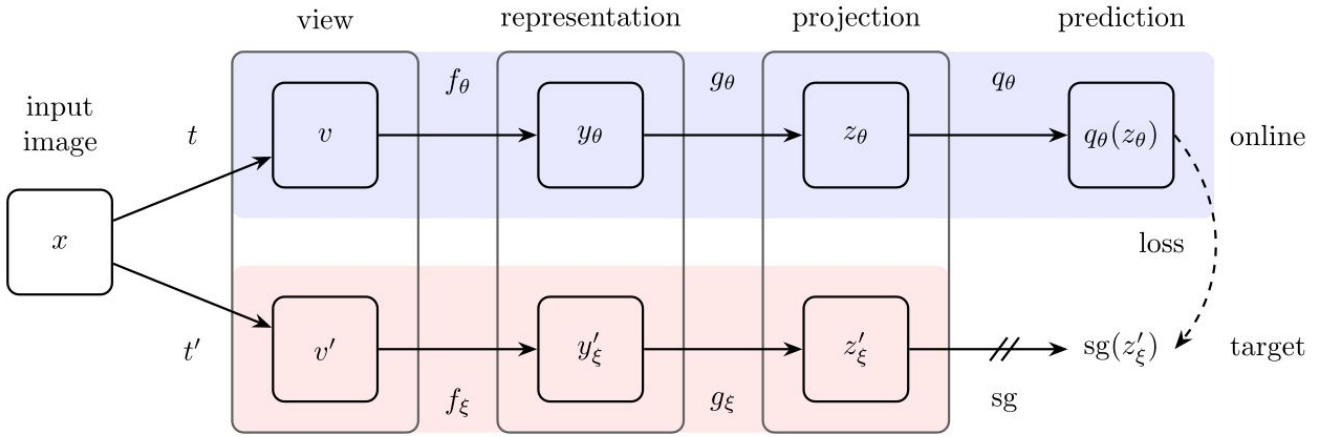
BYOL vs. Contrastive methods:

- BYOL is less sensitive to the choice of image transformations
- BYOL is more robust to smaller batch sizes

DeepMind

4 Building intuitions

Is BYOL optimizing a flawed objective?



$$\mathcal{L}_{\theta,\xi} := \|\bar{q}_\theta(z_\theta) - \bar{z}'_\xi\|^2 = 2 - 2 \cdot \frac{\langle q_\theta(z_\theta), z'_\xi \rangle}{\|q_\theta(z_\theta)\|_2 \cdot \|z'_\xi\|_2}$$

Is BYOL optimizing a flawed objective?

$$\mathcal{L}_{\theta,\xi} := \|\bar{q}_\theta(z_\theta) - \bar{z}'_\xi\|^2 = 2 - 2 \cdot \frac{\langle q_\theta(z_\theta), z'_\xi \rangle}{\|q_\theta(z_\theta)\|_2 \cdot \|z'_\xi\|_2}$$

This objective has **trivial global minima** in the form of **collapsed constant projections and predictions**. But

$$\xi_{t+1} = \underbrace{(1-\eta)\xi_t + \eta\theta_{t+1}}_{\text{EMA update}} \neq \underbrace{\xi_t - \alpha\nabla_\xi \mathcal{L}_{\theta_t,\xi_t}}_{\text{GD update}}$$

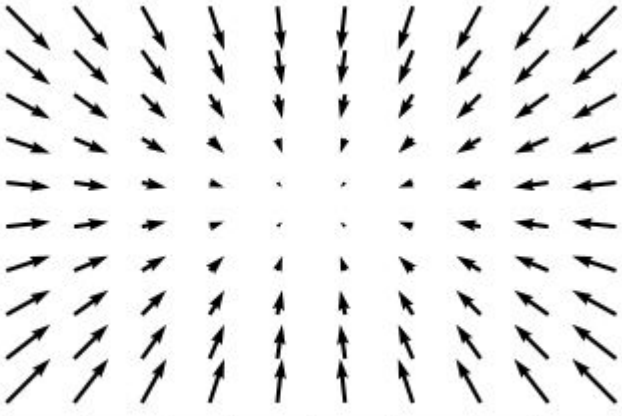
BYOL is not optimizing $\mathcal{L}_{\theta,\xi}$

Is BYOL optimizing a flawed objective?

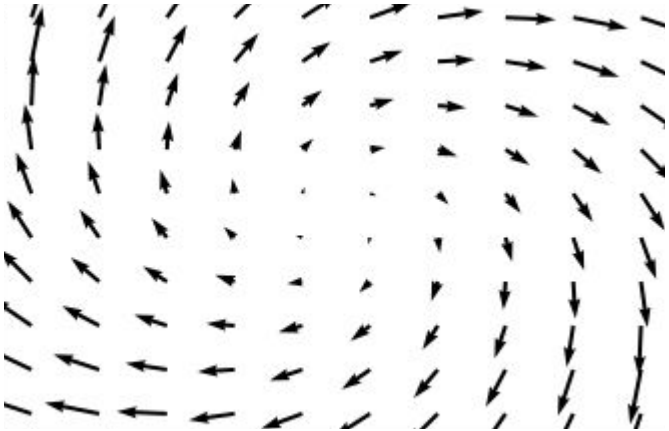
BYOL (probably) does **not solve an optimization problem**:

- No notion of **global/local optima** just **equilibria** of the dynamic (think GANs)
  → No convergence guarantees...

- Constant representations are **equilibria** but may not be **stable** or **attractive**.



V.S.



Is BYOL 'batch implicit contrastive'?

~~untitled blog~~

Understanding self-supervised and contrastive learning with "Bootstrap Your Own Latent" (BYOL)

Aug 24, 2020 • Abe Fetterman (email), Josh Albrecht (email)

Why batch normalization is imp
are compared to the mode

Unlike prior work like SimCLR and MoCo,

Removing all batch normalizat
unless at least one technique is used to prevent mode collapse.

work on reproducing BYOL:

(1) BYOL generally performs no better than random when batch normalization is removed, and

(2) the presence of batch normalization implicitly causes a form of contrastive learning.

These findings highlight the importance of contrast between positive and negative examples when learning representations and help us gain a more fundamental understanding of how and why self-supervised learning works.

The code used for this post can be found at https://github.com/untitled-ai/self_supervised.

Why does self-supervised learning matter?

Are batch statistics indeed crucial to make BYOL work?

UNDERSTANDING SELF-SUPERVISED LEARNING WITH DUAL DEEP NETWORKS

Yuandong Tian[1] **Lantao Yu**[2*] **Xinlei Chen**[1] **Surya Ganguli**[1,2]

[1]Facebook AI Research
{yuandong,xinleic}@fb.com

[2]Stanford University
{lantaoyu@cs.,sganguli@}standford.edu

elf-supervised learning
e.g., SimCLR, BYOL).
eights at each layer are
es initial random selec-
tivities that vary across data samples but survive averages over data augmentations,
which we show leads to the emergence of hierarchical features, if the input data
are generated from a hierarchical latent tree model. With the same framework,
we also show analytically that BYOL works due to an implicit contrastive term,
play between the zero-mean operation of BatchNorm and the extra predictor in
the online network. Extensive ablation studies justify our theoretical findings.

1 INTRODUCTION

While self-supervised learning (SSL) has achieved great empirical success across multiple domains, including computer vision (He et al., 2020; Goyal et al., 2019; Chen et al., 2020a; Grill et al., 2020; Misra and Maaten, 2020; Caron et al., 2020), natural language processing (Devlin et al., 2018), and speech recognition (Wu et al., 2020; Baevski and Mohamed, 2020; Baevski et al., 2019), its theoretical understanding remains elusive, especially when multi-layer nonlinear deep networks are involved. Unlike supervised learning (SL) that deals with labeled data, SSL learns meaningful

BYOL works even without batch statistics

Result 1: BYOL indeed performs very poorly when all BN are removed (projection + prediction + encoder).

Hypothesis: BN provides a good init, doubly crucial for BYOL, both for optim and for providing good initial targets.

Experiments to test hypothesis: Can we recover perf with better inits and no batch statistics.

Result 2: BYOL does not collapse and works well with **better initialization**.

Result 3: BYOL with **GroupNorm** and **WeightStandardization** (no batch stats) performs the same as BYOL with **BatchNorm**.

BYOL variant	Vanilla BN	No BN	Modified init.	GN + WS
Uses batch statistics	Yes	No	No	No
Top-1 accuracy (%)	74.3	0.1	65.7	73.9

What factors prevent collapsing?

ImageNet top-1 accuracy @300 epochs

Base BYOL	**72.5**
- Remove predictor*	**Collapse**
- Remove EMA of target network (and keep the stop gradient)	**Barely learns**
- Add explicit negative examples	**72.7**

Remark: BYOL **without predictor** → Mean Teacher[1] but without supervised signal.

[1]Mean Teacher: Tarvainen et al., *Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, 2017.*

How crucial is BYOL's predictor?

Results:

- BYOL **without predictor** → **Mean Teacher**, collapses without supervised signal.
- BYOL **with near optimal predictor** → Works without target network:
  - Optimal linear predictor on the batch → 45% top1 accuracy
  - Increased predictor learning rate (λ ratio of learning rates):

λ	Top-1
0	0.01
1	5.5
2	$62.8_{\pm 1.5}$
10	66.6
20	$66.3_{\pm 0.3}$
Baseline	72.5

$$\lambda = \frac{\text{predictor lr}}{\text{network lr}}$$

H: Near optimal predictor is key.

How crucial is BYOL's predictor?

Optimal predictor: Conditional expectation of targets w.r.t. online

$$q^*(z_\theta) = \mathbb{E}\left[z'_\xi \mid z_\theta\right]$$

Online projection objective: Conditional variance of targets w.r.t. Online

$$\mathcal{L}_{\theta,\xi} = \mathrm{Var}\left[z'_\xi \mid z_\theta\right]$$

z'_ξ: Target projection

z_θ: Online projection

To reduce conditional variance:

- **Collapse target** representation.
- **Increase** information in the online projection.

BYOL only plays on second one, (stop gradient in targets)

→ Always **increase the variability** of online projections!

DeepMind

Thank you!

The code and checkpoints are available:
https://github.com/deepmind/deepmind-research

Follow-up work on BYOL and BatchNorm:
https://arxiv.org/abs/2010.10241