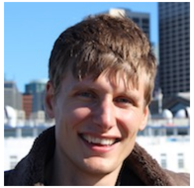
Barlow Twins: Self-Supervised Learning via Redundancy Reduction


Jure Zbontar*


Li Jing*


Ishan Misra


Yann LeCun


Stephane Deny

* equal contribution

ICML'21

FACEBOOK AI

Background

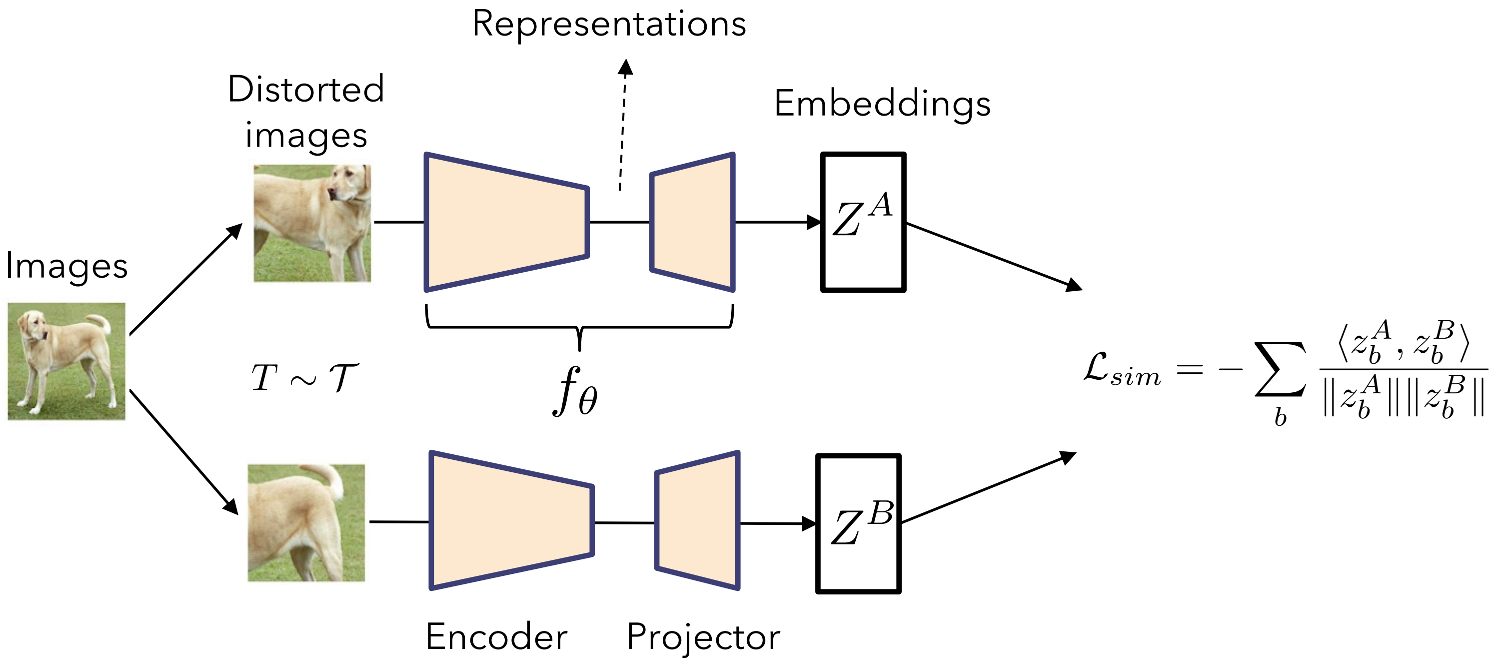
Supervised learning is limited by the amount of labeled data.

Self-supervised learning learns useful representations from intrinsically generated labels.

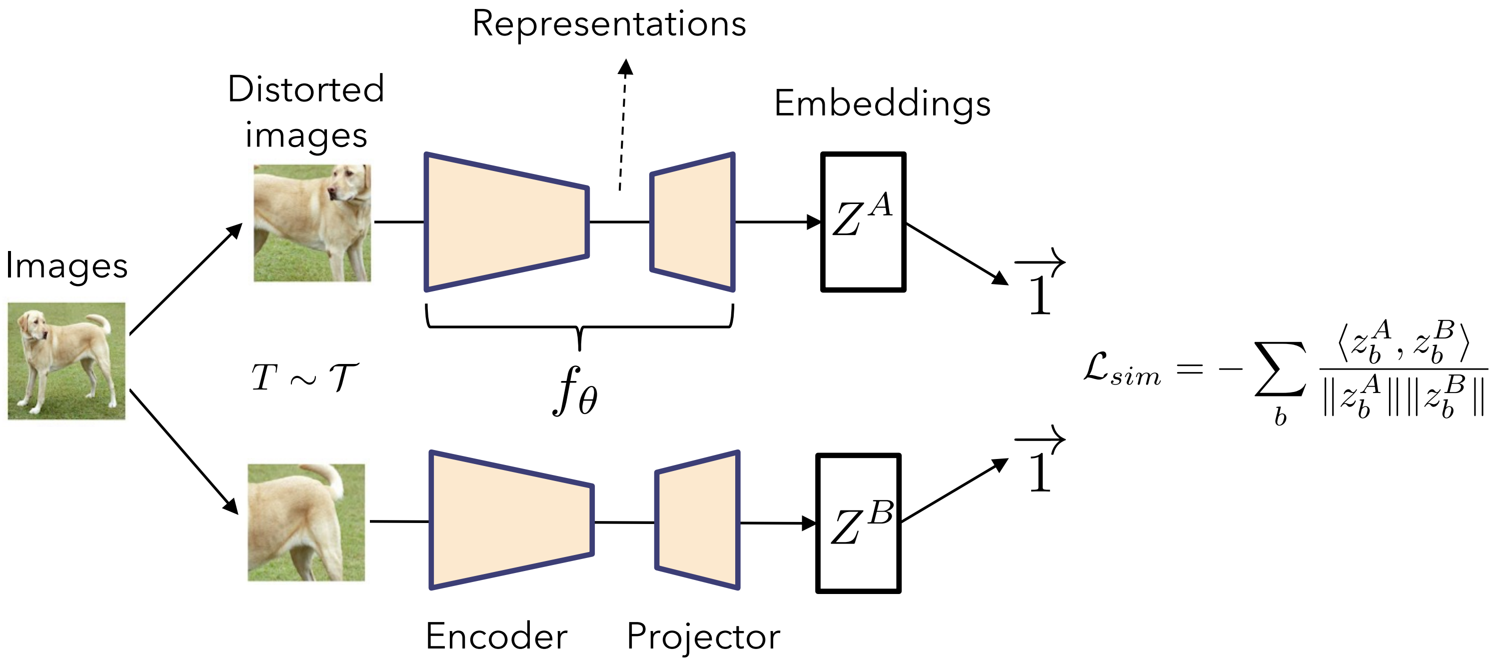
Self-supervised learning has shown success in the NLP domain

e.g. BERT, GPT3

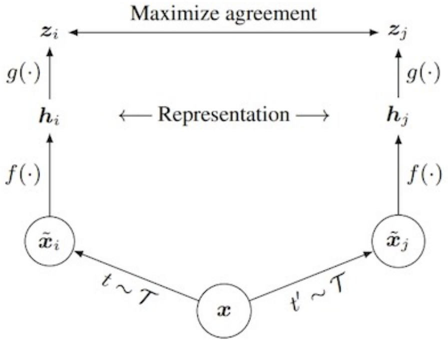
Self-supervised Learning via Joint Embedding Distortions



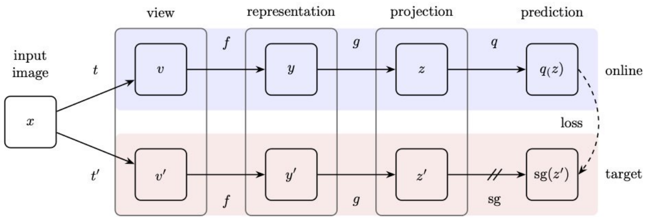
Collapsing Problem



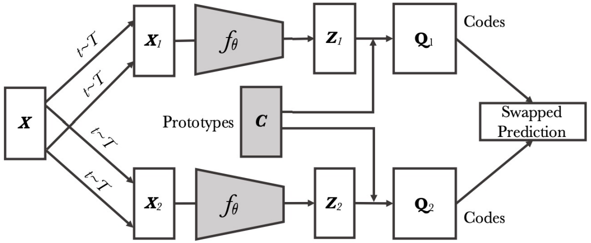
Existing Solutions



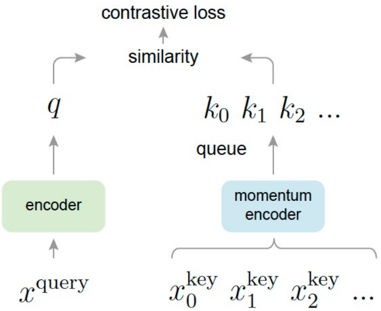
SimCLR (T. Chen 2020)



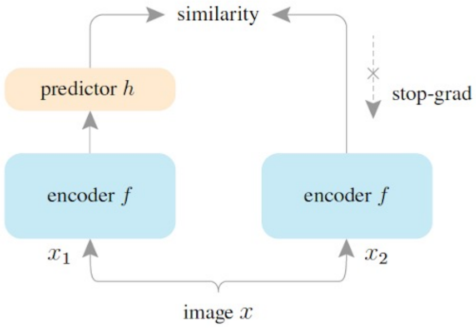
BYOL (J. Grill 2020)



SwAV (M. Caron 2020)

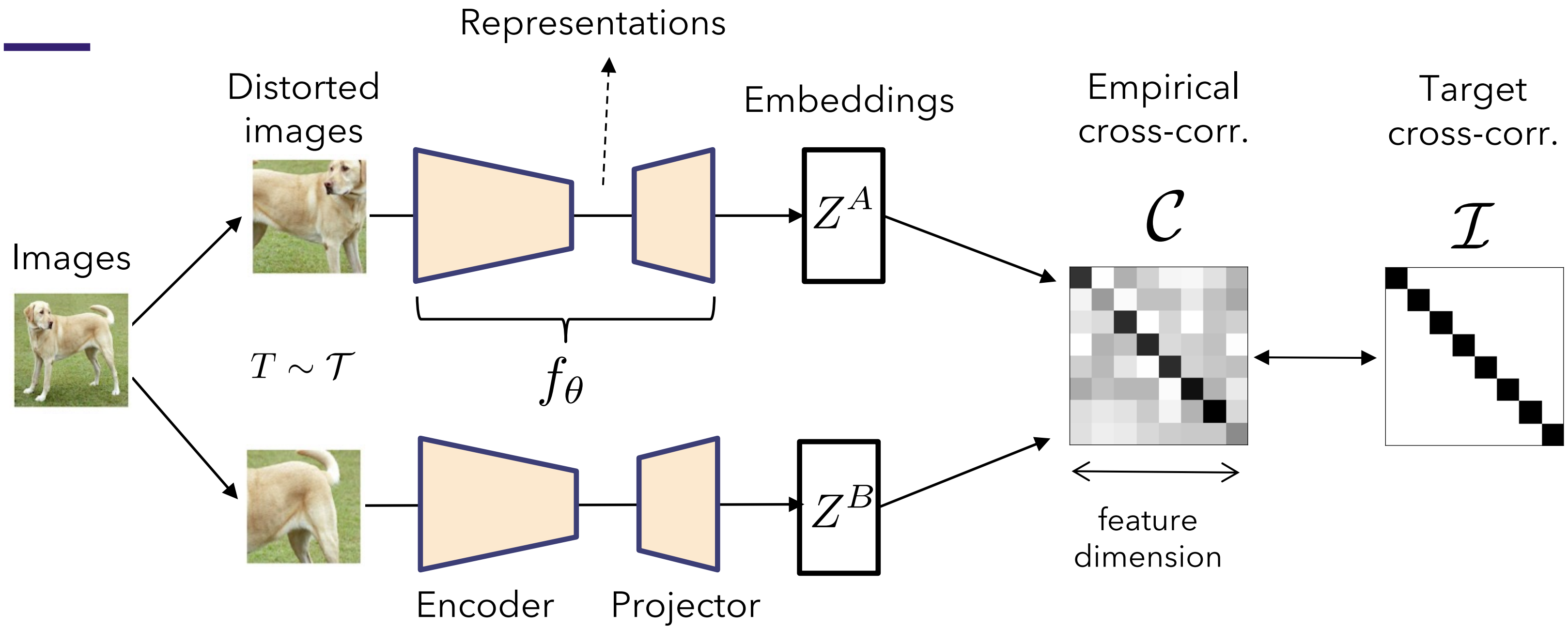


MoCo (K. He 2019)



SimSiam (X. Chen 2020)

Self-supervised Learning via Redundancy Reduction



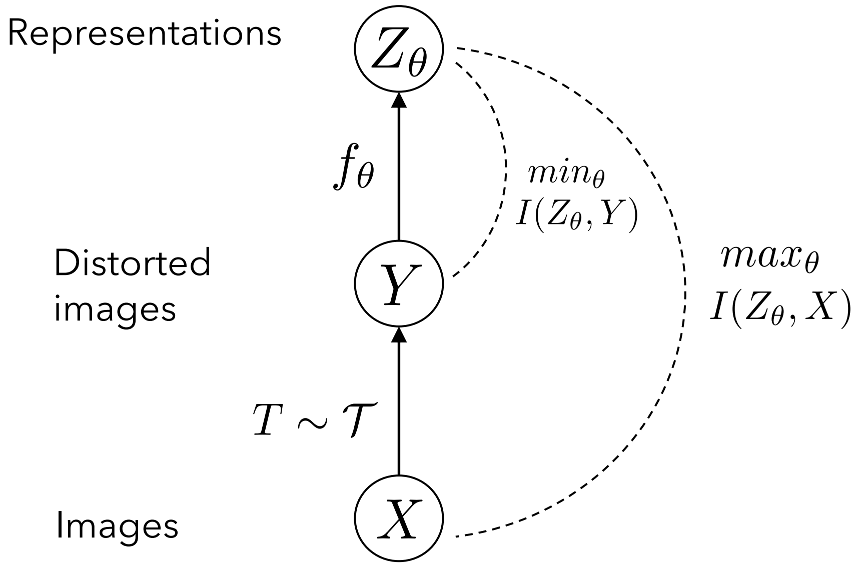
$$\mathcal{L}_{\mathcal{BT}} \triangleq \underbrace{\sum_i (1-\mathcal{C}_{ii})^2}_{\text{invariance term}} + \lambda \underbrace{\sum_i \sum_{j\neq i} {\mathcal{C}_{ij}}^2}_{\text{redundancy reduction term}}$$

$$\mathcal{C}_{ij} \triangleq \frac{\sum_b z^A_{b,i} z^B_{b,j}}{\sqrt{\sum_b \left(z^A_{b,i}\right)^2}\sqrt{\sum_b \left(z^B_{b,j}\right)^2}}$$

Remarks

Our method is related to information bottleneck: maximize invariant information while minimizing redundant part

$$\mathcal{IB}_\theta \triangleq I(Z_\theta, Y) - \beta I(Z_\theta, X)$$



ImageNet Classification

Table 1. **Top-1 and top-5 accuracies (in %) under linear evaluation on ImageNet.** All models use a ResNet-50 encoder. Top-3 best self-supervised methods are in underlined.

Method	Top-1	Top-5
Supervised	76.5	
MoCo	60.6	
PIRL	63.6	-
SimCLR	69.3	89.0
MoCo v2	71.1	90.1
SimSiam	71.3	-
SwAV	71.8	-
BYOL	<u>74.3</u>	91.6
SwAV (w/ multi-crop)	<u>75.3</u>	-
Barlow Twins (ours)	<u>73.2</u>	91.0

Linear Probe

Table 2. **Semi-supervised learning on ImageNet** using 1% and 10% training examples. Results for the supervised method are from (Zhai et al., 2019). Best results are in **bold**.

Method	Top-1		Top-5	
	1%	10%	1%	10%
Supervised	25.4	56.4	48.4	80.4
PIRL	-	-	57.2	83.8
SimCLR	48.3	65.6	75.5	87.8
BYOL	53.2	68.8	78.4	89.0
SwAV (w/ multi-crop)	53.9	**70.2**	78.5	**89.9**
Barlow Twins (ours)	**55.0**	69.7	**79.2**	89.3

Semi-supervised Learning

Transfer Learning

Table 3. **Transfer learning: image classification.** We benchmark learned representations on the image classification task by training linear classifiers on fixed features. We report top-1 accuracy on Places-205 and iNat18 datasets, and classification mAP on VOC07. Top-3 best self-supervised methods are underlined.

Method	Places-205	VOC07	iNat18
Supervised	53.2	87.5	46.7
SimCLR	52.5	85.5	37.2
MoCo-v2	51.8	<u>86.4</u>	38.6
SwAV	52.8	86.4	39.5
SwAV (w/ multi-crop)	<u>56.7</u>	<u>88.9</u>	<u>48.6</u>
BYOL	<u>54.0</u>	<u>86.6</u>	<u>47.6</u>
BARLOW TWINS (ours)	<u>54.1</u>	86.2	<u>46.5</u>

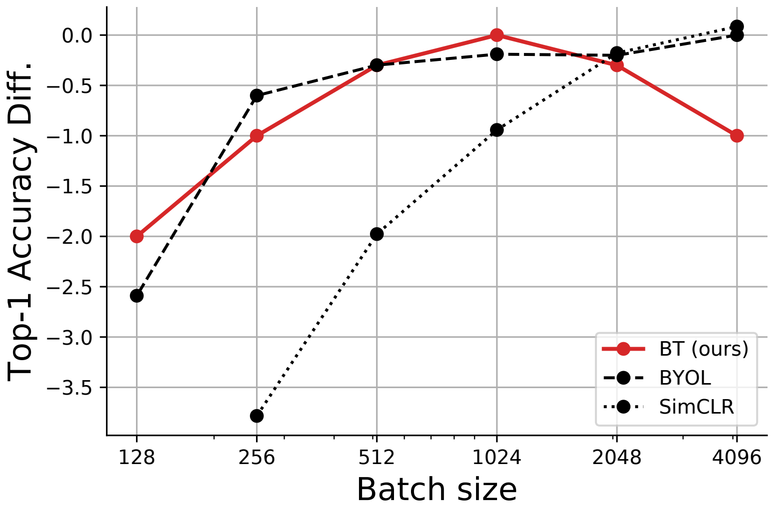
Image Classification

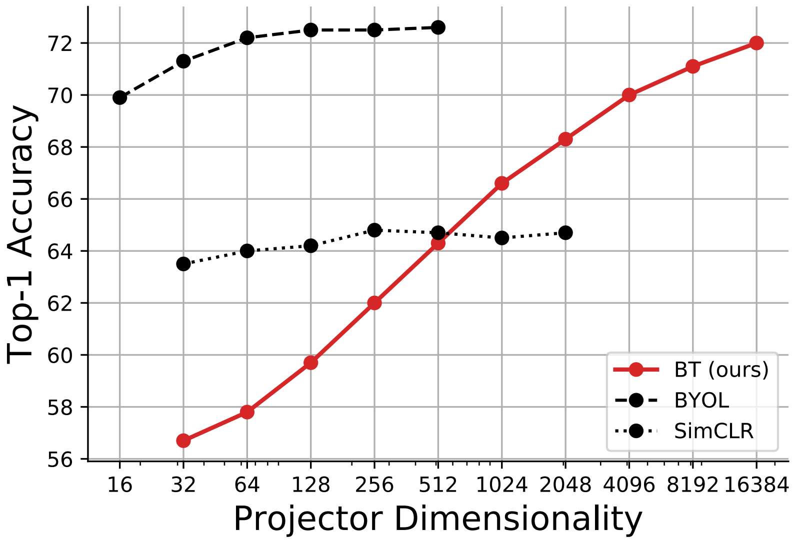
Table 4. **Transfer learning: object detection and instance segmentation.** We benchmark learned representations on the object detection task on VOC07+12 using Faster R-CNN (Ren et al., 2015) and on the detection and instance segmentation task on COCO using Mask R-CNN (He et al., 2017). All methods use the C4 backbone variant (Wu et al., 2019) and models on COCO are finetuned using the $1\times$ schedule. Best results are in **bold**.

Method	VOC07+12 det			COCO det			COCO instance seg		
	AP_{all}	AP_{50}	AP_{75}	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
Sup.	53.5	81.3	58.8	38.2	58.2	41.2	33.3	54.7	35.2
MoCo-v2	**57.4**	82.5	**64.0**	**39.3**	58.9	**42.5**	**34.4**	55.8	36.5
SwAV	56.1	**82.6**	62.7	38.4	58.6	41.3	33.8	55.2	35.9
SimSiam	57	82.4	63.7	39.2	**59.3**	42.1	**34.4**	**56.0**	**36.7**
BT (ours)	56.8	**82.6**	63.4	39.2	59.0	**42.5**	34.3	**56.0**	36.5

Object Detection

Ablation Study





Ablation Study

Table 5. **Loss function explorations**. We ablate the invariance and redundancy terms in our proposed loss and observe that both terms are necessary for good performance. We also experiment with different normalization schemes and a cross-entropy loss and observe reduced performance.

Loss function	Top-1	Top-5
Baseline	71.4	90.2
Only invariance term (on-diag term)	57.3	80.5
Only red. red. term (off-diag term)	0.1	0.5
Normalization along feature dim.	69.8	88.8
No BN in MLP	71.2	89.7
No BN in MLP + no Normalization	53.4	76.7
Cross-entropy with temp.	63.3	85.7

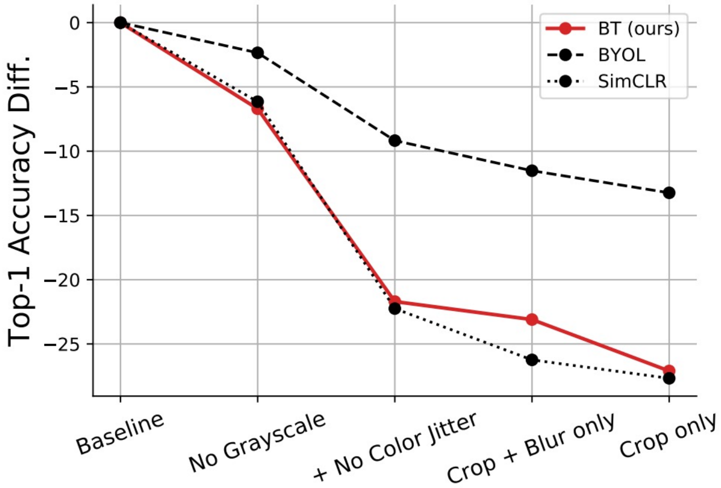


Figure 3. Effect of progressively removing data augmentations. Data for BYOL and SIMCLR (repro) is from (Grill et al., 2020) fig 3b.

Conclusion

We propose a new method for self-supervised learning based on the redundancy reduction principle

Our method does not depend on explicit negative terms or asymmetric architectures

Our method performs on par with other state-of-the-art methods

Thank you!