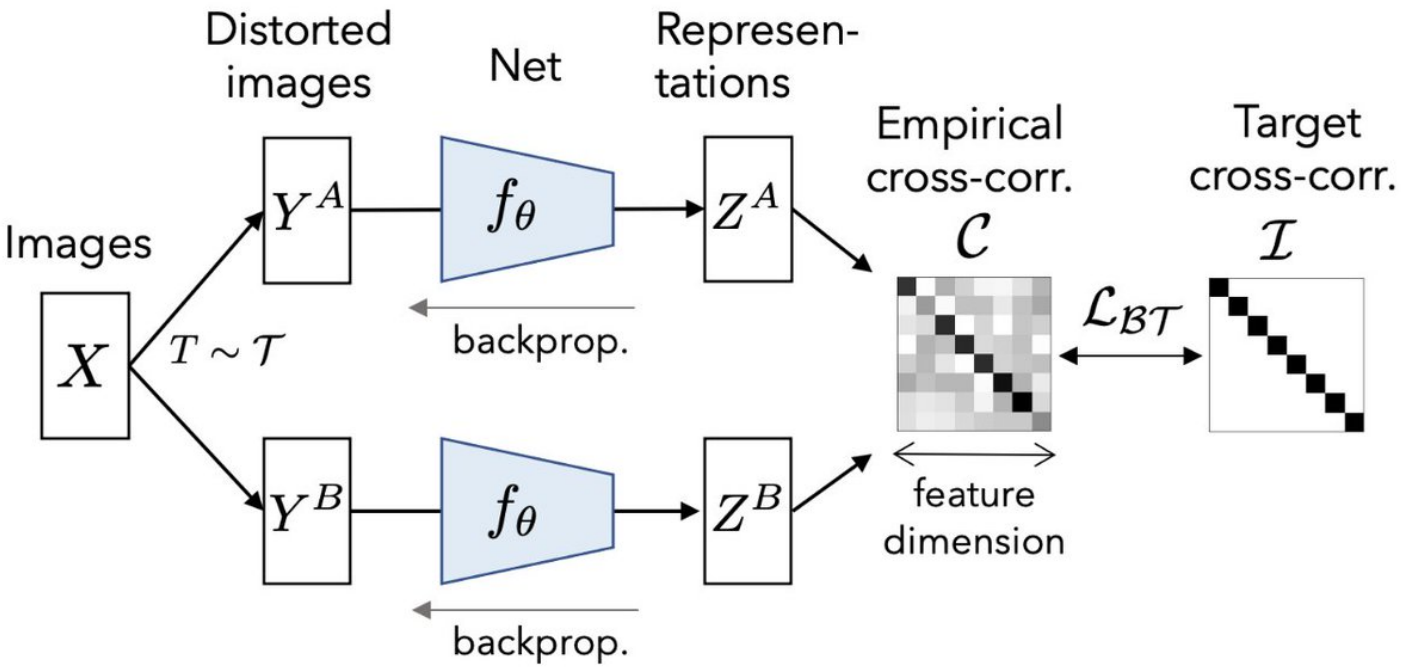
Twitter Thread by Yann LeCun

Yann LeCun
@ylecun

Barlow Twins: a new super-simple self-supervised method to train joint-embedding architectures (aka Siamese nets) non

Basic idea: maximize the normalized correlation between a variable in the left branch and the same var in the right branch, while making the normalized cross-correlation between one var in the left branch and all other vars in the right branch as close to zero as possible.
2/N



In short: the loss tries to make the normalized cross-correlation between the embedding vectors coming out of the left branch and the right branch as close to the identity matrix as possible.
3/N

The 2 branches are always fed with differently-distorted version of the same image, and there is no need for dissimilar training pairs.

The objective makes the embedding vectors of the two branches as similar as possible, while maximizing their information

content.
4/N

No contrastive samples, no huge batch size (optimal is 128), nor predictor, no moving-average weights, no vector quantization, nor cut gradients in one of the branches.
5/N

Competitive results on ImageNet with a linear classifier head.
Great results on semi-supervised ImageNet in the low labeled-data regime and on transfer tasks.
6/N

Results on ImageNet with linear classifier head
7/N

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 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 (w/o multi-crop)	71.8	-
BYOL	74.3	91.6
SwAV	75.3	-
Barlow Twins (ours)	73.2	91.0

Results with 1% and 10% of ImagNet labeled images
8/N

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	53.9	**70.2**	78.5	**89.9**
BARLOW TWINS (ours)	**55.0**	69.7	**79.2**	89.3

Results on transfer tasks.

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× 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

Arch is standard ResNet50 with 2048-D feature vec.
But contrary to others, the embedding size (projector output) is larger. The perf keeps going up as the embedding dim grows (we stopped at 16384).
Probably cause the feature vars are made independent, not just decorrelated.
10/N

Why Barlow? Horace Barlow was a pioneer of visual neuroscience who proposed the idea that the brain tries to minimize redundancy in representations.

By Jure Zbontar, Li Jing, Ishan Misra, yours truly, and Stéphane Deny.
All from FAIR.
To appear at ICML 2021
11/N

Don't you just hate slicing what would be a decent-size post into threaded thin tweets?
12/N

No, really. Don't you hate reading those long thread slices?
If you do, you could just read my Facebook post:
https://t.co/dQii7BEPQ5
13/N

N=13

Typo: optimal batch size is 1024, not 128.
14/13 (haha).