Solving Inefficiency of Self-supervised Representation Learning

Guangrun Wang[1,2] Keze Wang[4] Guangcong Wang[3] Philip H.S. Torr[2] Liang Lin[1*]

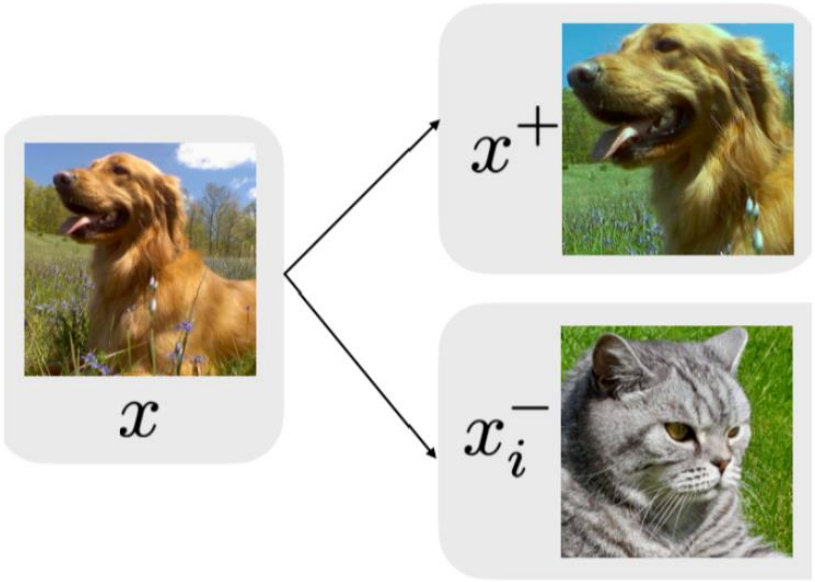
[1] Sun Yat-sen University [2] University of Oxford [3] Nanyang Technological University [4]DarkMatter AI Research

{wanggrun,wanggc3,kezewang}@gmail.com, philip.torr@eng.ox.ac.uk, linliang@ieee.org

ICCV2021

Introduce

contrastive learning



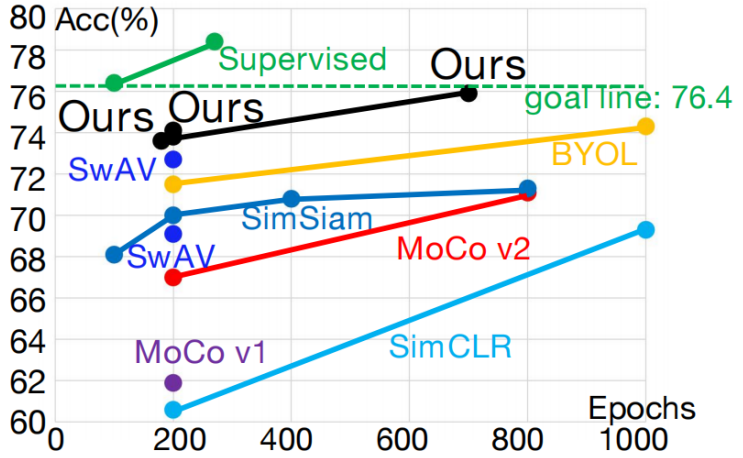


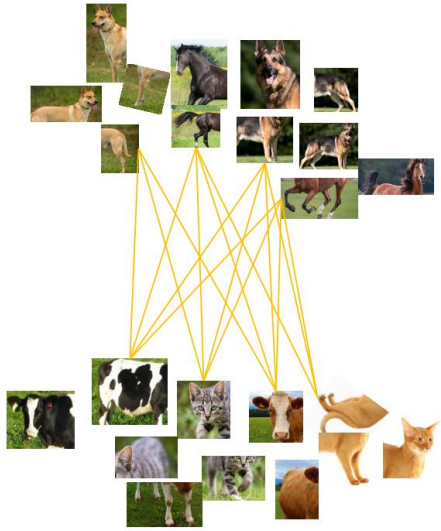
Figure 1. A comparison of learning efficiency among different SSL methods using ResNet-50. Here, the x-axis represents the training epochs of SSL, and the y-axis stands for the top-1 accuracy of ImageNet linear evaluation. All methods have lower learning efficiency than supervised learning, but our approach has a significantly higher learning efficiency than the existing SSL methods. (**best view in color**)

InfoNCE loss: $$-\log \frac{\exp(x^T x^+)/\tau}{\exp(x^T x^+)/\tau + \sum_{j=1}^{m} \exp(x^T x_j^-)/\tau}$$

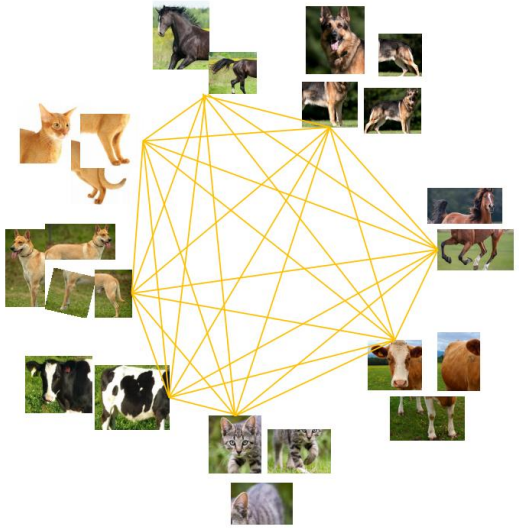
Introduce

under-clustering: lacking negative samples can make different object categories having overlaps

over-clustering: the model over-clusters samples of the same actual categories into different clusters



(a) under-clustering

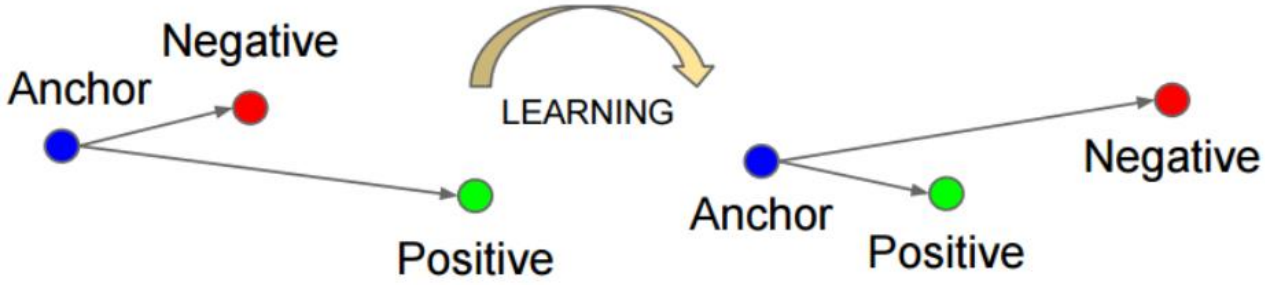


(c) over-clustering

(b) ideal clustering

Introduce

Triplet loss



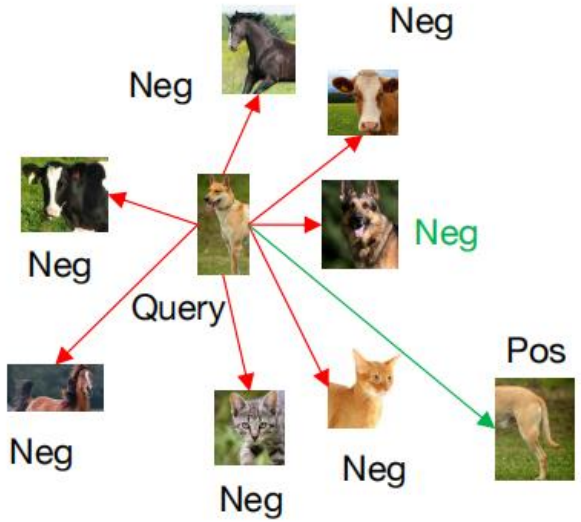
triplet set $\{(x_i, x_i^+, x_i^-)\}_{i=1,\cdots,m}$

$\{(x, x^+, x_i^-)\}_{i=1,\cdots,m}$

$$\mathcal{L}oss = \sum_{i=1}^{m} \max\left(d(x, x^+) - d(x, x_i^-), \mathcal{C}\right) \quad d(x, y) = -\frac{xy}{\|x\|_2 \|y\|_2}$$

C is a margin deciding whether or not to drop a triplet.

Method

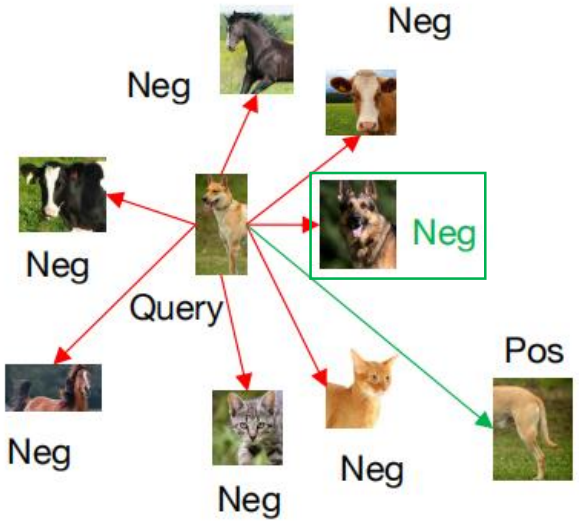


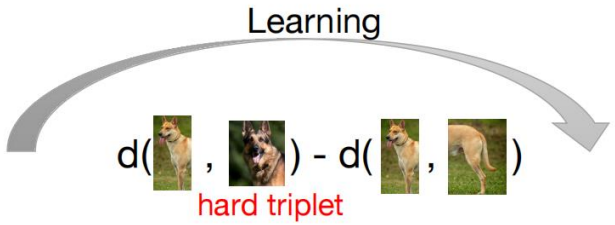
use the hardest triplet to represent the overall triplets, since $x_{hardest}^{-}$ is the hardest negative sample, we have d(x, $x_{hardest}^{-}$) ⩽ d(x, x_i^-) for all i.

$$\mathcal{L}oss = \max\left(d(x, x^+) - d(x, x_{hardest}^-), \mathcal{C}\right)$$

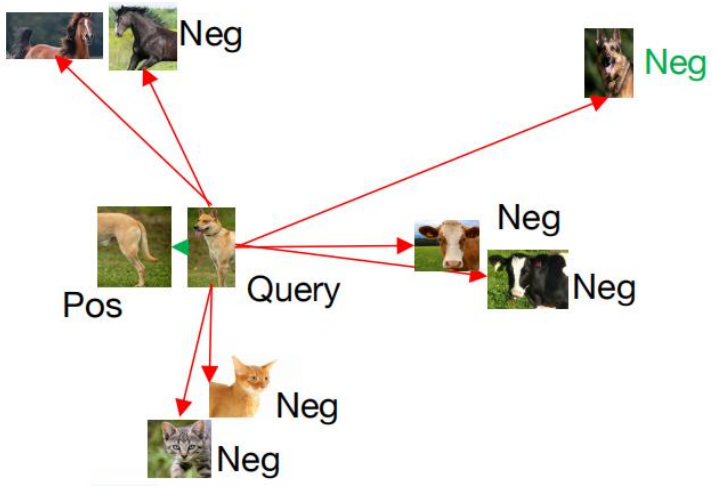
☑ under-clustering problem

Method





over-clustering problem

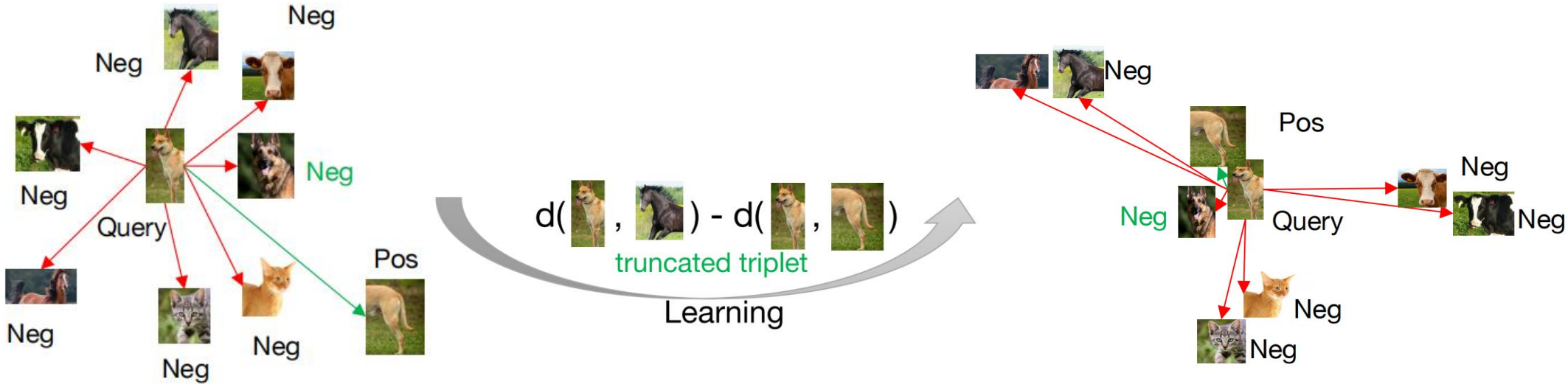


use the hardest triplet to represent the overall triplets, since $x_{hardest}^{-}$ is the hardest negative sample, we have d(x, $x_{hardest}^{-}$) ⩽ d(x, x_i^-) for all i.

$$\mathcal{L}oss = \max\left(d(x, x^+) - d(x, x_{hardest}^-), \mathcal{C}\right)$$

☑ under-clustering problem

Method



Truncated triplet loss $\mathcal{L}oss = \max\left(\gamma d(x, x^+) - d(x, x^-_{deputy}), \mathcal{C}\right)$

- rank-k triplet loss: the k-th element are selected from {d(x, x_i^-)}, yielding: d(x, x^-_{deputy}) = d(x, x^-_{rank-k})
- smoothed-rank-k triplet loss: selecting the top-2, top-3, ..., top-(2k+1) elements from {d(x, x_i^-)} and yielding: $d(x, x^-_{deputy}) = \frac{1}{2k}\sum_{j=2}^{2k+1} d(x, x^-_{rank-j})$ $\quad k = \frac{m}{2}$

Method

assumption: with a high probability, the image pairs from the same category have higher feature similarities than other pairs, and the distances between these pairs are smaller than other pairs.

event: the rank-k negative sample and the query belong to the same category.

Bernoulli Distribution model $\Pr = \sum_{j=k}^{m} \mathbf{C}_m^j p^j (1-p)^{m-j}$

p is used to denote the probability that a negative sample and the query belong to the same class

on ImageNet, we have $p = \frac{1}{1000}$

$m = 104, k = m/2 \quad Pr = 6.53e^{-121}$

$m = 104, k = 5 \quad Pr = 3.03e^{-94}$

Experiments

Linear evaluation on ImageNet:

Table 2. Top-1 accuracy and training epochs of state-of-the-art methods on ImageNet using linear classification for evaluation.

Method	top-1 acc.	train epochs
Random	4.4	0
Relative-Loc [13]	38.8	200
Rotation-Pred [19]	47.0	200
DeepCluster [5]	46.9	200
NPID [50]	56.6	200
ODC [53]	53.4	200
SimCLR [7]	60.6	200
SimCLR [7]	69.3	1000
MoCo [24]	61.9	200
MoCo v2 [8]	67.0	200
MoCo v2 [8]	71.1	800
SwAV [6] (single-crop)	69.1	200
SwAV [6] (multi-crop)	72.7	200
BYOL [22]	71.5	200
BYOL [22]	72.5	300
BYOL [22]	74.3	1000
SimSiam [9]	68.1	100
SimSiam [9]	70.0	200
SimSiam [9]	70.8	400
SimSiam [9]	71.3	800
truncated triplet	**73.6**	**180**
truncated triplet (smoothed)	**73.8**	**200**
truncated triplet (multi-crop)	**74.1**	**200**
truncated triplet	**75.9**	**700**
supervised	76.3	100
supervised + linear eval	74.1	100
supervised	78.4	270

Experiments

Transferring to downstream tasks:

(classification problem)

Table 3. Object detection results on COCO 2017 for Mask-RCNN.

Method	AP^{Box}	AP^{Mask}
Random	35.6	31.4
Relative-Loc [13]	40.0	35.0
Rotation-Pred [19]	40.0	34.9
NPID [50]	39.4	34.5
MoCo [24]	40.9	35.5
MoCo v2 [8]	40.9	35.5
SimCLR [7]	39.6	34.6
BYOL [22]	40.3	35.1
truncated triplet	**41.3**	**37.3**
supervised-100	40.0	34.7
supervised-270	42.0	37.7

Table 4. Object detection results on VOC07+12 for Faster-RCNN.

Method	$AP50^{Box}$	AP^{Box}	$AP75^{Box}$
Random	59.0	32.8	31.6
Relative-Loc [13]	80.4	55.1	61.2
Rotation-Pred [19]	80.9	55.5	61.4
NPID [50]	80.0	54.1	59.5
MoCo [24]	81.4	56.0	62.2
MoCo v2 [8]	**82.0**	**56.6**	**62.9**
SimCLR [7]	79.4	51.5	55.6
BYOL [22]	81.0	51.9	56.5
truncated triplet	81.8	56.4	**62.9**
supervised-100	81.6	54.2	59.8
supervised-270	82.2	56.9	63.1

Experiments

Transferring to downstream tasks:

(matching problem)

Table 5. Comparison with state-of-the-art methods on SYSU-30k.

supervision	method	rank-1
Transfer learning	DARI [44]	11.2
	DF [12]	10.3
	Local CNN [52]	23.0
	MGN [47]	23.6
Weakly supervised	W-Local CNN [46]	28.8
	W-MGN [46]	29.5
Self-supervised	SimCLR [7]	10.9
	MoCo v2 [8]	11.6
	BYOL [22]	12.7
	truncated triplet	**14.8**

Ablation study

event A: all the batch samplings

event B: If a batch contains at least two images belonging to the same actual category

event Ω: If (at least) these two images are mis-considered a false-negative pair

Table 6. Effect of avoiding over-clustering.

Training epoch	event	0	180
$k = 1$	$\Pr(\Omega\|\mathcal{A})$	0.1538	0.9656
	$\Pr(\Omega\|\mathcal{B})$	0.1618	0.9948
$k = 5$	$\Pr(\Omega\|\mathcal{A})$	0.1167	0.2105
	$\Pr(\Omega\|\mathcal{B})$	0.1230	0.2132
$k = 52$	$\Pr(\Omega\|\mathcal{A})$	0.1220	0.0110
	$\Pr(\Omega\|\mathcal{B})$	0.1288	0.0233

Ablation study

Table 7. Impact of margin.

Margin	$\mathcal{C} = -0.3$	$\mathcal{C} = -1.2$	$\mathcal{C} = -100$
Top-1 accuracy	28.3	29.8	30.0

Table 8. Impact of rank-k.

Rank-k	rank-1	rank-5	rank-52
Top-1 accuracy	28.9	29.5	30.0

Thank you