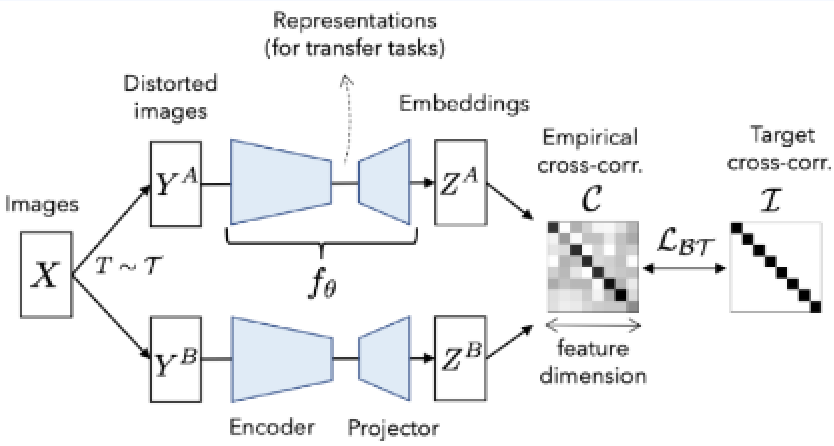
Barlow Twins

논문명 : Self-Supervised Learning via Redundancy Reduction

- Contrastive Learning
  - SimCLR
  - MoCo
  - Negative sampling에 대한 의존도가 큼.

- Non-contrastive learning
  - BYOL (momentum encoder + asymmetric)
  - Simsiam (Asymmetric)
  - 비대칭적인 구조를 이용해야 함. (다른 레이어에 적용할 때, prediction head를 추가해야 적용가능)

- Redundancy Reduction을 이용한 SSL
  - Redundancy Reduction : 중복되는(불필요한) 정보를 줄임.
  - Representation : 어떤 입력에 대해 표현하는 요소.
- Only Positive samples
- No large batch
- **Feature dimensions**이 커질 수록 성능이 향상됨.
  - 기존 simclr은 batch가 커야 성능 향상



기존 : feature dim으로 nomalization 진행

Barlow : batch dim으로 nomalization 진행

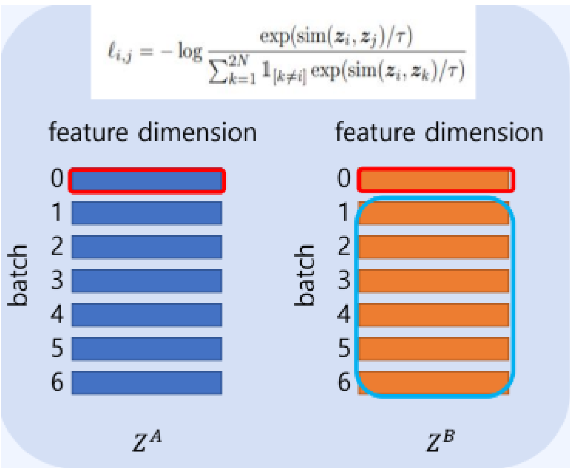
Cross correlation Matrix

$$\mathcal{C}_{ij} \triangleq \frac{\sum_b z_{b,i}^A z_{b,j}^B}{\sqrt{\sum_b \left(z_{b,i}^A\right)^2} \sqrt{\sum_b \left(z_{b,j}^B\right)^2}}$$

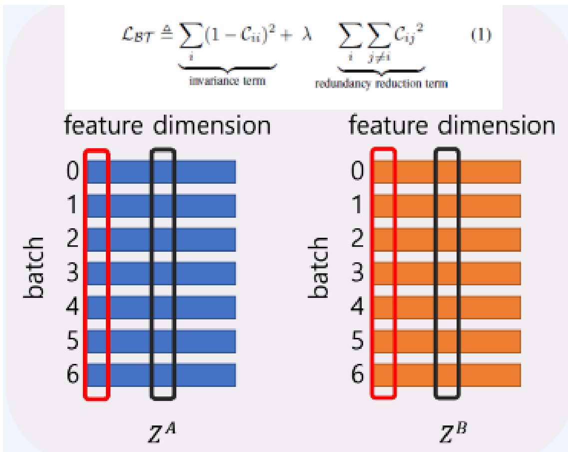
Barlow Twins loss

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

SimCLR



Barlow Twins



Barlow Twins loss는 학습되는 representation이 서로 달라 지도록 학습. (disentanglement)

실험결과

Experiments Barlow Twins

1. Linear evaluation on ImageNet

   기존의 방식과 유사

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

2. Semi-supervised learning on ImageNet

   더 높은 성능, 적은 데이터로 좋은 효과를 낼 것으로 추측

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

3. Transfer learning: image classification

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

4. Transfer learning: object detection and segmentation

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

Ablations

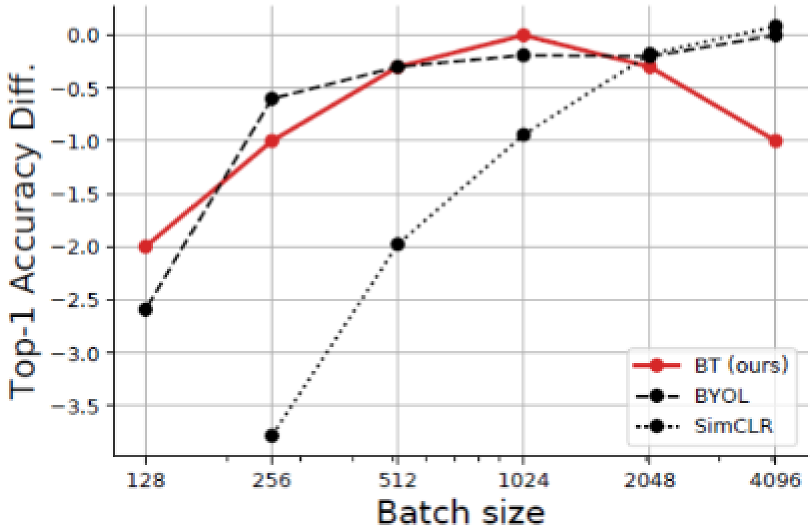
1. Loss function ablation

- on-diag term이 모델 collapse를 막아주는 주요 요소
- off-diag term은 모델의 성능을 향상.

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

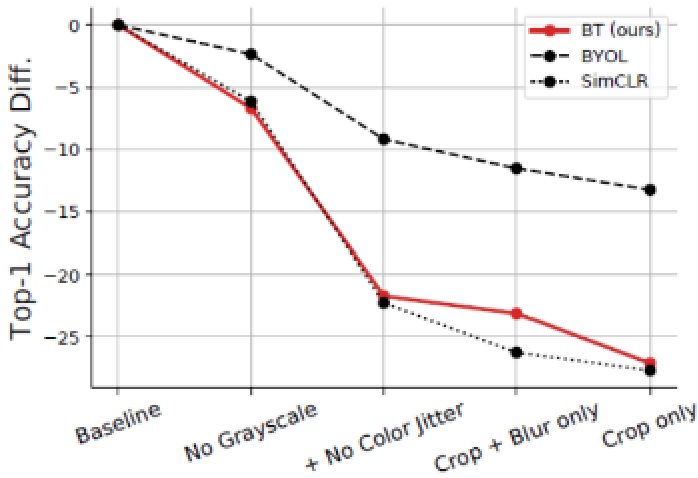
2. Robustness to batch size

BYOL과 유사하게 Batch size에 민감하지 않은 성능



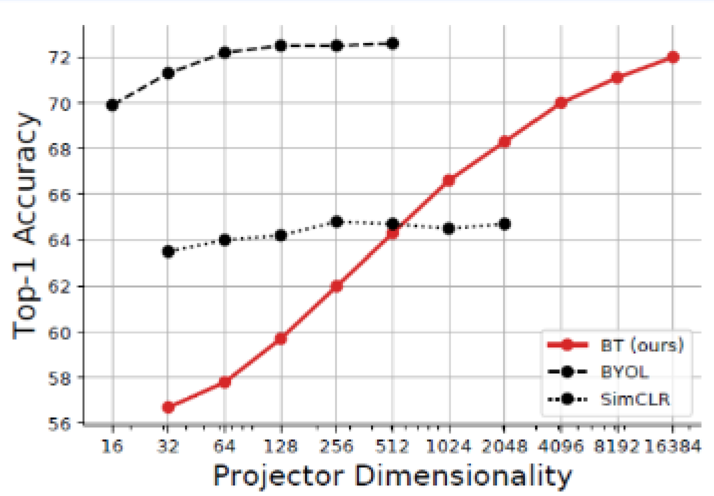
3. Effect of removing Augmentations

- Barlow Twins는 data augmentation을 지우는 것에 있어서 강인하지 않음



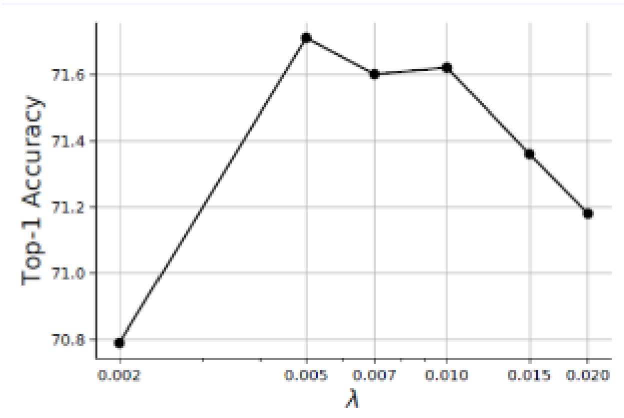
4. Projector dimensionality

- Barlow Twins는 projector의 차원이 커질 수록 성능이 향상함.
- Image Batch를 키우는 것보다 feature dim을 키우는 것이 훨씬 효율적.



5. Breaking Symmetric & sensitivity to hyperparameter
    - Stop-gradient의 사용 유무나 predictor의 사용유무에 상관 없이, Barlow Twins는 collapse problem이 일어나지 않음.
    - redundancy reduction term에 대해 민감도가 높지 않음(논문의 주장)

case	stop-gradient	predictor	Top-1	Top-5
Baseline	-	-	71.4	90.2
(a)	✓	-	70.5	89.0
(b)	-	✓	70.2	89.0
(c)	✓	✓	61.3	83.5



결과

- 자체 손실로 인해 문제가 무너지는 것을 방지합니다.

- 음수표본없음
- 배치 크기에 대한 견고성
- 기능이 더 어두울수록 성능이 향상됩니다.