Barlow Twins: Self-Supervised Learning via Redundancy Reduction

[ICML, 2021]

2023. 06. 30.

김성수

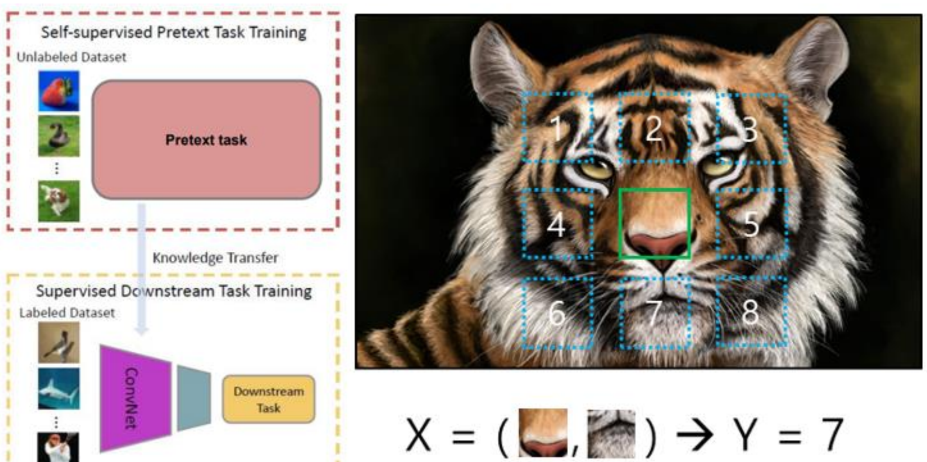
Data Mining and Quality Analytics

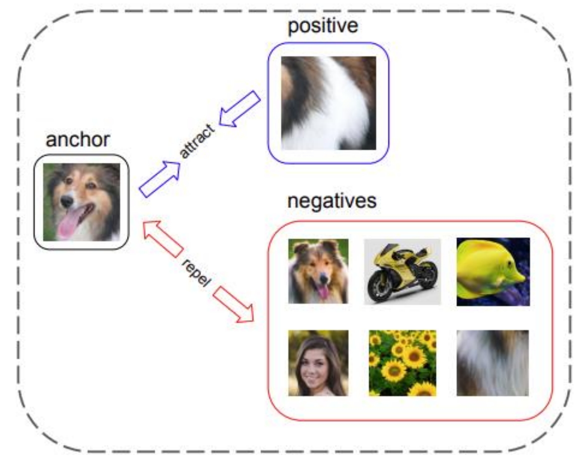
연구배경

- Task명

❖ Task: Self-supervised Learning

- Unlabeled 데이터만을 활용하여 풍부한 Representation을 학습
- Unlabeled 데이터는 Label을 갖고있지 않기에, 스스로 Supervised를 받을 수 있도록 학습





[Pretext Task]

[Contrastive Learning]

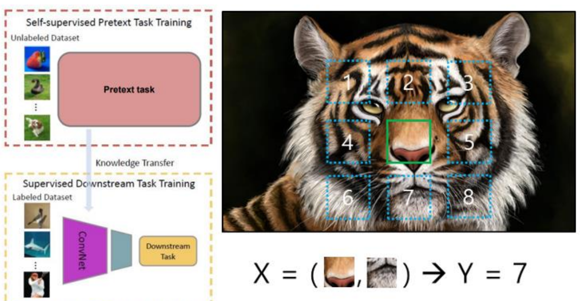
Data Mining Quality Analytics
고려대학교 KOREA UNIVERSITY

연구배경

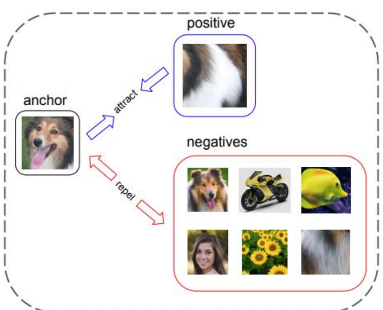
- 선행연구의 한계

❖ 선행연구들의 한계

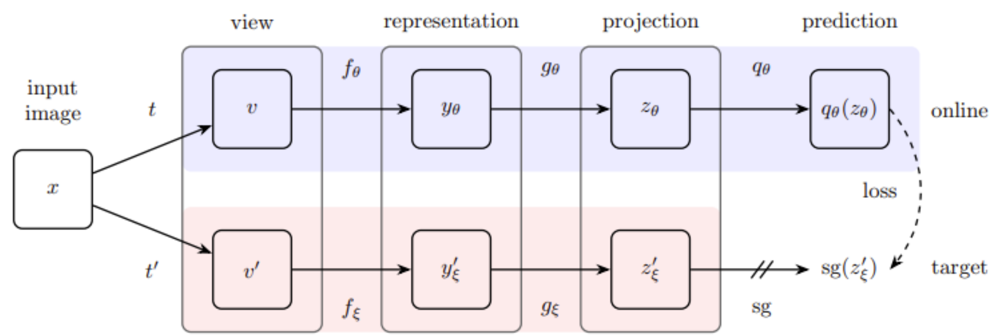
- Pretext Task: 일반화된 Feature를 학습하기 어려움
- Contrastive Learning: 큰 Batch Size가 요구되며, 많은 Computing Resource를 필요로 함
- Non-Contrastive Learning: 학습 불안정성(Collapse) + 복잡한 Technique(Stop Gradient, Momentum Update 등)



[Pretext Task]



[Contrastive Learning]



[Non-Contrastive Learning]

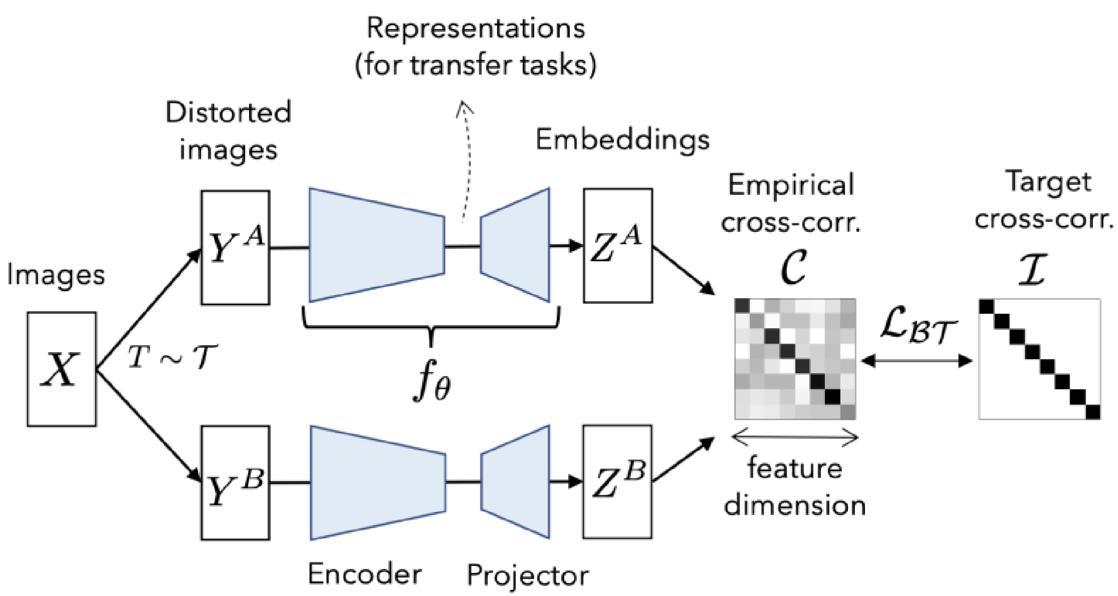
Data Mining Quality Analytics
고려대학교 KOREA UNIVERSITY

연구배경

- 선행연구의 한계를 극복과정 (Overview of Research)

❖ Overcome the Limitation

- 일반적인 Non-contrastive Learning의 개념처럼 두 이미지 간 유사도가 높아지도록 학습
- 이때, 각 Feature들이 독립적으로 학습되도록 중복을 제거 (Redundancy Reduction)
- Correlation Matrix가 Identity Matrix가 되도록 학습



Data Mining
Quality Analytics
고려대학교
KOREA UNIVERSITY

방법론

- 개요

❖ Barlow Twins

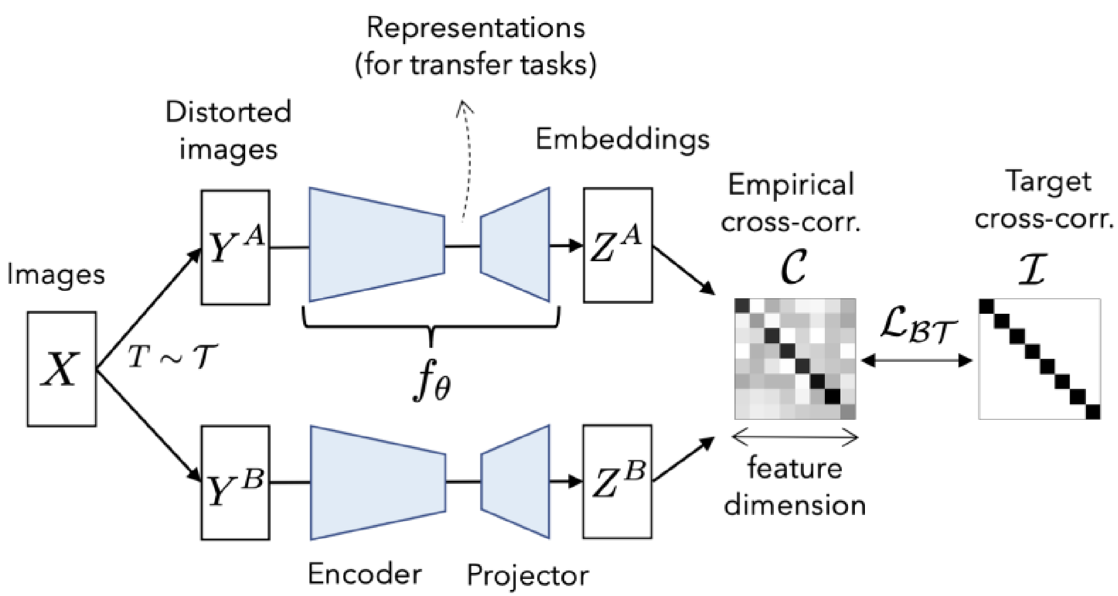
① 데이터 증강 2회

② Feature 추출 및 Projection (Symmetric)

③ Embedding Vector 정규화

④ Cross Correlation Matrix 생성

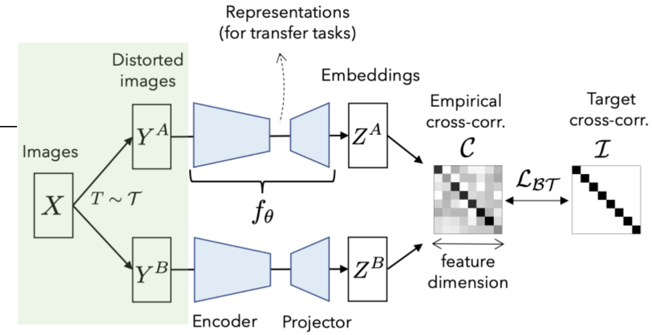
⑤ Loss 산출



Data Mining
Quality Analytics
고려대학교
KOREA UNIVERSITY

방법론

- Barlow Twins



❖ Step1. 데이터 증강 2회

- BYOL과 동일한 데이터 증강 적용
  - ➢ Random Cropping, Resize 224 x 224 (항상)
  - ➢ Horizontal Flipping, Color Jittering, Gray, Gaussian Blur, Solarization (확률적)
  - ➢ 데이터 증강에 민감

[Gray 변환]

[Color Jittering]

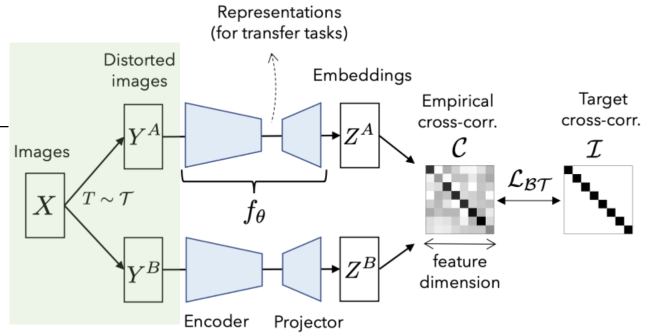
[Solarization]

[Gaussian Blur]

Data Mining Quality Analytics
고려대학교 KOREA UNIVERSITY

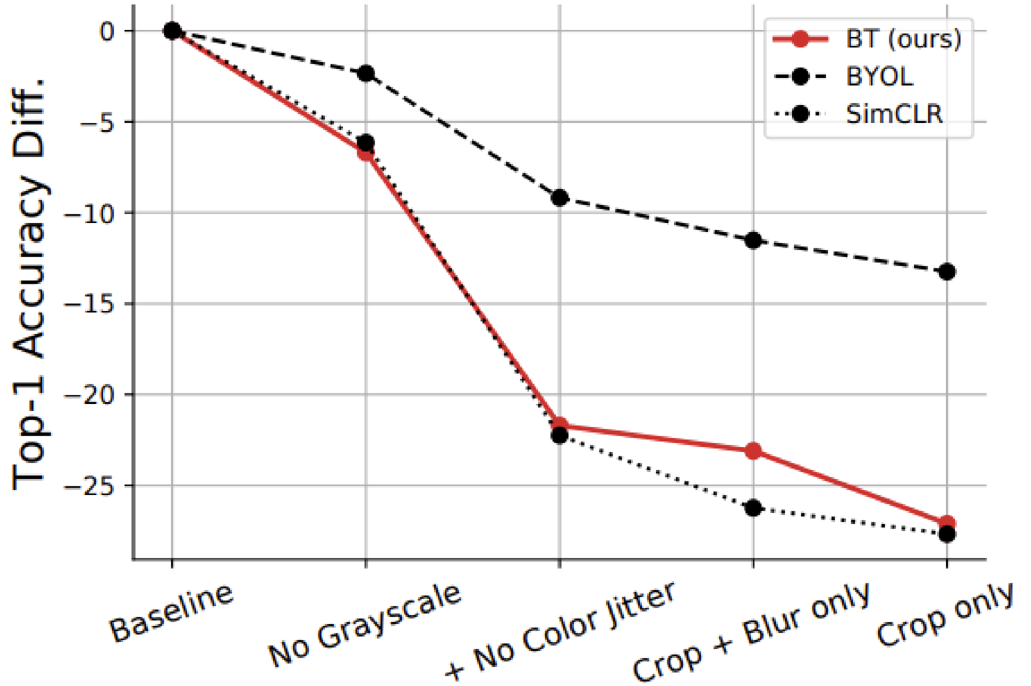
방법론

- Barlow Twins



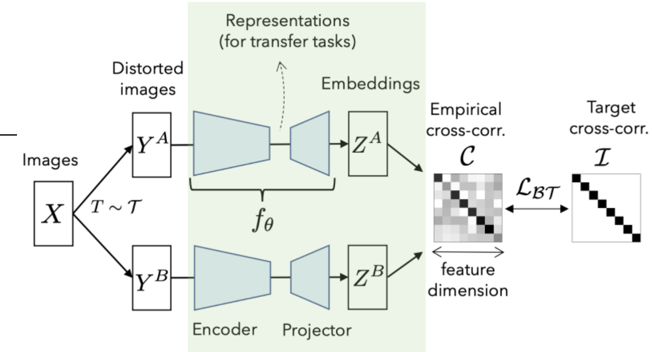
❖ Step1. 데이터 증강 2회

- BYOL과 동일한 데이터 증강 적용
  - ➢ Random Cropping, Resize 224 x 224 (항상)
  - ➢ Horizontal Flipping, Color Jittering, Gray, Gaussian Blur, Solarization (확률적)
  - ➢ <u>데이터 증강에 민감</u>



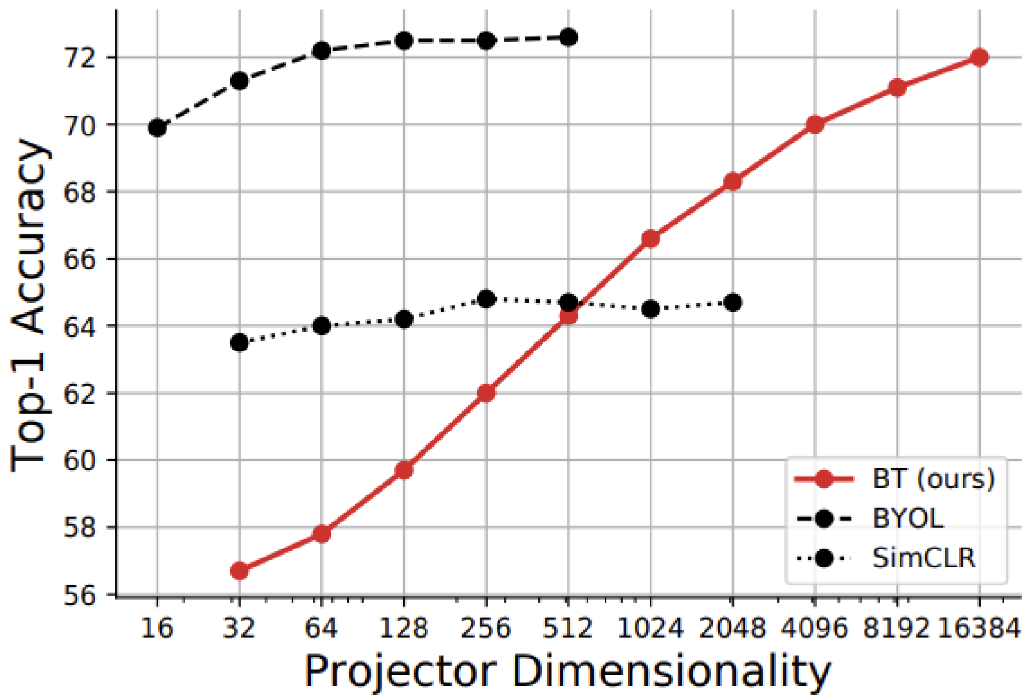
방법론

- Barlow Twins



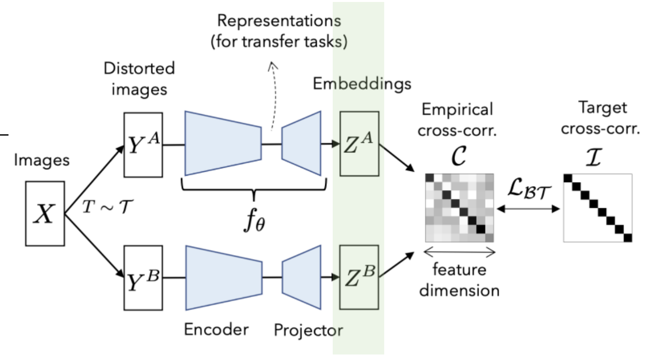
❖ Step2. Feature 추출 및 Projection

- Feature Extractor: ResNet-50 (Output Dim: 2,048)
- Projection: Linear – BN – ReLU - Linear – BN – ReLU – Linear (Output Dim: 8,192)
  - ➢ Output Layer의 개수가 Expansion하는 구조를 가지며, Output 차원이 클수록 좋은 성능을 보임



방법론

- Barlow Twins



❖ Step3. Embedding Vector 정규화

- 일반적인 Standard Scaler와 동일한 연산 적용
  - ➢ Batch 단위에서 각 Feature의 평균을 빼주고, 표준편차만큼 나누어 줌

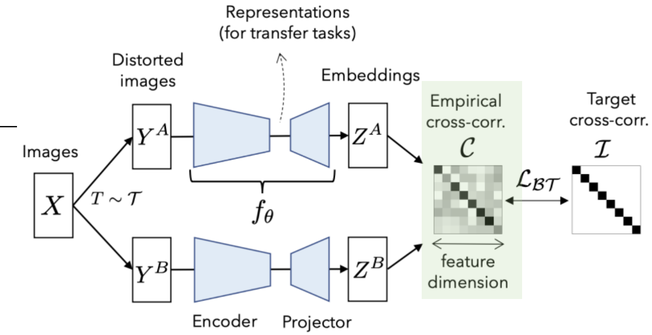
Algorithm 1 PyTorch-style pseudocode for Barlow Twins.

```
# f: encoder network
# lambda: weight on the off-diagonal terms
# N: batch size
# D: dimensionality of the embeddings
#
# mm: matrix-matrix multiplication
# off_diagonal: off-diagonal elements of a matrix
# eye: identity matrix

for x in loader: # load a batch with N samples
    # two randomly augmented versions of x
    y_a, y_b = augment(x)

    # compute embeddings
    z_a = f(y_a) # NxD
    z_b = f(y_b) # NxD

    # normalize repr. along the batch dimension
    z_a_norm = (z_a - z_a.mean(0)) / z_a.std(0) # NxD
    z_b_norm = (z_b - z_b.mean(0)) / z_b.std(0) # NxD

    # cross-correlation matrix
    c = mm(z_a_norm.T, z_b_norm) / N # DxD

    # loss
    c_diff = (c - eye(D)).pow(2) # DxD
    # multiply off-diagonal elems of c_diff by lambda
    off_diagonal(c_diff).mul_(lambda)
    loss = c_diff.sum()

    # optimization step
    loss.backward()
    optimizer.step()
```


Data Mining
Quality Analytics
고려대학교
KOREA UNIVERSITY

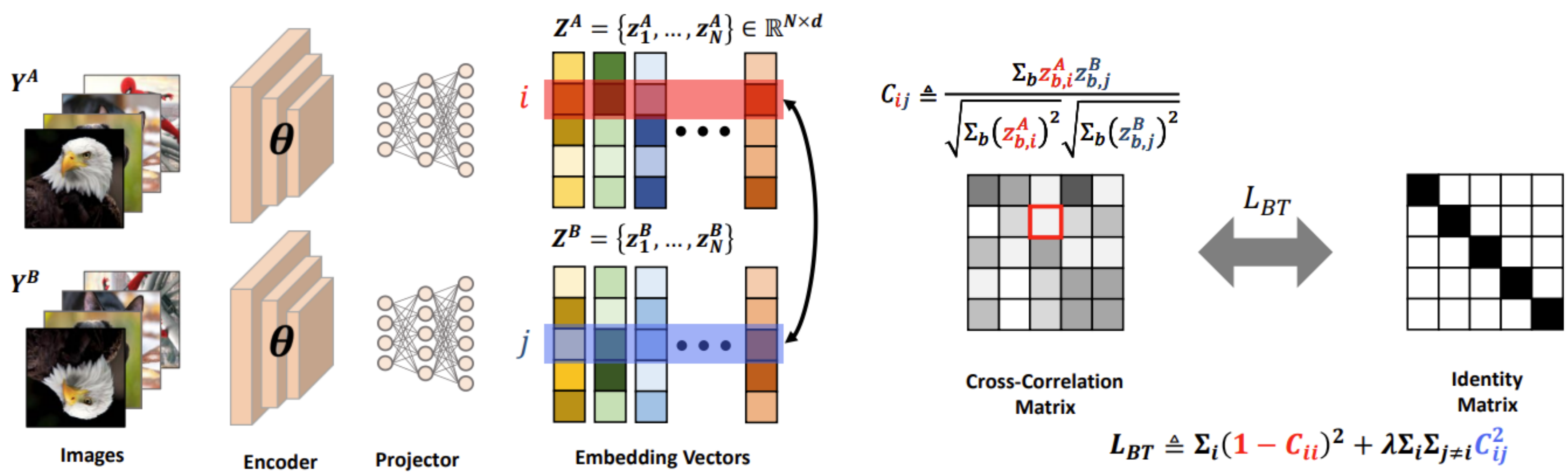
방법론

- Barlow Twins



❖ Step4. Correlation Matrix 산출

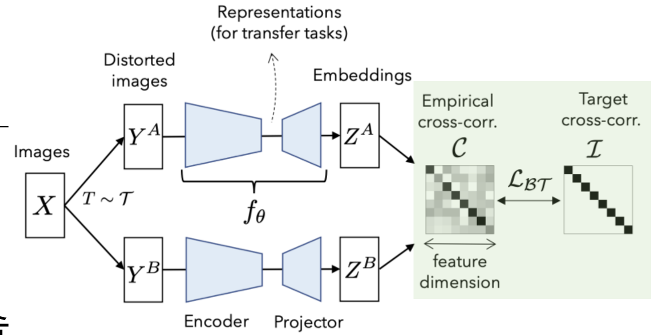
- 각 Feature가 {NxD}라면, DxD의 Correlation Matrix 생성
- 두 Feature Vector간 내적을 수행한 후, 각 크기로 나누어 줌



Data Mining Quality Analytics
고려대학교 KOREA UNIVERSITY

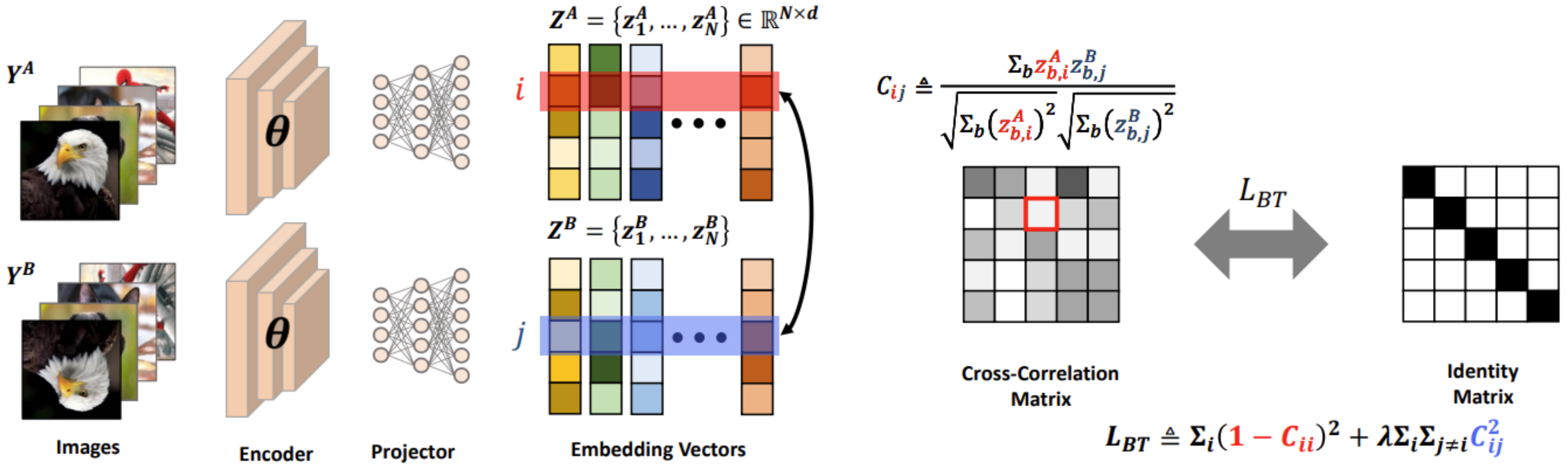
방법론

- Barlow Twins



❖ Step5. Loss 산출

- Cross Correlation Matrix가 Identity Matrix가 되도록 학습
- 이를 통해 동일한 Feature와 유사도는 크게, 다른 Feature와는 중복을 제거



Data Mining Quality Analytics
고려대학교 KOREA UNIVERSITY

실험결과

- Result

❖ Benchmark 데이터셋과 실험결과 비교

- 기존 방법론과 비교했을 때, SOTA급은 아님
- "이렇게도 SSL 접근이 가능하다.." 라는 것을 제안한 논문

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

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.

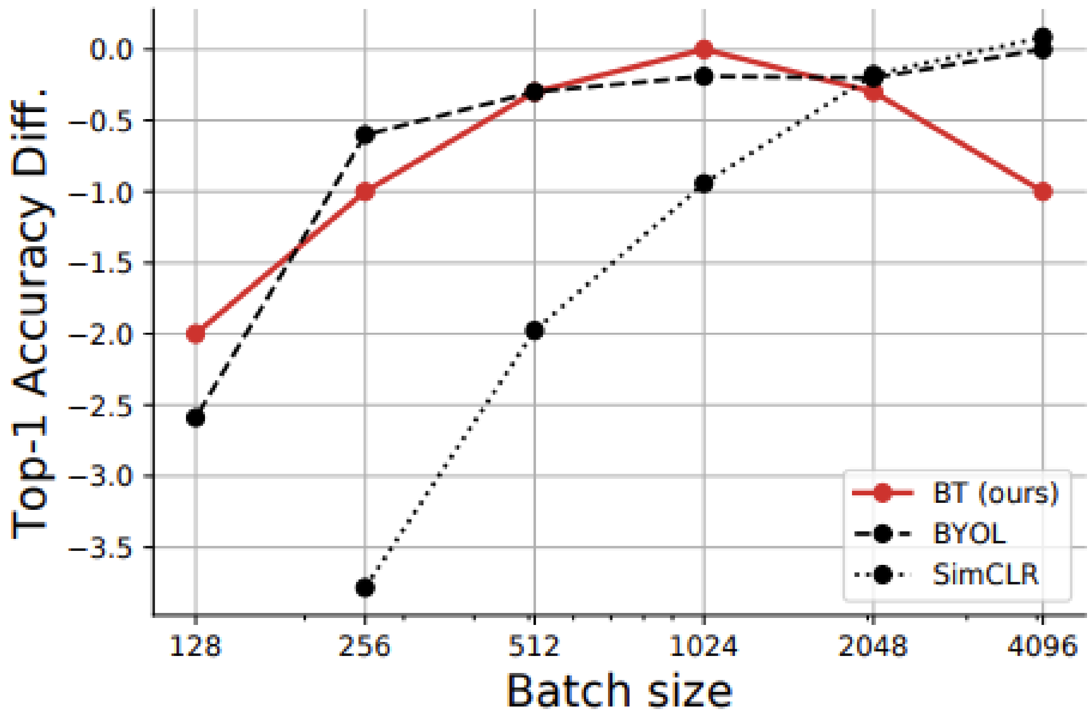
Method	Places-205	VOC07	iNat18
Supervised	53.2	87.5	46.7
SimCLR	52.5	85.5	37.2
MoCo-v2	51.8	86.4	38.6
SwAV (w/o multi-crop)	52.8	86.4	39.5
SwAV	56.7	88.9	48.6
BYOL	54.0	86.6	47.6
Barlow Twins (ours)	54.1	86.2	46.5

실험결과

- Result

❖ Ablation Study for Batch Size

- Batch Size에 강건
- 대조학습과 유사한 성능을 내면서, Resource는 덜 필요한 장점



결론

- 방법론의 장점 + 유의사항

❖ Contribution

- SSL에 Redundancy Reduction을 적용한 Information Maximization 계열의 최초 연구
- SOTA와 유사한 성능을 보여줌

❖ 장점

- 여러 Technique(Asymmetric구조, Stop Gradient 등)이 없는 간단한 구조지만 안정적인 학습
- Negative Sample을 활용하지 않기에, 큰 Batch Size 불필요 (Batch Size에 강건)

❖ 유의사항

- 큰 Projection Vector를 활용할 것 (클수록 성능은 선형적으로 증가)
- Embedding Vector의 정규화는 필수적이며, 배치방향으로 정규화의 방향을 잘 확인할 것
- Augmentation 정의 중요