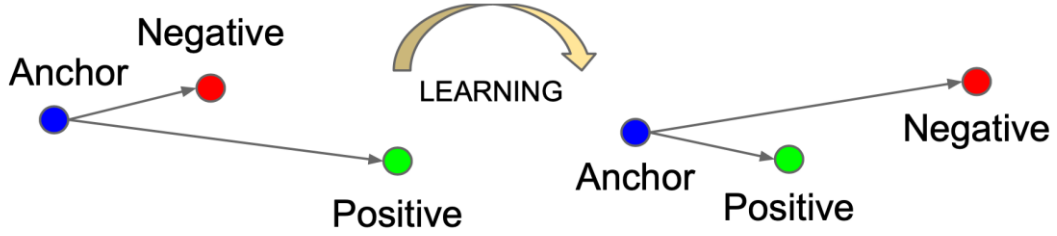
SimCLR

2026.01.02

Introduction

Contrastive learning

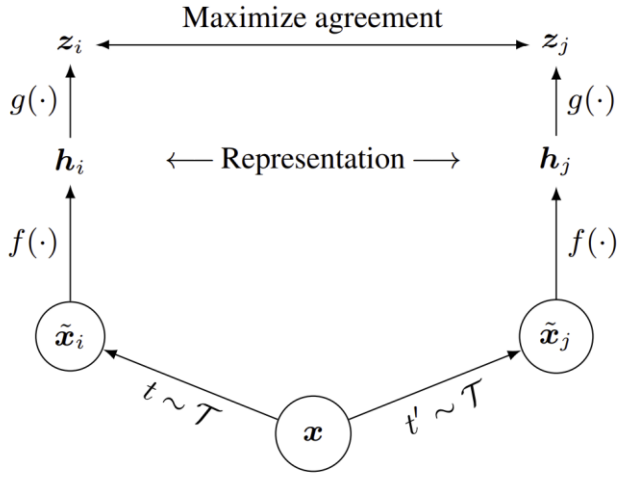
- '비슷해야 할 것을 가깝게, 달라야 할 것은 멀게' 표현 공간에 배치하도록 representation을 학습하는 방법
- 데이터에 정답 라벨은 없음
- 대신 pair를 만들어서 학습에 하용
  - Positive pair: 같은 의미를 가진 두 샘플
  - Negative pair: 의미가 다른 샘플들
- Positive pair는 유사도가 높게, Negative pair는 유사도가 낮게 학습 하는 것이 목표



Methodology

Definition

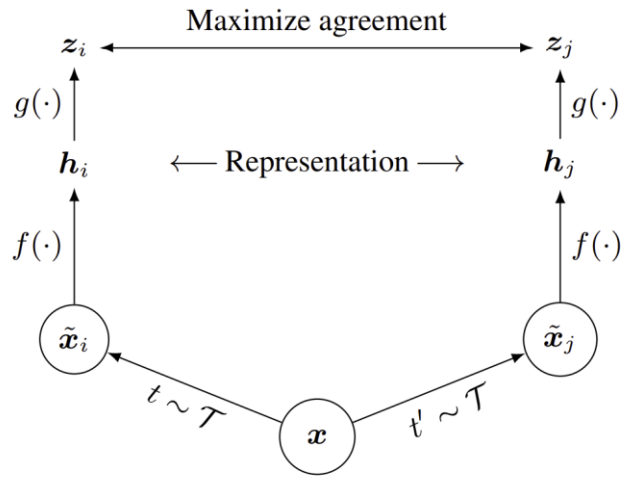
- Positive pair
  - 같은 이미지에 서로 다른 data augmentation을 적용한 두 view를 positive pair로 봄
  - $\widetilde{x_i} \sim \mathcal{T}(x), \quad \widetilde{x_j} \sim \mathcal{T}(x)$
- Representation
  - $h_i = f(\widetilde{x_i}) \in R^d$
  - ResNet50을 씀
- Projection head
  - $z_i = g(h_i) = W^{(2)}\sigma(W^{(1)}h_i)$
  - Contrastive learning에 사용



Methodology

Definition

- Mini-batch
  - 크기 N이고 각 샘플당 2개의 view → 총 2N개의 샘플 $\{\widetilde{x_k}\}_{k=1}^{2N}$
- Cosine Similarity
  - $sim(u, v) = \frac{u^\top v}{|u||v|}$
- Contrastive Loss (NT-Xent loss)
  - 자기 자신을 제외한 모든 샘플이 negative로 포함됨
  - # of negative pair: 2N-2
- 전체 Objective
  - $\mathcal{L} = \frac{1}{2N}\sum_{k=1}^{N}[\ell_{2k-1,2k} + \ell_{2k,2k-1}]$
  - I와 j가 각각 anchor가 됨



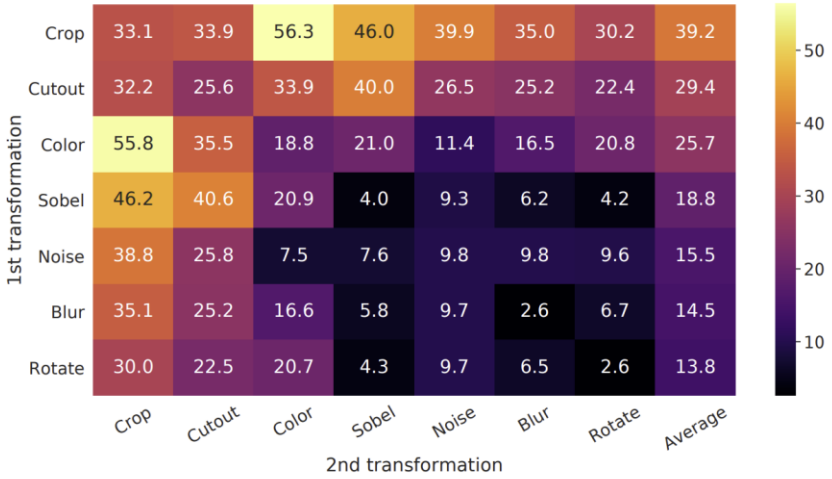
$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

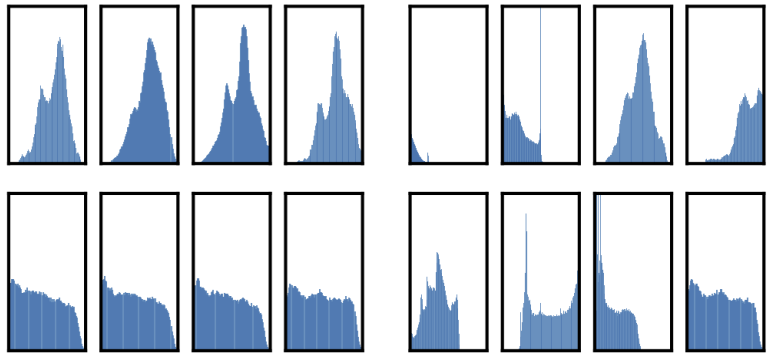
Experiment

Composition of data augmentation

- 단일 augmentation만으로는 좋은 표현을 학습하기 어려움
- contrastive task는 잘 풀 수 있지만 representation quality는 낮음
- 여러 augmentation을 조합할수록 contrastive prediction task가 더 어려워져서 representation 품질이 크게 향상됨
- Random Crop + Color Distortion 조합이 가장 좋음
  - Crop만 하면 patch간 color histogram이 유사해서 색 분포 만으로 positive pair 맞출 수 있음(short cut)
  - Color distortion은 이러한 shortcut을 무너뜨려서 모델이 색 정보에 의존 하지 않고 더 좋은 representation을 만들어낼 수 있도록 함

더 어려운 contrastive task → 더 좋은 representation





(a) Without color distortion. (b) With color distortion.

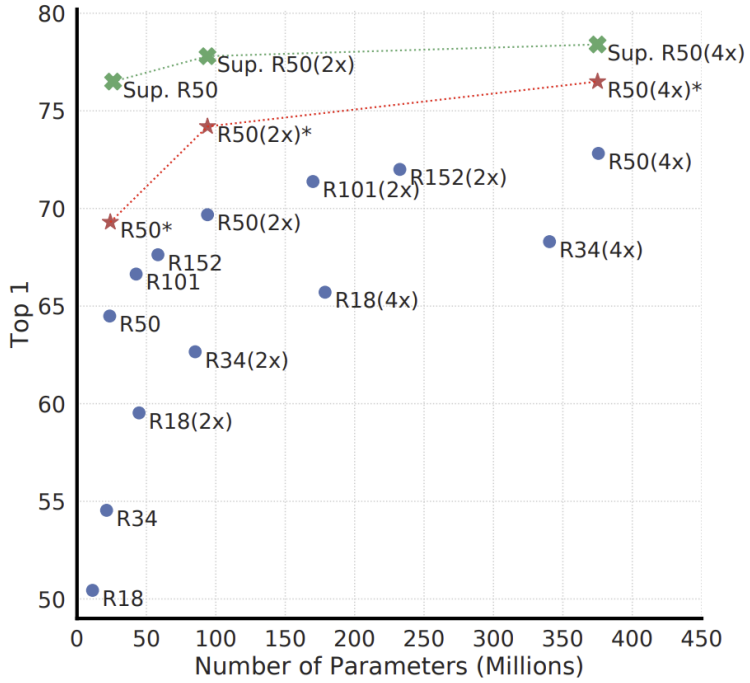
Experiment

Unsupervised contrastive learning benefits from bigger models

- 모델이 커질수록 성능 지속적으로 향상됨
  - depth 증가 (ResNet-18 → 34 → 50 → 101 → 152)
  - width 증가 (2×, 4× channel expansion)

→ 큰 모델일수록 unsupervised learning이 supervised에 더 근접

- Epoch 수 증가 → contrastive representation 품질 개선



Contrastive learning benefits disproportionately from larger and longer-trained models.

Experiment

projection head

- Contrastive learning은 z 공간에서 augmentation에 무관하게 positive pair를 맞추는 것이 목표임
  - z: augmentation에 강하게 invariant한 공간
  - h: 더 풍부한 semantic 정보를 유지함

- z는 contrastive loss에 직접 최적화되어 색, 방향, texture 같은 정보가 의도적으로 제거되지만 h는 projection head 이전 단계로 contrastive loss의 영향을 간접적으로만 받음 → general representation

- RotNet에서 두번째 conv.block의 semantic representation이 가장 높았던 것과 유사함

What to predict?	Random guess	Representation $\boldsymbol{h}$	Representation $g(\boldsymbol{h})$
Color vs grayscale	80	99.3	97.4
Rotation	25	67.6	25.6
Orig. vs corrupted	50	99.5	59.6
Orig. vs Sobel filtered	50	96.6	56.3

My Code

코드 구현

- Data Augmentation: RandomResizedCrop ,Flip, ColoJitter, GrayScale
- SimCLRDataset: $x \rightarrow x_i, x_j$
- SimCLRModel: encoder(ResNet18) + projection_head
- NT-Xent Loss
- Traning: 100 epoch
- knn-evaluation

구분	Accuracy
Original resnet18	61.07%
revised	76.51%,

https://colab.research.google.com/drive/1u431PdXTcrSZH7WvzcL841g7JWg4BggQ?usp=sharing

ELLab