OUTTA 부트캠프 딥러닝반 팀 과제 #1

<A Simple Framework for Contrastive Learning of Visual Representation>

24조 오민제, 좌대현

AlexNet, ResNet 등 CNN을 기반으로 한 다양한 지도학습 인공신경망 모델들은 컴퓨터 비전 분야에서 최근 강력한 성능을 입증하였다. 다만, 우리가 접속 가능한 데이터의 대부분은 label이 없는 데이터이기 때문에, 이런 데이터를 활용해 모델을 구축할 수 있는 비지도, 자기지도 학습 모델에 대한 연구가 최근 활발하다. 본 논문은 unlabeled 데이터에 대해 Contrastive Learning을 통해서 자기지도 학습 기반 컴퓨터 비전 모델을 구축하는 방법론을 최초로 제시한다.

본 논문에서 제시하는 SimCLR (Simple Framework for Contrastive Learning)은 representation learning의 일종으로, 복잡하고 고차원인 이미지 데이터에서 우리가 원하는 핵심 정보만 도출하고, 비슷한 이미지끼리 묶는 모델이다. 이때 이미지 데이터에 label이 없기 때문에 무엇이 비슷하고, 무엇이 다른지 우리도 모델도 알 수 없다. SimCLR에서는 이를 해결하기 위해 원본 사진 데이터를 변형해 비슷한 데이터를 생성한다. 이때, 하나의 원본을 바탕으로 두가지 서로 다른 변형을 통해 유사한 한 쌍을 만들고, 한 쌍끼리는 비슷한 데이터로, 다른 쌍끼리는 다른 데이터로 학습하도록 모델을 만든다. 본 논문은 NT-Xent라는 새로운 손실함수를 제시 및 사용한다. NT-Xent는 배치 내 유사한 쌍의 코사인 유사도에 softmax, -log를 취하고 모든 유사쌍에 대해 더한 함수로, 이를 최소화하는 과정에서 모델은 유사쌍 사이의 코사인 유사도는 최대로, 이외의 코사인 유사도는 최소로 만들도록 파라미터를 학습한다.

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

본 논문은 SimCLR 모델의 성능을 향상시키기 위한 여러가지 기술들을 제시한다. 우선 학습 과정에서 다른 데이터의 수가 많은 것이 성능 향상에 도움이 되기 때문에 배치 크기가 큰 것이 좋다. 또한 유사한 쌍 사이 차이가 최대한 큰 것이 좋다. 이는 모델이 학습 과정에서 풀기 어려운 문제를 풀어보아야 성능이 우수해진다고 정성적으로 이해가 가능하다. 차이가 큰 유사쌍을 만들기 위해서 색 변형의 사용이 필수적이며, 본 논문에 따르면 색변형과 crop을 각각 사용하는 것이 최적의 조합이다. 마지막으로 변형, 인코더 과정 이후에 헤드라는 추가적인 레이어를 형성해주는 것이 모델 성능을 향상시켜주며 이 레이어는 비선형(즉 Linear + ReLu)이 선형보다, 그리고 선형이 없는 것보다 우수한 성능을 보여준다.

이렇게 생성한 자기지도 학습 SimCLR모델은 대부분의 데이터셋에서 2020당시 SOTA모델이였던 지도학습 컴퓨터 비전 모델들과 유사한 성능을 보여주었으며, 사후 파인튜닝을 결합 시 새로운 SOTA를 달성하였다. 본 논문은 Non-labeled 데이터를 활용할 수 있는 방법을 제시함과 동시에 SOTA를 달성한 점에서 본 논문의 가치를 찾을 수 있다.