자기지도 학습에서 와서스타인 (Wasserstein) 거리의 손실함수로의 이용가능성 연구

구인화, 채동규
한양대학교 인공지능학과
ihkoo@hanyang.ac.kr, dongkyu@hanyang.ac.kr

A Research on Using Wasserstein Distance as a Loss Function in Self-Supervised Learning

Inhwa Koo, Dong-Kyu Chae
Dept. of Artificial Intelligence, Hanyang University, South Korea

요 약

딥러닝의 높은 예측 정확도를 위해서는 많은 양의 학습 데이터가 필요하다. 그러나 실세계에서 많은 양의 레이블이 붙은 데이터를 구하는 것은 어렵고 많은 비용이 든다. 때문에 레이블이 없이도 양질의 표현 학습이 가능한 자기지도학습이 각광을 받고 있다. 와서스타인 거리는 생성모델에도 쓰이지만 의사 레이블 (pseudo label) 을 만들어 레이블이 없는 데이터들을 분류 하는데도 좋은 성능을 보이고 있다. 따라서. 본 연구는 와서스타인 거리를 자기지도학습에 접목시키는 방법을 제안한다. 실험을 통해 연구의 가능성을 보인다.

1. 서론

기계학습으로 높은 성능을 얻기 위해서는 레이블이 있는 대량의 학습 데이터와 학습의 목적에 걸맞은 적절한 모델이 필요하다. 그러나 실세계에서 학습에 필요한 만큼 충분하게 레이블 된 데이터를 얻는 것은 쉽지 않다. 이러한 데이터 부족 문제를 해결하기 위한 연구들이 활발히 진행되고 있으며, 그 중 하나는 자기지도 학습 (self-supervised learning) 프레임워크이다. 자기지도학습은 레이블이 붙지 않은 대량의 데이터를 이용한다. 뿐만 아니라 데이터들의 구조를 파악하거나 예측함으로써 레이블을 생성해내고, 이를 지도학습 방식으로 학습한다. 자기지도학습을 이용하면 이미지의 문맥을 파악할 수 있는 좋은 표현 (representation) 들을 얻을 수 있다는 장점 때문에 이미지나 비디오 학습에 널리 쓰이고 있다.

와서스타인 거리 (Wasserstein distance) 는 두 확률분포 간의 거리의 기대값이 가장 작게 나오는 확률분포를 취하는 방법이다. 본 연구는 와서스타인 거리의 이러한 특성이 자기지도학습 모델의 학습에도 적합할 것이라고 가정하였다. 즉, 와서스타인 거리를 손실함수로 사용함으로써, 와서스타인 거리가 생성모델뿐만 아니라 자기지도학습에서도 효과적인 손실함수임을 밝히고자 한다. 이를 실험적으로 입증하기 위해 자기지도 학습 방식 중 대칭적 구조의 모델과 비대칭적 구조의 모델 각각의 손실 함수를 기존의 손실 함수에서 와서스타인 거리로 대체하여 실험을 진행하였다. 실험 결과 와서스타인 거리를 이용해도 기존과 비슷한 성능이 나왔으며, 이를 통해 와서스타인 거리 또한 손실 함수로써 사용 가능함을 알 수 있다.

2. 관련 연구

2.1. 자기지도학습

자기지도학습은 대조학습을 통해 발전되었다. 대조학습은 두 개의 이미지가 하나의 쌍이 되어 각각 네트워크에 입력되고 네트워크는 들어온 이미지가 비슷한 이미지(긍정 쌍)라면 서로 가까워지도록, 다른 이미지(부정 쌍)라면 서로 멀어지도록 표현을 학습하는 방법이다 [1]. 그러나 이 방법은 몇 가지 단점이 존재한다. 대조학습이 좋은 성능을 내려면 많은 양의 부정 쌍이 필요하며, 이는 종종 메모리 부족 문제를 일으킨다. 또한 이미지에 어떤 변형을 주느냐에 따라 모델의 성능이 달라진다는 단점이 있다. 이를 극복하기 위해 비대칭구조를 이용하여 긍정 쌍으로만 학습을 시키는 모델들이 제안되었으며, 대표적으로 SimSiam [2]이 있다.

2.2. 와서스타인 거리

와서스타인 거리는 두 확률분포 간의 거리의 기대값이 가장 작은 확률분포를 취하는 방법으로, 유명한 생성 모델 중 하나인 WGAN (Wasserstein Generative Adversarial Network)의 손실함수로 사용된다. 와서스타인 거리를 손실함수로 사용하면 GAN 을 학습시킬 때 발생하는 모드 붕괴 (mode collapse) 현상이 줄어드는 효과를 얻을 수 있다 [3].

또한 최근 WGAN 을 이용해 의사 레이블 (pseudo label) 을 만들어 자기지도학습 모델을 훈련시키는 연구가 발표되었다 [4]. 합성곱 신경망을 통해 레이블이 없는 데이터를 클러스터링한 후, WGAN 으로 각각의 클러스터들과 가장 비슷한 레이블을 갖고 있는 원래 레이블이 있던 데이터를 매치하여 유사 레이블을 만들어낸다. 유사 레이블이 있는 데이터를 사용하여 합성곱 모델을 훈련한다. 이 기법을 사용하여 모델의 오류가 기존 방법들보다 적어진 결과를 얻었다. 이는 의사 레이블링 분야에서 와서스타인 거리의 활용방안을 알아볼 수 있는 연구라고 할 수 있다.

3. 제안하는 방법

와서스타인 거리를 자기지도 학습에 도입하기 위해, 우리는 자기지도 학습 분야에서 좋은 성능을 내고 있는 모델들 중 대칭적 구조를 갖는 SimCLR 모델과 비대칭적 구조로 되어 있는 SimSiam 모델을 기본 베이스로 삼았다. 두 모델 다 유사성이 높은 이미지들, 즉 긍정 쌍들을 인코딩한 벡터 사이의 거리가 가까워지도록 학습하며, 이러한 특징은 와서스타인 거리를 사용하기에 알맞다.

그 후 우리는 각 모델의 기존 손실 함수를 와서스타인 거리로 변경하였다. 구체적으로, SimCLR 는 크로스 엔트로피 기반의 NT-Xent 함수를 사용하고 있으며 SimSiam 모델은 코사인 유사도 함수를 사용하였다. 이들을 모두 와서스타인 거리로 대체 하였으며, 다만 SimSiam 모델의 경우 스톱 그래디언트나 손실함수의 대칭성 등의 다른 요소들은 바꾸지 않았다.

두 모델들의 backbone 은 모두 ResNet-50 을 사용했다. 하이퍼파라미터는 모델 별로 서로 동일하게 세팅하였다. 학습을 위한 최적화 기법으로 SimCLR 모델은 LARS (layer-wise adaptive learning rate scaling) 를 사용했고, SimSiam 모델은 기본 SGD (stochastic gradient descent) 를 사용했다.

4. 실험 결과

제안한 방법의 정확도를 측정하기 위해 CIFAR-10 데이터 셋을 사용하여 훈련 및 평가를 진행하였다. CIFAR-10 데이터 셋은 6,000 개의 32x32 크기를 가진 이미지로 구성되어있다. CIFAR-10 으로 학습된 모델들은 feature 와 weight 를 고정시켜 linear evaluation 을 진행하였다. 표 1 은 실험의 결과를 나타낸다. 네트워크 구조의 대칭성과 상관없이 기존 손실함수를 와서스타인 거리로 대체해도 원래의 모델과 비슷한 성능을 얻을 수 있음을 보여준다.

<표 1> 실험 결과

Method	Accuracy
SimCLR [1]	87.2%
SimCLR(with W-distance)	86.6%
SimSiam [2]	90.9%
SimSiam(with W-distance)	90.5%

5. 결론 및 향후 연구

실험 결과를 통해 우리는 와서스타인 거리를 자기지도 학습 모델의 손실 함수로 사용하면 기존의 손실 함수와 비슷한 성능을 보일 수 있다는 것을 확인하였다. 아쉬운 점은 기존의 손실 함수를 대체할 수는 있겠지만 정확도 면에서 모델의 성능을 획기적으로 높여 주지는 못했다. 그럼에도 불구하고 추후 와서스타인 거리를 여러 학습에 사용한다면 안정적인 학습을 진행할 수 있다는 점에서 다른 자기지도학습 모델에 적용해볼 만하다는 것을 확인할 수 있었다.

향후 연구에서는 와서스타인 거리를 이용하여 유사 레이블링을 할 때도 소량의 레이블이 존재하는 데이터 없이도 레이블링이 가능한지 연구할 예정이다.

감사의 글

이 논문은 2022 년도 정부(과학기술정보통신부)의 재원으로 (1) 한국연구재단 바이오 의료기술개발사업의 지원 (No. NRF-2021M3E5D2A01021156)과 (2) 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2020-0-01373, 인공지능대학원지원(한양대학교))

참고문헌

[1] Ting Chen et al. "A Simple Framework for Contrastive Learning of Visual Representations" ICML, 2020

[2] Xinlei Chen et al. “Exploring Simple Siamese Representation Learning” CVPR, 2020

[3] Martin Arjovsky et al. “Wasserstein GAN”, PMLR, 2017

[4] Fariborz Taherkhani et al. “Self-Supervised Wasserstein Pseudo-Labeling for Semi-Supervised Image Classification” CVPR, 2021