Lab 6. Generative Adversarial Nets Summary

Summarized by: 202355514 강지원

Ian J. Goodfellow, Jean Pouget-Abadie , Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio
1 Department of Computer Science and Operational Research University of Montreal

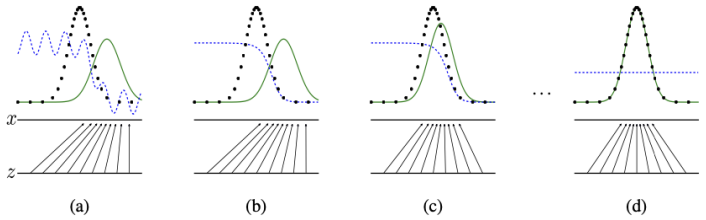

Figure 1: GAN 학습 과정 시각화. 검은 점선은 원 데이터의 확률 분포, 녹색 점선은 GAN이 만들어 내는 확률 분포, 파란 점선은 판별자의 확률 분포를 말한다.

본 논문에서는 생성 모델과 판별 모델을 동시에 훈련시키는 새로운 적대적 학습 프레임워크를 제안한다. 생성 모델은 데이터 분포를 복원하고, 판별 모델은 샘플이 훈련 데이터에서 왔는지 생성 모델에서 왔는지 구분한다. 이 프레임워크는 역전파로 훈련되며, 마르코프 체인이나 근사 추론 네트워크 없이도 작동한다. 이를 실험을 통해 생성된 샘플의 질적 및 양적 평가를 하고자 한다.

지금까지 딥러닝에서 가장 큰 성과는 판별 모델들에 의해 이루어졌으나, 딥 생성 모델은 계산을 근사하는 어려움과 생성적 맥락에서 선형 유닛을 활용하는 어려움이 있었다. 이에 본 논문은 새로운 생성 모델 추정 절차로 적대적 신경망(Adversarial Nets) 프레임워크를 제안한다. 이 프레임워크에서는 생성 모델과 판별 모델이 대결을 벌이며, 판별 모델은 샘플이 모델 분포에서 왔는지 데이터 분포에서 왔는지를 판별한다. 생성 모델과 판별 모델은 다층 퍼셉트론(Multilayer Perceptron)을 사용하여 훈련되며, 마르코프 체인이나 근사 추론을 필요로 하지 않는다.

적대적 신경망(GAN)은 생성 모델(Generator, G)과 판별기(Discriminator, D)간의 미니맥스(MiniMax) 게임을 통해 학습하는 모델이다. 생성 모델은 노이즈 z로 부터 가짜 데이터를 만들고, 판별자는 입력이 실제 데이터인지 생성된 데이터인지 구분한다. 이 과정은 다음 목적식을 통해 표현된다

$$V(D,G) = E_{x\sim p_{\text{data}}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1-D(G(z)))]$$

G(z)는 입력 노이즈인 z를 기반으로 데이터를 생성하는 다층 퍼셉트론(Multilayer Perceptron) 모델을 의미하며, D(x)는 주어진 데이터 x가 실제 데이터인지 생성 모델에서 온 것인지를 판별하는 또 다른 다층 퍼셉트론 모델을 의미한다.

위 학습 구조는 데이터 분포를 모방하도록 작동하는데, 그 이유는 모델의 용량이 충분하다고 가정할 때, 생성자 분포 p_g가 실제 데이터 분포 p_{data}와 같아질 때 이 minimax 게임의 전역 최적해가 존재한다. 또한 논문에서 제시한 알고리즘(Algorithm 1)이 실제로 목적식을 점진적이로 최적화하도록 설계되어 있어, 반복 학습을 통해 생성자가 실제 데이터 분포에 수렴하게 됨을 보인다. 이러한 기반 위에서, 고정된 생성자 G에 대해 최적 판별자 D^*의 형태를 유도하고, 이를 GAN 목적식에 대입ㅎ여 생성자의 새로운 비용 함수 C(G)를 도출한다.

본 실험에서는 GAN을 MNIST, TFD, CIFAR-10 등의 데이터셋에 적용해 생성 모델의 성능을 평가하였다. Generator는 ReLU와 sigmoid를 혼합해 구성하고, Discriminator에는 maxout과 dropout을 적용하여 학습 안정성을 확보하였다. 생성된 샘플의 품질은 Parzen window 기반 로그-우도(log-likelihood) 추정 방식으로 측정했으며, 그 결과 GAN은 MNIST와 TFD에서 기존 생성 모델들과 비교해 경쟁력 있는 성능을 보였다.

GAN의 주요 단점은 모델이 명시적으로 $P_9(x)$를 계산할 수 없고, 학습

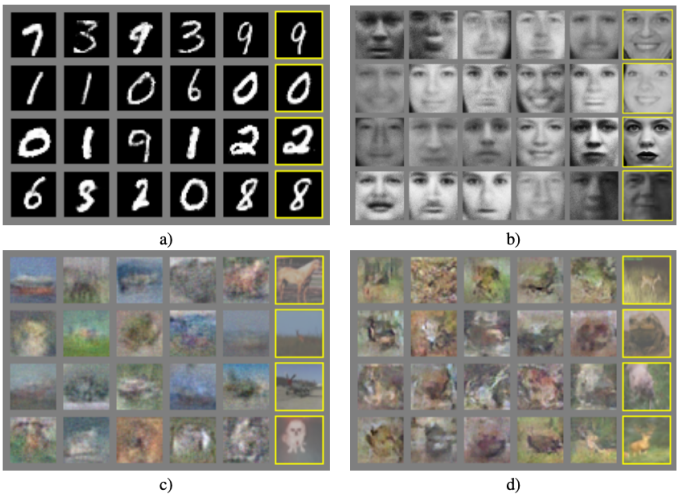

Figure 2: GAN 생성 결과 시각화한 모습. 각 샘플의 노란 박스 속 이미지는 실제 학습 데이터를 의미하며 나머지는 GAN이 새로 생성한 이미지이다. 이를 통해 모델이 단순히 학습 이미지를 외운 것이 아니라 새로운 이미지를 생성하고 있음을 확인할 수 있다.

과정에서 생성자와 판별자의 균형을 유지해야 한다는 점이다. 장점으로는 Markov chain 필요 없이, 오직 backprop으로 학습할 수 있으며, 추론 과정이 필요 없고 다양한 함수 구조를 자유롭게 사용 가능하다는 점이다. 또한 생성자는 데이터 자체를 직접 보지 않고 판별자를 통해 전달되는 gradien만 이용해 학습되므로, 모델이 더 날카로운 분포나 복잡한 모드를 표현할 수 있다는 통계적 이점도 존재한다.

결론적으로 본 논문은 적대적 학습(adversarial training)을 활용한 새로운 생성 모델(GAN)의 가능성을 제시하며, 다양한 확장 가능성을 보여준다. 특히 조건부 생성 모델, 보조 네트워크를 활용한 추론, 반지도 학습 학습 효율성 개선 등의 연구 방향을 제안하였다. 전체적으로 GAN 프레임워크는 실험을 통해 충분한 실용성과 확장 잠재력을 확인하였으며, 향후 다양한 응용 분야에서 발전 가능성이 크다는 점을 강조한다.

[1] *"[새로운 인공지능 기술 GAN] ① 스스로 학습하는 인공지능"* SAMSUNG SDS https://www.samsungsds.com/kr/insights/generative-adversarial-network-ai.html?referrer=https://www.google.com/

[2] *"[GAN] 생성적 적대 신경망(GAN) 쉽게 알아보기"* https://ebbnflow.tistory.com/167