GAN의 문제점과 평가지표들

GAN의 대표적인 문제점으로는 **(1)성능 평가** 와 **(2) 성능 개선**이 있다.

먼저 성능 평가부터 문제점을 살펴보겠다.

순서

1. 성능평가의 종류와 그 문제점

GAN은 그 모델을 평가하는 것이 어렵다.

Universal gold-standard discriminator가 존재하지 않기 때문이다.

GAN에게는 여러가지 평가지표가 존재하고, 모두 각각 장단점을 가지고 있다.

평가지표를 살펴보기 전에,

먼저 생성된 **이미지 평가의 중요한 두 가지 지표**는 다음과 같다 :

1. Fidelity : quality of images
2. Diversity : variety of images

따라서 평가 지표는 이 두가지 속성 모두에 대해 평가해야한다.

대표적인 평가지표로는 먼저 **FID(Fréchet Inception Distance)** 가 있다.

FID(Fréchet Inception Distance)

FID는 실제 이미지와 생성된 이미지 간의 특징거리 측정에 가장 널리 사용되는 메트릭 중 하나이다.

> **feature distance란?** 이미지들을 픽셀 단위로 비교하는 것이 아니라 특징들을 추출해서 비교하는 방법으로, pixel distance보다 위상변화에 대해 믿을만한 값을 얻을 수 있다.
> feature distance를 구하기 위해서는 특징을 나타내는 벡터가 연산을 위해 필요하다.

FID는 곡선을 따르는 점들의 위치와 순서를 고려하여 곡선 간의 유사성을 측정한다. 이는 두 분포 사이의 거리를 측정하는 데에도 사용된다.

수학적으로, FID는 두 다변량 정규분포 사이의 거리를 계산하는데 사용된다.

만약 일변량 정규분포의 경우, FID는 다음과 같이 계산된다 :

$$d(X,Y) = (\mu X - \mu Y)2 + (\sigma X - \sigma Y)2$$

여기서 μ와 σ는 정규분포의 평균 및 표준편차이며, X, Y는 두 개의 정규분포이다.

다변량 정규분포에 대한 FID는 다음과 같다.

$$FID = || \mu_X - \mu_Y || 2 - Tr(\sum_X + \sum_Y - 2\sqrt{\sum_X \sum_Y})$$

여기서 X와 Y는 두 개의 다변량 정규분포로, 실제와 가짜 분산을 나타낸다.

μ X 와 μY는 벡터 X와 Y의 크기이다.

Tr은 행렬의 대각합이며, ∑X와 ∑Y는 벡터의 공분산행렬이다.

FID가 낮다는 것은 두 분포의 거리가 가깝다는 것을 의미하며, 이는 즉 진짜와 가짜가 유사하다는 것을 뜻한다.

FID의 단점은 다음과 같다 :

1. 특정한 dataset으로 pre-trained된 경우, 다른 이미지를 다룰 때 원하는 특징을 포착하지 못할 수 있다.

2. pre-trained가 아닌 경우, 많은 수의 sample로 학습시키지 않으면 biased feature layer가 생기므로 FID score가 좋지 않을 수 있다.
3. 학습을 위한 시간이 오래 걸린다.
4. 표본의 분포가 정규분포가 아닌경우, 제한적인 통계량(평균, 분산) 만으로는 분포의 차이를 잘 못 설명할 수 있다.

또 다른 평가지표로는 FID 이전에 많이 사용되던 IS(Inception Score)가 있다

IS(Inception Score)

IS는 Inception 모델에서 영상이 픽별하기 쉬울수록, 또한 식별된 레이블의 편차가 풍부할수록 score가 높게 출력되도록 설계되었다.

> **엔트로피란?**
> 정보이론의 대표적인 개념으로 무작위성을 뜻한다.
> 예측하기 어렵다면 엔트로피는 높다.

실제 IS의 수식은 다음과 같다:

$$IS = exp(E_x KL(p(y|x)||p(y))) = exp(E_x E_{p(y|x)}[log(p(y|x)/p(y))])$$

좀더 간단하게 표현하면

$$IS = exp(E_{x\sim p_a} D_{KL}(p(y \mid x)||p(y)))$$

이다.

다음의 식에는 두가지 확률 $P(y \mid x)$와 $P(y)$가 존재한다.

먼저 $P(y \mid x)$는 조건부확률로, 생성된 이미지 x에 대해서 어떤 클래스y에 속하는지 예측하는 확률을 뜻한다.

또한 $P(y)$는 주변확률로, $P(y) = \int_z p(y|G(z))dz$ 다음과 같이 계산이 가능하다. 이는 여러 noise vector에 대해서 예측된 클래스들을 의미한다.

만약 GAN이 다양한 이미지를 생성한다면, P(y)는 높은 엔트로피를 가지게 될 것이다.

IS를 계산하기 위해서는 , 위의 식과 같이 조건부 확률과 주변확률의 KL-Divergence를 계산해준다.

> **KL divergence란?**
> KLD는 두 확률분포의 차이를 계산하는 데에 사용하는 함수로, 어떤 이상적인 분포에 대해 그 분포를 근사하는 다른 분포를 사용해 샘플링을 한다면, 그 때 발생할 수 있는 정보 엔트로피의 차이를 계산한다.
> KLD는 두 분포의 엔트로피 차이를 계산한다.
>
> 즉, p와 q의 cross entropy에서 p의 엔트로피를 뺀 값이다 :
> $KL(p||q) = H(p,q) - H(p)$
> 정확한 식은 다음과 같다 : $\int p(x) log(p(x)/q(x)) dx$

IS의 예시는 다음과 같다 :

fidelity와 diversity 모두 우수한 생성모델을 가정하자.

해당 모델에서 생성한 이미지들은 fidelity가 높으므로 판별기가 분류를 잘 할 것이다. 따라서 p(y|x)가 한 곳에서 높은 값을 갖는 분포가 된다.

또한 diversity가 높으므로 p(y)는 uniform distribution과 비슷한 형태의 분포가 될 것이다.

두 분포의 차이가 크므로, 두 분포의 KLD의 기댓값인 IS는 크다.

따라서 IS가 큰 값을 가진다면, fidelity와 diversity가 우수한 생성모델이라고 볼 수 있다.

IS의 단점은 다음과 같다 :

1. 생성자가 각 label마다 하나의 이미지만 반복해서 생성하는 경우 IS가 높지만 이는 inner diversity를 고려하지 못한 생성모델이 되어버린다.
2. fake image만을 이용하므로 real image와 비교하지 못한다
3. FID와 마찬가지로 학습된 이미지의 class와 다른 이미지를 다른 경우 원하는 특징을 포착하지 못할 수 있다.

또 다른 평가 방법으로는 HYPE가 있다.

HYPE(Human eYe Perceptual Evaluation)

HYPE는 사람이 진짜로부터 가짜를 구별하는데 필요한 시간 제한적인 인지 임계값을 결정하기 위해 시간 제약을 조정하며 이미지들을 보여준다.

신뢰할 수 있고, 일관된 측정방법이다.

Fidelity에 대해 평가할 수는 있지만, 다른 지표들(diversity, overfitting, train stability)은 평가하지 못한다.

HYPE(time)의 방법은 다음과 같다 :

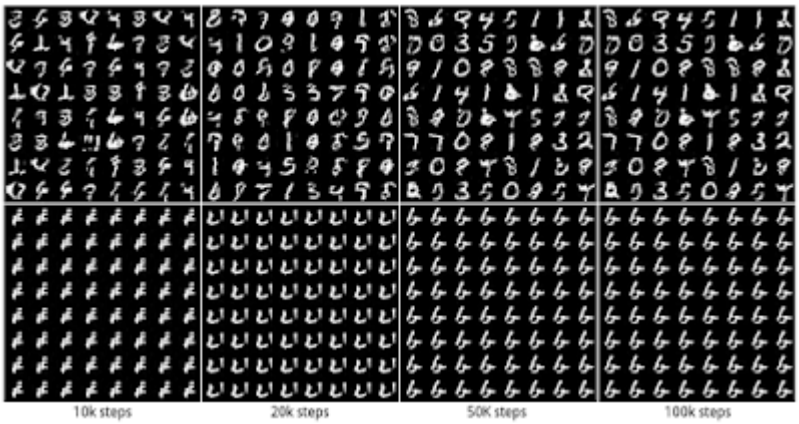
- 150개의 이미지(real:fake=1:1)들을 노출시간을 500ms으로 시작해서 가짜 진짜를 판별시킨다.
- 맞춘다면 노출시간을 줄이고, 틀린다면 늘린다.
- 각각 평가자의 마지막 값을 계산해서 평균을 구한 것이 점수가 된다.
- 높은 점수는 진짜와 가짜를 구분하는데 더 긴 시간이 필요하다는 것을 의미한다. 즉, 좋은 생성모델인 것을 뜻한다.

2. 성능 개선의 문제

Mode collapse

mode collapse란 generator가 서로 다른 data가 뮌 input z를 같은 output으로 mapping 하고자 할 때 발생한다.

partial mode collapse가 흔하게 발생하는데, 이는 generator가 동일한 색이나 텍스처를 가지고 여러 이미지를 만들려고 할 때 발생한다.



이를 알 수 있는 대표적인 예시로 다음의 MNIST 문제가 있다.

위가 unrolled GAN(발전된 GAN)이고, 아래가 vanilla GAN이다. vanilla GAN의 경우 여러 종류의 데이터가 아니라 한 가지의 모양만 나타내고 있음을 알 수 있다.

이러한 mode collapse가 생기는 원인은 무엇일까?

우리가 풀고자 하는 GAN 문제는 다음과 같은 minimax problem이다

$$G^* = min_G max_D V(G, D).$$

그런데 우리가 실제 학습을 할 때는 G와 D에 대한 update를 번갈아가며 해주기 때문에 NN의 입장에서는 minmax문제와 maxmin 문제가 구분되지 않는다 :

$$G^* = max_D min_G V(G, D).$$

문제는 이와 같은 maxmin problem의 경우에서 생긴다.

수식의 안쪽부터 살펴보면, G에 대한 minimization문제가 먼저 있기 때문에 generator의 입장에서는 현재 고정되어있는 discriminator가 가장 헷갈려 할 수 있는 sample 하나만, 즉 value V를 가장 최소화할 수 있는 mode 하나만을 내보내면 된다.

Discriminator와 generator가 서로 상호작용하며 학습이 진행되어야 하는데, 이렇게 학습에서 불균형이 일어날 시mode collapse가 발생하게 된다.

discriminator가 너무 학습이 잘 된 경우 : 완벽하게 generate된 이미지를 구분할 수 있으므로 generator는 어떠한 이미지를 내더라도 discriminator를 속일 수 없을 것이므로 학습이 진행되지 않을 것이다.

generator가 너무 학습이 잘 된 경우 : 이 경우 역시 다른쪽은 더이상 학습이 진행되지 않고 멈추어버린다. 이 경우 generator는 한 종류의 이미지만 계속 생성하게된다.