기존 GAN의 경우 D는 T,F에 대한 확률 model이다.

확률 model은 0,1 만 내뱉는다.

D를 energy function으로 사용하여 다른 각도로 보기 함.

G는 minimum E로 contrastive sample을 생성하려하고, D는 생성된 sample의 E가 크도록 학습됨.

* Introduction.

Energy-based model의 필수 요소는 input space에서 scalar로 mapping 되는 energy function이다.

훈련은 data로부터 맞는 것은 low energy로, 틀린 것은 high energy로 하는 energy surface를 구성하는 과정.

Supervised 에서는 (X,Y) pair가 맞으면 E↓. 틀리면 E↑.

Unsupervised 에서는 data manifold에는 낮은 E

Contrastive sample은 E를 높이는 data point나 낮은 data density 영역을 의미.

EBGAN 에서는 D를 energy function으로 본다.

Energy function (이하 D)는 G에 대한 cost function으로 볼 수 있다.

ㄴ data 밀도가 높은 곳에 낮은 값, 낮은 곳에 높은 값

G는 낮은 E 영역에서 sample을 생성하는 trainable parameterized function.

D는 AE고, energy는 reconstruction error.

* The EBGAN MODEL

P_z에서 P_{data}에 대한 분포를 갖는 sample을 생성하고,

D는 생성된 이미지나, real image를 받고, $E \in R$에 해당하는 energy를 측정

Objective function.

D는 E에 대해 objective function을 가져, 실제에는 낮은 E를, 생성에는 높은 E를 할당.

$$\mathcal{L}_D(x, z) = D(x) + [m - D(G(z))]^+$$
$$\mathcal{L}_G(z) = D(G(z))$$

m은 positive margin 이고, $[\cdot]^+$는 $max(0, \cdot)$을 의미.

$\mathcal{L}_D$, $\mathcal{L}_G$를 maximize

Optimality of the Solution.

P_G는 G로 생성된 image의 data distribution.

$V(G, D) = \int_{x,z} \mathcal{L}_D(x, z) P_{data}(x) P_z(z) dx dz$

$U(G, D) = \int_z \mathcal{L}_G(z) P_z(z) dz$

일 때, Nash equilibrium을 만족하는 (G^*, D^*)에 대해서,

$V(G^*, D^*) \leq V(G^*, D) \quad \forall D$

$U(G^*, D^*) \leq U(G, D^*) \quad \forall G$

Theorem 1. (D^*, G^*)가 nash equilibrium에 해당할 때, 거의 항상 $P_{G^*} = P_{data}$고, $V(D^*, G^*) = m$ 이다.

Theorem 2. Nash equilibrium은 존재하고, i) $P_{G^*} = P_{data}$, ii) $D^*(x) = \gamma$ 인 $\gamma \in [0, m]$이 존재한다.

$$\begin{aligned}
U(G^*, D^*) - U(G_0, D^*) &= \int_x (p_{data} - p_{G_0}) D^*(x) \mathrm{d}x &(21)\\
&= \int_x (p_{data} - p_{G_0})(D^*(x) - C) \mathrm{d}x &(22)\\
&= \int_S (p_{data} - p_{G_0})(D^*(x) - C) \mathrm{d}x + \int_{\mathcal{R}^N \setminus S} (p_{data} - p_{G_0})(D^*(x) - C) \mathrm{d}x &(23)\\
&> 0 &(24)
\end{aligned}$$

Using Auto encoder.

D는 AE로 구성된다.

$D(x) = \| Dec(Enc(x)) - x \|$

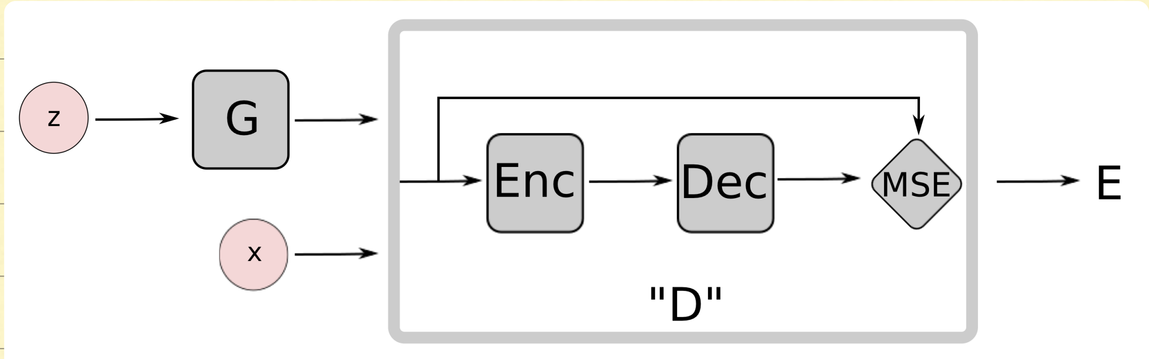

Figure 1: EBGAN architecture with an auto-encoder discriminator.

0과 1의 D보다 recon-based의 경우 D에 다양한 target을 제공

Binary의 경우 target이 2개 이므로, minibatch에 대해서,

서로 다른 sample의 gradient가 orthogonal에서 멀다.

↳ 두 sample의 기울기가 와 같이 서로 상충될 수 있다.

반면 recon-base의 경우 여러 방향의 기울기를 생성하여 효용성 있는 더 큰 minibatch를 허용한다.

그리고 AE는 energy-based에 사용되고, supervision이나 negative sample 없이 manifold 학습이 가능하다.

↳ D가 자체적으로 data manifold를 발견할 수 있다.

↳ D가 G의 관점에서 더 자유로움

Connection to the regularized AE

한가지 문제는 AE가 identity function만 학습할 수도 있다

ㄴ data manifold 밖에 더 높은 energy를 push해야함.

ㄴ latent space regularization.

ㄴ recon에 제약을 걸어, input일 때 낮은 복원이 낮은 E

D가 contrasive sample을 생성하여 regularization.

ㄴ 유연성

ㄴ G의 안정한 훈련

ㄴ sample 생성과, E function 학습의 직접적인 상호작용

Repelling regularizer

repelling regularizer는 P_{data}의 하나 또는 몇으로 몰려된 생성된 sample로부터 model을 유지하기 위한 목적으로

EBGAN의 AutoEncoder에 쓰인다.

Minibatch D는 같은 의도로 개발됨

repelling regularizer는 representation level에서 수행되는 Pulling-away Term(PT)이다

$S \in \mathbb{R}^{s \times N}$ 를 Encoder의 output에서 batch의 sample representation이라고 하면,

$$f_{PT}(S) = \frac{1}{N(N-1)} \sum_i \sum_{j \neq i} \left(\frac{S_i^\top S_j}{\|S_i\| \|S_j\|} \right)^2.$$

ㄴ sample들을 직교화하려고 함.

ㄴ cosine similiarity는 below boundary로 invariant scale

ㄴ G에만 적용됨.

* Experiment.

semi-supervised 에서는 m을 점점 줄여나간다.

P_G가 data manifold에 가까워지면. D로부터의 loss가 줄어들도록

$P_G = P_{data}$ 일때. $m=0$ 으로.

m은 너무 크면 불안정, 작으면 mode drop

ㄴ Appendix D

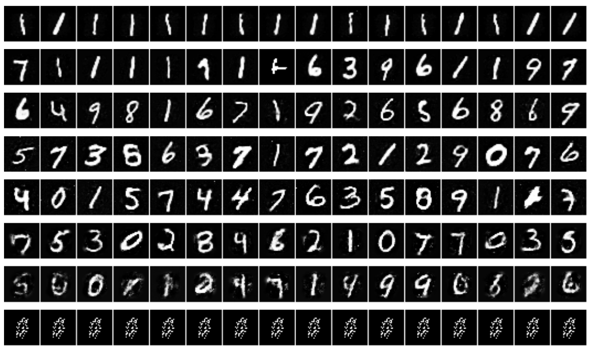

Figure 9: Generation from the EBGAN auto-encoder model trained with different m settings. From top to bottom, m is set to 1, 2, 4, 6, 8, 12, 16, 32 respectively. The rest setting is `nLayerG`=5, `nLayerD`=2, `sizeG`=1600, `sizeD`=1024, `dropoutD`=0, `optimD`=ADAM, `optimG`=ADAM, `lr`=0.001.

Contrasive sample은 reg에 효과적이고, classifier에 대해 더 많은 sample

* Appendix

A.

Lemma 1. *Let $a, b \geq 0$, $\varphi(y) = ay + b\,[m-y]^+$. The minimum of φ on $[0, +\infty)$ exists and is reached in m if $a < b$, and it is reached in 0 otherwise (the minimum may not be unique).*

Proof. The function φ is defined on $[0, +\infty)$, its derivative is defined on $[0, +\infty)\backslash\{m\}$ and $\varphi'(y) = a - b$ if $y \in [0, m)$ and $\varphi'(y) = a$ if $y \in (m, +\infty)$.
So when $a < b$, the function is decreasing on $[0, m)$ and increasing on $(m, +\infty)$. Since it is continuous, it has a minimum in m. It may not be unique if $a = 0$ or $a - b = 0$.
On the other hand, if $a \geq b$ the function φ is increasing on $[0, +\infty)$, so 0 is a minimum. □

Lemma 2. *If p and q are probability densities, then $\int_x \mathbb{1}_{p(x)<q(x)}\mathrm{d}x = 0$ if and only if $\int_x \mathbb{1}_{p(x)\neq q(x)}\mathrm{d}x = 0$.*

Proof. Let's assume that $\int_x \mathbb{1}_{p(x)<q(x)}\mathrm{d}x = 0$. Then

$$
\begin{aligned}
& \int_x \mathbb{1}_{p(x)>q(x)}(p(x) - q(x))\mathrm{d}x && (15)\\
= & \int_x (1 - \mathbb{1}_{p(x)\leq q(x)})(p(x) - q(x))\mathrm{d}x && (16)\\
= & \int_x p(x)\mathrm{d}x - \int_x q(x)\mathrm{d}x + \int_x \mathbb{1}_{p(x)\leq q(x)}(p(x) - q(x))\mathrm{d}x && (17)\\
= & 1 - 1 + \int_x \left(\mathbb{1}_{p(x)<q(x)} + \mathbb{1}_{p(x)=q(x)}\right)(p(x) - q(x))\mathrm{d}x && (18)\\
= & \int_x \mathbb{1}_{p(x)<q(x)}(p(x) - q(x))\mathrm{d}x + \int_x \mathbb{1}_{p(x)=q(x)}(p(x) - q(x))\mathrm{d}x && (19)\\
= & 0 + 0 = 0 && (20)
\end{aligned}
$$

So $\int_x \mathbb{1}_{p(x)>q(x)}(p(x) - q(x))\mathrm{d}x = 0$ and since the term in the integral is always non-negative, $\mathbb{1}_{p(x)>q(x)}(p(x) - q(x)) = 0$ for almost all x. And $p(x) - q(x) = 0$ implies $\mathbb{1}_{p(x)>q(x)} = 0$, so $\mathbb{1}_{p(x)>q(x)} = 0$ almost everywhere. Therefore $\int_x \mathbb{1}_{p(x)>q(x)}\mathrm{d}x = 0$ which completes the proof, given the hypothesis. □