Theoretical analysis for GAN

Speaker: Yi-Shan Wu

Institute of Information Science

Academia Sinica

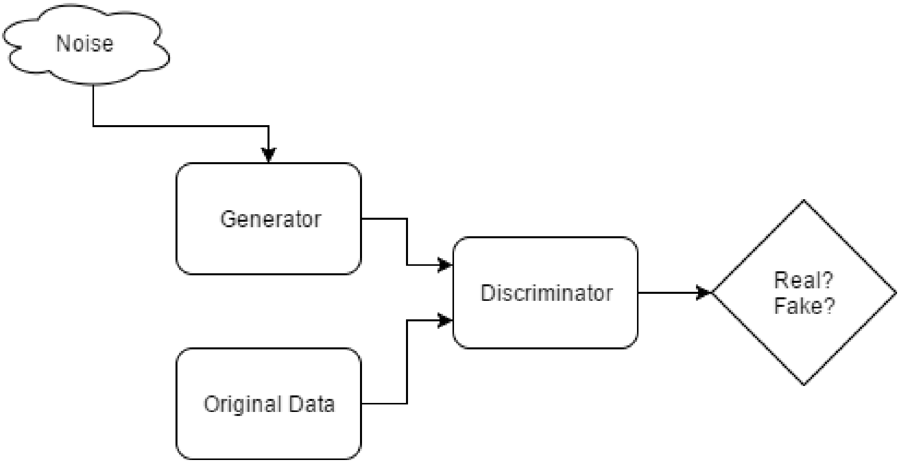
Taiwan

01 Dec, 2017

Outline

1. Basic ideas of GAN
2. Generalization and Equilibrium in GANs (ICML2017)
3. Do GANs actually learn the distribution? An empirical study
4. Basic ideas of Encoder-Decoder GAN (biGAN)
5. Theoretical Limitations of Encoder-Decoder GAN

architecture of GAN



Notations

- $\{D_v, v \in V\}$ is a class of discriminators.
- $\{G_u, u \in U\}$ is a class of generator.
- z is the seed distribution, which is usually assumed to be spherical Gaussian.
- D_{real} is the real distribution. D_{synth} is the distribution by generator
- p is the capacity of the discriminator, namely, the number of trainable parameters
- $L-$Lipschitz w.r.t the parameters of G means $\forall u, u' \in U$ and any input z, we have

$$\|G_u(z) - G_{u'}(h)\| \leq L\|u - u'\|$$

Objective for GAN

- GAN
  - objective function:

$$\min_{u \in U} \max_{v \in V} \mathbb{E}_{x \sim D_{real}}[log(D_v(x))] + \mathbb{E}_z[log(1 - D_v(G_u(z)))]$$

    where $\{D_v : v \in V\}$ can be any function
  - If the ideal discriminator is obtained, the objective is to minimize the JS-divergence between D_{real} and D_{synth}
- Wasserstein GAN
  - objective function:

$$\min_{u \in U} \max_{v \in V} \mathbb{E}_{x \sim D_{real}}[D_v(x)] - \mathbb{E}_z[D_v(G_u(z))]$$

    where $\{D_v : v \in V\}$ is limited to 1-Lipschitz functions.
  - The goal here is then to minimize the Wasserstein distance between D_{real} and D_{synth}

Empirical loss

Both divergence can be minimized under two strong assumptions that

- We have enough samples.
- The networks has enough capacity to approximate the ideal discriminator.

Unfortunately, both of them are impratical. Furthermore, there is no way to generalize to real world, since

- For two uniform Gaussian distributions, μ and ν with m samples

$$d_{JS}(\mu, \nu) = 0, \quad d_{JS}(\hat{\mu}, \hat{\nu}) = log(2)$$

$$d_W(\mu, \nu) = 0, \quad d_W(\hat{\mu}, \hat{\nu}) \geq 1.1$$

F-distance

To avoid the gap between empirical distance and the ideal distance when obtaining few samples, we define another distance

Definition (F-distance)

For two distribution μ, ν, we define the F-distance w.r.t $\{D_v : v \in V\}$ where V is a compact set,

$$d_F(\mu,\nu) = \sup_{v \in V} |\mathbb{E}_{x \sim \mu}[D_v(x)] - \mathbb{E}_{x \sim \nu}[D_v(x)]|$$

Theorem

Suppose the discriminator has capacity p. Suppose we have at least m samples for both distribution μ and ν. When $m \geq O(plog(p/\epsilon)/\epsilon^2)$, we have w.h.p. that

$$|d_F(\hat{\mu},\hat{\nu}) - d_F(\mu,\nu)| \leq \epsilon$$

Finite Discriminators have limited power

Theorem 2

Suppose the discriminator has capacity p. Let $\hat{\mu}$ be the empirical version of distribution μ with m samples. When $m \geq O(\frac{plog(p/\epsilon)}{\epsilon^2})$, we have with probability at least $1 - exp(-p)$ such that

$$d_{F,\phi}(\mu, \hat{\mu}) \leq \epsilon$$

That is, the F-distance for discriminator with p parameters cannot distinguish between a distribution μ and a distribution with support $O(\frac{plog(p/\epsilon)}{\epsilon^2})$.
Thus, a finite-capacity discriminator is unable to even enforce that D_G has large diversity, let alone enforce that $D_G \approx D_{real}$.

Empirical studies

However, the above theorem is merely a theoretical result. This does not settle the question of what happens with realistic sample sizes and realistic training. Thus, we need some empirical test.
Ideally, we hope that the "support" will be infinite to have large diversity on images.

- Results on CelebA dataset

DCGAN	MIX+DCGAN	ALI(BiGAN)
400^2	400^2	1000^2

- Results on CIFAR-10 dataset by stacked GAN

Aeroplane	Auto-Mobile	Bird	Cat	Deer	Frog
500^2	50^2	500^2	100^2	500^2	50^2

Birthday paradox test

Suppose there are k people in a room. How large must k be before we have a high ($> 50\%$) likelihood of having two people with the same birthday?

- 100% probability if $k > 366$
- $> 50\%$ probability even when k is as small as 23

(proof:)

$$\begin{aligned}
& Pr[\text{there is at least a collision among n samples}] \\
= & 1 - Pr[\text{there is no collision}] \\
= & 1 - 1 \cdot (1 - \frac{1}{365}) \cdot (1 - \frac{2}{365}) \cdots (1 - \frac{n-1}{365}) \\
\geq & 1 - (e^{-n/730})^{n-1} \\
\geq & 1/2 \text{ , when } n > 23
\end{aligned}$$

If for samples of size N have "collision" with good probability, then suspect that the distribution has support size about N^2.

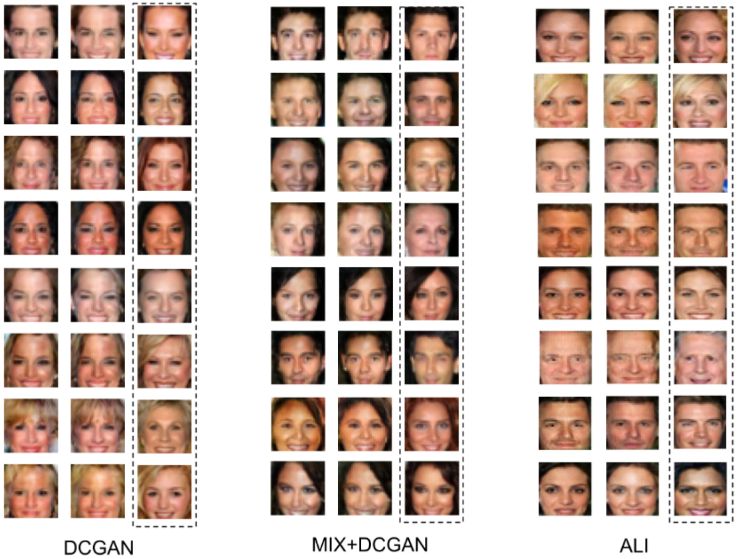
Birthday paradox test for support size

1. Pick a sample of size N from the generated distribution.
2. Use an automated measure of image similarity to flag the 20 most similar pairs in the sample.
3. Visually inspect the flagged pairs and check for duplicates.
4. Repeat.

We look for “near-duplicates” for this continuous distribution. Thus, it is necessary to some heuristic metric for step 2.

- CelebA: Euclidean distance in pixel space.
  It works probably because the samples are centered and aligned.
- CIFAR-10: pre-train a discriminative CNN for classification problem. Use the top layer as an embedding of the image, and measure euclidean distance in embedding space.

CelebA



DCGAN

MIX+DCGAN

ALI

CIFAR-10

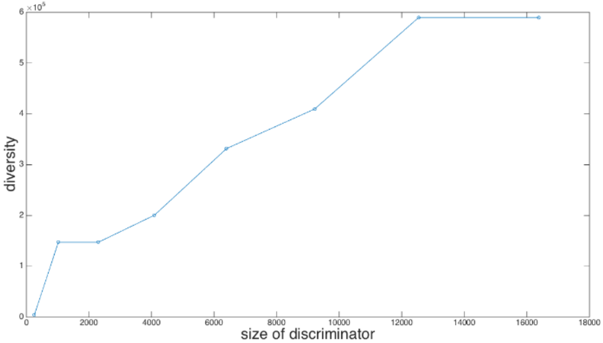
This heuristic similarity test quickly becomes useless if the samples display noise artifacts, and thus was effective only on the very best GANs that generate the most real-looking images.
For CIFAR-10 this led us to Stacked GAN, (Inception Score 8.59) Its diversity is measured within each class separately.

Aeroplane	Auto-Mobile	Bird	Cat	Deer
500^2	50^2	500^2	100^2	500^2
Dog	Frog	Horse	Ship	Truck
300^2	50^2	200^2	500^2	100^2

CIFAR-10

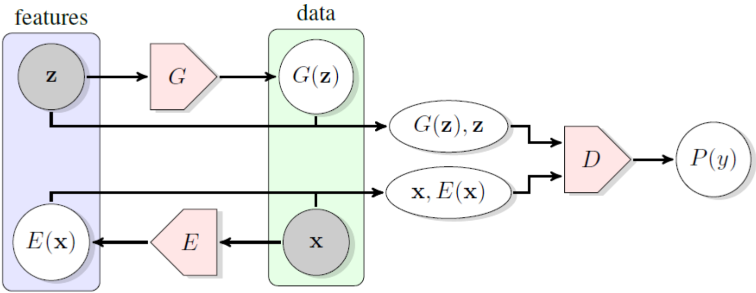
diversity v.s. discriminator size

- Theoretical
  - discriminator size : p
  - theoretical support size : $O(\frac{plog(p)}{\epsilon^2})$
- Experimental
  - Collision batch size : s
  - empirical support size (diversity) : s^2



framework of BiGAN

- Generative player:
  - generator G
  - encoder E
- Discriminator player: D



Objective of biGAN

biGAN objective:

$$\min_{G,E} \max_{D} |\mathbb{E}_{x\sim\hat{\mu}}\phi(D(x, E(x))) - \mathbb{E}_{z\sim\hat{\nu}}\phi(D(G(z), z))|$$

- $\hat{\mu}$: the empirical distribution over data samples x
- μ: noised data distribution
- $\hat{\nu}$: distribution over random seed z (standard Gaussian)
- ν: spherical zero-mean Gaussian
- ϕ: concave “measuring” function.

Limitation of encoder-decoder GAN architectures

Some assumption:

- ϕ is assumed to be in the range $[-\Delta, \Delta], \Delta \geq 1$ and L_ϕ-Lipschitz
- $F = \{D_v : v \in V\}$ discriminators are L-lipschitz w.r.t v
- Domain(μ)= $\mathbb{R}^d$
- Domain(ν)= $\mathbb{R}^{\tilde{d}}$, $\tilde{d} < d$
- encoder net is smaller than the discriminator

Main theorem

$\exists$ a generator G of support $\frac{p\Delta^2 log^2(p\Delta L L_\phi/\epsilon)}{\epsilon^2}$ and an encoder E with at most $\tilde{d}$ non-zero weights, s.t. $\forall D$ that are $L-$Lipschitz and have capacity less than p

$$|\mathbb{E}_{x\sim\mu}\phi(D(x, E(x))) - \mathbb{E}_{z\sim\nu}\phi(D(G(z), z))| \leq \epsilon$$

That is, an encoder E with small complexity and a generator G with small-support distribution are able to fool all discriminators.

Precise noise model

- $\tilde{\mu}$: the distribution of unnoised images
- ν : the distribution of seeds (zero-mean spherical Gaussian)

To produce a noised image x in μ take a $\tilde{x} \sim \tilde{\mu}$ and a $z \sim \nu$ and output $x = \tilde{x} \otimes z$, which is defined as

$$x_i = \begin{cases} z_{\frac{i}{\lfloor \frac{d}{\tilde{d}} \rfloor}}, & \text{if } i \equiv 0 (mod \lfloor \frac{d}{\tilde{d}} \rfloor) \\ \tilde{x}_i, & otherwise \end{cases}$$

Proof (1/4)

Main idea: show the existence of generator/encoder pair via a probabilistic construction that success w.h.p.

- encoder E : $E(x = \tilde{x} \otimes z) = z$.
  It can be captured by connecting the i-th output to the $(i\lfloor \frac{d}{\tilde{d}} \rfloor)$-th input.
- generator G :
  1. Define a partition of Domain(ν)= $\mathbb{R}^{\tilde{d}}$ into m equal-measure blocks
  2. Sample m samples $x_1^*, x_2^*, \cdots, x_m^*$ from image distribution
  3. For a sample z select corresponding x^* and output $G(z) = x^*_{cor(z)} \otimes z$

Since the set of samples $\{x\}$ is random, this species a distribution over generators. We prove that with high probability, one of these generators satisfies the statement of Theorem.

Proof (2/4)

Thus, we have

$$\mathbb{E}_{x\sim\mu}\phi(D(x,E(x))) = \mathbb{E}_{x\sim\tilde{\mu},z\sim\nu}\phi(D(x\otimes z,z)) = \mathbb{E}_G\mathbb{E}_{z\sim\nu}\phi(D(G(z),z))$$

The remaining is to show that $\mathbb{E}_{z\sim\nu}\phi(D(G(z),z))$ concentrates enough around $\mathbb{E}_G\mathbb{E}_{z\sim\nu}\phi(D(G(z),z))$, so that we can say w.h.p over G the objective is small. That is

Lemma (Concentration of good generators)

With high probability over the choice of G,

$$|\mathbb{E}_{z\sim\nu}\phi(D(G(z),z)) - \mathbb{E}_G\mathbb{E}_{z\sim\nu}\phi(D(G(z),z))| \leq \epsilon$$

for all D of capacity at most p

Proof (3/4)

We further define:

- Non-colliding: A set T of samples from ν is "non-colliding" if no two lie in the same block.
- T_{nc} is the distribution over non-colliding sets $\{z_1, \cdots z_m\}$

Thus, for a fixed G and a fixed D

$$\mathbb{E}_{z\sim\nu}\phi(D(G(z),z)) = \mathbb{E}_{T\sim T_{nc}}\mathbb{E}_{z\sim T}\phi(D(G(z),z))$$

Consider $\mathbb{E}_{z\sim T}\phi(D(G(z),z))$ as a random variable in $T \sim T_{nc}$ and G for a fixed D. Then we can write

$$\mathbb{E}_{z\sim T}\phi(D(G(z),z)) = f(x_i^*, z_i, i \in [m])$$

for some f. Note that the variables are all independent, we can apply McDiarmid's inequality.

Proof (4/4)

Therefore, we have

$$Pr_{T,G}(|f(x_i^*, z_i, i \in [m]) - \mathbb{E}_{T,G} f(x_i^*, z_i, i \in [m])| \geq \epsilon) \leq exp(-\Omega(m\epsilon^2))$$

Since this only hold for a fixed D, we can build a ϵ/LL_ϕ-net for the discriminators as in the previous paper, and union bounding over the points in the net. We then prove the theorem that w.h.p

$$|\mathbb{E}_{x\sim\mu}\phi(D(x, E(x))) - \mathbb{E}_{z\sim\nu}\phi(D(G(z), z))| \leq \epsilon$$

holds for all discriminators D