Generalization and Equilibrium in Generative Adversarial Nets (GANs

Sanjeev Arora[1] Rong Ge[2] Yingyu Liang[1] Tengyu Ma[1] Yi Zhang[1]

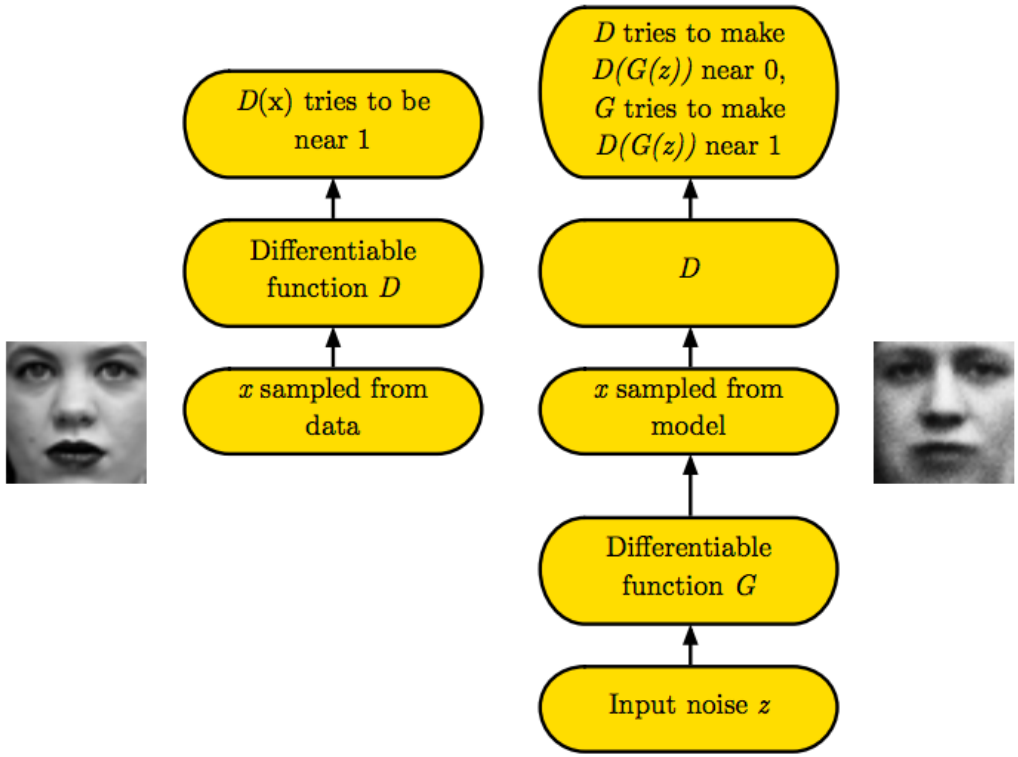
[1]Princeton University,

[2]Duke University

ICML,2017

Presenter: Arshdeep Sekhon

Background: GANs



Background: GANs

1. Given a sample from noise (a uniform or a Gaussian distribution), generate a sample from the true data distribution

2. Training is continued until the generator wins, meaning that the discriminator can do no better than random guessing when deciding whether or not a particular sample came from D or D_{real}.

$$\min_{u \in \mathcal{U}} \max_{v \in \mathcal{V}} \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{real}} [\log D_v(x)] + \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{G_u}} [\log(1 - D_v(x))].$$

Background: GAN Nash Equilibrium

1. "a solution concept of a non-cooperative game involving two or more players in which each player is assumed to know the equilibrium strategies of the other players, and no player has anything to gain by changing only his own strategy"
2. Just as a zero gradient is a necessary condition for standard optimization to halt, the corresponding necessary condition in a two-player game is an equilibrium.

Motivation

1. A standard analysis shows that when the discriminator capacity (= number of parameters) and number of samples is large enough, then a win by the generator implies that D is very close to D_{real}
2. D_{real} can be very complicated
3. Curse of dimensionality: : in d dimensions there are $exp(d)$ directions whose pairwise angle exceeds say $\pi/3$, and each could be the site of a peak(mode).
4. sufficiently large in this analysis may need to be $exp(d)$.

Defining Generalization for GANs

1. In supervised classification, if the training and test error are close to each other it is said to generalize well

Generalizability for GANs

A divergence or distance $d(\cdot,\cdot)$ between distributions generalizes with m training examples and error ϵ if for the learnt distribution D_G:

1. $$|d(D_{real}, D_G) - d(\hat{D}_{real}, \hat{D}_G)| \leq \epsilon \quad (1)$$
2. $\hat{D_{real}}, \hat{D}_G$ are the empirical distribution of D_{real}, D_G

Wasserstein Distance and JS divergence don't generalize

Lemma 2 (Lemma 1 restated). *Let μ be uniform Gaussian distributions $\mathcal{N}(0, \frac{1}{d}I)$ and $\hat{\mu}$ be an empirical versions of μ with m examples. Then we have*

$$d_{JS}(\mu, \hat{\mu}) = \log 2$$
$$d_W(\mu, \hat{\mu}) \geq 1.1$$

Theorem A.1. *Let μ, ν be uniform Gaussian distributions $\mathcal{N}(0, \frac{1}{d}I)$. Suppose $\hat{\mu}, \hat{\nu}$ are empirical versions of μ, ν with m samples. Then with probability at least $1 - m^2 \exp(-\Omega(d))$ we have*

$$d_{JS}(\mu, \nu) = 0, d_{JS}(\hat{\mu}, \hat{\nu}) = \log 2.$$
$$d_W(\mu, \nu) = 0, d_W(\hat{\mu}, \hat{\nu}) \geq 1.1.$$

Further, let $\tilde{\mu}, \tilde{\nu}$ be the convolution of $\hat{\mu}, \hat{\nu}$ with a Gaussian distribution $N(0, \frac{\sigma^2}{d}I)$, as long as $\sigma < \frac{c}{\sqrt{\log m}}$ for small enough constant c, we have with probability at least $1 - m^2 \exp(-\Omega(d))$.

$$d_{JS}(\tilde{\mu}, \tilde{\nu}) > \log 2 - 1/m.$$

$\mathbb{F}$-distance

$\mathbb{F}$-distance

Let F be a class of functions from $R_d \to [0, 1]$ and ϕ be a concave measuring function. Then the F-divergence with respect to ϕ between two distributions μ and ν supported on R_d is defined as

$$d_{F,\phi} = \sup_{D \in F} |E_{x \in \mu}[\phi(D(x))] + E_{x \in \nu}[\phi(1 - D(x))]| - 2\phi(\frac{1}{2}) \qquad (2)$$

For example: When $\phi(t) = log(t)$ and F = all functions from R_d to [0, 1], we have that $d_{F,\phi}$, is the same as JS divergence.
When $\phi(t) = t$ and F = all 1-Lipschitz functions R_d to [0, 1], then $d_{F,\phi}$, is the Wasserstein distance.

Theorem 1: Generalization holds for Neural net distance

, let μ, ν be two distributions and $\hat{\mu}, \hat{\nu}$ be empirical versions with at least m samples each. There is a universal constant c such that when $m \geq \frac{cp^2 log(LL\phi p/\epsilon)}{\epsilon^2}$, we have with probability at least $1 - exp(-p)$ over the randomness of μ and ν,

$$|d_{F,\phi}(\hat{\mu}, \hat{\nu}) - d_{F,\phi}(\mu, \nu)| \leq \epsilon \tag{3}$$

Theorem 1 shows that :

1. the neural network divergence (and neural network distance) has a much better generalization properties than Jensen-Shannon divergence or Wasserstein distance.
2. If the GAN successfully minimized the neural network divergence between the empirical distributions, $d(\hat{D_{real}}, \hat{D_G})$, then we know the neural network divergence $d(D_{real}, D_G)$ between the distributions D_{real} and D_G is also small

The problem: Generalization vs Diversity

Low-capacity discriminators cannot detect lack of diversity

Let $\hat{\mu}$ be the empirical version of distribution μ with m samples. There is a some universal constant c such that when m $m \geq \frac{cp^2 log(LL\phi p/\epsilon)}{\epsilon^2}$ we have that with high probability, $d_{F,\phi}(\mu, \hat{\mu}) \leq \epsilon$

The neural network distance for nets with p parameters cannot distinguish between a distribution μ and a distribution with support $\tilde{O}(\frac{p}{\epsilon^2})$.

Expressive Power and Equilibrium

1. Question: Why does the generator always win so that in the end t the discriminator is unable to do much better than random guessing ?
2. Was it sheer luck that so many real-life distributions D_{real} turned out to be close in neural-net distance to a distribution produced by a fairly compact neural net?

Expressive Power and Equilibrium

1. A mixture of infinite gaussians can approximate any density.
2. Say a generator is an infinite mixture of deep nets
3. Now this generator can approximate any density.
4. Will always win against simple or powerful discriminators

Expressive Power and Equilibrium

1. Now say we have a limited number of deep Nets
2. Now the generator can not approximate all possible densities
3. Can this generator still fool the discriminator into thinking it's samples are actually from the real distribution?
4. This paper: YES

Informal Theorem:

Informal Theorem

If the discriminator is a deep net with p parameters, then a mixture of $O(plog(p/\epsilon)\epsilon^2)$ generator nets can produce a distribution D that the discriminator will be unable to distinguish from D_{real} with probability more than ϵ

Pure Strategy and Pure Equilibrium

1. A pure strategy determines all your moves during the game (and should therefore specify your moves for all possible other players' moves).
2. A mixed strategy is a probability distribution over all possible pure strategies

General ϕ: mixed equilibrium

1. For a class of generators $\{G_u, u \in U\}$ and a class of dicriminators $\{D_v, v \in V\}$
2. The payoff is defined as:

$$F(u, v) = E_{x \sim D_{real}}[\phi(D_v(x))] + E_{x \sim D_G}[1 - \phi(D_v(x))] \tag{4}$$

3. Pure equilibrium may not exist for pure strategies
4. The well-known min-max theorem (v. Neumann, 1928) in game theory shows if both players are allowed to play mixed strategies then the game has a min-max solution.

Von Neuman MiniMax Game

Von Neuman MiniMax Game

There exists value V , and a pair of mixed strategies (S_u, S_v) s.t. $\forall v, E_{u\sim S_u}[F(u,v)] \leq V, \forall u, E_{v\sim S_v}[F(u,v)] \geq V$.

1. The payoff is generated by the generator first sample $u \sim S_u, h \sim D_h$, and then generate an example $x = G_u(h)$.
2. Pay off for Discriminator:

$$F(u,v) = E_{x\sim D_{real}, v\sim S_v}[\phi(D_v(x))] + E_{h\sim D_h, v\sim S_v}[1 - \phi(D_v(G_u(h)))] \tag{5}$$

3. this equilibrium involving an infinite mixture makes little sense in practice: define a ϵ **-approximate equilibrium**

ϵ -approximate equilibrium

ϵ -approximate equilibrium

A pair of mixed strategies (S_u, S_v) is an ϵ- approximate equilibrium, if for some value V $\forall v \in V$, $EuSu[F(u,v)]V+;8u2U, EvSv[F(u,v)]V$. If the strategies S_u, S_v are pure strategies, then this pair is called an *epsilon*-approximate pure equilibrium.

finite mixture of D and G

In the settings above, there is a universal constant C ¿ 0 such that for any ϵ, there exists $T = \frac{C\Delta^2 plog(LL^{'}L\phi p/\epsilon)}{\epsilon^2}$ generators $Gu1, ..., GuT$ and T discriminators $Dv1, ..., DvT$, let S_u be a uniform distribution on u_i and S_v be a uniform distribution on v_i, then (S_u, S_v) is an ϵ-approximate equilibrium. Furthermore, in this equilibrium the generator wins, meaning discriminators cannot do better than random guessing.

MIX+GANs

1. using a mixture of (not too many) generators and discriminators guarantees existence of approximate equilibrium: more stable training.
2. use a mixture of T components,
3. train a mixture of T generators $\{G_{u_i}, i \in [T]\}$
4. T discriminators $D_{v_i}, i \in [T])$ which share the same network architecture but have their own trainable parameters.
5. Maintaining a mixture means maintaining a weight w_{u_i} for the generator G_{u_i} which corresponds to the probability of selecting the output of G_{u_i}
6. Final objective:

$$min_{u_i,\alpha_{u_i}} max_{v_j,\alpha_{v_j}} E_{i,j\in T}(F(u_i, v_j)) = min_{u_i,\alpha_{u_i}} max_{v_j,\alpha_{v_j}} \sum_{i,j\in[T]} w_{u_i} w_{v_j} (F(u_i \quad (6)$$

Experiments: Qualitative results

(a)

(b)

(a)

(b)

Experiments: Quantitative results

Method	Score
SteinGAN (Wang & Liu, 2016)	6.35
Improved GAN (Salimans et al., 2016)	8.09±0.07
AC-GAN (Odena et al., 2016)	8.25 ± 0.07
S-GAN (best variant in (Huang et al., 2017))	8.59± 0.12
DCGAN (as reported in (Wang & Liu, 2016))	7.37
DCGAN (best variant in (Huang et al., 2017))	7.16±0.10
DCGAN (5x size)	7.34±0.07
MIX+DCGAN (with 5 components)	7.72±0.09
WASSERSTEINGAN	3.82±0.06
MIX+WASSERSTEINGAN (with 5 components)	4.04±0.07
Real data	11.24±0.12