The analysis of GANs from mathematics perspective

model part

2017-7-14

Table

Generative Adversarial Networks[1]

Generator: G, Parameterization θ_g, The true distribution of data $P_r(PDF: \ p_r(x))$, Generated distribution P_g $(PDF: \ p_g(x))$

Discriminator D, Parameterization θ_d, discriminate the probability of a data comes from P_g

The objective function is a classical saddle point problem

$$\min_G \max_D V(D, G) = \mathbb{E}_{x\sim P_r}[\log D(x)] + \mathbb{E}_{x\sim P_g}[\log(1 - D(x))] \tag{1}$$

using variational method to D, it is not hard to get $D^*(x) = \dfrac{p_r(x)}{p_g(x) + p_r(x)}$, substitute it into the objective function, we have:

$$\min 2JSD(P_r||P_g) - 2\log 2 \tag{2}$$

FACT: The optimization of the primal problem is equivalent to the optimization of the JS divergence between P_r and P_g

[1] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, NIPS, 2014

The Generalization of JS divergence[2]

The methods to measure the distance between two distributions are various, some work focus on the generalization of the JSD.

f-divergence family(Ali-Silvey distances):

$$D_f(P_r||P_g) = \int_x p_g(x) f(\frac{p_r(x)}{p_g(x)}) dx \tag{3}$$

in which, generative function $f : \mathbb{R}_+ \to \mathbb{R}$ is convex, lower-semicontinuous, with $f(1) = 0$

Based on the powerful theorem that if f is convex and lower semicontinuous, then $f^{**} = f$, it can be rewritten as

$$f(u) = \sup_{t \in dom_{f^*}} \{tu - f^*(t)\} \tag{4}$$

Thus, the lower bound for the divergence of P_r and P_g comes as:

$$D_f(P_r||P_g) \geq \sup_{T \in \mathcal{T}} (\mathbb{E}_{x \sim P_r}[T(x)] - \mathbb{E}_{x \sim P_g}[f^*(T(x))]) \tag{5}$$

[2] f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization.

Still by variational method

$$T^*(x) = f'(\frac{p_r(x)}{p_g(x)}) \tag{6}$$

The primal objective function can be rewritten as：

$$\min_{\theta_g}\max_{w} F(\theta_g, w) = \min_{\theta_g}\max_{w} \mathbb{E}_{x\sim P_r}[T_w(x)] - \mathbb{E}_{x\sim P_g}[f^*(T_w(x))] \tag{7}$$

Name	$D_f(P\|Q)$	Generator $f(u)$	$T^*(x)$
Kullback-Leibler	$\int p(x)\log\frac{p(x)}{q(x)}\,\mathrm{d}x$	$u\log u$	$1+\log\frac{p(x)}{q(x)}$
Reverse KL	$\int q(x)\log\frac{q(x)}{p(x)}\,\mathrm{d}x$	$-\log u$	$-\frac{q(x)}{p(x)}$
Pearson χ^2	$\int\frac{(q(x)-p(x))^2}{p(x)}\,\mathrm{d}x$	$(u-1)^2$	$2(\frac{p(x)}{q(x)}-1)$
Squared Hellinger	$\int\left(\sqrt{p(x)}-\sqrt{q(x)}\right)^2\mathrm{d}x$	$(\sqrt{u}-1)^2$	$(\sqrt{\frac{p(x)}{q(x)}}-1)\cdot\sqrt{\frac{q(x)}{p(x)}}$
Jensen-Shannon	$\frac{1}{2}\int p(x)\log\frac{2p(x)}{p(x)+q(x)}+q(x)\log\frac{2q(x)}{p(x)+q(x)}\,\mathrm{d}x$	$-(u+1)\log\frac{1+u}{2}+u\log u$	$\log\frac{2p(x)}{p(x)+q(x)}$
GAN	$\int p(x)\log\frac{2p(x)}{p(x)+q(x)}+q(x)\log\frac{2q(x)}{p(x)+q(x)}\,\mathrm{d}x-\log(4)$	$u\log u-(u+1)\log(u+1)$	$\log\frac{p(x)}{p(x)+q(x)}$

it is not hard to prove that if f is taken as the last function in the chart, the objective function is exactly same as that in vanilla GAN. By the way, for *Pearson*χ^2 divergence, there is also a corresponding paper [3]

[3] Least square GAN

The disadvantages of using f-divergence family

- the optimization process is unstable[4]
- it may lead to mode collapse [5]
- Generator didn't learn anything useful, just memorize some samples [6]

Possible reasons

- it isn't appropriate to assume that both Generator and Discriminator has infinite modeling capacity

FACT:

It isn't appropriate to use the JS divergence to measure the distance of two lower-dimensional distributions if both of them lie on a high-dimensional space[7]

[4] Improved techniques for training GANS

[5] Loss-Sensitive Generative Adversarial Networks on Lipschitz Densities

[6] Generalization and Equilibrium in Generative Adversarial Nets

[7] Towards Principle Methods for Training GAN

Why JSD is not good?

Lemma 1

True distribution P_r and generated distribution P_g lie on two lower-dimensional submanifolds of a high-dimensional space.

Lemma 2

if A and B are two lower-dimensional, compact submanifolds, A doesn't intersect with B, then, there is always a perfect discriminator D (smooth, continuous, and takes two different constants on A and B)that can separate A and B.

（Elegant Proof：Urysohn's lemma）

Lemma 3: A weaker condition

As long as two lower-dimensional submanifolds doesn't perfect align (the probability for such circumstance is 1), there is always a perfect discriminator can separate these two submanifolds (almost everywhere)

Theorem 1

As long as theses two lower-dimensional submanifolds don't perfect align, their JS divergence always takes a constant, which is $\log 2$, neglecting the real distance of them.

Theorem 2:

As long as the distribution P_g is generated by a continuous differentiable function, what's more P_g, P_r lies on two lower-dimensional submaniflods, which don't perfect align, then as Discriminator converges to the perfect one, the gradient for the generator converges to 0

FACT:

$JSD(P_r||P_{g_\theta})$ is a constant almost everywhere thus not continuous w.r.t P_{g_θ}, the back propagated gradient for G is close to 0

Solution: Find a distance measurement ρ that satisfies,

- Distribution P_{g_θ} is continuous w.r.t g_θ
- Measurement $\rho(P_r||P_{g_\theta})$ is continuous w.r.t P_{g_θ}

Wasserstein Distance[8]

$$W(P_r, P_g) = \inf_{\gamma \in \Pi(P_r, P_g)} \mathbb{E}_{(x,y)\sim\gamma}[||x-y||] \tag{8}$$

Properties

1. if g_θ is continuous w.r.t θ, then $W(P_r||P_{g_\theta})$ is continuous w.r.t θ
2. if g_θ is local Lipschitz, then $W(P_r||P_{g_\theta})$ is continuous almost everywhere, differentiable almost everywhere.

How weak

- TV Distance: $\delta(P_r||P_g) = \sup_{A\in\sum} |p_r(A) - p_g(A)|$
- KL Divergence: $KL(P_r||P_g) = \int \log(\frac{p_r(x)}{p_g(x)})p_r(x)d\mu(x)$
- JS Divergence: $JSD(P_r||P_g) = KL(P_r||\frac{1}{2}(P_r+P_g)) + KL(P_g||\frac{1}{2}(P_r+P_g))$

[8] Wasserstein GAN

- $\delta(P_r||P_g) \to 0 \iff JSD(P_r||P_g) \to 0$ (the two norms induced by TV divergence and JS divergence are euqivalent)
- $KL(P_g||P_r) \to 0 \Longrightarrow JSD(P_r||P_g) \to 0 \Longrightarrow W(P_r||P_g) \to 0$
- the topology induced by KL divergence is the strongest one, then comes JSD and TV, the weakest one is induced by Wasserstein

Why it is weak

- $W(P_r||P_{g_\theta}) \to 0$ means P_{g_θ} convergence in distribution to P_r
- JSD and TV induce the strongest topology in $C_b(\chi)^*$, while the Wasserstein distance is the *weak** topology corresponding to that 。

How to use?

$$W(P_r||P_{g_\theta}) = \sup_{||f||_L \leq 1} \mathbb{E}_{x\sim P_r}[f(x)] - \mathbb{E}_{x\sim P_{g_\theta}}[f(x)] \tag{9}$$

Minimize $W(P_r||P_{g_\theta})$ w.r.t θ.

The objective function change into

$$\min_{\theta} \max_{w} \mathbb{E}_{x \sim P_r}[f_w(x)] - \mathbb{E}_{z \sim P(z)}[f_w(g_\theta(z))] \tag{10}$$

How to realize $||f||_L \leq 1$?

a trick used is called weight clipping, just force all the parameters of f fall into $w \in [-t, t]$。

Disadvantages

- This trick greatly reduce the set for Lipschitz functions.
- the experiment results show that all parameters of f are around t and $-t$

Some subsequent work are came up with [9] to solve these two problems

[9] Improved Training of Wasserstein GANs

Perspective of Statistics

Lemma

$P_r = P_g \iff \forall \Phi \in C, ||\mathrm{E}_{x\sim P_r}\Phi(x) - \mathrm{E}_{x\sim P_g}\Phi(x)||^2 = 0$

GAN: view Φ as the Discriminator D, the first step is to maximize the discrepancy by adjusting D:

$$\max_{w} ||\mathrm{E}_{x\sim P_r}\Phi_w(x) - \mathrm{E}_{x\sim P_g}\Phi_w(x)||^2 \tag{11}$$

then minimize the maximum w.r.t Generator:

$$\min_{\theta_g}\max_{w} ||\mathrm{E}_{x\sim P_r}\Phi_w(x) - \mathrm{E}_{x\sim P_g}\Phi_w(x)||^2 \tag{12}$$

MMD[10]: using the samples mean after nolinear transformation as a substitute for Expectation:

$$||\frac{1}{N}\sum_{i=1}^{N}\Phi(x_i) - \frac{1}{M}\sum_{j=1}^{M}\Phi(x_j)||^2 \tag{13}$$

Make using of kernel trick, no Discriminator, optimize the objective function w.r.t Generator

[10] Training generative neural networks via Maximum Mean Discrepancy optimization

Perspective of Info. Theory[11 12]

True distribution P_r, generated distribution P_{g_θ}, $Z \sim Ber(\pi)$, the random variable X satisfies:

$$P(X|Z=0) = P_g,\ P(X|Z=1) = P_r \tag{14}$$

Generator G_θ hopes to inference the value of π, try to independent with Z, it can be achieved by minimizing the mutual information:

$$I(X,Z) = KL(p(x,z)||p(x)p(z)) \tag{15}$$

minimize mutual information $\iff$ X, Z are independent, $\iff$ $P_r = P_g$

$$\begin{aligned} I(X,Z) &= H(Z) + \mathbb{E}_X \mathbb{E}_{Z|X} \log q(z|x) + \mathbb{E}_X KL[p(z|x)||q(y|x)] \\ &= \max_q H(Z) + \mathbb{E}_X \mathbb{E}_{Z|X} \log q(z|x) \end{aligned} \tag{16}$$

$$\begin{aligned} I(X,Z) &\geq H(Z) + \max_\Psi \mathbb{E}_{X,Z} \log q(z|x;\Psi) \\ &= H(Z) + \max_\Psi \pi \mathbb{E}_{P_r} \log q(1|x;\Psi) + (1-\pi)\mathbb{E}_{P_g} \log q(0|x;\Psi) \end{aligned} \tag{17}$$

$$\min I(X,Z) \Longrightarrow \min_{g_\theta} \max_\Psi \pi \mathbb{E}_{P_r} \log q(1|x;\Psi) + (1-\pi)\mathbb{E}_{P_g}(1 - \log q(1|x;\Psi))$$

Conclusion

The Non-parametric Estimation of the distance between P_r and P_g[13]:

Definition

Integral Probability Metrics:

$$\gamma_F(P_r, P_g) := \sup_{f \in F} \left| \int_M f dP_r - \int_M f dP_g \right| \tag{18}$$

in which F is a set which contains all real valued, bounded measurable functions on M

$\boxed{\text{All the foregoing models differs in the choice of } F}$

- Wasserstein distance: $F = \{f : ||f||_L \leq 1\}$
- TV distance or Kolmogorov distance: $F = \{f : ||f||_\infty \leq 1\}$
- MMD:$F = \{f : ||f||_H \leq 1\}$

[13] Non-parametric Estimation of Integral Probability Metrics