Friday, April 2, 2021 3:10 PM

Lecture 19

Problems with GANs

- Convergence: min-max unstable

Vanishing gradients: if discriminator is too powerful, can overwhelm generator

conversely, if discriminator is completely fooled, then just guessing randomly → gradients no info.

generator cannot learn → generator will start to behave badly.

If GAN achieves good performance → quickly destabilize

Friday, April 2, 2021 3:28 PM

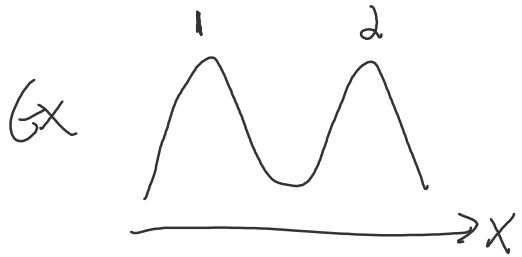
- Metrics for success
  - D & G losses
    accuracy → not very correlated w/ image quality.

↓

hard to know when to stop training

- Mode collapse

G – D get stuck in a vicious cycle



G can learn to produce very realistic 1's.
D will be fooled for a while. → G will switch to making 2's.
then D will learn 2's real, guess on 1's.

Friday, April 2, 2021 3:39 PM

Ways to improve GANs

Many issues w/ GANs related to nature of the loss fn.

Vanilla GAN loss $\min_G \max_D L$

$\min_G JSD(p_{data}, p_G)$

JSD = 1 whenever p_{data} & p_G are disjoint

vanishing gradients!

Vanilla GAN
loss not very sensitive to diff btw p_{data} & p_G. (kind of explains mode collapse)

Friday, April 2, 2021 3:44 PM

Wat: measure of similarity of dist'n that is unbounded from above,

one promising approach that has been applied to HEP:

Wasserstein GANs (WGANs)

loss based "Earth Mover's Distance" or "Wasserstein Distance" (optimal transport theory)

$$W(P,Q) = \min_{\gamma \in \Pi(P,Q)} \int dx\,dy\, \gamma(x,y)\, \|x-y\| = \min_{\gamma \sim \Pi(P,Q)} \mathbb{E}_{(x,y)\sim\gamma}\left[\|x-y\|\right]$$

$x \in X$, $y \in X$

γ: joint probability dist'n on $(x,y) \in X \times X$.

$W = 0$ iff $P = Q$

unbounded from above for $P \neq Q$.

$\Pi(P,Q)$: space of all γ whose marginals are $P(x)$ & $Q(y)$.

$\cap \longleftrightarrow \cap$ $W \to \infty$ JSD = 1.

Friday, April 2, 2021 3:52 PM

W completely intractable — how to use as loss fcn?

Brilliant idea: use amazing result Kantorovich–Rubinstein duality

$$W(P,Q) = \max_{\|h\|_L \le 1} \left[\mathop{E}_{x\sim p}[h(x)] - \mathop{E}_{y\sim q}[h(y)] \right]$$

↑ space of 1-Lipschitz fns

$$|h(x_1) - h(x_2)| \le |x_1 - x_2|$$

K-Lipschitz fn

$$|h(x_1) - h(x_2)| \le \|x_1 - x_2\| \cdot K$$

$\forall\, x_1, x_2$ and for some const K.

Continuity condition ↔ equiv. to saying $|h'(x)| \le K \;\; \forall\, x$.

parametrize h by $\in$ NN! "critic" network

$\max_{\|h\|_L \le 1}(\cdots)$ ↳ training the critic!

Friday, April 2, 2021 4:06 PM

partial proof of K-R formula

Idea: use Lagrange multipliers

$$L(\gamma, f, g) = \int \|x-y\| \gamma(x,y)\,dx\,dy + \int \left(\rho(x) - \int \gamma(x,y)\,dy\right) f(x)\,dx$$

$$+ \int \left(Q(y) - \int \gamma(x,y)\,dx\right) g(y)\,dy .$$

$$= \underset{x\sim\rho}{E}[f(x)] + \underset{y\sim Q}{E}[g(y)] + \int dx\,dy\ \gamma(x,y)\left(\|x-y\| - f(x) - g(y)\right)$$

$$W = \min_{\gamma} \max_{f,g} L(\gamma, f, g)$$

↖ why max? f, g : if constraints satisfied → Lag. mult. terms give 0
not satisfied → " " give $+\infty$ under max$_{f,g}$

Friday, April 2, 2021 4:13 PM

Reverse min max order

$$W = \max_{f,g} \min_{\gamma} \left(\mathop{\mathbb{E}}_{x\sim p}[f(x)] + \mathop{\mathbb{E}}_{y\sim Q}[g(y)] + \int dx\,dy\, \gamma(x,y)\left[\|x-y\| - f(x)-g(y)\right] \right)$$

γ ↓ everywhere

if $f(x)+g(y) \le \|x-y\|$ then $\min_{\gamma}(\cdots) = 0$

$f(x)+g(y) > \|x-y\|$ somewhere $\min_{\gamma}(\cdots) = -\infty$ the max spoiled.

$$= \max_{\substack{f,g \\ f(x)+g(y)\le\|x-y\|}} \left(\mathop{\mathbb{E}}_{x\sim p}[f(x)] + \mathop{\mathbb{E}}_{y\sim Q}[g(y)] \right)$$

→ last step (see literature)

prove $\max_{f,g} = \max_{f=-g}$

□.

Friday, April 2, 2021 4:20 PM

Back to WGAN

loss: $L = \sum_{x \in \text{real}} h(x) - \sum_{z \in \text{latent}} h(G(z))$

objective

$\min_G \max_h L$

h being Lipschitz.

(compare w/ vanilla GAN $\min_G \max_D (-BCE)$)

How to enforce Lipschitz constraint?

use weight clipping: after gradient update

clip weights to some fixed range $w \to \text{clip}(w, -c, c)$ (hyperparameter)

idea: weight clipping $h(x; w)$ ↳ lives in a compact space. so $\nabla_x h$ have to have a maximum somewhere.

Friday, April 2, 2021 4:27 PM

weight clipping is very suboptimal — severely restricts the capacity or expressivity of model.

→ poor performance

↓

better way to implement Lipschitz: Gradient Penalty

K-Lipschitz $\Leftrightarrow \max |\nabla_x h| = K$.

So add regularizer to loss

$$L \mathrel{+}= \lambda \sum_x \left(|\nabla_x h|^2 - 1 \right)^2$$ gradient penalty term

"WGAN-GP"

(check this)

x is sampled from a mish-mash of $x \sim p_{data}(x)$ $y \sim p_{gen}(y)$

$\hat{x} = t x + (1-t) y$ $t \in (0,1)$ random from uniform