Wasserstein GAN

Yixuan Sun

What is GAN?

Generative adversarial networks (GAN)

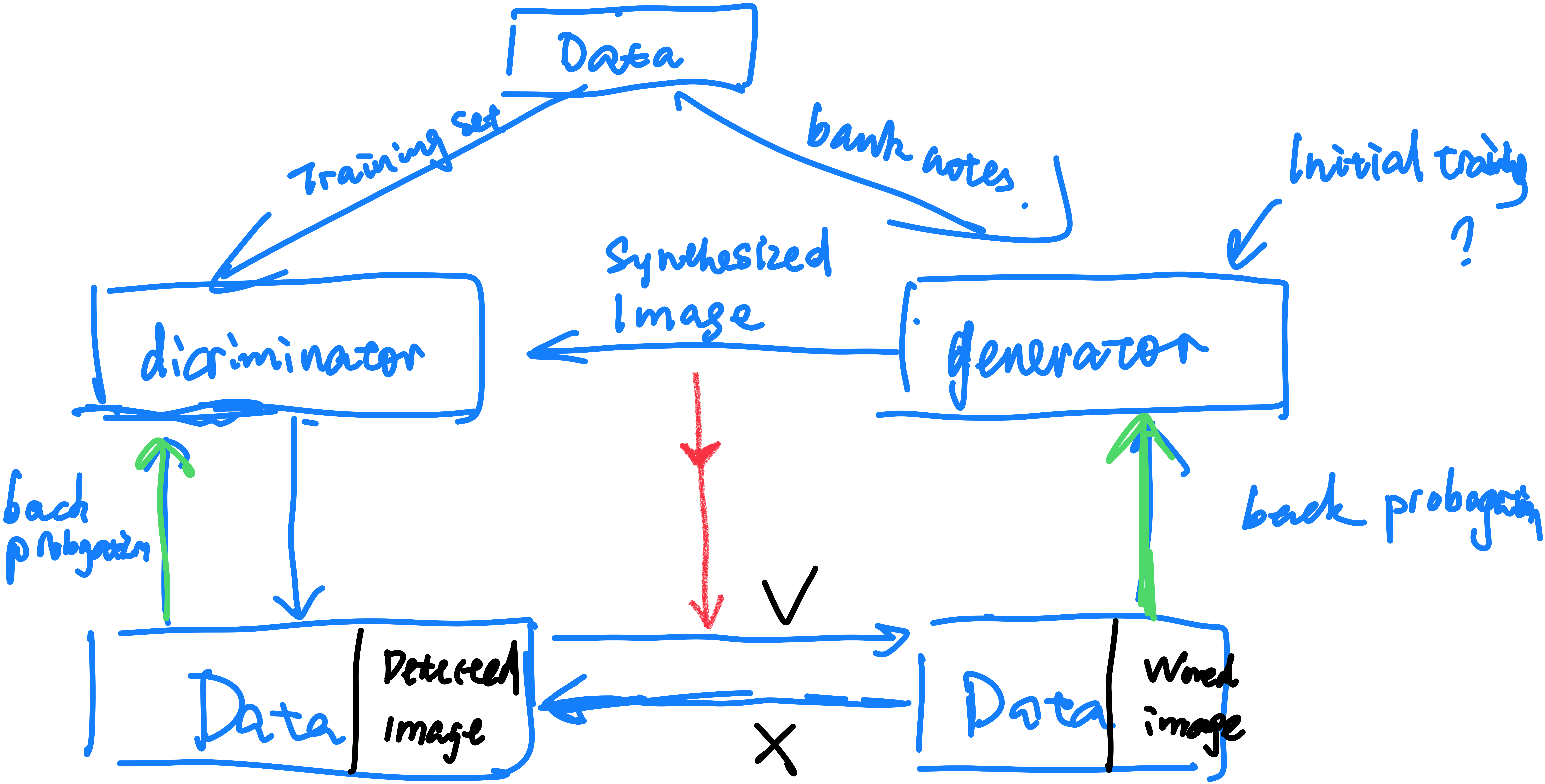
1. machine learning frameworks designed by Ian Goodfellow (2014)

2. Two neural networks contests w/ each other in a zero-sum game. (generative and discriminative)

3. Indirect training through the discriminator

How does it work?

1. Train the discriminator w/ training dataset, to achieve acceptable accuracy.
2. Generator is trained based on how well it fools the discrimator.
3. Independent backprobagation procedure applied to both generator and discrimator to produce better synthesized image and better discriminator.


Data
Training set
bank notes.
Initial traing
?
dicriminator
Synthesized
Image
generator
back
probagation
back probagation
V
Data
Detected
image
X
Data
Wored
image

Based on the paper

Wasserstein Gan - Martin Arjovsky,

\- Soumith Chintala

Léon Bottou.

---

unsupervised learning,

What does it mean to learn a probability distribution?

The classical answer is to learn a probability density.

→ Define a parametric family of densities $(P_\theta)_{\theta \in \mathbb{R}^d}$ and finding the one that maximize the likelihood on our data: If we have real data $\{x^{(i)}\}_{i=1}^{m}$, we would solve the problem

$$\max_{\theta \in \mathbb{R}^d} \frac{1}{m} \sum_{i=1}^{m} \log P_\theta(x^{(i)})$$

If the real data distribution $\mathbb{P}_r$ admits a density and $\mathbb{P}_\theta$ is the distribution of the parametrized density P_θ, then, asymptotically, this amounts to minimizing the Kullback-Leibler divergence $KL(\mathbb{P}_r \| \mathbb{P}_\theta)$.

Problem: KL distance is not defined for distributions supported by low dimensional manifolds.

One remedy is to add noise term to the model distribution.

Rather than estimating the density of $\mathbb{P}_r$, we can define a random variable z w/ a fixed distribution $p(z)$ and pass it through a parametric function $g_\theta: Z \to X$ (neural network for ex) to generate samples following a certain distribution $\mathbb{P}_\theta$.

Pros: 1. this approach can represent distributions confined to a low dimensional manifold.

2. the ability to easily generate samples is often more useful than knowing the numerical values of the density

Variational Auto-Encoders (VAEs) and Generative Adversarial Networks (GANs) are well-known examples of this approach.

GANs Pros: 1. No need to fiddle w/ additional noise term (VAEs has to)
2. flexibility in the def of the objective fun.

GANs Cons: training GANs is delicate and unstable

TV distance.

$$\delta(\mathbb{P}_r, \mathbb{P}_g) = \sup_{A \in \Sigma} |\mathbb{P}_r(A) - \mathbb{P}_g(A)|$$

$\mathcal{X} := [0,1]^d$ Σ be the set of all Borel subsets of $\mathcal{X}$.

$\mathbb{P}_r, \mathbb{P}_g \in \text{Prob}(\mathcal{X})$ $\text{Prob}(\mathcal{X}) \sim$ space of probability measures defined on $\mathcal{X}$.

KL divergence (Kullback-Leibler distance)

$$KL(\mathbb{P}_r \| \mathbb{P}_g) = \int \log \frac{P_r(x)}{P_g(x)} P_r(x)\, d\mu(x).$$

μ - measure $\mathcal{X}$.

$\forall A \in \Sigma$, $\mathbb{P}_r(A) = \int_A P_r(x)\, d\mu(x)$

$\mathbb{P}_g(A) = \int_A P_g(x)\, d\mu(x)$.

Jensen-Shannon (JS) divergence (distance)

$$JS(\mathbb{P}_r, \mathbb{P}_g) = KL(\mathbb{P}_r \| \mathbb{P}_m) + KL(\mathbb{P}_g \| \mathbb{P}_m),$$

where $\mathbb{P}_m = (\mathbb{P}_r + \mathbb{P}_g)/2$, symmetric

Earth-Mover (EM) distance or (Wasserstein-1)

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma}[\|x-y\|]$$

$\Pi(\mathbb{P}_r, \mathbb{P}_g)$ denotes all joint distributions $\gamma(x,y)$ whose marginals are respectively $\mathbb{P}_r$ and $\mathbb{P}_g$.

Thm 1. Let $\mathbb{P}_r$ be a fixed dist over $\mathcal{X}$. Let Z be a random variable (e.g. Gaussian) over another space $\mathcal{Z}$. Let $g: \mathcal{Z} \times \mathbb{R}^d \to \mathcal{X}$ be a func; denoted by $g_\theta(z)$ w/ z the first coordinate and θ the second. Let $\mathbb{P}_\theta$ denote the dist of $g_\theta(Z)$. Then,

1. If g is continuous in θ, so is $W(\mathbb{P}_r, \mathbb{P}_\theta)$

2. If g is locally Lipschitz and satisfies regularity assumption 1. then $W(\mathbb{P}_r, \mathbb{P}_\theta)$ is continuous everywhere, and differentiable a.e.

Wasserstein GAN

Instead of find inf ~~of~~ for $W(\mathbb{P}_r, \mathbb{P}_\theta)$ we use Kantorovich-Rubinstein duality.

$$W(\mathbb{P}_r, \mathbb{P}_\theta) = \sup_{\|f\| \le 1} E_{x \sim \mathbb{P}_r}[f(x)] - E_{x \sim \mathbb{P}_\theta}[f(x)]$$

$\|f\|_L \le 1 \sim$ 1-Lipschitz functions $f: \mathcal{X} \to \mathbb{R}$.

Thm 3. Let $\mathbb{P}_r$ be any dist. Let $\mathbb{P}_\theta$ be the dist of $g_\theta(Z)$ w/ Z a random variable w/ density p and g_θ a fun satisfying assumption 1. Then there is a soln $f: \mathcal{X} \to \mathbb{R}$ to the problem

$$\max_{\|f\|_L = 1} \mathbb{E}_{x \sim \mathbb{P}_r}[f(x)] - \mathbb{E}_{x \sim \mathbb{P}_\theta}[f(x)]$$

and we have

$$\nabla_\theta W(\mathbb{P}_r, \mathbb{P}_\theta) = -\mathbb{E}_{z \sim p(z)}[\nabla_\theta f(g_\theta(z))]$$

when both terms are well-defined.