Deep Learning

Vazgen Mikayelyan

December 19, 2020

FAST DISCOVERING THE FUTURE

cGANs

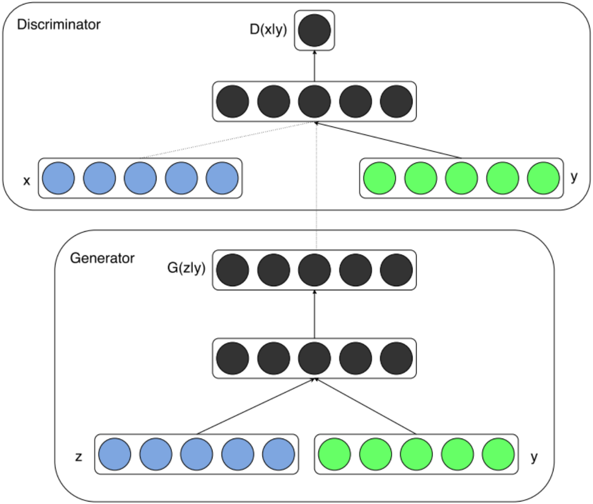
These GANs use extra label information and result in better quality images and are able to control how generated images will look.

cGANs

These GANs use extra label information and result in better quality images and are able to control how generated images will look. In this case we will do the following optimization

$$\min_G \max_D \left(\mathbb{E}_{x\sim p_{data}}\left[\log D\left(x|y\right)\right] + \mathbb{E}_{z\sim p_z}\left[\log\left(1 - D\left(G\left(z|y\right)\right)\right)\right]\right)$$

cGANs



CycleGANs

- Image-to-image translation is a class of vision and graphics problems where the goal is to learn the mapping between an input image and an output image using a training set of aligned image pairs.

CycleGANs

- Image-to-image translation is a class of vision and graphics problems where the goal is to learn the mapping between an input image and an output image using a training set of aligned image pairs.
- However, for many tasks, paired training data will not be available.

CycleGANs

- Image-to-image translation is a class of vision and graphics problems where the goal is to learn the mapping between an input image and an output image using a training set of aligned image pairs.
- However, for many tasks, paired training data will not be available.
- So our task is to learn how to translate an image from a source domain X to a target domain Y in the absence of paired examples.

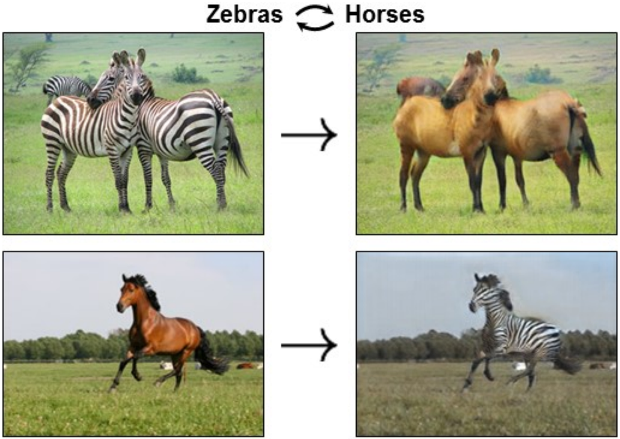
CycleGANs

- Image-to-image translation is a class of vision and graphics problems where the goal is to learn the mapping between an input image and an output image using a training set of aligned image pairs.
- However, for many tasks, paired training data will not be available.
- So our task is to learn how to translate an image from a source domain X to a target domain Y in the absence of paired examples.
- Our goal is to learn a mapping $G : X \to Y$ such that the distribution of images from $G(X)$ is indistinguishable from the distribution Y.

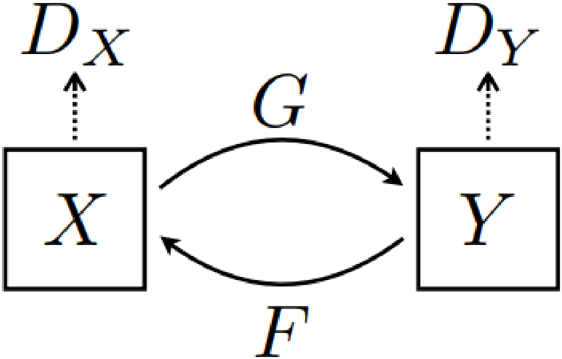
CycleGANs

- Image-to-image translation is a class of vision and graphics problems where the goal is to learn the mapping between an input image and an output image using a training set of aligned image pairs.
- However, for many tasks, paired training data will not be available.
- So our task is to learn how to translate an image from a source domain X to a target domain Y in the absence of paired examples.
- Our goal is to learn a mapping $G : X \to Y$ such that the distribution of images from $G(X)$ is indistinguishable from the distribution Y.
- We will couple it with an inverse mapping $F : Y \to X$ and introduce a cycle consistency loss to enforce $F(G(X)) \approx X$.

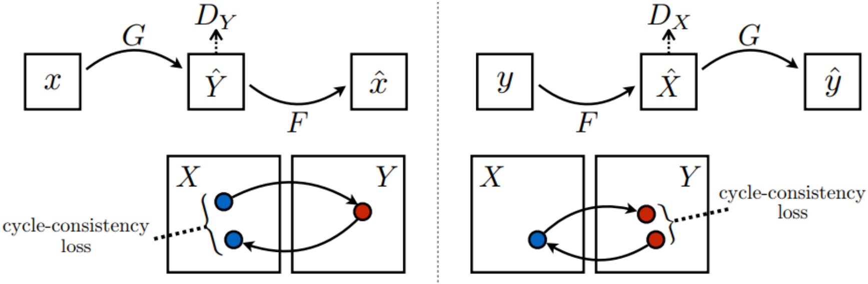
CycleGANs



CycleGANs



CycleGANs



D_Y
G
x
$\hat{Y}$
$\hat{x}$
F
X
Y
cycle-consistency loss
D_X
G
y
$\hat{X}$
$\hat{y}$
F
X
Y
cycle-consistency loss

CycleGANs

Our loss function will be the following

$$L(G, F, D_X, D_Y) = L_{GAN}(G, D_Y, X, Y) + L_{GAN}(F, D_X, Y, X) + \lambda L_{cyc}(G, F),$$

CycleGANs

Our loss function will be the following

$$L(G, F, D_X, D_Y) = L_{GAN}(G, D_Y, X, Y) + L_{GAN}(F, D_X, Y, X) + \lambda L_{cyc}(G, F),$$

where

$$L_{GAN}(G, D_Y, X, Y)$$
$$= \mathbb{E}_{y \sim p_{data}(y)}[\log D_Y(y)] + \mathbb{E}_{x \sim p_{data}(x)}[\log(1 - D_Y(G(x)))]$$

CycleGANs

Our loss function will be the following

$$L(G, F, D_X, D_Y) = L_{GAN}(G, D_Y, X, Y) + L_{GAN}(F, D_X, Y, X) + \lambda L_{cyc}(G, F),$$

where

$$L_{GAN}(G, D_Y, X, Y)$$
$$= \mathbb{E}_{y \sim p_{data}(y)}[\log D_Y(y)] + \mathbb{E}_{x \sim p_{data}(x)}[\log(1 - D_Y(G(x)))]$$

$$L_{GAN}(F, D_X, Y, X)$$
$$= \mathbb{E}_{x \sim p_{data}(x)}[\log D_X(x)] + \mathbb{E}_{y \sim p_{data}(y)}[\log(1 - D_X(F(y)))]$$

CycleGANs

Our loss function will be the following

$$L(G, F, D_X, D_Y) = L_{GAN}(G, D_Y, X, Y) + L_{GAN}(F, D_X, Y, X) + \lambda L_{cyc}(G, F),$$

where

$$L_{GAN}(G, D_Y, X, Y)$$
$$= \mathbb{E}_{y \sim p_{data}(y)}[\log D_Y(y)] + \mathbb{E}_{x \sim p_{data}(x)}[\log(1 - D_Y(G(x)))]$$

$$L_{GAN}(F, D_X, Y, X)$$
$$= \mathbb{E}_{x \sim p_{data}(x)}[\log D_X(x)] + \mathbb{E}_{y \sim p_{data}(y)}[\log(1 - D_X(F(y)))]$$

$$L_{cyc}(G, F) = \mathbb{E}_{x \sim p_{data}(x)}[\|F(G(x)) - x\|_1] + \mathbb{E}_{y \sim p_{data}(y)}[\|G(F(y)) - y\|_1]$$

CycleGANs

Our loss function will be the following

$$L(G,F,D_X,D_Y) = L_{GAN}(G,D_Y,X,Y) + L_{GAN}(F,D_X,Y,X) + \lambda L_{cyc}(G,F),$$

where

$$L_{GAN}(G,D_Y,X,Y)$$
$$= \mathbb{E}_{y\sim p_{data}(y)}[\log D_Y(y)] + \mathbb{E}_{x\sim p_{data}(x)}[\log(1-D_Y(G(x)))]$$

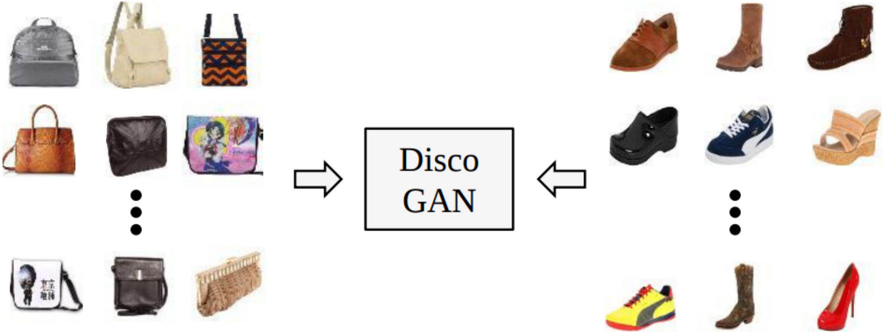
$$L_{GAN}(F,D_X,Y,X)$$
$$= \mathbb{E}_{x\sim p_{data}(x)}[\log D_X(x)] + \mathbb{E}_{y\sim p_{data}(y)}[\log(1-D_X(F(y)))]$$

$$L_{cyc}(G,F) = \mathbb{E}_{x\sim p_{data}(x)}[\|F(G(x))-x\|_1] + \mathbb{E}_{y\sim p_{data}(y)}[\|G(F(y))-y\|_1]$$

We aim to solve

$$G^*,F^* = \arg\min_{G,F}\max_{D_X,D_Y} L(G,F,D_X,D_Y)$$

DISCO GAN



DISCO GAN

DISCO GAN

Huber Loss

$$L_\delta\left(y, f(x)\right) = \begin{cases} \frac{1}{2}\left(y - f(x)\right)^2, & \text{for } |y - f(x)| \leq \delta \\ \delta\left|y - f(x)\right| - \frac{1}{2}\delta^2, & \text{otherwise.} \end{cases}$$

WGANs

Let $\mathcal{X}$ be a compact metric set (e.g $[0, 1]^d$) and let Σ denote the set of all Borel subsets of $\mathcal{X}$.

WGANs

Let $\mathcal{X}$ be a compact metric set (e.g $[0,1]^d$) and let Σ denote the set of all Borel subsets of $\mathcal{X}$.

- Total Variation (TV) distance

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

Let $\mathcal{X}$ be a compact metric set (e.g $[0,1]^d$) and let Σ denote the set of all Borel subsets of $\mathcal{X}$.

- Total Variation (TV) distance

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

- Kullback-Leibler (KL) divergence

$$KL(\mathbb{P}_r \| \mathbb{P}_g) = \int_{\mathbb{R}} \log\left(\frac{p_r(x)}{p_g(x)}\right) p_r(x)\, d\mu(x).$$

WGANs

Let $\mathcal{X}$ be a compact metric set (e.g $[0,1]^d$) and let Σ denote the set of all Borel subsets of $\mathcal{X}$.

- Total Variation (TV) distance

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

- Kullback-Leibler (KL) divergence

$$KL(\mathbb{P}_r \| \mathbb{P}_g) = \int_{\mathbb{R}} \log\left(\frac{p_r(x)}{p_g(x)}\right) p_r(x)\, d\mu(x).$$

- Jensen-Shannon (JS) distance

$$JS(\mathbb{P}_r \| \mathbb{P}_g) = \frac{1}{2}\left(KL(\mathbb{P}_r \| \mathbb{P}_m) + KL(\mathbb{P}_g \| \mathbb{P}_m)\right)$$

where $\mathbb{P}_m = \dfrac{\mathbb{P}_r + \mathbb{P}_g}{2}$.

WGANs

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

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

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

Example

Let $Z \sim U[0,1]$, $\mathbb{P}_0$ be the distribution of points $(0, Z) \in \mathbb{R}^2$ and $g_\theta(z) = (\theta, z)$, then

Example

Let $Z \sim U[0,1]$, $\mathbb{P}_0$ be the distribution of points $(0, Z) \in \mathbb{R}^2$ and $g_\theta(z) = (\theta, z)$, then

- $\delta(\mathbb{P}_0, \mathbb{P}_\theta) = \begin{cases} 1, & \text{if } \theta \neq 0 \\ 0, & \text{if } \theta = 0, \end{cases}$

Example

Let $Z \sim U[0,1]$, $\mathbb{P}_0$ be the distribution of points $(0, Z) \in \mathbb{R}^2$ and $g_\theta(z) = (\theta, z)$, then

- $\delta(\mathbb{P}_0, \mathbb{P}_\theta) = \begin{cases} 1, & \text{if } \theta \neq 0 \\ 0, & \text{if } \theta = 0, \end{cases}$
- $KL(\mathbb{P}_0 \| \mathbb{P}_\theta) = \begin{cases} \infty, & \text{if } \theta \neq 0 \\ 0, & \text{if } \theta = 0, \end{cases}$

Example

Let $Z \sim U[0,1]$, $\mathbb{P}_0$ be the distribution of points $(0, Z) \in \mathbb{R}^2$ and $g_\theta(z) = (\theta, z)$, then

- $\delta(\mathbb{P}_0, \mathbb{P}_\theta) = \begin{cases} 1, & \text{if } \theta \neq 0 \\ 0, & \text{if } \theta = 0, \end{cases}$
- $KL(\mathbb{P}_0 \| \mathbb{P}_\theta) = \begin{cases} \infty, & \text{if } \theta \neq 0 \\ 0, & \text{if } \theta = 0, \end{cases}$
- $JS(\mathbb{P}_0 \| \mathbb{P}_\theta) = \begin{cases} \log 2, & \text{if } \theta \neq 0 \\ 0, & \text{if } \theta = 0, \end{cases}$

Example

Let $Z \sim U[0,1]$, $\mathbb{P}_0$ be the distribution of points $(0, Z) \in \mathbb{R}^2$ and $g_\theta(z) = (\theta, z)$, then

- $\delta(\mathbb{P}_0, \mathbb{P}_\theta) = \begin{cases} 1, & \text{if } \theta \neq 0 \\ 0, & \text{if } \theta = 0, \end{cases}$
- $KL(\mathbb{P}_0 \| \mathbb{P}_\theta) = \begin{cases} \infty, & \text{if } \theta \neq 0 \\ 0, & \text{if } \theta = 0, \end{cases}$
- $JS(\mathbb{P}_0 \| \mathbb{P}_\theta) = \begin{cases} \log 2, & \text{if } \theta \neq 0 \\ 0, & \text{if } \theta = 0, \end{cases}$
- $W(\mathbb{P}_0, \mathbb{P}_\theta) = |\theta|$

WGANs

Note that

WGANs

Note that

- All distances other than EM are not continuous.

Note that

- All distances other than EM are not continuous.
- When $\theta_t \to 0$, the sequence $(\mathbb{P}_{\theta_t})_{t\in\mathbb{N}}$ converges to $\mathbb{P}_0$ under the EM distance, but does not converge at all under either the JS, KL, reverse KL or TV divergences.

Note that

- All distances other than EM are not continuous.
- When $\theta_t \to 0$, the sequence $(\mathbb{P}_{\theta_t})_{t\in\mathbb{N}}$ converges to $\mathbb{P}_0$ under the EM distance, but does not converge at all under either the JS, KL, reverse KL or TV divergences.
- Only EM distance has informative gradient.

Lipschitz functions

Definition 1

Let X and Y are normed vector spaces. A function $f : X \to Y$ is called

- *K-Lipschitz if there exists a real constant $K > 0$ such that, for all x_1 and x_2 in X*
$$\|f(x_1) - f(x_2)\| \leq K\|x_1 - x_2\|.$$
- *local Lipschitz if for every $x \in X$ there exists a neighbourhood U of x such that f is Lipschitz on U.*

Lipschitz functions

Definition 1

Let X and Y are normed vector spaces. A function $f : X \to Y$ is called

- *K-Lipschitz if there exists a real constant $K > 0$ such that, for all x_1 and x_2 in X*
$$\|f(x_1) - f(x_2)\| \leq K\|x_1 - x_2\|.$$
- *local Lipschitz if for every $x \in X$ there exists a neighbourhood U of x such that f is Lipschitz on U.*

Theorem 1

If function $f : \mathbb{R}^n \to \mathbb{R}$ has bounded gradient, then f is a Lipschitz function.