Deep Learning

Vazgen Mikayelyan

YSU, Krisp

December 10, 2019

Outline

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.

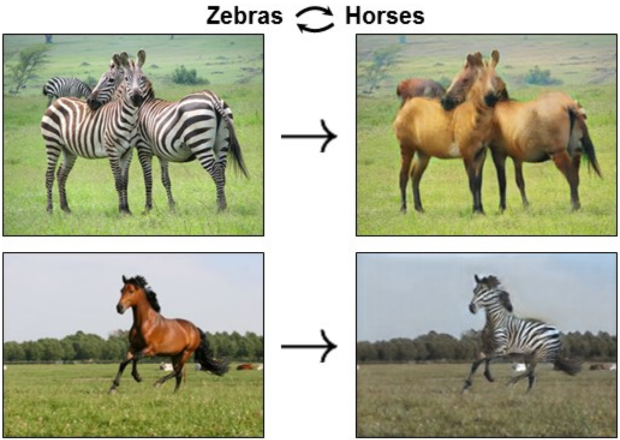
CycleGANs

- 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.

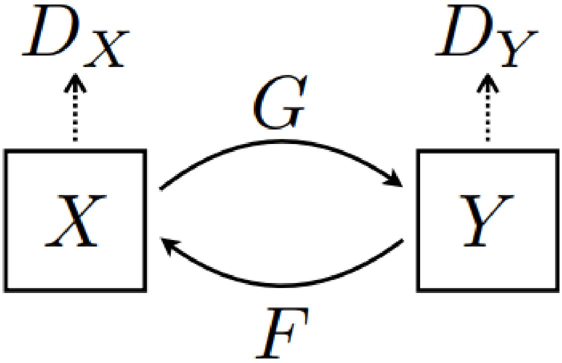
CycleGANs

- 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 $G : X \rightarrow Y$ such that the distribution of images from $G(X)$ is indistinguishable from the distribution Y.
- We will couple it with an inverse mapping $F : Y \rightarrow X$ and introduce a cycle consistency loss to enforce $F(G(X)) \approx X$.

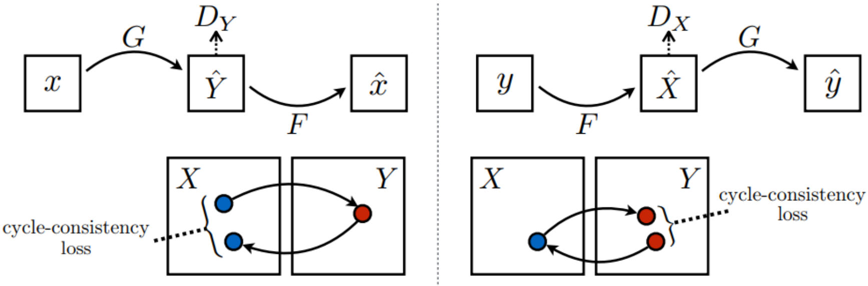
CycleGANs



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

Our loss function will be the following

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

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

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

where

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

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

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

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

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

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 $\mathcal{X}$.

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

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

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

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.

EM properties

Theorem 1. *Let $\mathbb{P}_r$ be a fixed distribution 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 function, that will be denoted $g_\theta(z)$ with z the first coordinate and θ the second. Let $\mathbb{P}_\theta$ denote the distribution 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 almost everywhere.*
3. *Statements 1-2 are false for the Jensen-Shannon divergence $JS(\mathbb{P}_r, \mathbb{P}_\theta)$ and all the KLs.*

EM properties

Theorem 1. *Let $\mathbb{P}_r$ be a fixed distribution 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 function, that will be denoted $g_\theta(z)$ with z the first coordinate and θ the second. Let $\mathbb{P}_\theta$ denote the distribution 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 almost everywhere.*
3. *Statements 1-2 are false for the Jensen-Shannon divergence $JS(\mathbb{P}_r, \mathbb{P}_\theta)$ and all the KLs.*

Corollary 1. *Let g_θ be any feedforward neural network[4] parameterized by θ, and $p(z)$ a prior over z such that $\mathbb{E}_{z\sim p(z)}[\|z\|] < \infty$ (e.g. Gaussian, uniform, etc.). Then assumption 1 is satisfied and therefore $W(\mathbb{P}_r, \mathbb{P}_\theta)$ is continuous everywhere and differentiable almost everywhere.*

EM properties

Theorem 2. *Let $\mathbb{P}$ be a distribution on a compact space $\mathcal{X}$ and $(\mathbb{P}_n)_{n\in\mathbb{N}}$ be a sequence of distributions on $\mathcal{X}$. Then, considering all limits as $n \to \infty$,*

1. *The following statements are equivalent*
   - $\delta(\mathbb{P}_n, \mathbb{P}) \to 0$ *with δ the total variation distance.*
   - $JS(\mathbb{P}_n, \mathbb{P}) \to 0$ *with JS the Jensen-Shannon divergence.*
2. *The following statements are equivalent*
   - $W(\mathbb{P}_n, \mathbb{P}) \to 0$.
   - $\mathbb{P}_n \xrightarrow{\mathcal{D}} \mathbb{P}$ *where $\xrightarrow{\mathcal{D}}$ represents convergence in distribution for random variables.*
3. $KL(\mathbb{P}_n \| \mathbb{P}) \to 0$ *or* $KL(\mathbb{P} \| \mathbb{P}_n) \to 0$ *imply the statements in (1).*
4. *The statements in (1) imply the statements in (2).*

EM properties

Summary 1

Wasserstein (or EM) loss for neural networks is continuous and differentiable almost everywhere. Moreover, Convergence in KL implies convergence in TV and JS which implies convergence in EM.

Training

Kantorovich-Rubinstein duality 2

$$W(\mathbb{P}_r, \mathbb{P}_\theta) = \sup_{\|f\|_L \leq 1} \left(\mathbb{E}_{x \sim \mathbb{P}_r}[f(x)] - \mathbb{E}_{x \sim \mathbb{P}_\theta}[f(x)] \right)$$

where the supremum is over all the 1-Lipschitz functions $f : \mathcal{X} \to \mathbb{R}$.

Training

Kantorovich-Rubinstein duality 2

$$W\left(\mathbb{P}_r, \mathbb{P}_\theta\right) = \sup_{\|f\|_L \leq 1} \left(\mathbb{E}_{x\sim\mathbb{P}_r}\left[f\left(x\right)\right] - \mathbb{E}_{x\sim\mathbb{P}_\theta}\left[f\left(x\right)\right]\right)$$

where the supremum is over all the 1-Lipschitz functions $f : \mathcal{X} \to \mathbb{R}$.

Note that if we replace $\|f\|_L \leq 1$ for $\|f\|_L \leq K$ (consider K-Lipschitz for some constant K) then we end up with $K \cdot W\left(\mathbb{P}_r, \mathbb{P}_g\right)$.

Training

Kantorovich-Rubinstein duality 2

$$W\left(\mathbb{P}_r, \mathbb{P}_\theta\right) = \sup_{\|f\|_L \leq 1} \left(\mathbb{E}_{x \sim \mathbb{P}_r}\left[f\left(x\right)\right] - \mathbb{E}_{x \sim \mathbb{P}_\theta}\left[f\left(x\right)\right]\right)$$

where the supremum is over all the 1-Lipschitz functions $f : \mathcal{X} \to \mathbb{R}$.

Note that if we replace $\|f\|_L \leq 1$ for $\|f\|_L \leq K$ (consider K-Lipschitz for some constant K) then we end up with $K \cdot W\left(\mathbb{P}_r, \mathbb{P}_g\right)$.
Therefore, if we have a parameterized family of functions $\{f_w\}_{w \in \mathcal{W}}$ that are all K-Lipschitz for some K, we could consider solving the problem

$$\max_{w \in \mathcal{W}} \left(\mathbb{E}_{x \sim \mathbb{P}_r}\left[f_w\left(x\right)\right] - \mathbb{E}_{z \sim p(z)}\left[f_w\left(g_\theta\left(z\right)\right)\right]\right)$$

for estimating $W\left(\mathbb{P}_r, \mathbb{P}_\theta\right)$.