Wasserstein Generative Adversarial Nets

Presented by Yang Xue

Sichuan University

April 21, 2018

Outline

KL-散度的优化

KL-散度:

$$KL\left(\mathbb{P}_r \parallel \mathbb{P}_{g_\theta}\right) = \int_\chi p_r(x) log \frac{p_r(x)}{p_g(x)}$$

不对称性:

$$KL\left(\mathbb{P}_r \parallel \mathbb{P}_g\right) \neq KL\left(\mathbb{P}_g \parallel \mathbb{P}_r\right)$$

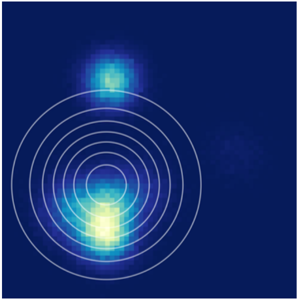
影响: $\nabla_\theta KL\left(\mathbb{P}_r \parallel \mathbb{P}_{g_\theta}\right) =$

$$= \int_\chi \nabla_\theta \left[p_r(x) log(p_r(x)) - p_r(x) log(p_{g_\theta}(x))\right] dx$$

$$= -\int_\chi \nabla_\theta \left[p_r(x) log(p_{g_\theta}(x))\right] dx = -\int_\chi \left[\frac{p_r(x)}{p_g(x)} \nabla_\theta p_{g_\theta}(x)\right] dx$$

$$= -\int_{\chi_A} \left[\frac{p_r(x)}{p_g(x)} \nabla_\theta p_{g_\theta}(x)\right] dx - \int_{\chi_B} \left[\frac{p_r(x)}{p_g(x)} \nabla_\theta p_{g_\theta}(x)\right] dx$$

例子:



完美判别器

真实的数据分布情况：

1. $\mathbb{P}_r$ 分布在 χ 的低维流形上。
2. $\mathbb{P}_g$ 也在 χ 的低维流形上。

结论：

1. $\mathbb{P}_r$ 和 $\mathbb{P}_{g\theta}$ 在 χ 是离散的。
2. $\mathbb{P}_r$ 和 $\mathbb{P}_{g\theta}$ 在 χ 几乎不重叠的。

定理 (完美判别器)

存在一个光滑的判别器 D^* 使得 $\mathbb{P}_r[D^*(x) = 1] = 1, \mathbb{P}_g[D^*(x) = 0] = 1$，且 $\nabla_x D^*(x) = 0$。

实验的情况：

1. 训练一开始 loss 迅速降为 0。
2. 判别器趋于“完美判别器”，由 $loss = \mathbb{E}_{x\sim\mathbb{P}_r}[logD(x)] + \mathbb{E}_{x\sim\mathbb{P}_g}[log(1 - D(x))]$)
3. <u>一旦出现完美判别器情况，就无法训练。</u>

生成器梯度消失

定理

在之前的假设下：

$$\lim_{\|D-D^*\|\to 0} \nabla_\theta \mathbb{E}_{z\sim p(z)}\left[log(1-D(g_\theta(z)))\right] = 0$$

$$\begin{aligned}
&\left\|\nabla_\theta \mathbb{E}_{z\sim p(z)}\left[log(1-D(g_\theta(z)))\right]\right\|_2^2 \\
&\quad < \mathbb{E}_{z\sim p(z)}\left[\frac{\left(\|\nabla_\theta D^*(g_\theta(z))\|_2+\epsilon\right)^2 \|J_\theta g_\theta(z)\|_2^2}{\left(|1-D^*(g_\theta(z))|-\epsilon\right)^2}\right] \\
&\quad = \mathbb{E}_{z\sim p(z)}\left[\frac{\epsilon^2 \left\|J_{\theta(z)}\right\|_2^2}{(1-\epsilon)^2}\right] \\
&\quad \leq M^2 \frac{\epsilon^2}{(1-\epsilon)^2}
\end{aligned}$$

梯度消失-举例

例 (分布的支撑集不重叠导致梯度消失)

对于数据空间中的任意一点 x 只可能有下面 4 种情况:

1. $p_r(x) = 0$ 且 $p_g(x) = 0$
2. $p_r(x) \neq 0$ 且 $p_g(x) \neq 0$
3. $p_r(x) \neq 0$ 且 $p_g(x) = 0$
4. $p_r(x) = 0$ 且 $p_g(x) \neq 0$

上面 4 项对 JS-散度计算的贡献:

1. 1项对 JS-散度计算的贡献为 0
2. 3和4项表示的是分布 $\mathbb{P}_r$ 和 $\mathbb{P}_g$ 不重叠的部分对 JS-散度计算的贡献为常数 $log2$
3. 2项表示的是分布 $\mathbb{P}_r$ 和 $\mathbb{P}_g$ 重叠的部分，只有这一部分才能对生成器提供梯度

上例说明：如果分布 $\mathbb{P}_r$ 和 $\mathbb{P}_g$ 重叠的部分可以忽略不计则会出现梯度消失。而训练不稳定的根因正是梯度时而会消失。

根本原因：

1. 实际情况糟糕：分布基本不重叠，容易使判别器达到“完美”
2. JS-散度，KL-散度在上面情况下不可靠（可能无法提供有效的梯度）

带来的问题：

1. 生成器和判别器的训练需要平衡
2. 网络结构对结果影响很大（实验现象）

解决办法：

1. 添加噪声，使添加噪声后的分布重叠，再逐渐退火
2. 更换另一种距离，它能处理分布不重叠的情况

替换目标函数的模型坍塌以及训练不稳定

替代目标函数：

$$\mathbb{E}_{z\sim p(z)}\left[-log(D(g_\theta(z)))\right]$$

特点：

1. 相同的不动点。
2. 不再等价于优化 JS 散度。

定理

令 $\mathbb{P}_r, \mathbb{P}_{g_\theta}$ 是两个连续的概率分布，它们的概率密度分别为 p_r, p_{g_θ}。固定 $\theta = \theta_0$，令 $D^* = \frac{p_r}{p_r + p_{g_{\theta_0}}}$ 为最优的判别器，则：

$$\mathbb{E}_{z\sim p(z)}\left[-\nabla_\theta log D^*(g_\theta(z))\mid_{\theta=\theta_0}\right] = \nabla_\theta\left[KL(\mathbb{P}_{g_\theta}\parallel\mathbb{P}_r) - 2JSD(\mathbb{P}_{g_\theta}\parallel\mathbb{P}_r)\right]\mid_{\theta=\theta_0}$$

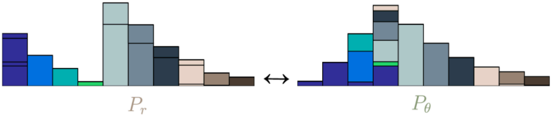
可以看出：

1. 梯度更新不稳定（梯度服从柯西分布，期望和方差无限大）。
2. 造成模型坍塌的问题。
3. 倾向于生成高质量的图像。

Outline

什么是 EMD 距离

最优传输距离：



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

优点：

1. 适用于测量支撑集不重叠的分布的距离。
2. $W(\mathbb{P}_r, \mathbb{P}_\theta)$ 对 θ 连续。

缺点：

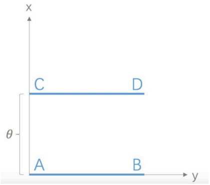
1. 计算复杂度高。

一个例子

例 (支撑集不重叠的分布之间的散度)

令 $Z \sim U[0,1]$，$\mathbb{P}_0 \sim (0, Z)$，$\mathbb{P}_\theta \sim (\theta, Z)$，则:

1. $W(\mathbb{P}_r, \mathbb{P}_\theta) = |\theta|$
2. $JS(\mathbb{P}_r, \mathbb{P}_\theta) = \begin{cases} log2 & \theta \neq 0 \\ 0 & \theta = 0 \end{cases}$
3. $KL(\mathbb{P}_r, \mathbb{P}_\theta) = \begin{cases} +\infty & \theta \neq 0 \\ 0 & \theta = 0 \end{cases}$



上面的例子可以看出:

1. KL 与 JS 散度不连续，无意义的梯度
2. W 距离能提供梯度

定理

在一定的条件下， $W(\mathbb{P}_r, \mathbb{P}_\theta)$ 对 θ 是连续的，且几乎处处可微。

W-距离的优化

W-距离的对偶形式：

$$
\begin{aligned}
W(\mathbb{P}_r, \mathbb{P}_\theta) &= \sup_{\|f\|_L \le 1} \mathbb{E}_{x \sim \mathbb{P}_r}[f(x)] - \mathbb{E}_{x \sim \mathbb{P}_\theta}[f(x)] \\
&= \sup_{\|f\|_L \le 1} \mathbb{E}_{x \sim \mathbb{P}_r}[f(x)] - \mathbb{E}_{z \sim \mathbb{P}_{g_\theta}}[f(g_\theta(z))]
\end{aligned}
$$

算法 1 WGAN

Sample $\left\{x^{(i)})\right\}_{i=1}^m \sim \mathbb{P}_r$

Sample $\left\{z^{(i)})\right\}_{i=1}^m \sim \mathbb{P}_z$

$\Delta\omega \leftarrow \nabla_\omega \left[\frac{1}{m}\sum_{i=1}^m f_\omega(x^{(i)}) - \frac{1}{m}\sum_{i=1}^m f_\omega(g_\theta(z^{(i)}))\right]$

$\omega \leftarrow \omega + \eta \cdot \Delta\omega$

$\omega \leftarrow clip(\omega, -c, c)$

Sample $\left\{z^{(i)})\right\}_{i=1}^m \sim \mathbb{P}_z$

$\Delta\theta \leftarrow -\nabla_\theta \left[\frac{1}{m}\sum_{i=1}^m f(g_\theta(z^{(i)}))\right]$

$\theta \leftarrow \theta - \eta \cdot \Delta\theta$

与原始 GAN 的区别

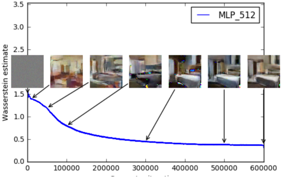
算法区别：

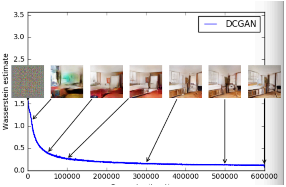
1. 去掉了 D 最后一层的非线性 Sigmoid 函数。
2. 去掉了目标函数里的 Log。
3. D 的权重被截断在 $[-c, c]$ 之间。

提升：

1. 训练稳定，不需要平衡 G 和 D 的能力。
2. 没有模型坍塌，图片效果更好。
3. 有意义的 loss 测量。

有意义的 loss:





The End