Generative model with gradient...

Two ingredients:
→ Langevin dynamics.
→ Score Matching.

① Definition of score:

$\nabla_x \log p(x)$.

② Langevin dynamics: $\tilde{x}_0 \sim \pi(x)$. π, a piror distribution.

$$\tilde{x}_t = \tilde{x}_{t-1} + \frac{\varepsilon}{2}\nabla_x \log p(\tilde{x}_{t-1}) + \sqrt{\varepsilon}\, z_t.$$

$z_t \sim N(0,1)$.

recursively compute. when $\varepsilon \to 0$ and $T \to \infty$

when ε is small and T is large, $\tilde{x}_t \sim p(x)$.

Instead of compute $p(x)$, only requires score function.

First train the score network. $S_\theta(x) \approx \nabla_x \log p(x)$.

③ Score Matching: (分数匹配). Basic score matching.

Objective function: $L = \frac{1}{2} E[\| S_\theta(x) - \nabla_x \log p_{data}(x) \|_2^2]$.

which is equivalent to:

$$L = E_{p(x)}[\mathrm{tr}(\nabla_x S_\theta(x)) + \tfrac{1}{2}\|S_\theta(x)\|_2^2].$$

To avoid jacobian of $S_\theta(x)$.

use denoising score matching.

推导如下: $\frac{1}{2} E[\| \nabla_x \log p_{data}(x) - S_\theta(x) \|_2^2]$

$= \frac{1}{2} E[\| \nabla_x \log p_{data}(x) \|^2 - 2 \nabla_x \log p_{data}(x) \cdot S_\theta(x) - \|S_\theta(x)\|^2]$.

constant. L_2 L_3

$$L_2 = \int p_{data}(x) \nabla_x \log p_{data}^T(x) \cdot s_\theta(x)\, dx. \Rightarrow x^T y = \sum_i^n x_i y_i.$$

$$= \int p_{data}(x) \sum_{x_i} \nabla_{x_i} \log p_{data}(x_i) \cdot \sum_{x_i} s_\theta(x_i) \cdot dx.$$

$$= \sum_{x_i} \int p_{data}(x) \nabla_{x_i} \log p_{data}(x_i)\, s_\theta(x_i)\, dx. \quad \text{求导.}$$

$$= \sum_{x_i} \int p_{data}(x) \frac{\log p_{data}(x_i)}{p_{data}(x_i)} s_\theta(x_i)\, dx.$$

$$= \sum_{x_i} \int \log p_{data}(x_i)\, s_\theta(x_i)\, dx. \rightarrow \text{分步积分: } \int u\, dv = uv - \int v\, du$$

$$= \sum_{x_i} \left(p_{data}(x)\, s_\theta(x_i) - \int p_{data}\, d\, s_\theta(x) \right).$$

$$= E_{p_{data}}\left[\| \mathrm{tr}(\nabla_x s_\theta(x)) + \frac{1}{2} \| s_\theta(x) \|_2^2 \right].$$

$\psi_{\theta,i}$ 已经.

④. Denoising score Matching:

First perturb x with a pre-specified noise distribution $p_\sigma(\tilde{x})$, then use score matching.

可以理解为 diffusion model?

$$p_\sigma(\tilde{x}) \triangleq \int p_{data}(x)\, p(\tilde{x}|x)\, dx. \leftarrow \text{固定 } q(x_t|x_{t+1}) \text{ 为啥??}$$

Objective function:

$$\frac{1}{2} E_{q_\sigma(\tilde{x}|x)\, p(x)} \left[\| s_\theta(\tilde{x}) - \nabla_{\tilde{x}} \log q_\sigma(\tilde{x}|x) \|_2^2 \right]$$

$s_\theta \quad s_{\theta^*}(x) \approx \nabla_x \log q_\sigma(x).$

$s_{\theta^*}(x) \approx \nabla_x \log p(x)$

Only when noise very small.

在数据密度较低的地方，梯度不准确。

⑧ Noise conditional score Network (NCSN)

给密度较低的地方加较大幅度噪声，但加太大会失去数据原来特征。

所以用一个 conditional 的 Network，with multiple scale noise of perturbations simultaneously.

When using langevin dynamics to generate samples, first uses score corresponding to large noise, the gradually anneal down the noise level.

Add isotropic Gaussian noise: $\{\sigma_i\}_{i=1}^{L}$ $\sigma_1 < \sigma_2 \cdots < \sigma_L$.

$p_{\sigma_i} \sim N(0, \sigma_i^2 I)$ $i = 1, 2 \cdots L$

$$p_{\sigma_i}(x) = \int p_{data}(x)\, N(x|t, \sigma^2 I)\, dt$$

Train a Network to jointly estimate the score of all perturbed data distributions s_θ. conditional score Network (CCSN).