LSGM

Title	Score-based Generative Modeling in Latent Space
日期	2021.09
发表单位	NVIDIA
github	https://nvlabs.github.io/LSGM
上次编辑	@2025年5月2日 19:50
状态	In progress
重要程度	⭐⭐⭐

Score-based Generative Modeling in Latent Space
The Cross Entropy Term
Mixing Normal and Neural Score Functions
Training with Different Weighting Mechanisms
Variance Reduction
Variance reduction for likelihood weighting
Variance reduction for unweighted and reweighted objectives

Score-based Generative Modeling in Latent Space

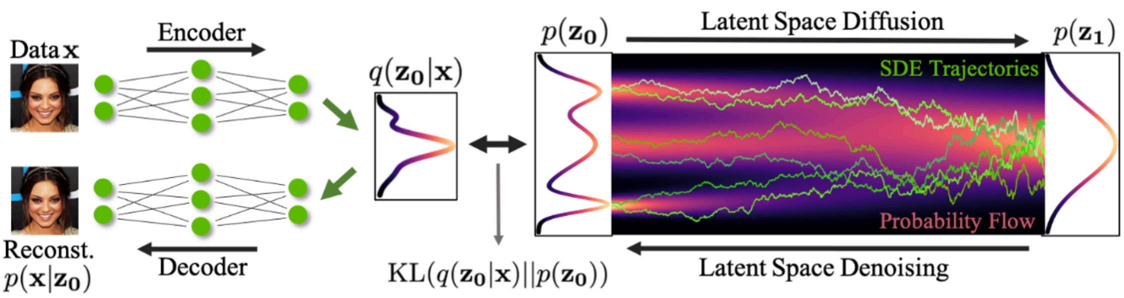

Figure 1: In our latent score-based generative model (LSGM), data is mapped to latent space via an encoder $q(\mathbf{z}_0|\mathbf{x})$ and a diffusion process is applied in the latent space ($\mathbf{z}_0 \to \mathbf{z}_1$). Synthesis starts from the base distribution $p(\mathbf{z}_1)$ and generates samples in latent space via denoising ($\mathbf{z}_0 \leftarrow \mathbf{z}_1$). Then, the samples are mapped from latent to data space using a decoder $p(\mathbf{x}|\mathbf{z}_0)$. The model is trained end-to-end.

Encoder $q_\phi(z_0|x)$, Decoder $p_\psi(x|z_0)$, SGM $p_\theta(z_t)$

其中

$$\nabla_{z_t} \log p_\theta(z_t) \approx \nabla_{z_t} \log q_t(z_t|z_0)$$

总的Loss如下

$$
\begin{aligned}
\mathcal{L}(\mathbf{x}, \phi, \theta, \psi) &= \mathbb{E}_{q_\phi(z_0|x)}\left[-\log p_\psi(\mathbf{x}|z_0)\right] + \mathrm{KL}(q_\phi(z_0|x)||p_\theta(z_0)) \\
&= \underbrace{\mathbb{E}_{q_\phi(z_0|x)}\left[-\log p_\psi(\mathbf{x}|z_0)\right]}_{\text{reconstruction term(Decoder)}} + \underbrace{\mathbb{E}_{q_\phi(z_0|x)}\left[\log q_\phi(z_0|x)\right]}_{\text{negative encoder entropy}} + \underbrace{\mathbb{E}_{q_\phi(z_0|x)}\left[-\log p_\theta(z_0)\right]}_{\text{cross entropy(DPM)}}
\end{aligned}
$$

The Cross Entropy Term

直接最小化上式的KL散度是Intractable的(见原文Sec 3.1)，因此作者提出将第三项交叉熵作如下转化

Theorem 1. *Given two distributions $q(\mathbf{z}_0|\mathbf{x})$ and $p(\mathbf{z}_0)$, defined in the continuous space $\mathbb{R}^D$, denote the marginal distributions of diffused samples under the SDE in Eq. 1 at time t with $q(\mathbf{z}_t|\mathbf{x})$ and $p(\mathbf{z}_t)$. Assuming mild smoothness conditions on $\log q(\mathbf{z}_t|\mathbf{x})$ and $\log p(\mathbf{z}_t)$, the cross entropy is:*

$$
CE(q(\mathbf{z}_0|\mathbf{x})||p(\mathbf{z}_0)) = \mathbb{E}_{t\sim\mathcal{U}[0,1]}\left[\frac{g(t)^2}{2}\mathbb{E}_{q(\mathbf{z}_t,\mathbf{z}_0|\mathbf{x})}\left[||\nabla_{\mathbf{z}_t}\log q(\mathbf{z}_t|\mathbf{z}_0) - \nabla_{\mathbf{z}_t}\log p(\mathbf{z}_t)||_2^2\right]\right] + \frac{D}{2}\log\left(2\pi e\sigma_0^2\right),
$$

with $q(\mathbf{z}_t, \mathbf{z}_0|\mathbf{x}) = q(\mathbf{z}_t|\mathbf{z}_0)q(\mathbf{z}_0|\mathbf{x})$ and a Normal transition kernel $q(\mathbf{z}_t|\mathbf{z}_0) = \mathcal{N}(\mathbf{z}_t; \boldsymbol{\mu}_t(\mathbf{z}_0), \sigma_t^2\mathbf{I})$, where $\boldsymbol{\mu}_t$ and σ_t^2 are obtained from $f(t)$ and $g(t)$ for a fixed initial variance σ_0^2 at $t = 0$.

A proof with generic expressions for $\boldsymbol{\mu}_t$ and σ_t^2 as well as an intuitive interpretation are in App. A.

将交叉熵转化为与Score Matching Loss一样的形式，此时不仅可用于优化$p_\theta(z_0)$，还可用于优化编码分布$q(z_0|x)$

Mixing Normal and Neural Score Functions

假设t时刻z_t的先验为$p(z_t) \propto \mathcal{N}(z_t; 0, 1)^{1-\alpha} p'_\theta(z_t)^\alpha$，其中$p'_\theta(z_t)$为可训练的先验，$\alpha$为可训练的标量，于是，该先验对应的Score Function为

$$
\nabla_{z_t}\log p(z_t) = -(1-\alpha)z_t + \alpha\nabla_{z_t}\log p'_\theta(z_t)
$$

于是，用$\epsilon_\theta(z_t, t)$参数化$\nabla_{z_t}\log p(z_t)$

$$
\nabla_{z_t}\log p(z_t) = -\frac{\epsilon_\theta(z_t, t)}{\sigma_t}
$$

其中

$$
\epsilon_\theta(z_t, t) = \sigma_t(1-\alpha)\odot z_t + \alpha\odot\epsilon'_\theta(z_t, t)
$$

考虑到$\nabla_{z_t}\log q(z_t|z_0) = -\frac{\epsilon}{\sigma_t}$，于是可将定理1的交叉熵化为

$$
\mathrm{CE}(q_\phi(\mathbf{z}_0|\mathbf{x})||p_\theta(\mathbf{z}_0)) = \mathbb{E}_{t\sim\mathcal{U}[0,1]}\left[\frac{w(t)}{2}\mathbb{E}_{q_\phi(\mathbf{z}_t,\mathbf{z}_0|\mathbf{x}),\varepsilon}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t)\|_2^2\right]\right] + \frac{D}{2}\log\left(2\pi e\sigma_0^2\right)
$$

其中

$$
w(t) = g(t)^2/\sigma_t^2
$$

Training with Different Weighting Mechanisms

利用上式训练Diffusion时，将权重$w(t)$全设为1可以提高生成的样本质量，但这仅可以用于训练Diffusion Prior，在更新Encoder ϕ时，仍需要使用完整的权重$w(t)$来保证$q_\phi(z_0|x)$能拟合后验$p_\psi(z_0|x)$

以下为作者实验的三种加权机制

Table 1: Weighting mechanisms

Mechanism	Weights
Weighted	$w_{\mathrm{ll}}(t) = g(t)^2/\sigma_t^2$
Unweighted	$w_{\mathrm{un}}(t) = 1$
Reweighted	$w_{\mathrm{re}}(t) = g(t)^2$

$$\min_{\phi,\psi} \mathbb{E}_{q_\phi(\mathbf{z}_0|\mathbf{x})}\left[-\log p_\psi(\mathbf{x}|\mathbf{z}_0)\right] + \mathbb{E}_{q_\phi(\mathbf{z}_0|\mathbf{x})}\left[\log q_\phi(\mathbf{z}_0|\mathbf{x})\right] + \mathbb{E}_{t,\boldsymbol{\epsilon},q(\mathbf{z}_t|\mathbf{z}_0),q_\phi(\mathbf{z}_0|\mathbf{x})}\left[\frac{w_{\mathrm{ll}}(t)}{2}||\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{z}_t,t)||_2^2\right]$$

VAE,Diffusion联合训练

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{t,\boldsymbol{\epsilon},q(\mathbf{z}_t|\mathbf{z}_0),q_\phi(\mathbf{z}_0|\mathbf{x})}\left[\frac{w_{\mathrm{ll/un/re}}(t)}{2}||\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{z}_t,t)||_2^2\right] \quad \text{with} \quad q(\mathbf{z}_t|\mathbf{z}_0) = \mathcal{N}(\mathbf{z}_t;\boldsymbol{\mu}_t(\mathbf{z}_0),\sigma_t^2\mathbf{I}),$$

Variance Reduction

Variance reduction for likelihood weighting

考虑VPSDE

$$\mathrm{d}z = -\frac{1}{2}\beta(t)z\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}w$$
$$\beta(t) = \beta_0 + (\beta_1 - \beta_0)t$$

前面的Loss(最大似然对应的完整权重的Loss)包含对t均匀采样，会带来很大的方差，下面讨论如何减小采样方差

考虑

$q(z_0) = p(z_0) = \mathcal{N}(z_0; 0, \mathbf{I})$，其中q(z_0)为边缘分布$\mathbb{E}_{p_{\mathrm{data}}}[q(z_0|x)]$，可以证明

$$\mathrm{CE}(q(z_0)\|p(z_0)) = \frac{D}{2}\mathbb{E}_{t\sim\mathcal{U}[0,1]}[\mathrm{d}\log\sigma_t^2/\mathrm{d}t] + c$$

- **Geometric VPSDE**
  设计SDE，使得
  $\mathrm{d}\log\sigma_t^2/\mathrm{d}t$为常数，如
  $\sigma_t^2 = \sigma_{\min}^2(\sigma_{\max}^2/\sigma_{\min}^2)^t, \beta(t) = \log(\sigma_{\max}^2/\sigma_{\min}^2)\frac{\sigma_t^2}{1-\sigma_t^2}$
- **Importance sampling**
  保持$\beta(t), \sigma_t^2$不变，使用重要性采样减少方差。可以证明，如下Proposal会带来最小采样方差

  $$r(t) \propto \mathrm{d}\log\sigma_t^2/\mathrm{d}t$$

  可以证明，对r(t)采样等价于对如下逆变换采样

  $$t = \mathrm{var}^{-1}((\sigma_1^2)^\rho(\sigma_0^2)^{1-\rho})$$

  其中$\mathrm{var}^{-1} = (\sigma_t^2)^{-1}$
  这种方法对任何形式$\beta(t)$的VPSDE均有效

Variance reduction for unweighted and reweighted objectives

对w_{re}，最优重要性采样分布为

$$r(t) \propto \mathrm{d}\sigma_t^2/\mathrm{d}t$$

等价于

$$t = \mathrm{var}^{-1}((1-\rho)\sigma_0^2 + \rho\sigma_1^2)$$

证明及更多细节见App. B