Deep Learning 2024011303 Liu Hanzuo

HW5

True or False

P1 True; in the noise adding step, we could replace the word with [MASK] until a all-mask sentence. If we take a relatively large dimension of word embedding, For the position embedding part, we remain them unchanged(such as cosine embedding). Thus, the final embedding sequence still contain position information. $H_t = [h_t, e_t]$, where $h_t = \sqrt{\alpha_t}h_0 + \sqrt{1-\alpha_t}\epsilon_t$. And the denoising step is to sample $P(H_{t-1}|H_t) = Transformer(H_t)$, and finally, after sample H_0, use nearest neighbor search to find the most similar word in the dictionary.

Q&A

P2 **1.**

$$q(x_t|x_{t-1}) = N(x_t|\sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I)$$

$$x_t = \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}\epsilon_t$$

We prove the equation by induction.
if x_{t-1} holds, for x_t, we have:

$$x_t = \sqrt{\alpha_t}(\sqrt{\overline{\alpha}_{t-1}}x_0 + \sqrt{1-\overline{\alpha}_{t-1}}\epsilon') + \sqrt{1-\alpha_t}\epsilon_t$$

$$= \sqrt{\overline{\alpha}_t}x_t + \sqrt{1-\overline{\alpha}_t}\epsilon$$

Where $\overline{\alpha}_t = \overline{\alpha}_{t-1}\alpha_t$, and $\sqrt{\alpha_t - \overline{\alpha}_t}\epsilon' + \sqrt{1-\alpha_t}\epsilon_t$ is a unit Gaussian with coefficient $\sqrt{1-\overline{\alpha}_t}$, thus equals to $\sqrt{1-\overline{\alpha}_t}\epsilon$, where $\epsilon \sim N(0,1)$
2. Note that

$$q(x_{t-1}|x_t, x_0) = \frac{q(x_t|x_{t-1}, x_0)\cdot q(x_{t-1}|x_0)}{q(x_t|x_0)} = \frac{q(x_t|x_{t-1})\cdot q(x_{t-1}|x_0)}{q(x_t|x_0)} \sim q(x_t|x_{t-1})\cdot q(x_{t-1}|x_0)$$

Also note that:

$$q(x_t|x_{t-1}) = N(x_t|\sqrt{\alpha_t}x_t, (1-\alpha_t)I), \quad q(x_{t-1}|x_0) = N(x_{t-1}|\sqrt{\overline{\alpha}_{t-1}}x_0, \sqrt{1-\overline{\alpha}_{t-1}}I)$$

Thus, we have:

$$\overline{\mu}_t = \frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})x_{t-1} + \sqrt{\overline{\alpha}_{t-1}}(1-\alpha_t)x_0}{\alpha_t(1-\overline{\alpha}_{t-1}) + \overline{\alpha}_{t-1}(1-\alpha_t)x_0}$$
$$= \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon)$$

Here we use the fact that $x_0 = \frac{x_t - \sqrt{1-\overline{\alpha}_t}\epsilon}{\sqrt{\overline{\alpha}_t}}$.

3. Note that $q(x_{1:T}|x_0) \sim q(x_{0:T})$

$$\mathbb{E}_{q(x_0)} - \log p_\theta(x_0) \le \mathbb{E}_{q(x_0)} - \log p_\theta(x_0) + D_{KL}(q(x_{1:T}|x_0)\|p_\theta(x_{1:T}|x_0))$$

$$= \sum -q(x_0)\log p_\theta(x_0) + \sum q(x_{0:T})\log\frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T}|x_0)}$$

$$= \sum -q(x_0)\log p_\theta(x_0) + \sum q(x_{0:T})\log\frac{q(x_{1:T}|x_0)p_\theta(x_0)}{p_\theta(x_{0:T})}$$

$$= \sum q(x_{0:T})\log\frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})} - \mathbb{E}_{q(x_0)}\log p_\theta(x_0) + \mathbb{E}_{q(x_{0:T})}\log p_\theta(x_0)$$

Note that $q(x_{0:T}) \sim q(x_0)$ and the expectation form does not contain any other form exclude x_0, thus the later two forms cancel each other.

$$= \sum q(x_{0:T})\log\frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})}$$

Thus the first half is proven.
For the later half, note that:

$$\log\frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})} = \log\frac{q(x_T|x_0)}{p_\theta(x_0|x_1)} + \sum_{t=2}^{T}\log\frac{q(x_{t-1}|x_t,x_0)}{p_\theta(x_{t-1}|x_t)} - \log p_\theta(x_0|x_1)$$

take expectation on both sides, we have:

$$\mathbb{E}_{q(x_{0:T})}\log\frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})} = \mathbb{E}_{q(x_{0:T})}\left[\log\frac{q(x_T|x_0)}{p_\theta(x_0|x_1)} + \sum_{t=2}^{T}\log\frac{q(x_{t-1}|x_t,x_0)}{p_\theta(x_{t-1}|x_t)} - \log p_\theta(x_0|x_1)\right]$$

$$= \mathbb{E}_q\left[D_{KL}(q(x_T|x_0)\|p_\theta(x_0|x_1)) + \sum_{t=2}^{T}\mathbb{E}_q + \sum_{t=2}^{T}D_{KL}(q(x_{t-1}|x_t,x_0)\|p_\theta(x_{t-1}|x_t)) - \log p_\theta(x_0|x_1)\right]$$

Thus we have show the total question.

4. Note that:

$$L_t = \mathbb{E}_{x_0,\epsilon}\left[\frac{1}{2\|\Sigma_0\|_2^2}\|\tilde{\mu}_t(x_t,x_0) - \mu_\theta(x_t,t)\|\right]$$

$$\tilde{\mu}_t - \mu_\theta = \frac{1}{\sqrt{\alpha_t}}(\frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}(\epsilon_0 - \epsilon_\theta)), \|\tilde{\mu}_t - \mu_\theta\|^2 = \frac{(1-\alpha_t)^2}{\alpha_t(1-\overline{\alpha}_t)}\|\epsilon_0 - \epsilon_\theta\|^2$$

Thus we have (bring in $x_t = \sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon_t$):

$$L_t = \mathbb{E}_{x_0,\epsilon}\left[\frac{1}{2\|\Sigma_0\|_2^2}\frac{(1-\alpha_t)^2}{\alpha_t(1-\overline{\alpha}_t)}\|\epsilon_0 - \epsilon_\theta(\sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon_t, t)\|^2\right]$$

And thus the initial statement is proven.

P3 From the definition of fisher divergence, we have:

$$F(p_{data}\|p_\theta) = \mathbb{E}_{x\sim p_{data}}\left[\frac{1}{2}\|\nabla_x \log p_{data}(x_0) - \nabla_x \log p_\theta(x_0)\|^2\right]$$

$$= \mathbb{E}_{x\sim p_{data}}\left[\frac{1}{2}\|\nabla_x \log p_x(x)\|^2 + \frac{1}{2}\|\nabla_x \log p_\theta(x)\|^2 - \nabla_x \log p_{data}(x)\cdot\nabla_x \log p_\theta(x_0)\right]$$

$$= \mathbb{E}_{x\sim p_{data}}\left[\frac{1}{2}\|\nabla_x \log p_x(x)\|^2 - \nabla_x \log p_{data}(x)\cdot\nabla_x \log p_\theta(x)\right] + Const$$

Thus we only need to show that:

$$\mathbb{E}_{x\sim p_{data}}\left[\nabla_x \log p_{data}(x)\cdot\nabla_x \log p_\theta(x)\right] = - - \mathbb{E}_{x\sim p_{data}}\left[tr(\nabla_x^2 \log p_\theta(x))\right]$$

Proof.

$$\mathbb{E}_{x\sim p_{data}}\left[\nabla_x \log p_{data}(x)\cdot\nabla_x \log p_\theta(x)\right]$$

$$= \int_x p_{data}(x)\nabla_x \log p_{data}(x)\cdot\nabla_x \log p_\theta(x)dx$$

$$= \int_x \nabla_x p_{data}(x)\cdot\nabla_x \log p_\theta(x)dx = \int_x \nabla_x \log p_\theta(x)dp_{data}(x)$$

$$= \nabla_x \log p_\theta(x)p_{data}(x)|_{-\infty}^{+\infty} - \int_x p_{data}(x)d\nabla_x \log p_\theta(x)$$

$$= -\int_x p_{data}(x)tr(\nabla_x^2 \log p_\theta(x))dx = -\mathbb{E}_{x\sim p_{data}}\left[tr(\nabla_x^2 \log p_\theta(x))\right]$$

$\square$

Thus the initial statement is proven.

P4

$$\mathbb{E}_{x\sim p_{data},\tilde{x}\sim q_\sigma(\cdot|x)}\left[\nabla_{\tilde{x}} \log q_\sigma(\tilde{x}|x)^T s_\theta(\tilde{x})\right]$$

$$= \int p_{data}(x)q_\sigma(\tilde{x}|x)\nabla_{\tilde{x}} \log q_\sigma(\tilde{x}|x)^T s_\theta(\tilde{x})d\tilde{x}dx$$

$$= \int p_{data}(x)\nabla_{\tilde{x}} q_\sigma(\tilde{x}|x)^T s_\theta(\tilde{x})d\tilde{x}dx$$

do the integral over x first, we could obtain that the equation equals to:

$$\int \nabla_{\tilde{x}} q_\sigma(\tilde{x})^T s_\theta(\tilde{x}) d\tilde{x} = \int q_\sigma(\tilde{x}) \nabla_{\tilde{x}} \log q_\sigma(\tilde{x})^T s_\theta(\tilde{x}) d\tilde{x}$$

Thus the initial statement holds.

P5 We begin with the process and the similarity between NCSN and DDPM.

Process of NCSN and DDPM:

1. NCSN $\sigma_1 > \cdots > \sigma_T$, learn the probability distribution $s_\theta(x, \sigma_t) = \nabla_x \log p_{\sigma_l}(x)$
2. DDPM predict the denoising step, the forwarding step is defined as: $q(x_t|x_{t-1}) = N(x_t; \sqrt{\alpha_t} x_{t-1}, (1-\alpha_t)I)$, and learn the denoising model $p_\theta(x_{t-1}|x_t)$

Similarity between NCSN and DDPM:

1. Denoising step similarity between NCSN and DDPM is that DDPM adjust the noise level latently ($\sqrt{1-\overline{\alpha}_t}$) and NCSN adjust the noise level manually.
2. Same optimization goal between NCSN and DDPM: NCSN is to minimize
$$\mathbb{E}\left[\|s_\theta(x, \sigma_t) - \nabla_x \log p_{\sigma_t}(x)\|^2\right]$$
While DDPM is to minimize
$$\mathbb{E}\left[\|\epsilon_\theta(x_t, t) - \epsilon_t\|^2\right]$$

Note that

$$q(x_t|x_0) = N(x_t; \sqrt{\overline{\alpha}_t} x_0, (1-\overline{\alpha}_t)I)$$

The score function of diffusion process could be defined as:

$$\nabla_{x_t} \log q(x_t|x_0) = -\frac{x_t - \sqrt{\overline{\alpha}_t} x_0}{1-\overline{\alpha}_t} = -\frac{\epsilon_t}{\sqrt{1-\overline{\alpha}_t}}$$

Thus the denoising step of diffusion could be approximated as:

$$s_\theta(x_t, t) = \nabla_{x_t} \log q(x_t|x_0) = -\frac{x_t - \sqrt{\overline{\alpha}_t} x_0}{1-\overline{\alpha}_t} = -\frac{\epsilon_t}{\sqrt{1-\overline{\alpha}_t}}$$