GLIDE(Guided Language to Image Diffusion for Generation and Editing)

ソース

- GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models

概要

- テキストから画像の生成の精度を上げた
- classifier-free guidance, clip guidanceという既存の手法に対して大量の学習を行って実験した
- classifier-free guidanceの方が精度良かった

前提知識

diffusion model

- 詳しくはDDPM, DDIM, Improved DDPMなどを見てください。

$$
\begin{aligned}
q\left(x_t|x_{t-1}\right) &:= \mathcal{N}\left(x_t; \sqrt{\alpha_t}x_{t-1}, \left(1-\alpha_t\right)I\right) \\
p_\theta\left(x_{t-1}|x_t\right) &:= \mathcal{N}\left(\mu_\theta\left(x_t\right), \Sigma_\theta\left(x_t\right)\right) \\
L_{\text{simple}} &:= E_{t\sim[1,T],x_0\sim q(x_0),\epsilon\sim\mathcal{N}(0,I)}\left[\left|\left|\epsilon-\epsilon_\theta\left(x_t,t\right)\right|\right|^2\right]
\end{aligned}
\tag{1}
$$

guided diffusion model

- 詳しくはADMを見てください

$$
\hat{\mu}_\theta\left(x_t|y\right) = \mu_\theta\left(x_t|y\right) + s\cdot\Sigma_\theta\left(x_t|y\right)\nabla_{x_t}\log p_\phi\left(y|x_t\right)
$$

classifier-free guidance

- 詳細はClassifier-Free Diffusion Guidance(https://openreview.net/forum?id=qw8AKxfYbI)
- conditional diffusion modelで条件付けを行う
- GLIDEではcaptionで重み付ける

- 特徴としては、diffusion modelの学習に条件づけを入れているために異なるguidanceをする際に再学習が必要になる

$$\hat{\epsilon}_\theta\left(x_t|y\right) = \epsilon_\theta\left(x_t|\emptyset\right) + s \cdot \left(\epsilon_\theta\left(x_t|y\right) - \epsilon_\theta\left(x_t|\emptyset\right)\right)$$

- $\epsilon_\theta\left(x_t|\emptyset\right)$は無条件ノイズのこと
- sは1以上の値をとる
- GLIDEではcaption cを用いて以下

$$\hat{\epsilon}_\theta\left(x_t|c\right) = \epsilon_\theta\left(x_t|\emptyset\right) + s \cdot \left(\epsilon_\theta\left(x_t|c\right) - \epsilon_\theta\left(x_t|\emptyset\right)\right)$$

clip guidance

- classifier modelの損失勾配で重み付けるというADMのideaを利用
- ただし、CLIPの潜在空間での類似度を利用
  - $\log p_\phi\left(y|x_t\right)$ではなく$f\left(x_t\right) \cdot g\left(c\right)$
- 2つのモデルが必要だが、同時に学習する必要はない
- CLIPが、きれいな画像だけではなくノイズ付きの画像に対しても学習されていないといけない

$$\hat{\mu}_\theta\left(x_t|y\right) = \mu_\theta\left(x_t|y\right) + s \cdot \Sigma_\theta\left(x_t|y\right) \nabla_{x_t}\left(f\left(x_t\right) \cdot g\left(c\right)\right)$$

- fはimage encoder
- gはtext encoder
- cはcaption

学習条件

classifier-free guidance

- 35億のパラメータから成るテキスト条件付け拡散モデル(text conditional diffusion model)に対して64×64の解像度の画像を学習

clip guidance

- 64×64の解像度でVit-L CLIPで学習