Unsupervised Representation Learning from Pre-trained Diffusion Probabilistic Models

Denoising Diffusion Probabilistic Models

Forward process:

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t} x_{t-1}, \beta_t I),$$
$$q(x_{1:T} \mid x_0) = \prod_{t=1}^{T} q(x_t \mid x_{t-1})$$

Reverse process:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$
$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t)$$

Objective:

$$\mathcal{L}_{simple}(\theta) = \mathbf{E}_{x_0,t,\epsilon} \left[\left\| \epsilon - \epsilon_\theta \left(\sqrt{\overline{\alpha}_t} x_0 + \sqrt{1-\overline{\alpha}_t} \epsilon, t \right) \right\|^2 \right]$$

Classifier-guided Sampling Method

1. Train a classifier $p_\phi(y \,|\, x_t)$ on noisy data
2. Use $\nabla_{x_t} \log p_\phi(y \,|\, x_t)$ to guide pretrained unconditional DDPM to sample from a class y:
   - Get $p_t(x_t \,|\, y)$ using classifier:

     $$p_t(x_t \,|\, y) = \frac{p(y \,|\, x_t) p_t(x_t)}{p(y)}$$

   - Get $\nabla_{x_t} \log p_t(x_t \,|\, y)$ using classifier:

     $$\nabla_{x_t} \log p_t(x_t \,|\, y) = \nabla_{x_t} \log p(y \,|\, x_t) + \nabla_{x_t} \log p_t(x_t)$$

   - Then we have

     $$p_{\theta,\phi}(x_{t-1} \,|\, x_t, y) = \mathcal{N}(x_t; \mu_\theta(x_t, t) + \Sigma_\theta(x_t, t) \cdot \nabla_{x_t} \log p_\phi(y \,|\, x_t), \Sigma_\theta(x_t, t))$$

Motivation

Observation (Posterior mean gap)

1. There is a gap between $p_\theta(x_{t-1} \,|\, x_t)$ and the posterior $q(x_{t-1} \,|\, x_t, x_0)$ for fully trained DDPM.
2. If Σ_θ is set as untrained time dependent constants, this is equivalent as the mean gap, i.e.

$$\|\mu_\theta(x_t, t) - \tilde{\mu}_t(x_t, x_0)\|$$

3. This gap is smaller for class-conditional DPMs, i.e.

$$\|\mu_\theta(x_t, t) - \tilde{\mu}_t(x_t, x_0)\| > \|\mu_\theta(x_t, y, t) - \tilde{\mu}_t(x_t, x_0)\|$$

Motivation

Conjecture

1. The posterior gap is caused by the information loss in the forward process.
2. The label y contains *some* information about x_0 reducing the gap.
3. If y contains *all* information about x_0, the the gap will be filled, and x_0 can be recovered.
4. Conversely, if we train a model to predict mean shift according to an encoded latent z and train it to fill the gap as much as possible, then z will learn as much information as possible from x_0.

Components

- Encoder: $z = E_\varphi(x_0)$
- Decoder: pre-trained unconditional DPM

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

- Gradient-estimator: $G_\psi(x_t, z, t) \simeq \nabla_{x_t} \log p(z \mid x_t)$

Algorithm

Algorithm 1 Training

Given:

$p_{data}(x_0)$, pretrained DPM $(\epsilon_\theta, \Sigma_\theta)$, encoder E_φ, gradient-estimator G_ψ

while not converge **do**

$x_0 \sim p_{data}(x_0)$

$t \sim \text{Unif}(1, 2, \cdots, T)$

$\epsilon \sim \mathcal{N}(0, I)$

$x_t \leftarrow \sqrt{\overline{\alpha}_t} x_0 + \sqrt{1 - \overline{\alpha}_t} \epsilon$

$\mathcal{L}(\varphi, \psi) \leftarrow$

$$\lambda_t \left\| \epsilon - \epsilon_\theta(x_t, t) + \frac{\sqrt{\alpha_t}\sqrt{1-\overline{\alpha}_t}}{\beta_t} \cdot \Sigma_\theta(x_t, t) \cdot G_\psi(x_t, E_\varphi(x_0), t) \right\|^2$$

$\varphi \leftarrow \varphi - \eta \nabla_\varphi \mathcal{L}$

$\psi \leftarrow \psi - \eta \nabla_\eta \mathcal{L}$

end while

P2 Weighting

What information does the model learn at each step during training?

- $\text{SNR} < 10^{-2}$ (large t): coarse features
- $10^{-2} \leq \text{SNR} < 10^{0}$ (middle t): content
- $\text{SNR} \geq 10^{0}$ (small t): imperceptible details (denoising)

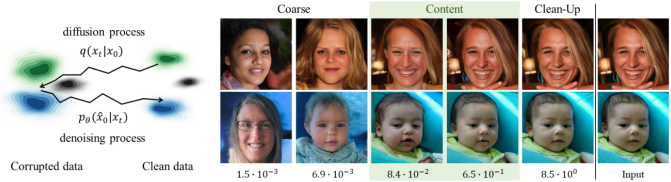

Figure 1: Stochastic reconstruction

P2 Weighting

Compensating previous observation, one can *redesign* the training weight λ_t satisfying

- Assign minimal weight to the clean-up stage
- Emphasize training on the content stage

$\implies$ *P2 Weighting*

$$\lambda'_t = \frac{\lambda_t}{(k + \mathsf{SNR}(t))^\gamma},$$

where γ and k are hyperparameters.

Weighting Scheme Redesign

Similarly, authors experience different effects of classifier guidance (or mean shift) for different time stage. Compensating such observation, authors redesign the weighting as

$$\lambda_t = \left(\frac{1}{1+\mathsf{SNR}(t)}\right)^{1-\gamma} \cdot \left(\frac{\mathsf{SNR}(t)}{1+\mathsf{SNR}(t)}\right)^{\gamma}$$

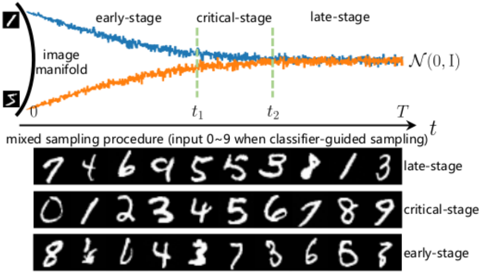

Figure 2: Effect of classifier guidance on different stage of sampling.

Experiments

Is posterior mean gap really filled?

1. Average posterior mean gap is smaller for PDAE than for pretrained DPM

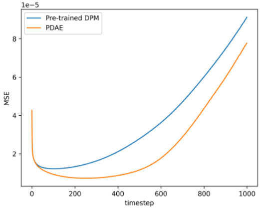

Figure 3: Average posterior mean gap

2. x_0 is well reconstructed from x_t with only one-step denoising.

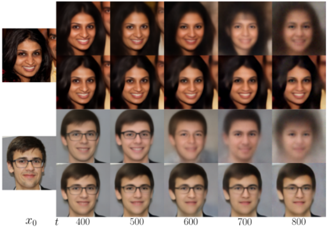

Figure 4: One step reconstruction of x_0 from x_t

Experiments

Figure 5: Autoencdoer reconstruction

Experiments

Model	Latent dim	SSIM ↑	LPIPS ↓	MSE ↓
StyleGAN2 ($\mathcal{W}$ inversion) [22]	512	0.677	0.168	0.016
StyleGAN2 ($\mathcal{W}$+ inversion) [1, 2]	7,168	0.827	0.114	0.006
VQ-GAN [10]	65,536	0.782	0.109	3.61e-3
VQ-VAE2 [37]	327,680	0.947	0.012	4.87e-4
NVAE [47]	6,005,760	0.984	**0.001**	**4.85e-5**
Diff-AE @130M (T=100, random $\boldsymbol{x}_T$) [36]	512	0.677	0.073	0.007
PDAE @64M (T=100, random $\boldsymbol{x}_T$)	512	0.696	0.094	0.005
DDIM @130M (T=100) [44]	49,152	0.917	0.063	0.002
Diff-AE @130M (T=100, inferred $\boldsymbol{x}_T$) [36]	49,664	0.991	0.011	6.07e-5
PDAE @64M (T=100, inferred $\boldsymbol{x}_T$)	49,664	**0.993**	0.008	5.48e-5

Figure 6: Autoencoder reconstruction quality of different models

Experiments

One can interpolate smoothly between image, by interpolating the guidance in one of the following ways:

- $G_\psi(x_t, Lerp(z^1, z^2; \lambda), t)$ (First row)
- $Lerp(G_\psi(x_t, z^1, t), G_\psi(x_t, z^2, t); \lambda)$ (Second row)

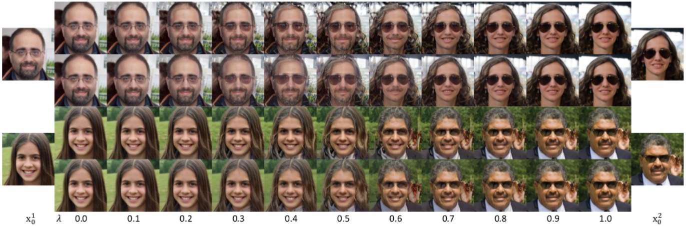

Figure 7: Interpolation

Experiments

For a given attribute c, train a classifier $w^\intercal z + b$ that outputs probability of a latent z having positive c. Then by taking

$$z' = z + sw,$$

with $s > 0$, we expect more c and with $s < 0$, we expect less c.

Figure 8: Attribute manipulation

Thank You

Q & A