Simple Diffusion

Мещанинов Вячеслав

Centre of Deep Learning and Bayesian Methods
HSE University

Background

Diffusion process:

$$q(z_t|x) = N(z_t|\alpha_t x, \sigma_t^2)$$

$$q(z_t|z_s) = N(z_t|\alpha_{ts} z_s, \sigma_{ts}^2)$$

$$\alpha_{ts} = \frac{\alpha_t}{\alpha_s}$$

$$\sigma_{ts}^2 = \sigma_t^2 - \alpha_{t|s}^2 \sigma_s^2$$

Denoising:

$$q(z_s|z_t, x) = N(z_t|\mu_{t\to s}, \sigma_{t\to s}^2)$$

$$\mu_{t\to s} = \frac{\alpha_{ts}\sigma_s^2}{\sigma_t^2} z_t + \frac{\alpha_s \sigma_{ts}^2}{\sigma_t^2} x$$

$$\sigma_{t\to s} = \frac{\sigma_{ts}^2 \sigma_s^2}{\sigma_t^2}$$

$$\hat{x} = f_\theta(z_t, t)$$

Parametrization

DDPM: $\hat{\varepsilon} = f_\theta(z_t, t)$ $\hat{x} = \frac{z_t - \sigma_t \hat{\varepsilon}}{\alpha_t}$

При $t = 1 \Rightarrow \alpha_t = 0$ предсказание получается неустойчивым

Авторы предлагают решение:

$$v_t = \alpha_t \varepsilon - \sigma_t x$$

$$\hat{x} = \alpha_t z_t - \sigma_t \hat{v}_t$$

$$\hat{\varepsilon} = \sigma_t z_t + \alpha_t \hat{v}_t$$

Optimization

Авторы, предложившие v-параметризацию, используют следующую функцию потерь

$$L_{v_t} = ||v_t - \hat{v}_t||_2^2$$

$$L_{v_t} = (1 + \frac{\alpha_t^2}{\sigma_t^2})L_x = \frac{1}{\alpha_t^2}L_\varepsilon$$

Авторы Simple Diffusion используют

$$L_\varepsilon = \alpha_t^2 L_{v_t}$$

в этом случае моменты близкие к 1 входят в функцию потерь с коэффициентом близким к нулю

Noise Schedule for High Resolutions

SNR for High Resolutions

Пересчет SNR для уменьшенных изображений:

$$SNR^{d/s \times d/s}(t) = SNR^{d \times d}(t) \cdot s^2$$

SNR для расписания DDPM для изображений разрешения 64 x 64:

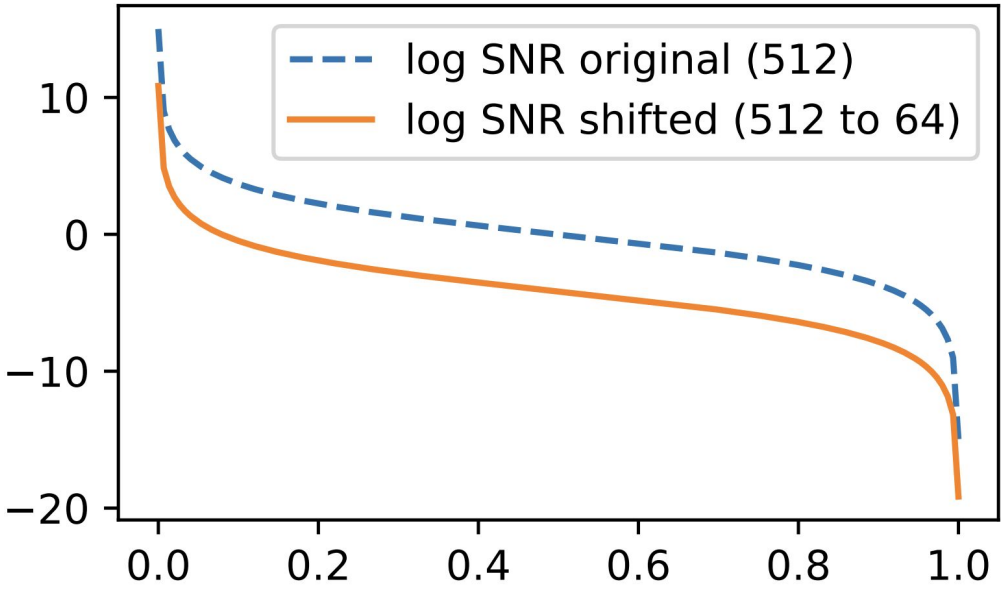
$$SNR^{64 \times 64}_{DDPM}(t) = tanh(\frac{\pi t^2}{2})^{-2}$$

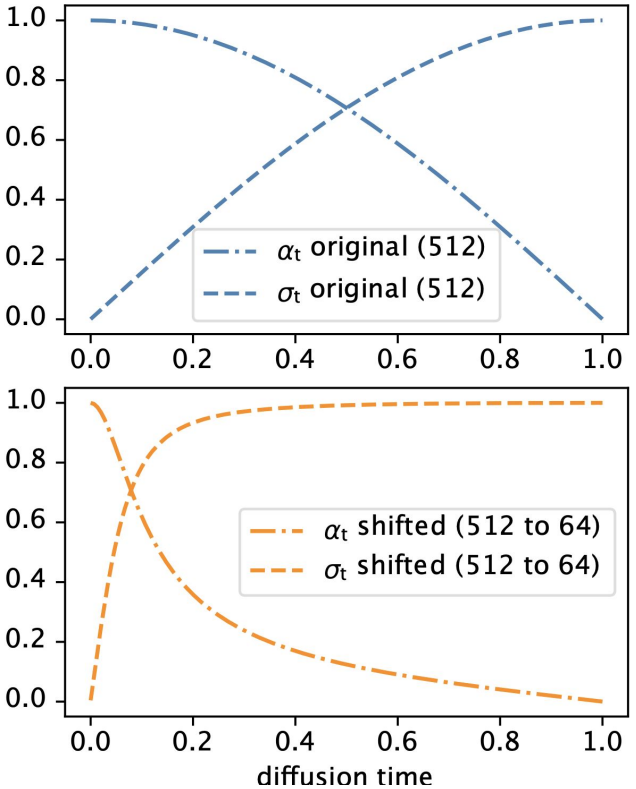
Тогда можно пересчитать SNR и коэффициенты расписания для увеличенных изображений

$$\alpha_t^2 = sigmoid(logSNR(t))$$

$$\sigma_t^2 = sigmoid(-logSNR(t))$$

Shifted SNR





Noise Schedule on ImageNet

Noise Schedule	FID train	FID eval
128 × 128 resolution		
cosine (original at 128)	2.96	3.38
cosine (shifted to 64)	2.41	3.03
cosine (shifted to 32)	**2.26**	**2.88**
256 × 256 resolution		
cosine (original at 256)	7.65	6.87
cosine (shifted to 128)	5.05	4.74
cosine (shifted to 64)	3.94	3.89
cosine (shifted to 32)	**3.76**	**3.71**

Multiscale Training Loss

$$\widetilde{L}^{d\times d} = \sum_{s\in\{32,64,128,\ldots,d\}} \frac{1}{s} L^{s\times s}$$

$$L^{s\times s} = \frac{1}{s^2} \left|\left| D^{s\times s}[\varepsilon] - D^{s\times s}[\hat{\varepsilon}] \right|\right|$$

Resolution	FID train	FID eval	IS
256	**3.76**	**3.71**	**171.6**
+ multiscale loss (32)	4.00	3.89	171.0
512	4.85	4.58	156.1
+ multiscale loss (32)	**4.30**	**4.28**	**171.0**

D[] — операция уменьшения разрешения

d — изначальное разрешение изображения

Comparison to Generative Models

Method	FID train	FID eval	IS
128 × 128 resolution			
ADM (Dhariwal & Nichol, 2021)	5.91		
CDM (32, 64, 128) (Ho et al., 2022)	3.52	3.76	128.8 ± 2.51
RIN (Jabri et al., 2022)	2.75		144.1
simple diffusion (U-Net) (ours)	2.26	**2.88**	137.3 ± 2.03
simple diffusion (U-ViT, L) (ours)	**1.94**	3.23	**171.9** ± 3.24
256 × 256 resolution			
BigGAN-deep (no truncation)	6.9		171.4 ± 2
MaskGIT (Chang et al., 2022)	6.18		182.1
DPC* (full 5) (Anonymous, 2023)	4.45		**244.8**
Denoising diffusion models			
ADM (Dhariwal & Nichol, 2021)	10.94		
CDM (32, 64, 256) (Ho et al., 2022)	4.88	4.63	158.71 ± 2.26
LDM-4 (Rombach et al., 2022)	10.56		103.49
RIN (Jabri et al., 2022)	4.51		161.0
DiT-XL/2 (Peebles & Xie, 2022)	9.62		121.5
simple diffusion (U-Net) (ours)	3.76	**3.71**	171.6 ± 3.07
simple diffusion (U-ViT, L) (ours)	**2.77**	3.75	211.8 ± 2.93
512 × 512 resolution			
MaskGIT (Chang et al., 2022)	7.32		156.0
DPC (U)* (Anonymous, 2023)	3.62		**249.4**
Denoising diffusion models			
ADM (Dhariwal & Nichol, 2021)	23.24		
DiT-XL/2 (Peebles & Xie, 2022)	12.03		105.3
simple diffusion (U-Net) (ours)	4.30	**4.28**	171.0 ± 3.00
simple diffusion (U-ViT, L) (ours)	**3.54**	4.53	205.3± 2.65

Generated Images

p(B|A)yesgroup.ru

512 x 512

256 x 256

128 x 128