Latent Diffusion

Reuben Brasher

May 4, 2023

Diffusion models

Diffusion objective function

$$L_{DM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1)}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right]$$

Latent diffusion objective function

Encode x into a *Latent space* as $z = \mathcal{E}(x)$.

$$L_{LDM} = \mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1)}\left[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2\right]$$

Cross-attention mechanism

If y is a modal input, such a text description compute Q, K, V by

$$Q = W_Q \varphi(z_t)$$
$$K = W_K \tau(y)$$
$$V = W_V \tau(y)$$

Then $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V$.

Conditional latent diffusion objective function

$$L_{CLDM} = \mathbb{E}_{\mathcal{E}(x), \epsilon \sim \mathcal{N}(0,1)} \left[\|\epsilon - \epsilon_\theta(z_t, t, \tau(y))\|_2^2 \right]$$

Futher reading

Rombach et al., 2022
Zhang and Agrawala, 2023

References I

Rombach, Robin et al. (2022). "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695.

Zhang, Lvmin and Maneesh Agrawala (2023). "Adding conditional control to text-to-image diffusion models". In: *arXiv preprint arXiv:2302.05543*.