Semantic Latent Space in Diffusion Models

Vinayak Goyal   Bhavya Singh   Prateek Garg   Dadhichi Telwadkar

Department of Electrical Engineering

Indian Institute of Technology Bombay

May 3,2023

Table of Contents

Task Description

- Controlling the generative process in diffusion models to get desirable changes.

(a) "Dog"

(b) "Similing Dog"

Figure: Generate the image (b) given (a) with attribute "smiling"

Table of Contents

Denoising Diffusion Implicit Model

- DDIM conditions the original process on the original image x_o

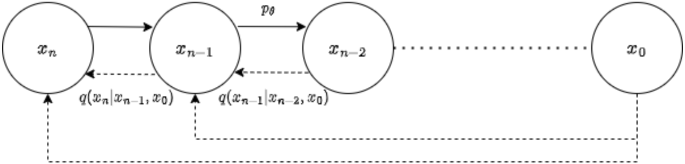


Figure: Non-Markovian process which trains faster then DDPM.

$$x_{t-1} = \sqrt{\alpha_{t-1}} \left(\frac{x_t - \sqrt{1-\alpha_t}\epsilon_t^\theta(x_t)}{\sqrt{\alpha_t}} \right) + \sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \epsilon_t^\theta(x_t) + \sigma_t z_t, \quad (1)$$

Denoising Diffusion Implicit Model

- $\boldsymbol{P}_t(\epsilon_t^\theta(x_t)) := \left(\frac{x_t - \sqrt{1-\alpha_t}\epsilon_t^\theta(x_t)}{\sqrt{\alpha_t}}\right)$,
  $\boldsymbol{D}_t(\epsilon_t^\theta(x_t)) := \sqrt{1 - \alpha_{t-1} - \sigma_t^2} \cdot \epsilon_t^\theta(x_t)$
- Now, (1) becomes

$$x_{t-1} = \sqrt{\alpha_{t-1}}\boldsymbol{P}_t(\epsilon_t^\theta(x_t)) + \boldsymbol{D}_t(\epsilon_t^\theta(x_t)), \tag{2}$$

- Can we manipulate $\epsilon_t^\theta(x_t)$ to get desired changes?
  - If yes, How do we know the direction of the Δ_ϵ change?
  - Ans: Use CLIP

CLIP

- CLIP embeds images and texts, whose similarity indicates semantic similarity between them

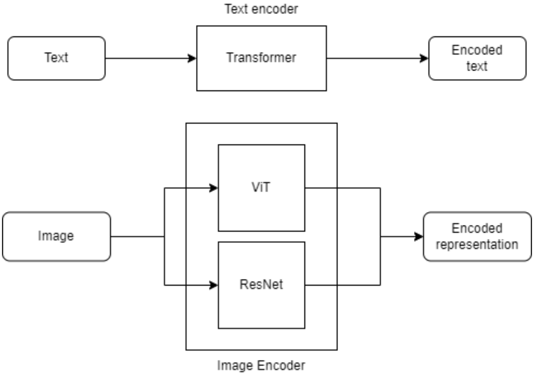


Figure: Architecture of CLIP

- E_T : $text \to \mathbb{R}^N$, E_I : $Image \to \mathbb{R}^N$

CLIP

- **Idea**: Maximise cosine similarity between target description and edited image.
- Directional loss with cosine distance achieves homogeneous editing without mode collapse

$$\mathcal{L}(x^{edit}, y^{target}; x^{source}, y^{source}) := 1 - \frac{\Delta I \cdot \Delta T}{||\Delta I|| \cdot ||\Delta T||}, \qquad (3)$$

  where
  $\Delta I := E_I(x^{edit}) - E_I(x^{source}), \Delta T := E_T(y^{target}) - E_T(y^{source})$
- For example: Given a face image x^{source} and attribute `smiling` y^{target} := "smiling face" y^{source} := "face", we optimise x^{edit} on 3

Learned Perceptual Image Patch Similarity

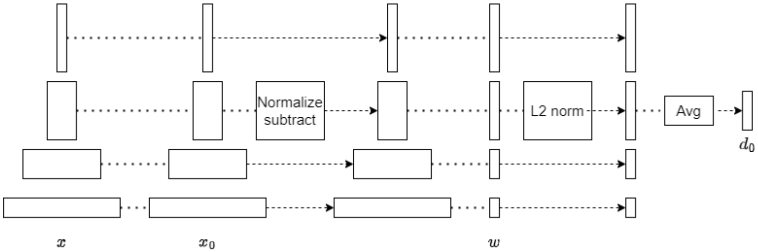


Figure: LPIPS is used to measure similarity in between images x and x_0

- LPIPS computes the similarity between the activations of two image patches for some pre-defined network.
- This measure has been shown to match human perception well.
- A low LPIPS score means that image patches are perceptually similar.

Table of Contents

Discovering Latent Space

- **Update** x_T to optimize the directional CLIP loss given text prompts: leads to distortion or incorrect manipulation

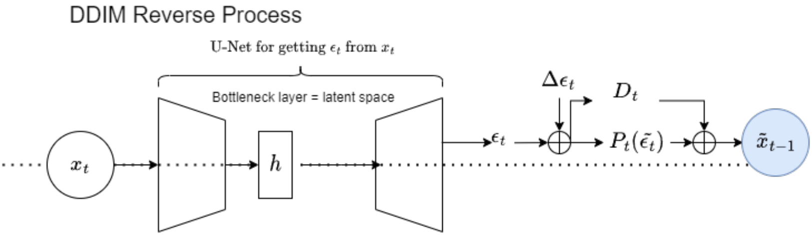


Figure: Comparison of DDIM and Asyrp

- **Shift the noise vector** ϵ_t^θ at each sampling step: changes in $\boldsymbol{P}_t$ and $\boldsymbol{D}_t$ cancel out similar to destructive interference

Table of Contents

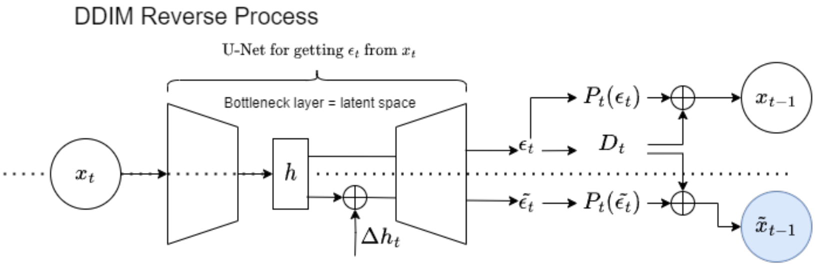
Asymmetric Reverse Process

- Asyrp applies change to only $\boldsymbol{P}$ term

$$\tilde{x}_{t-1} = \sqrt{\alpha_{t-1}}\boldsymbol{P}_t(\tilde{\epsilon}_t^{\theta}(x_t)) + \boldsymbol{D}_t(\epsilon_t^{\theta}(x_t)), \qquad (4)$$

where

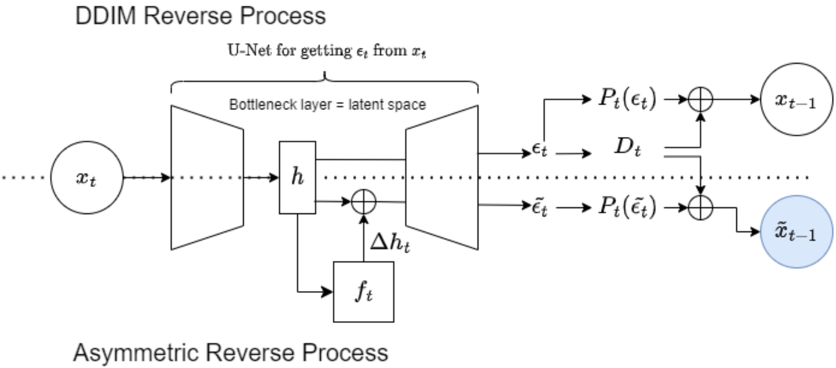
$$\Delta\epsilon_t = \tilde{\epsilon}_t^{\theta} - \epsilon_t^{\theta}$$



- Changes are applied to bottleneck of U-net

Asymmetric Reverse Process

- Additionally, instead of optimizing Δh at every time, a neural f_t is trained on clip loss to predict Δh_t



$\mathcal{H}$-space

The h-space, represented by h_t, has smaller spacial resolutions and high-level semantics than ϵ_t^θ

$$x_{t-1} = \sqrt{\alpha_{t-1}}\boldsymbol{P}_t(\epsilon_t^\theta(x_t|\Delta h_t)) + \boldsymbol{D}_t(\epsilon_t^\theta(x_t)) + \sigma_t z_t \quad (5)$$

Optimizing Δh_t requires a lot of training examples and training time and thus we define a function $f_t(h_t)$(implemented as a neural network) which produces Δh_t for a given h_t and t. f_t converges faster then learning all Δh_t.

Editing Intervals and Quality Boosting

In a diffusion model, early time-steps generate high-level context. Later time steps generate fine details.

Editing Interval: the smallest early interval $[T, t_{edit}]$ that brings enough distinguishable changes to the images. *Editing Strength* is used to determine t_{edit}:

$$\xi_t = LPIPS(x, \boldsymbol{P}_T) - LPIPS(x, \boldsymbol{P}_t)$$
$$LPIPS(x, x_{t_{edit}}) = 0.33$$

Boost Interval: Injecting stochastic noise improves image quality but longer intervals may cause modifications to the image. *Quality Deficiency*, which measures the noise in x_t compared to original image x, is used to determine t_{boost}:

$$\lambda_t = LPIPS(x, x_t)$$
$$\lambda_t = LPIPS(x, P_{t_{boost}}) = 1.2$$

Generative Process

$$p_\theta^{(t)}(x_{t-1}|x_t) = \begin{cases} \mathcal{N}\left(\sqrt{\alpha_{t-1}}\boldsymbol{P}_t(\epsilon_t^\theta(x_t|\boldsymbol{f}_t)) + \boldsymbol{D}_t, \sigma_t^2 \boldsymbol{I}\right), \eta = 0 & [T, t_{edit}] \\ \mathcal{N}\left(\sqrt{\alpha_{t-1}}\boldsymbol{P}_t(\epsilon_t^\theta(x_t)) + \boldsymbol{D}_t, \sigma_t^2 \boldsymbol{I}\right), \eta = 0 & (t_{edit}, t_{boost}] \\ \mathcal{N}\left(\sqrt{\alpha_{t-1}}\boldsymbol{P}_t(\epsilon_t^\theta(x_t)) + \boldsymbol{D}_t, \sigma_t^2 \boldsymbol{I}\right), \eta = 1 & (t_{boost}, 0] \end{cases} \tag{6}$$

Results

These are the results that we got by implementing changes for a h-space in a DDIM:

(a) (b) (c) (d)

Figure: Top is the input. Bottom is output

Table of Contents

Work Distribution

1. Coding and Theory: Dadhichi Telwadkar and Prateek Garg
2. Presentation and Report: Vinayak Goyal, Bhavya Singh

Table of Contents

References

- Kwon, M., Jeong, J. & Uh, Y. Diffusion Models Already Have A Semantic Latent Space. *The Eleventh International Conference On Learning Representations* . (2023), https://openreview.net/forum?id=pd1P2eUBVfq
- Song, J., Meng, C. & Ermon, S. Denoising Diffusion Implicit Models. *International Conference On Learning Representations*. (2021), https://openreview.net/forum?id=St1giarCHLP
- Radford, A., Kim, J., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. & Sutskever, I. Learning Transferable Visual Models From Natural Language Supervision. (2021)
- Zhang, R., Isola, P., Efros, A., Shechtman, E. & Wang, O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. (2018)

The End

Questions? Comments?