Image Process - Individual Study Report
New Age Digital Art
310551067 吳岱容

- **Introduction**

  Image generation has been a very hot topic among deep learning applications. Recently, in the artistic creation area, this kind of technology has triggered a heated debate, since several business or non-business applications are able to generate art of different styles, and also, in specific situations, whose generated results look better than creations from human-being, such as Midjourney, Stable Diffusion, DALL-E 2, NovelAI. We might think of GAN when we talk about image generation, but the applications mentioned above are using a very different concept called diffusion model. This model is not a very new method, however, its application has had an eye-catching result recently.

Picture of Stable Diffusion's Result (from their GitHub Page)

Picture of Midjourney's Result (from their Gallery)

In this paper report, we will focus on the papers related to the diffusion model to have a glimpse into the theory behind these applications. The first paper presents high quality image synthesis results using diffusion probabilistic model, of which the concept is nonequilibrium thermodynamics.

The second paper is the previous work of Stable Diffusion. The third one is the research paper of DALL-E 2.

- **Denoising Diffusion Probabilistic Models**

  This paper presents high quality image synthesis results using diffusion probabilistic model, which is a class of latent variable models inspired by considerations from nonequilibrium thermodynamics[1]. A diffusion probabilistic model (which we will call a “diffusion model” for brevity) is a parameterized Markov chain trained using variational inference to produce samples matching the data after finite time[2].

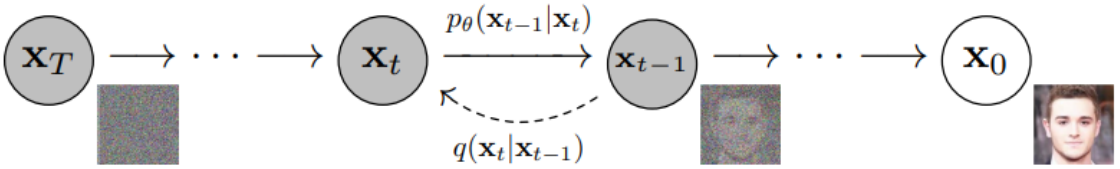
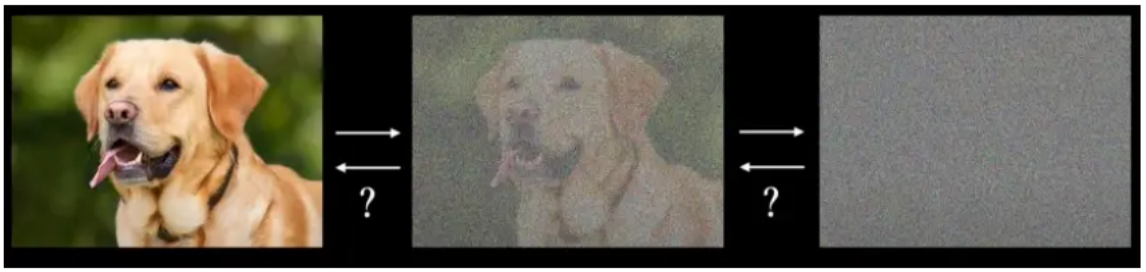


Figure 2: The directed graphical model considered in this work.

[3]

  - Concept

    The core of the diffusion model is to let the model learn how to denoise. That means when we are training the model, we will first add noise to the picture step by step, and then we train the model to denoise the picture.



[4]

    - Markov Chain

      Diffusion probability model is a parameterized Markov Chain. To be simple, the Markov chain defines a series of phenomena that will evolve into different possibilities as time goes on, and at every different time, every different state or phenomena is called state space. If there are k different states, we can turn them into a k by k matrix. Every image in the training state can be taken as a Markov Chain, and the learning target is to denoise (stochastic matrix).

  - Model

---

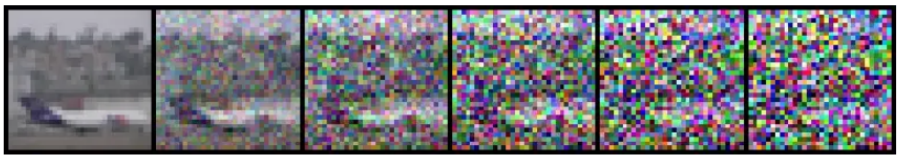
[1] Denoising Diffusion Probabilistic Models (CVPR 2020) - Abstract
[2] Denoising Diffusion Probabilistic Models (CVPR 2020) - Introduction

[3] Figure from the paper (Denoising Diffusion Probabilistic Models)
[4] Picture from https://www.youtube.com/watch?v=fbLgFrlTnGU&t=208s&ab_channel=AriSeff

- Diffusion process

  In the diffusion process, we let the image be the form of a Markov chain to add the Gaussian noise step by step.

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) := \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \qquad q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$$

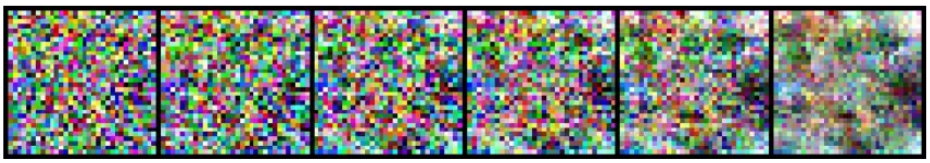


[5]

- Reverse process

  In the reverse process, the task we want to achieve is to denoise the image. The image can still be presented as Markov Chain form. We can take it as a Gaussian distribution, and therefore, the model's target is to learn the mean value and the variance of the distribution.

$$p_\theta(\mathbf{x}_{0:T}) := p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \qquad p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$



[6]

- Algorithm

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \text{Uniform}(\{1, \dots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$$
6: **until** converged

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \dots, 1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

[7]

- Experiment Results

[5] https://d246810g2000.medium.com/%E5%9F%BA%E6%96%BC-diffusion-models-%E7%9A%84%E7%94%9F%E6%88%90%E5%9C%96%E5%83%8F%E6%BC%94%E7%AE%97%E6%B3%95-984212710610

[6] https://d246810g2000.medium.com/%E5%9F%BA%E6%96%BC-diffusion-models-%E7%9A%84%E7%94%9F%E6%88%90%E5%9C%96%E5%83%8F%E6%BC%94%E7%AE%97%E6%B3%95-984212710610

[7] Denoising Diffusion Probabilistic Models (CVPR 2020)

- Rate Distortion

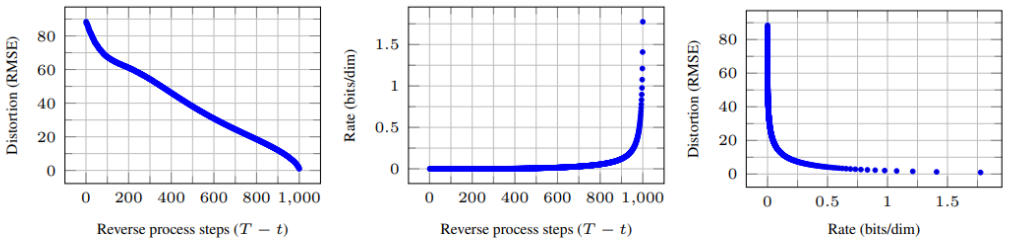


Figure 5: Unconditional CIFAR10 test set rate-distortion vs. time. Distortion is measured in root mean squared error on a $[0, 255]$ scale. See Table 4 for details.

8

- Progressive generation

Figure 6: Unconditional CIFAR10 progressive generation ($\hat{\mathbf{x}}_0$ over time, from left to right). Extended samples and sample quality metrics over time in the appendix (Figs. 10 and 14).

9

- Connection to autoregressive decoding

Figure 8: Interpolations of CelebA-HQ 256x256 images with 500 timesteps of diffusion.

10

- **High-Resolution Image Synthesis with Latent Diffusion Models**

  The target of the paper is to decrease the computing complexity of Diffusion models. The main contribution of this work is significantly reducing the computational requirements compared to pixel-based Diffusion models by applying the training in the latent space.[11] While trying to decrease computational requirements, the model still can have a comparable image quality and also flexibility. Another contribution is that the paper introduced the cross-attention layers into the model architecture, and this enabled tasks such

[8] Denoising Diffusion Probabilistic Models (CVPR 2020)
[9] Denoising Diffusion Probabilistic Models (CVPR 2020)
[10] Denoising Diffusion Probabilistic Models (CVPR 2020)
[11] High-Resolution Image Synthesis with Latent Diffusion Models (CVPR 2022)

as class-condition, text-to-image, layout-to-image are possible.

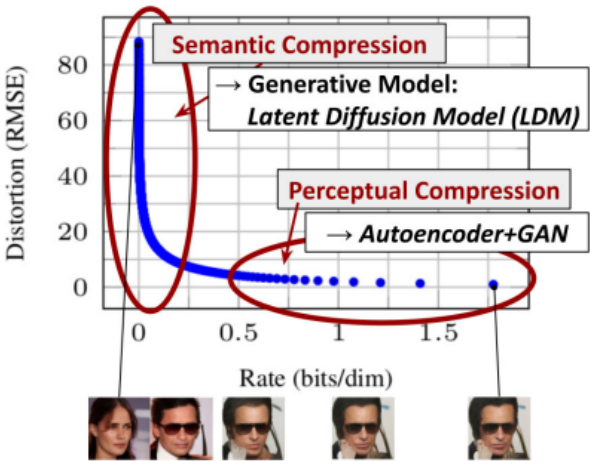


Figure 2. Illustrating perceptual and semantic compression: Most bits of a digital image correspond to imperceptible details. While DMs allow to suppress this semantically meaningless information by minimizing the responsible loss term, gradients (during training) and the neural network backbone (training and inference) still need to be evaluated on all pixels, leading to superfluous computations and unnecessarily expensive optimization and inference. We propose *latent diffusion models (LDMs)* as an effective generative model and a separate mild compression stage that only eliminates imperceptible details. Data and images from [30].

[12]

- Concept
  - Departure to Latent Space

    Learning of likely-based model conventionally can be divided into two parts:

    (1) Perceptual Compression

    Removes high-frequency details but still learns little semantic variation.[13]

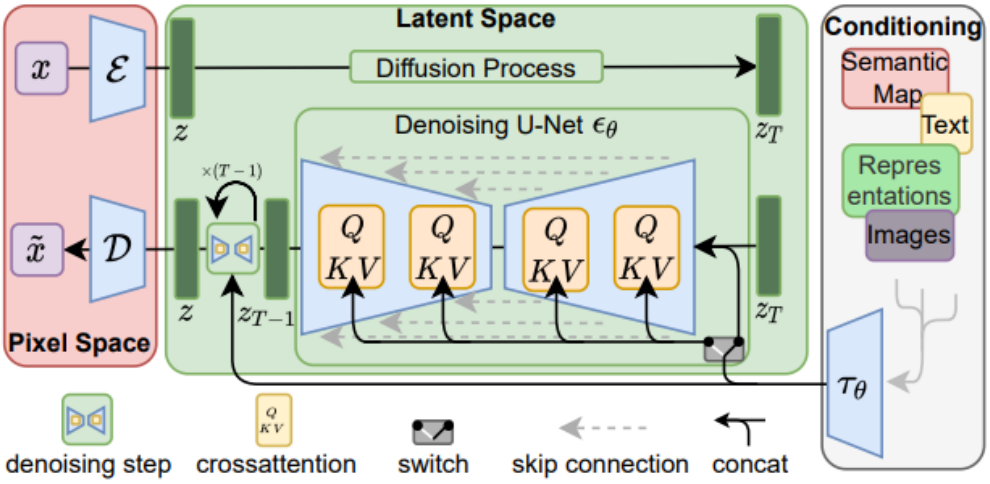
    (2) Semantic Compression

    The actual generative model learns the semantic and conceptual composition of the data.[14]

    Following this practice, the paper separated the training process into 2 phases. First, they trained the autoencoder which provides a lower-dimensional representational space which is perceptually equivalent to the data space, and then trained the Diffusion Models in the learned latent space.[15]

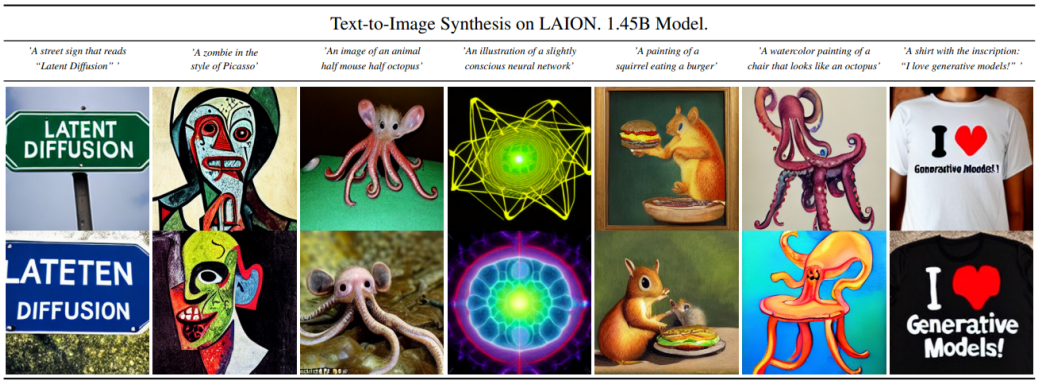
[12] High-Resolution Image Synthesis with Latent Diffusion Models (CVPR 2022)
[13] High-Resolution Image Synthesis with Latent Diffusion Models (CVPR 2022)
[14] High-Resolution Image Synthesis with Latent Diffusion Models (CVPR 2022)
[15] High-Resolution Image Synthesis with Latent Diffusion Models (CVPR 2022)

- Model Architecture



16

- Experiment Results
  - Text-to-image



17

- **Hierarchical Text-Conditional Image Generation with CLIP Latents**

  This paper proposed a two-stage model: a prior that generates a CLIP (Contrastive Language-Image Pre-Training) image embedding given a text caption, and a decoder (a diffusion model) that generates an image conditioned on the image embedding. The main contribution of this paper is that image representations improve image diversity with minimal loss in photorealism and caption similarity. Moreover, the joint embedding space of CLIP enables language-guided image manipulations in a zero-shot fashion.[18]

  - Concept

    In this work, they combined CLIP and diffusion models for the problem of text-conditional image generation.

---

[16] High-Resolution Image Synthesis with Latent Diffusion Models (CVPR 2022)
[17] High-Resolution Image Synthesis with Latent Diffusion Models (CVPR 2022)
[18] Hierarchical Text-Conditional Image Generation with CLIP Latents (CVPR 2022)

- Contrastive Language-Image Pre-Training (CLIP)

  CLIP embeddings have a number of desirable properties: they are robust to image distribution shift, have impressive zero-shot capabilities, and have been fine-tuned to achieve state-of-the-art results on a wide variety of vision and language tasks.[19]

- Model

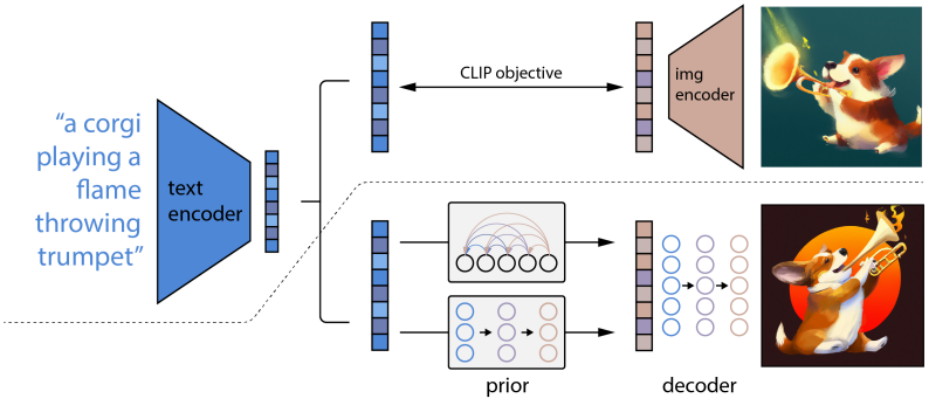


Figure 2: A high-level overview of unCLIP. Above the dotted line, we depict the CLIP training process, through which we learn a joint representation space for text and images. Below the dotted line, we depict our text-to-image generation process: a CLIP text embedding is first fed to an autoregressive or diffusion prior to produce an image embedding, and then this embedding is used to condition a diffusion decoder which produces a final image. Note that the CLIP model is frozen during training of the prior and decoder.

[20]

- Experiment Results
  - Text Diffs

a photo of a cat → an anime drawing of a super saiyan cat, artstation

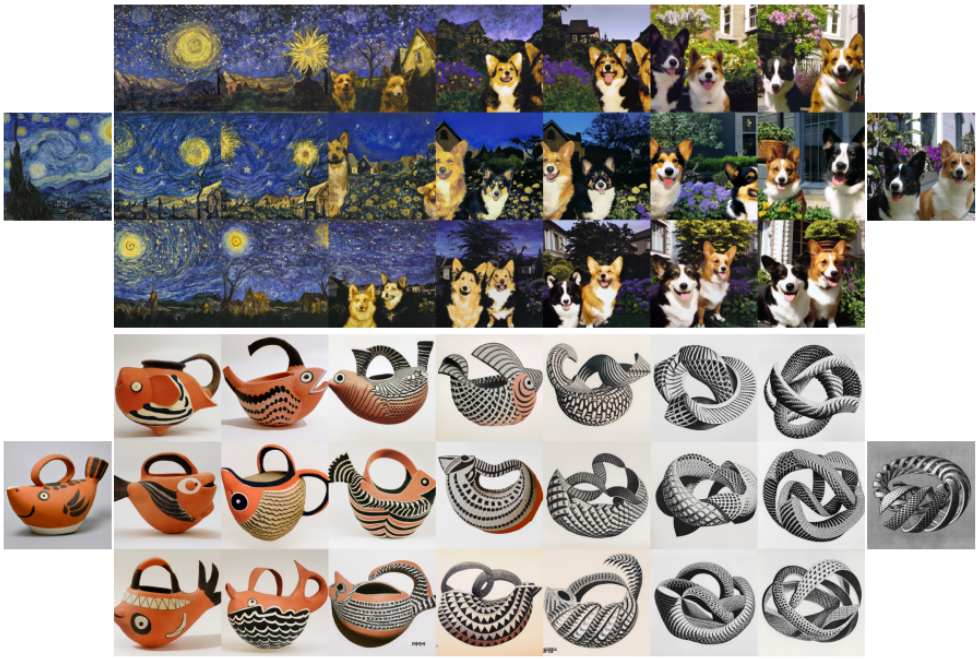
a photo of a victorian house → a photo of a modern house

a photo of an adult lion → a photo of lion cub

a photo of a landscape in winter → a photo of a landscape in fall

[19] Hierarchical Text-Conditional Image Generation with CLIP Latents (CVPR 2022)
[20] Hierarchical Text-Conditional Image Generation with CLIP Latents (CVPR 2022)

- Variations (between images)



- **Conclusion**

  In this report, we look over some of the famous papers related to image generation which are based on diffusion models. Recently, AI digital art has sparked lots of discussion and attracted everyone's attention, and this is not only among the computer science community but also artists and further normal people are not in the related area, since there are many commercial or non-commercial artwork generation AI services based on diffusion models have been popping up. The services in the market now are able to generate amazing artwork, but there are still some considerations. First, from my point of view, semantic understanding still has space to improve. The services are not friendly toward everyone (the prompt mechanism). The models are able to create wonderful works, but it is still difficult to "work with people". Second, the copyright issues will become a hot potato. How to protect the artists but not over-limited the research might be the thing that everyone should take into consideration carefully.