智能计算系统

实验第八章

大模型实验

张欣

中国科学院计算技术研究所

目录

实验目的

熟悉潜在扩散模型的算法原理，能够在DLP平台上使用Python语言基于Stable diffusion实现图生图以及文生图，具体包括：

（1）介绍基于潜在扩散模型的Stable diffusion的基本原理、计算步骤，及其在图像领域生成中的应用；

（2）深入了解潜在扩散模型的关键概念及底层逻辑，包括模型加载、模型推理的关键步骤等；

（3）了解Stable diffusion图像生成的基本流程，在DLP平台上实现图生图、文生图、图像修复功能，为后续针对不同需求使用不同的图像生成大模型奠定基础。

实验工作量：约 70 行代码，5 小时。

Rombach R , Blattmann A , Lorenz D ,et al.High-Resolution Image Synthesis with Latent Diffusion Models[J]. 2021.DOI:10.48550/arXiv.2112.10752.

背景介绍

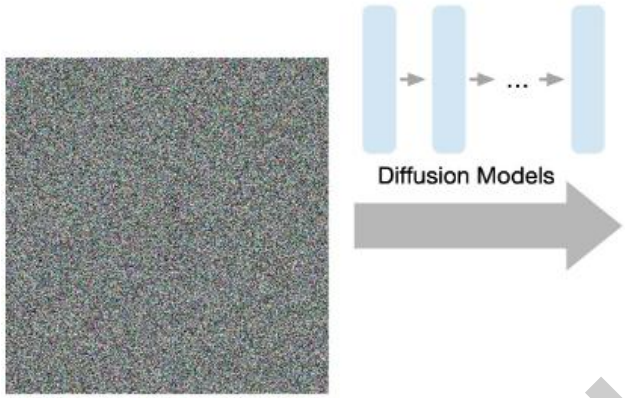
Stable Diffusion是StabilityAI于2022年发布的基于潜在扩散模型的生成式大模型。

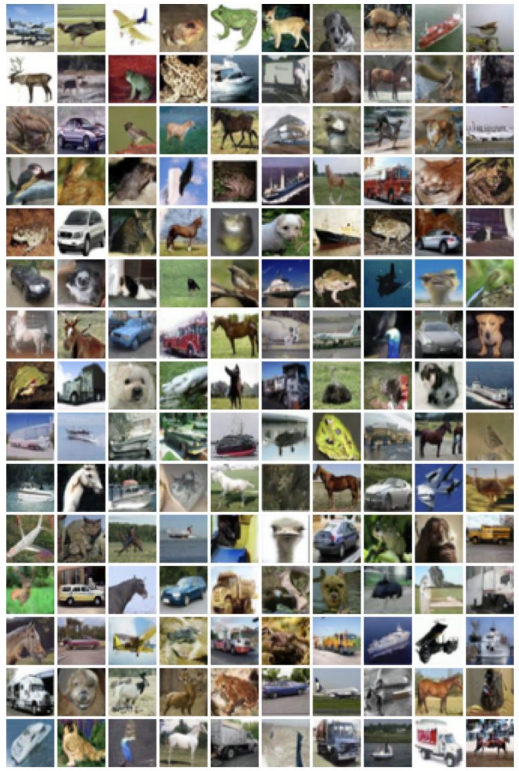
其通过对海量图像数据进行预训练的方式，让模型能够根据任意输入(文本/图像)生成高质量的图像，主要应用包含文本生成图像、图像生成图像以及图像补全等。与传统基于GAN的图像生成模型相比，它能根据需求更高效地生成高质量的图像，是 AI 图像生成领域的里程碑。

根据预训练版本以及应用任务的不同，目前Stable Diffusion的模型包括Stable Diffusion V1.5、Stable Diffusion V2.1、Stable Diffusion Inpainting、Stable Diffusion x4-Upscaler等多种模型。

扩散模型

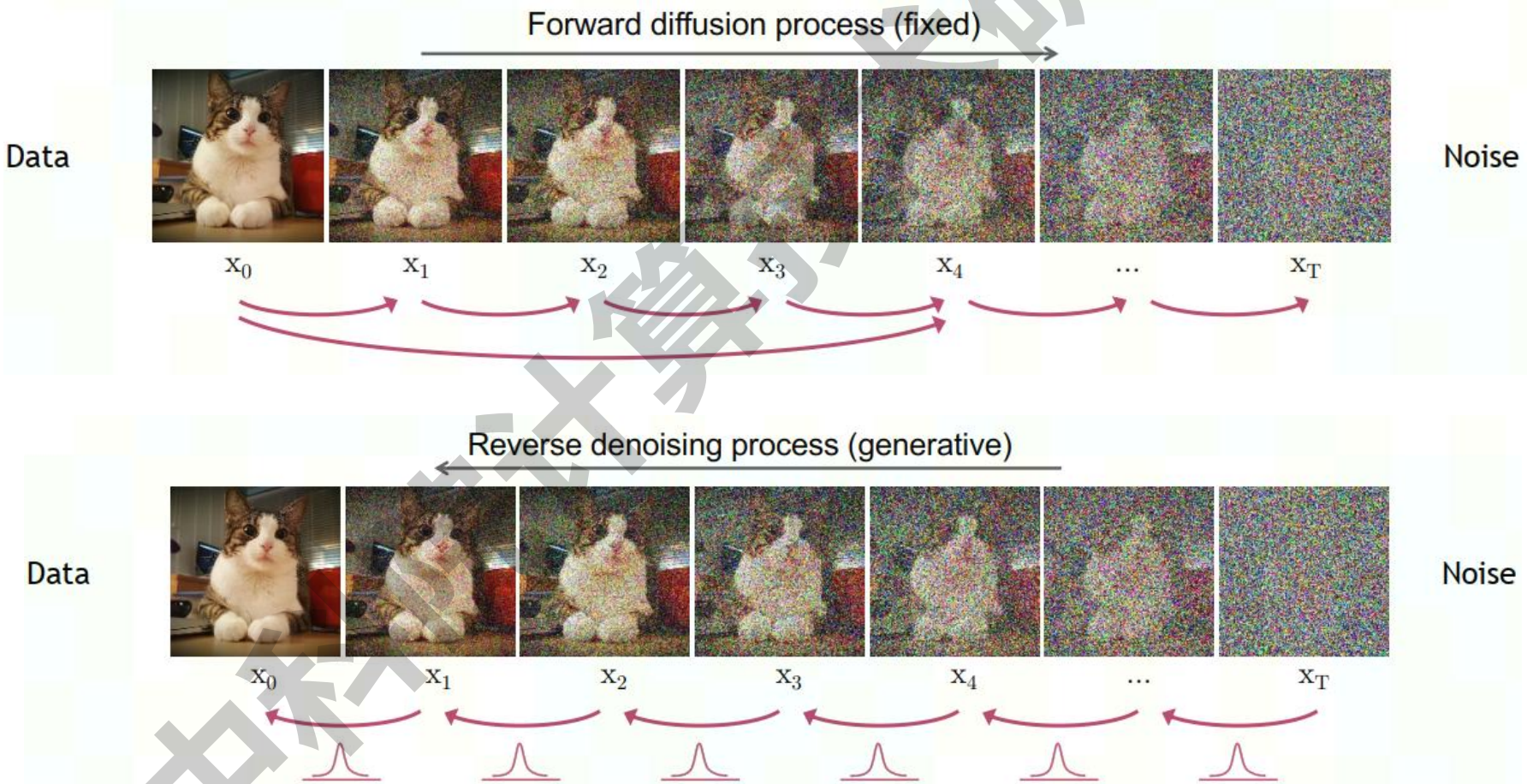
- J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models[C]. NeurIPS, 2020.
- DDPM在图像合成方面击败了GAN





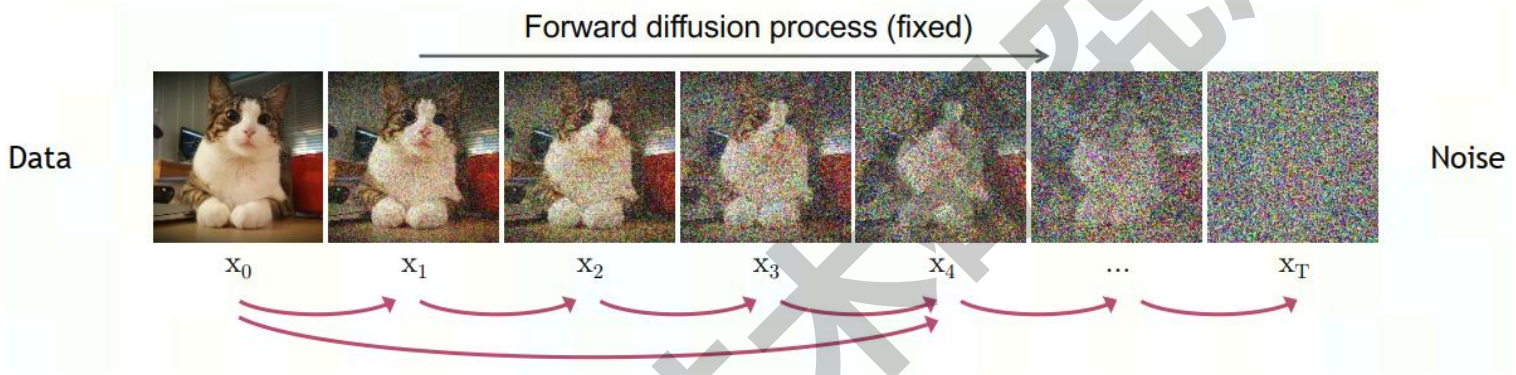
扩散模型

- **基本原理**
  - 正向过程和反向过程



扩散模型

- 正向过程



- 给定输入图像x_0，不断在前一时刻的图像上添加高斯噪声，将输入数据分布转换为近似标准正态分布（即高斯分布）
- 马尔科夫链，在第t时间步加噪时， 满足以下条件概率

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}) \qquad \beta_t \in [0,1]$$

- 推导出

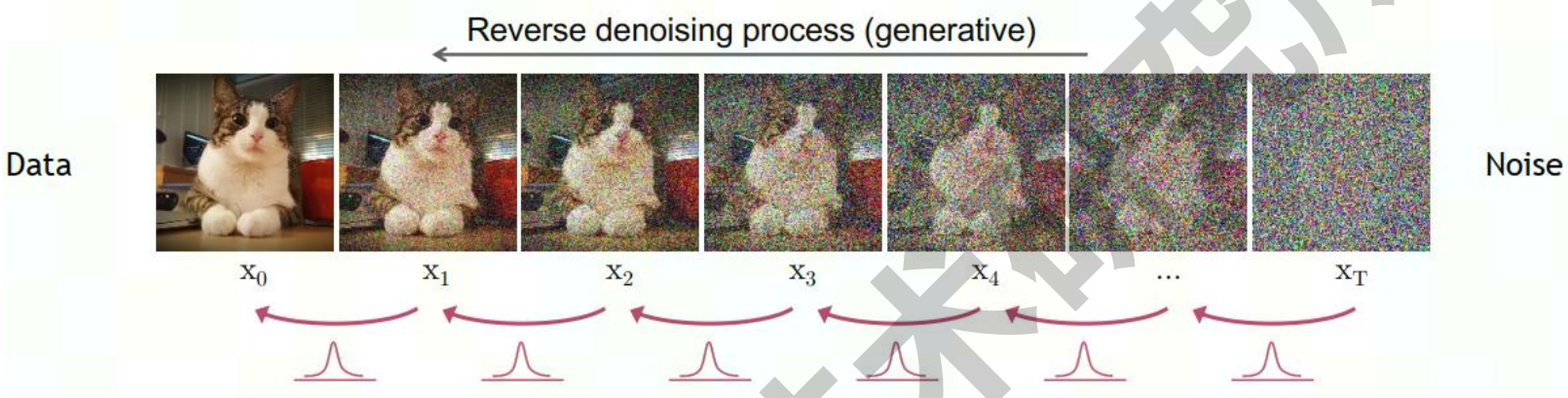
$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\boldsymbol{I}) \qquad \alpha_t = 1-\beta_t, \bar{\alpha}_t = \prod_{i=1}^{T} \alpha_i$$

- 采样获得样本

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{(1-\bar{\alpha}_t)}\epsilon, \quad \epsilon \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$$

扩散模型

- 反向过程



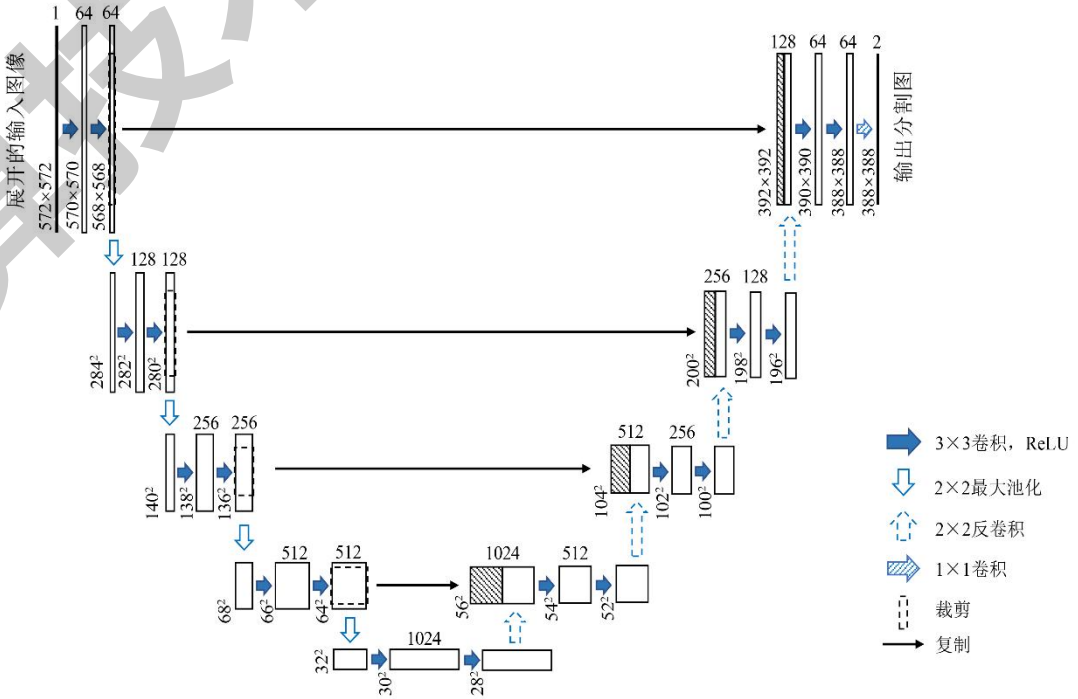
- 对于高斯噪声图像x_T，通过逐步去噪，最终还原出与输入数据分布接近的数据分布，即重构出图像x_0
- 马尔可夫链，根据当前时间步和图像数据，对噪声进行采样，进而得到前一时间步的图像数据
- 初始采样 $p(x_T) = \mathcal{N}(x_T; 0, I)$
- 其他时刻的迭代采样 $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I)$

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right) + \sigma_t z$$

扩散模型

- 噪声预测网络
  - 扩散模型的关键在于构建一个噪声预测网络，能在反向过程的每一步预测合理的去噪参数
  - 需要进行像素级的预测，因此采用常用于图像分割任务的U-Net

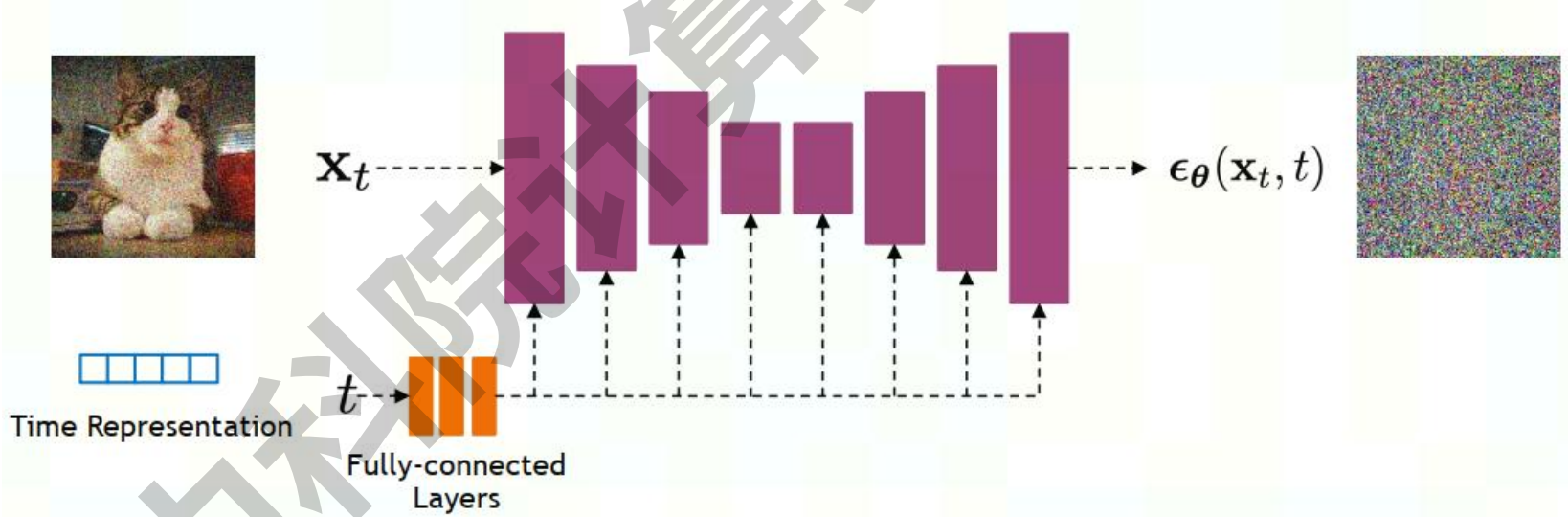
- U-Net
  - 完全对称的全卷积网络，由左侧的编码器（encoder）和右侧的解码器（decoder）组成
  - 编码器通过一系列的下采样，去除输入图像中的冗余信息，并压缩转换到小尺寸的图像特征
  - 解码器通过对称的上采样（反卷积），将图像特征还原到原图尺寸



Ronneberger O, Fischer P, and Brox T. U-Net: Convolutional networks for biomedical image segmentation. MICCAI, 2015

扩散模型

- 噪声预测网络
  - 扩散模型的关键在于构建一个噪声预测网络，能在反向过程的每一步预测合理的去噪参数
  - 需要进行像素级的预测，因此采用常用于图像分割任务的U-Net
  - DDPM改进了原始U-Net
    - 将每个尺度的卷积块替换为残差块，并增加自注意力层来增强关系感知能力
    - 将当前时间 t 编码为向量，作为网络的条件输入，来为不同的时间步预测噪声



扩散模型

- 训练过程
  - 损失函数

$$L_{\text{simple}} = \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t \sim \mathcal{U}(1,T)} \left[||\epsilon - \epsilon_\theta(\underbrace{\sqrt{\bar{\alpha}_t}\, \mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon}_{\mathbf{x}_t}, t)||^2 \right]$$

算法 3.1 训练过程[99]

1: **repeat**
2: $\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0)$
3: $t \sim \mathcal{U}(1, T)$
4: $\epsilon \sim \mathcal{N}(0, \boldsymbol{I})$
5: 进行梯度下降 $\nabla_\theta ||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)||^2$
6: **until** 收敛

- 推理过程

算法 3.2 采样过程[99]

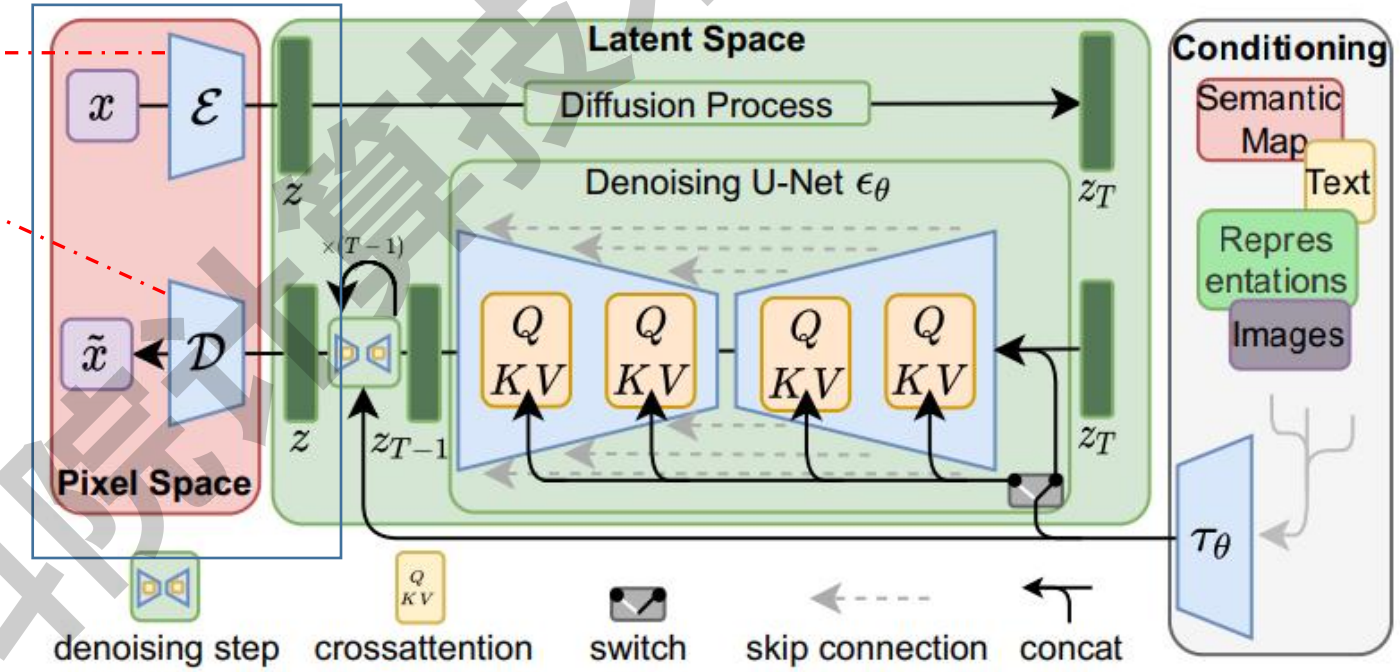
1: $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
2: **for** $t = T, ..., 1$ **do**
3: $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ 如果 $t > 1$， 否则 $\boldsymbol{z} = \mathbf{0}$
4: $\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\boldsymbol{x}_t,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

Stable Diffusion

- R Rombach, A Blattmann, D Lorenz, P Esser, B Ommer, High-Resolution Image Synthesis with Latent Diffusion Models, CVPR'2022.
- 训练和推理的扩散和逆扩散过程都在隐空间中进行，大大减小了显存使用量和计算量

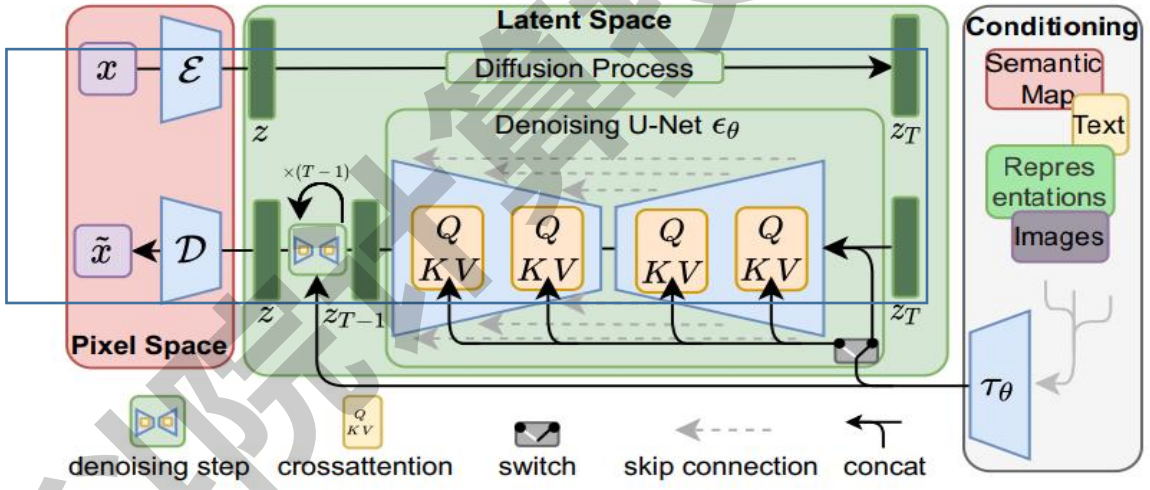
VAE的编码器和解码器

主要用于将原始图片转换到隐空间中
比如：512x512x3 转换后变成 64x64x4



Stable Diffusion

- 训练
  - 损失函数 $L_{LDM} := \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon-\epsilon_\theta(z_t,t,\tau_\theta(y))\|_2^2\right]$
  - 1、给一张图像x，经过VAE编码器得到隐空间表示z_0
  - 2、经过随机时间步t 的加噪ϵ，得到z_t
  - 3、将z_t输入到U-Net模型，得到预测噪声$\hat{\epsilon}$，训练模型使预测噪声$\hat{\epsilon}$更接近真实噪声ϵ



- 推理
  - 给一个随机噪声z_T，“减去”U-Net模型预测的噪声$\hat{\epsilon}$，不断去噪得到不带噪声的z_0，再经过VAE解码器得到真实图像x

实验内容

本实验使用Stable Diffusion模型实现**图生图**、**图生文**以及**图像修复功能**。在工程实现中，首先依据推理阶段的底层逻辑，实现潜在空间反向过程的特征采样；然后根据潜在扩散模型的基本原理，补全推理阶段的关键步骤；最后完成包含模型加载等一系列的基本适配工作，实现**由图像生成图像**、**由文字生成图像**、**图像修复**等功能。由于本实验只涉及Stable Diffusion模型的推理过程，因此本实验并不包括潜在扩散模型噪声预测网络的训练部分。另外，本实验暂不涉及文字生成图像推理过程所采用的采样算法，有兴趣的可进行额外研究。

实验步骤

运行实验

运行图生图推理程序

运行文生图推理程序

运行图像修复推理程序

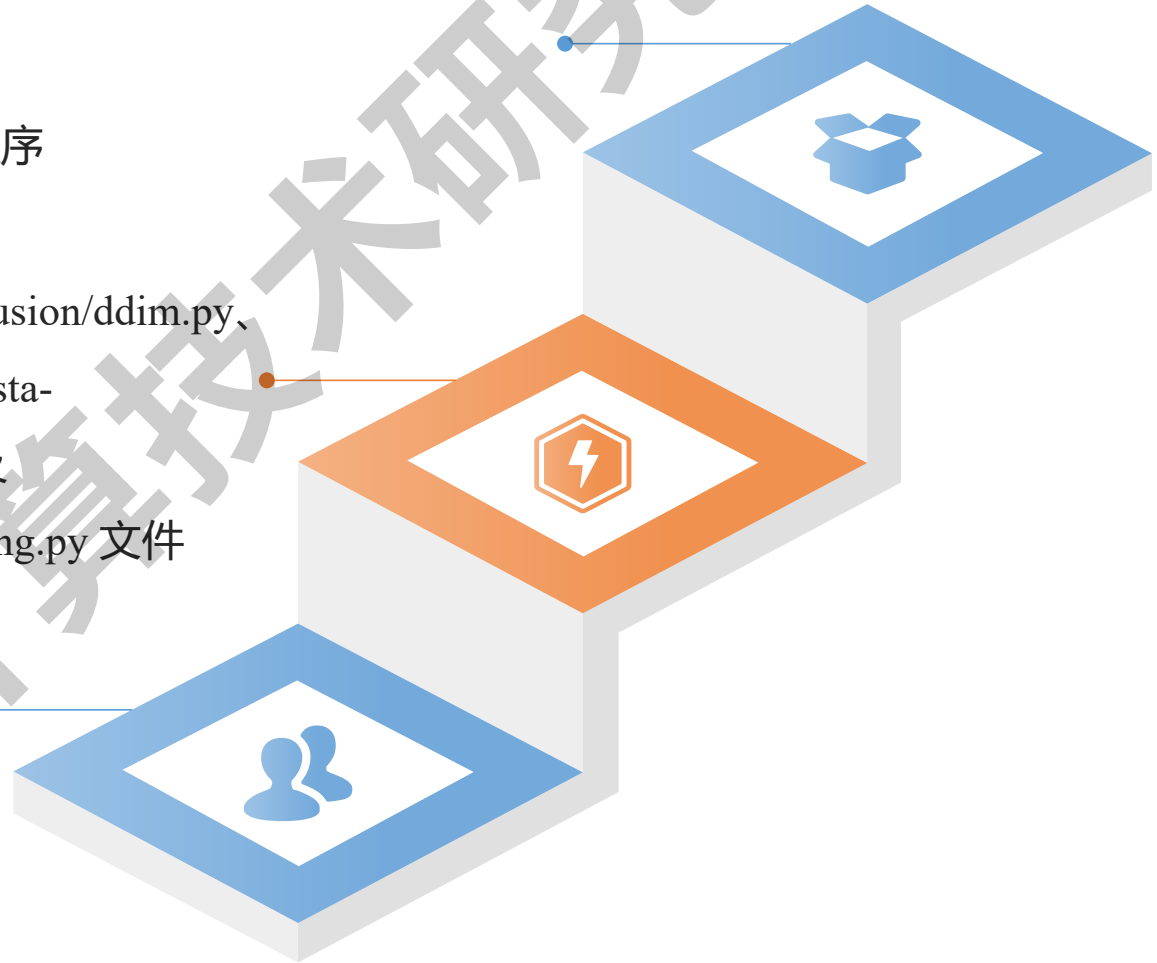
代码实现

补全 stable_diffusion/ldm/models/diffusion/ddim.py、stable_diffusion/scripts/img2img.py、stable_diffusion/scripts/txt2img.py、以及stable_diffusion/scripts/gradio/inpainting.py 文件

环境安装

安装accelerate安装包

安装transformer安装包



实验评估

- 60 分标准：能够正确实现 make schedule 运算以及 stochastic encode 模块，完成扩散模型推理阶段的基础运算。
- 70 分标准：在 60 分标准基础上，能够正确实现 p sample ddim 模块、decode 模块以及主函数模块，完成推理阶段反向过程的所有函数。
- 80 分标准：在 70 分标准基础上，能够正确实现图像生成图像模块以及文字生成图像模块。
- 90 分标准：在 80 分标准基础上，能够正确实现文字生成图像模块。
- 100 分标准：在 90 分标准基础上，能够正确实现图像修复推理模块。

敬请指正！

课程官网：http://novel.ict.ac.cn/aics

MOOC网址：

https://space.bilibili.com/494117284/video