一、前言

背景

DIT 是官方的 Pytorch 版 Diffusion Transformer 模型，这是两年前的一个项目，来自 facebook。DiT 方法是在 2022 年底被提出的，它的主要目标是优化图像生成，并未太多涉及视频领域。之后，也有一些针对 D+T 在视频领域的研究，比如：T2V，Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation。效果虽然不如 Sora，但方法可以借鉴。因为 OpenAI 并没有公开技术细节，有些具体方法我们只好脑补一下。具体的核心模块使用的 DIT。

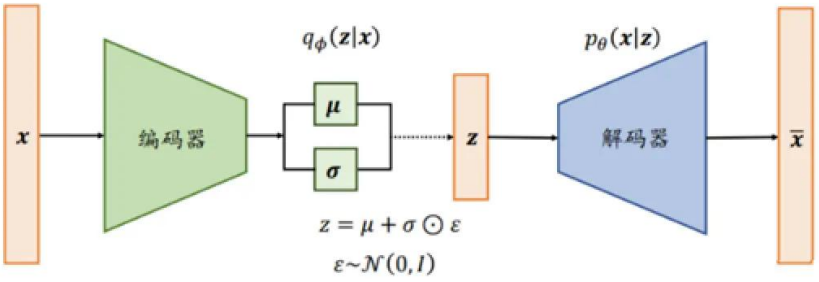
DIT

paper: Scalable Diffusion Models with Transformers
链接: http://arxiv.org/abs/2212.09748v2
代码: https://github.com/facebookresearch/DiT
DiT 代码不算多，共 1415 行 Python 代码模型结构在 models.py 文件中；diffusion 部分修改自 openai 的 diffusion 代码；其它代码还包含下载，训练模型等。

二、核心模块

2.1 变分自编码器 VAE

VAE 变分自编码（Variational Autoencoder）是一种生成模型，它结合了自编码器和概率图模型的思想。其目标是：解决对复杂性高，且量大的数据难以拟合的问题。



2.1.1 自编码器

自编码器通常由编码器和解码器两部分组成，其中编码器将原始数据映射到低维表示，解码器则将低维表示映射回原始数据空间。即：原始数据为 x，将其输入编码器降维后，变成数据 z，再经过编码器还原成数据 x’。它常用于如：图像去噪，修复图片，生成高分辨率图片等。

2.1.2 变分自编码器

变分自编码器在中间加了一层逻辑，它假设中间过程的数据 z 每个维度都是正态分布的，可以使用：均值 μ 和 方差 σ 表示。由此，就变成了变分自编码器：训练编码器和解码器网络，可将图片 x 分布压缩后再拆分成多个高斯分布的叠加，如上图所示。

因此可以认为：变分自编码器从图像的像素层面提取出了更多性质。

2.2 扩散模型 Diffusion

扩散模型由加噪 q 和去噪 p 两部分组成，如图 -2 所示，先从右往左看下边部分加噪 q，x0 是原始图像，经过 T 步逐渐加噪变为纯高斯噪声 XT，其中每一步的图像 xt 根据上一步的 xt-1 通过加少量高斯噪声得到；再看上边部分去噪 pθ，它是 q 的逆过程，每一步通过 xt 得到 xt-1，最终还原图像 x0，p 由神经网络实现，θ 是神经网络参数，最后得到的深度学习模型就是可用噪声生成真实图像的网络。

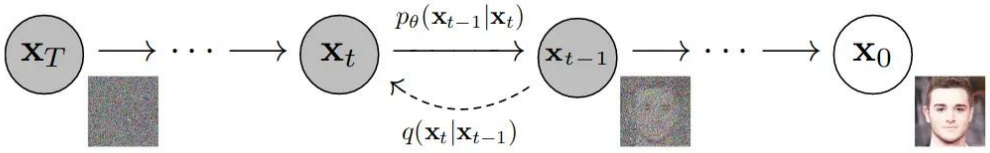
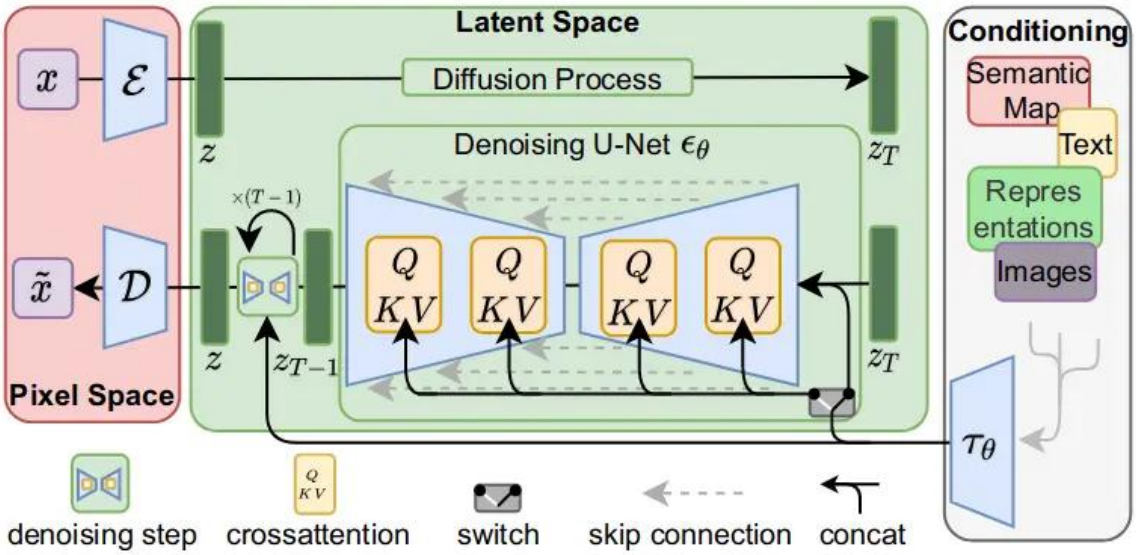


Figure 2: The directed graphical model considered in this work.

由此可见，扩散模型原理与 VAE 很类似，通过加高斯噪声和去噪来重建图像，不过它是分多步完成的。其原理也是训练模型参数学习图像内部的关系。

2.3 潜空间扩散模型 LDM

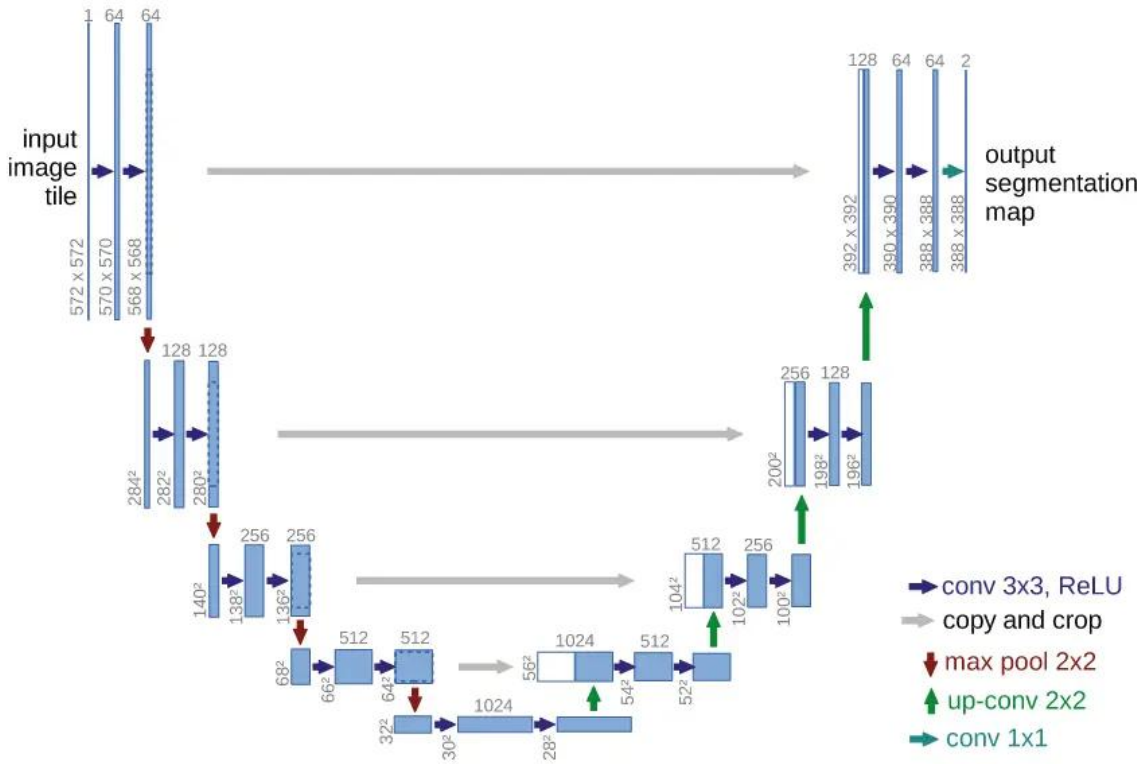


Latent Diffusion Models（LDMs）是基于潜空间的扩散模型。之前的扩散模型运算都在像素层面，优化通常会消耗数百个 GPU 天，且评估和推理成本也很高。LDMs 大量自编码器的运算基于潜空间数据，降低了计算复杂度，从而大幅节省了算力。其应用场景包含有条件（根据文本或图像生成图像）和无条件（去噪/着色/根据涂鸦合成）的图像生成。

主逻辑分成三部分，第一部分是像素空间与潜空间之间的转换，即感知图像压缩（粉色）；第二部分是在潜空间操作的扩散模型（绿色）；第三部分是用文本描述或其它图片作为条件，控制图像生成（白色）。

2.4 U 型网络结构 U-Net

LDM 的绿色部分包含 U-Net，它最早被应用于医学影像领域，用于识别病灶。其原理如下图所示：

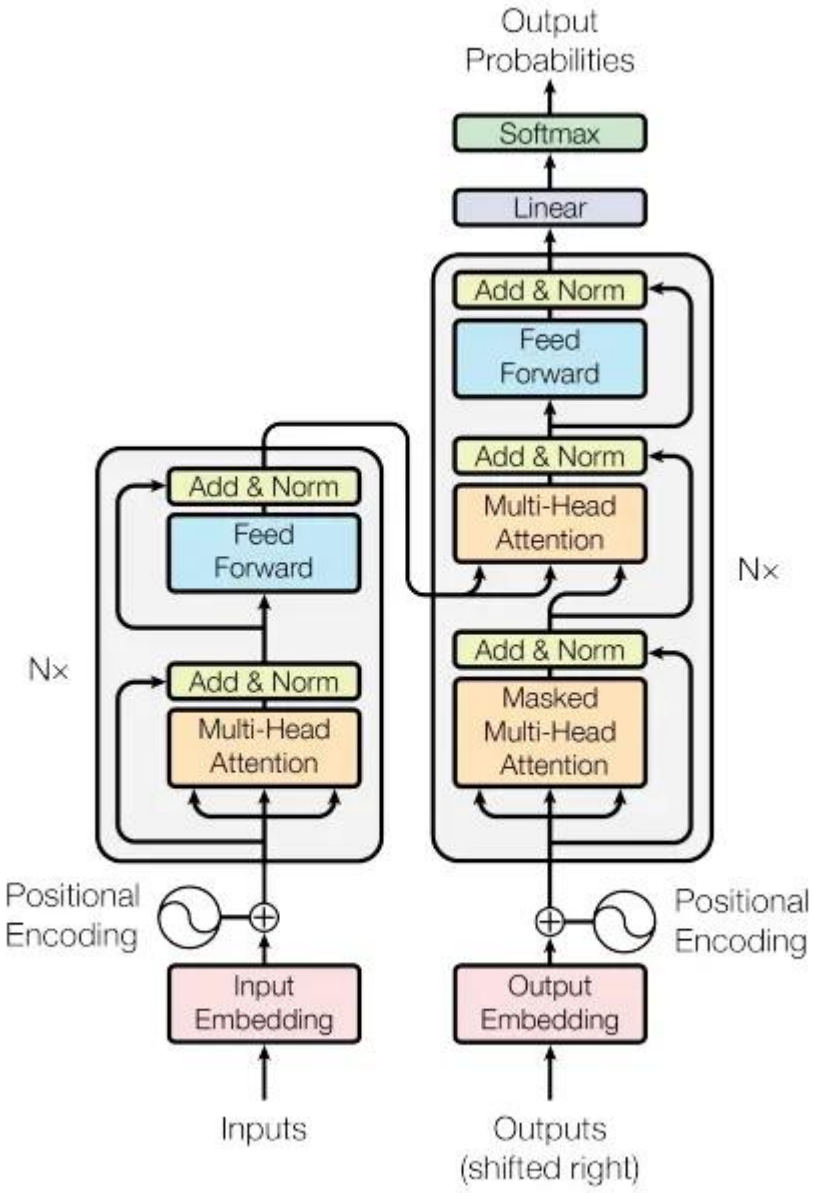


它用 U 型堆叠卷积网络对图像进行压缩再恢复，实验也证明了 UNet 架构的归纳偏差，对具有空间结构（上下左右的相关性）的数据特别有效。

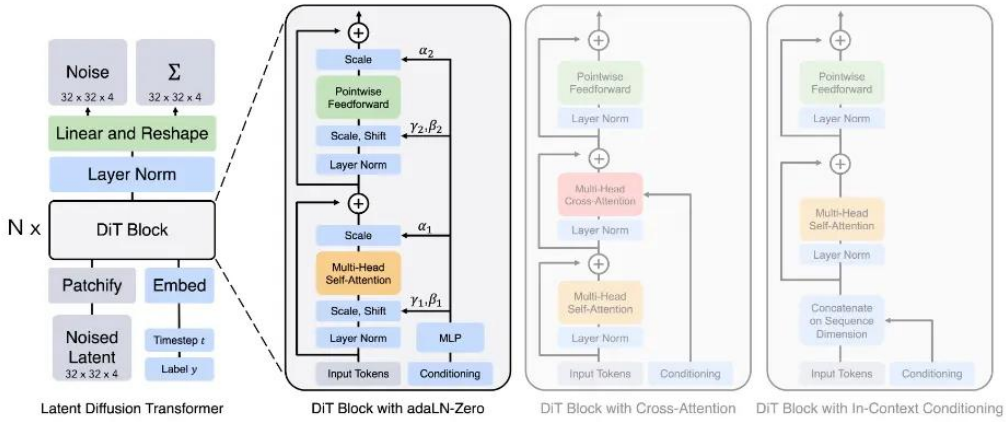
也就是说之前的扩展模型内部都是基于 U-Net 网络构建的。

2.5 Transformer 框架

Transformer 是 Google 团队在 2017 年提出的自然语言处理（NLP）框架，其论文名为：Attention is all you need。可见注意力机制在其中的关键性作用。它使用注意力 Attention 算法计算序列中各个元素之间的关系。
Transformer 近年被广泛应用于自然语言，图像，音频等领域；可以看到，LDM 图中的 QKV 部分使用的也是注意力机制。
如图所示：它由一个编码器和一个解码器组成。它善于处理序列数据，避免了递归和卷积计算，相比之前模型，训练速度快，模型效果好。



2.6 结合 Diffusion 与 Transformer DiT



DIT 的核心是提出用基于潜空间 patch 的 Transformer 模型替换之前的 U-Net 模型。

把图片切成小块 patch（可将其视为 LLM 中的 token，简单地讲就是语言模型中的词），然后输入 Transformer 结构，使模型学习每帧图像内部小块之间的空间关系。

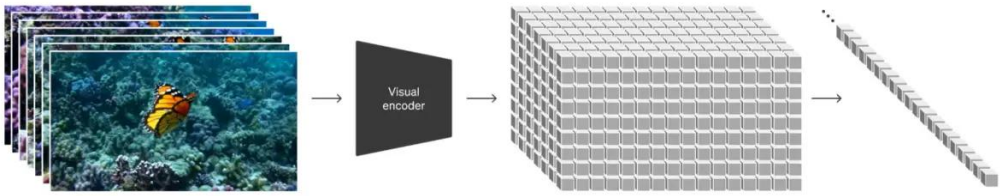
使用 DDIM 采样器来生成中间图像，如狗与网球的混合体，通过与之前模型 BigGAN 的对比，可以看到 DIT 似乎学到了更多的关于世界的一般性知识。

DiT “Dog-Ball”　　BigGAN “Dog-Ball”

从论文的实验部分可以看到：作者使用 ImageNet 来训练模型，没有提及视频数据训练。可以说两年前的 DIT 模型只对单图（即空间）进行了优化，而并为训练视频来捕捉时间变化的规律。

DIT 被提出时，只是用 Transformer 替换了 U-Net 主干网络，用途是图像生成，证明了 D+T 的可用性：即使不使用 U-Net 的归纳偏差，Transformer 在空间上也能表现很好。而在效果上并没有突飞猛进。所以很长一段时间主流图像生成仍延用之前的方法，去年 Stable Diffusion 的最新模型 SDXL 内部仍使用 U-Net 架构。那时候大家可能都觉得在图像生成领域，除了更好地与语言对齐，短时间在架构上不太需要大改。

2.7 Sora



Sora 加入了对时序的建模，也就是图中纵伸的部分。把图像切分成小块，一次可以结合更长的上下文；除了一帧画面内部上下左右的结构关系，还能学习帧与帧之间的关系，也就是运动的方向和速度，以及物理学中不同物体的变化规律。并且使用了大量视频数据训练模型。

使用 patch 的潜空间的表示，不仅学习表面的关系，还学习内在规律，进行性质的分解（简单地讲，除了光影以外，还能识别物体是什么）。几相结合：内在规律，空间规律，时间规律，再加上 OPENAI 系的图像模型（如：DALL-E3）相对其它模型能更好地理解自然语言，于是成就了 Sora 在视频领域的质变。

三、一些思考

Sora 展示的不只是生成视频对当前领域的冲击，可能更多的是一种基于时空建模的思路：结合潜空间，空间，时间，学习内部规律，什么该静止，什么该变化；再加入对自然的理解，通过自然语言引入现有知识。
所以说 Sora 不仅展示了视频领域的突破，还展示了目前更加完善的工具链，和更多的可能性。
当然它也不仅是 OpenAI 一家的成果，也是这些年技术累积的结果，也算是一种涌现，一种必然趋势吧。