Diffusion Model

常见图像生成模型：

1.VAE

变分自编码器(Variational Autoencoder)通过编码器将图像压缩到潜在空间,再通过解码器重构图像,学习数据的概率分布。

2.GAN (生成对抗网络)

由Ian Goodfellow提出,通过生成器和判别器的对抗训练生成逼真图像。生成器试图创造假图像,而判别器试图区分真假,两者相互博弈不断提升。

3.Diffusion Model (扩散模型)

通过逐步向数据添加噪声,再学习逆向去噪过程来生成高质量图像。代表模型包括DDPM、Stable Diffusion等,在图像生成质量和多样性方面表现出色。

4.flow-based模型

基于可逆神经网络的normalizing flow模型,通过一系列可逆变换将简单分布映射到复杂数据分布。优点是可以精确计算似然,训练稳定,但模型架构设计较为复杂。

一、核心思想（最重要的一段）

一句话定义：

> Diffusion Model 是一种通过“逐步加噪—再逐步去噪”来学习数据分布的生成模型。

本质：

- 训练阶段：学习如何从带噪声的数据中预测噪声
- 推理阶段：从纯噪声出发，一步步去噪生成样本

二、整体框架

Diffusion Model 分为两个过程：

1. Forward Process（前向扩散）

- 已知原始数据 x0x_0x0

- 不断往里加高斯噪声
- 最终变成近似纯噪声 xTx_TxT

---

2. Reverse Process（逆向扩散）

- 训练一个神经网络
- 学习从 xtx_txt 恢复 xt−1x_{t-1}xt−1
- 最终从噪声重建出清晰样本

三、前向扩散的数学建模

3.1 单步加噪

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t}\epsilon, \quad \epsilon \sim N(0, I)$$

含义：

- $alpha_t in (0,1)$：噪声控制系数
- 越往后时间步：
  - 信号权重 $\sqrt{\alpha_t}$ 越来越小
  - 噪声权重 $\sqrt{1-\alpha_t}$ 越来越大

3.2 累积公式（关键！）

通过递推可以得到：

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$$

其中：

$$\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$$

意义：

- 可以**一步到位**从原图 x_0 直接采样任意时间步 t 的噪声图 x_t
- 极大简化训练过程，无需逐步递推

四、训练目标

Diffusion 的核心思想：

> 不直接预测图像，而是预测噪声

4.1 神经网络输入输出

网络： $\hat{\epsilon}_\theta(x_t, t)$

输入：

- 带噪声图像 x_t
- 时间步 t

输出：

- 对真实噪声 ϵ 的预测

4.2 损失函数

$$L(\theta) = \mathbb{E}_{x_0,t,\epsilon}\left[\|\epsilon - \hat{\epsilon}_\theta(x_t, t)\|^2\right]$$

特点：

- 极其简单：均方误差（MSE）
- 训练稳定
- 类似去噪自编码器的思想

五、逆向扩散（生成阶段）

目标：

> 学习条件分布 $p(xt-1 \mid xt)p(x_{t-1}|x_t)p(xt-1 \mid xt)$

5.1 采样公式

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

其中：

- z sim N(0,I)：随机噪声
- sigma_t：控制生成的随机性程度

5.2 生成流程

1. **初始化：** 采样纯噪声 $x_T \sim N(0, I)$
2. **迭代去噪：** for t = T, T-1, ..., 1:
   - 用神经网络预测噪声：$hatepsilon_theta(x_t, t)$
   - 根据采样公式计算 x_{t-1}
3. **输出：** 得到最终生成样本 x_0

六、为什么 Diffusion 比 GAN 稳?

维度	GAN	Diffusion
训练目标	对抗训练	MSE 去噪
稳定性	很差	很稳定
模式坍塌	常见	几乎没有
收敛难度	高	低

七、网络结构

7.1 常用结构

- U-Net（最常见）
- Transformer
- 混合结构

7.2 时间步编码

需要将 t 输入网络：

常用：

emb(t)=[sin,cos] 的位置编码$emb(t) = [\sin, \cos]$ 的位置编码

emb(t)=[sin,cos] 的位置编码

八、条件生成（Conditional Diffusion）

为了实现：

- 文生图
- 图生图
- 类别条件生成

网络变为：

$\epsilon^{\theta}(xt, t, c)\hat{\epsilon}_{\theta}(x_t, t, c)$

$\epsilon^{\theta}(xt, t, c)$

其中 c是条件信息：

- 文本 embedding
- 类别标签
- 图像特征

九、Stable Diffusion 的关键改进

9.1 潜空间 Diffusion

不在像素空间做扩散，而是：

```
image → VAE Encoder → latent z
在 latent 上做 diffusion
z → VAE Decoder → image
```

优点：

- 计算量更小
- 显存需求更低

- 分辨率更高

9.2 文本条件

加入 CLIP 文本编码：

$\epsilon^{\theta}(xt, t, text)\hat{\epsilon}_{\theta}(x_t, t, text)$

$\epsilon^{\theta}(xt, t, text)$

实现：

> Text → Image

十、常见 Diffusion 变体

模型	特点
DDPM	经典扩散模型
DDIM	加速采样
Score-based	基于梯度场
Latent Diffusion	潜空间扩散
Classifier-free guidance	条件增强

十一、算法伪代码

训练：

```
for eachimage x0:
    sample t ~Uniform(1,T)
    sample noise ε ~N(0,I)
    xt =sqrt(α_bar_t) * x0 +sqrt(1-α_bar_t) * ε
    loss =MSE(ε,model(xt, t))
    update θ
```

生成：

```
x_T ~N(0,I)
for t = T ...1:
    predict ε =model(x_t, t)
    compute x_{t-1}
return x_0
```

十二、与其他生成模型对比

模型	优点	缺点
GAN	快	不稳定
VAE	有概率	模糊
Flow	精确似然	表达力弱
Diffusion	质量最高	采样慢

十三、总结

记住这三句话：

1. **Diffusion = 加噪声 + 学去噪**
2. 训练目标：预测噪声
3. 生成过程：从噪声一步步去噪

十四、最小 PyTorch 实现思路

一个最简结构：

```
for x0in dataset:
    t = random.randint(1, T)
    noise = torch.randn_like(x0)
    xt = sqrt_alpha_bar[t] * x0 + sqrt_one_minus[t] * noise
    pred = model(xt, t)
    loss = mse(pred, noise)
```