Diffusion Model 介绍

INSIS论文分享会

林彦

2022年9月26日

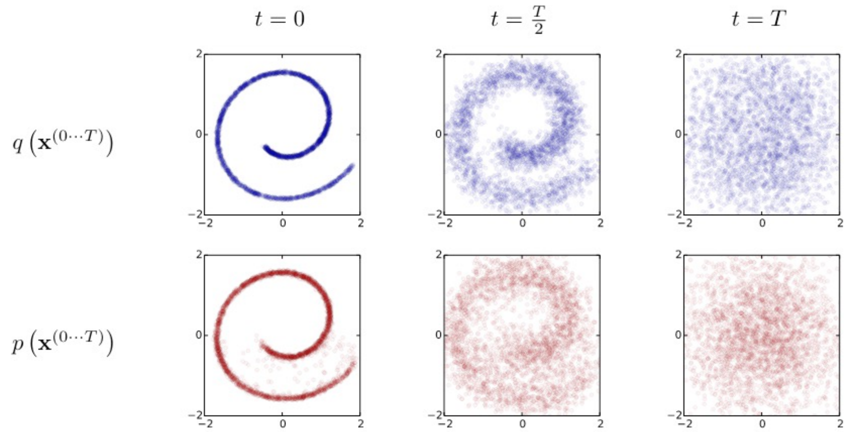
论文历史

➢ 首次提出

- Jarzynski C等 [1] 建立理论基础
- Sohl-Dickstein J等 [2] 初次提出Diffusion Generation

➢ Diffusion Model的优化与普及

- 首次应用于图像生成：DDPM模型[3]
- 随后被大量用于图像生成、文本生成[4]、时间序列插补[5]等领域



Sohl-Dickstein J等构建的Diffusion Generation

1. Jarzynski C. Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach[J]. Physical Review E, 1997, 56(5): 5018.
2. Sohl-Dickstein J, Weiss E, Maheswaranathan N, et al. Deep unsupervised learning using nonequilibrium thermodynamics[C]//International Conference on Machine Learning. PMLR, 2015: 2256-2265.
3. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in Neural Information Processing Systems, 2020, 33: 6840-6851.
4. Li X L, Thickstun J, Gulrajani I, et al. Diffusion-LM Improves Controllable Text Generation[J]. arXiv preprint arXiv:2205.14217, 2022.
5. Bansal A, Borgnia E, Chu H M, et al. Cold diffusion: Inverting arbitrary image transforms without noise[J]. arXiv preprint arXiv:2208.09392, 2022.

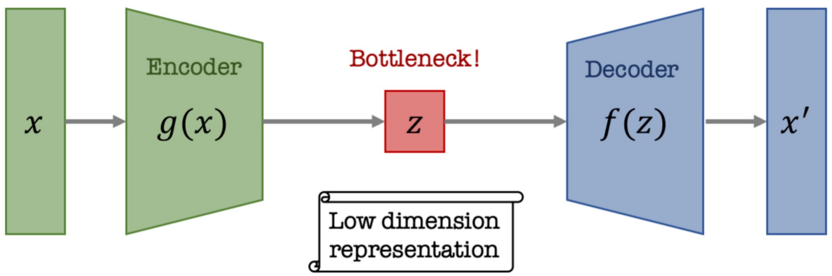
生成模型

➢ 和其他生成模型的共性与差异

- 类似Normalizing Flows, GAN和VAE [1]模型，将符合简单分布的噪音数据转换为理想中的数据分布
- 借助马尔科夫过程和神经网络，将噪音数据逐步降噪为原始数据分布

➢ Diffusion Model的生成过程

- 一个预定义的Forward diffusion process q，逐步向图片中添加噪音，直到得到一个纯粹的噪音
- 一个可学习的Reverse denoising diffusion process p，从纯粹噪音开始逐步降噪，直到得到一张无噪音的图片
- 两个过程均为马尔科夫过程，过程状态的步骤可用下标t表示



VAE的编码与解码过程

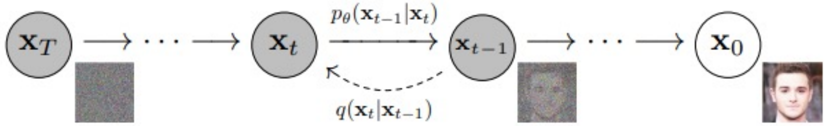


Figure 2: The directed graphical model considered in this work.

Diffusion Model的加噪与降噪过程

1. Kingma D P, Welling M. Auto-encoding variational bayes[J]. arXiv preprint arXiv:1312.6114, 2013.

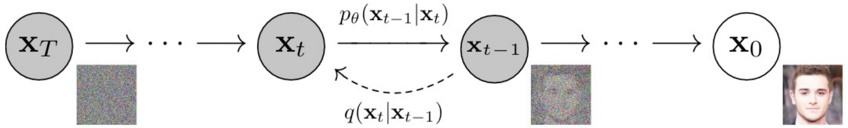
Forward Diffusion

➢ 原始图像分布

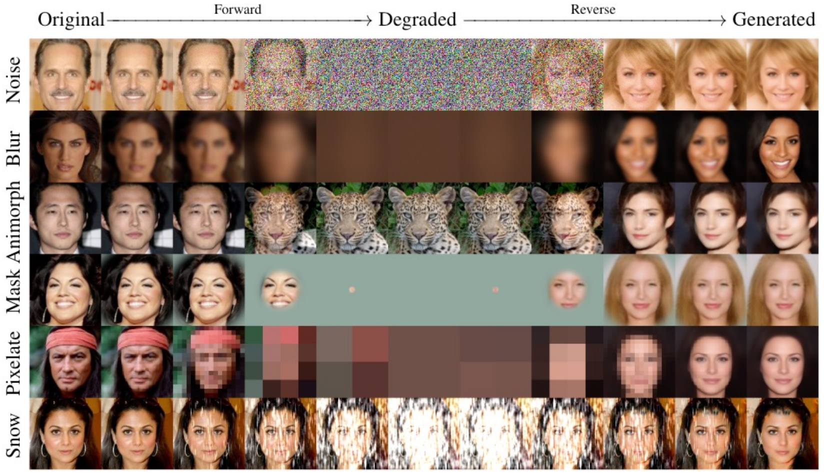
- $q(x_0)$为原始图像分布
- 数据集中的原始图像x_0看做分布的采样

➢ 前向扩散过程

- 条件概率$q(x_t|x_{t-1})$定义为Forward diffusion process
- $q(x_t|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}\right)$，相当于从一个多维高斯分布中采样
- 理想状态下，最终的x_T是纯粹的高斯噪音



Diffusion Model的加噪与降噪过程[1]



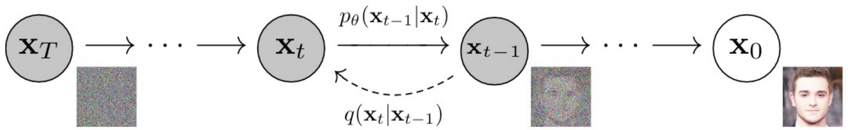
Diffusion Model的加噪与降噪步骤可视化[2]

1. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in Neural Information Processing Systems, 2020.
2. Bansal A, Borgnia E, Chu H M, et al. Cold diffusion: Inverting arbitrary image transforms without noise[J]. arXiv preprint arXiv:2208.09392, 2022.

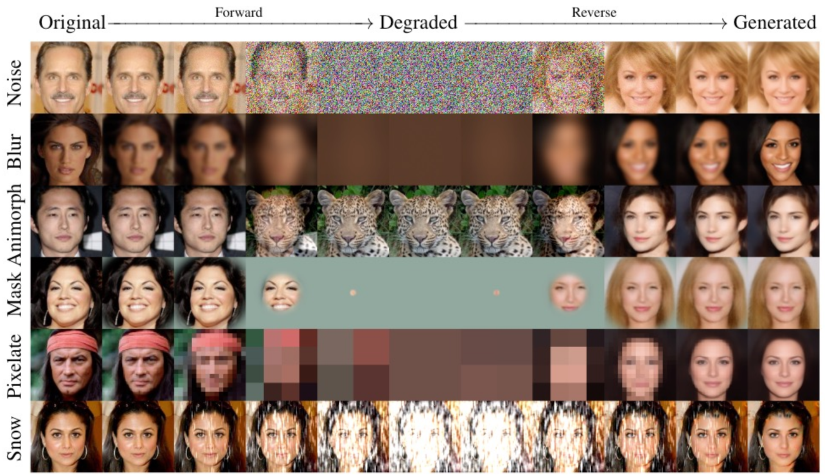
Forward Diffusion

➢ 简化的前向扩散过程

- 多次添加高斯噪音的过程**等价于在x_0上添加一次高斯噪音**，因为高斯分布之和依然是高斯分布
- $q(x_t|x_0) = \mathcal{N}\left(x_t; \sqrt{\overline{\alpha_t}}x_0, (1-\overline{\alpha_t})\mathbf{I}\right)$，其中$\alpha_t = 1 - \beta_t, \overline{\alpha_t} = \prod_{s=1}^{t}\alpha_s$
- 从x_0直接得到任意一步的噪音数据分布 [1]： $x_t = \sqrt{\overline{\alpha_t}}x_0 + \sqrt{1-\overline{\alpha_t}}\epsilon$，其中$\epsilon \sim \mathcal{N}(0, \mathbf{I})$



Diffusion Model的加噪与降噪过程



Diffusion Model的加噪与降噪步骤可视化

1. Sohl-Dickstein J, Weiss E, Maheswaranathan N, et al. Deep unsupervised learning using nonequilibrium thermodynamics[C]//International Conference on Machine Learning. PMLR, 2015: 2256-2265.

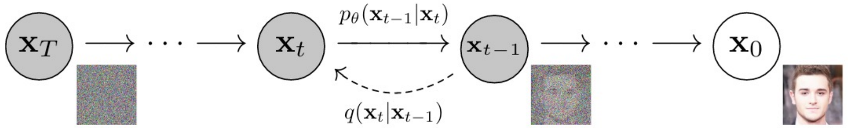
Reverse Denoising Diffusion

➢ 理论情况

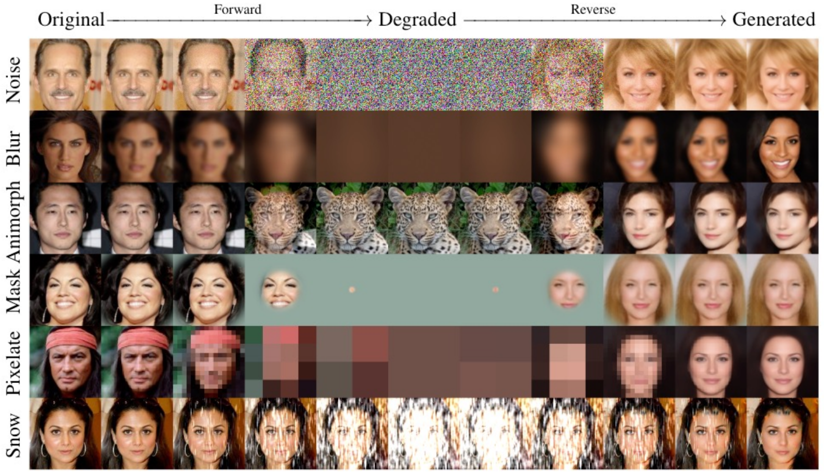
- 假如条件概率$p(x_{t-1}|x_t)$已知，则可以直接运行反向去噪过程
- 将高斯噪音x_T逐步去噪，最终得到无噪音的原始图像x_0

➢ 实际情况

- 条件概率$p(x_{t-1}|x_t)$未知，除非已知所有可能的x_t否则无法计算
- 借助可学习的神经网络，对条件概率进行估计
- 神经网络表示为$p_\theta(x_{t-1}|x_t)$，其中θ是可学习参数



Diffusion Model的加噪与降噪过程[1]



Diffusion Model的加噪与降噪步骤可视化[2]

1. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in Neural Information Processing Systems, 2020.
2. Bansal A, Borgnia E, Chu H M, et al. Cold diffusion: Inverting arbitrary image transforms without noise[J]. arXiv preprint arXiv:2208.09392, 2022.

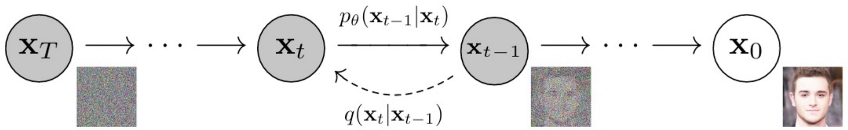
Reverse Denoising Diffusion

➢ 参数化去噪过程

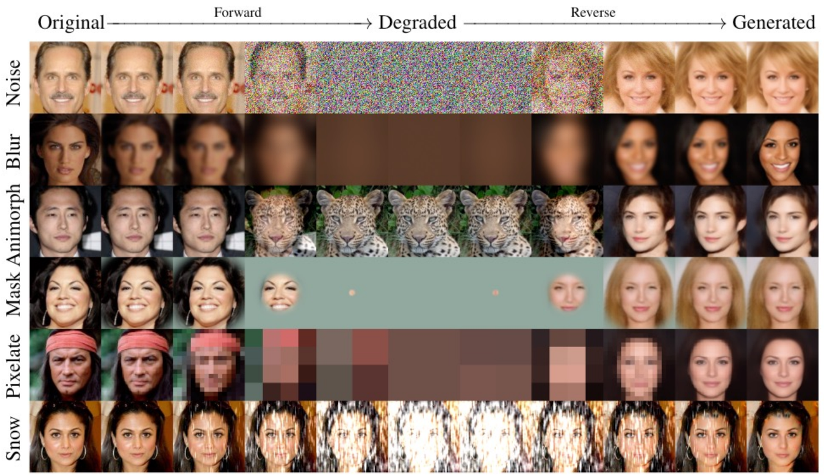
- 基础假设：条件概率$p(x_{t-1}|x_t)$也符合高斯分布
- 参数化表示： $p(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$
- 和VAE类似，神经网络需要计算高斯分布的均值和方差

➢ 简化实现

- 将方差设定为固定值，神经网络仅计算均值 [1]



Diffusion Model的加噪与降噪过程



Diffusion Model的加噪与降噪步骤可视化

1. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in Neural Information Processing Systems, 2020.

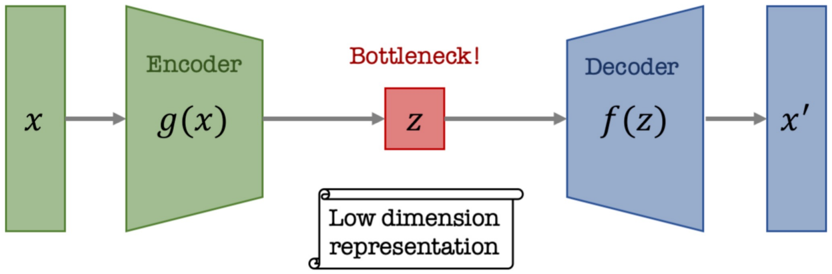
目标函数构造

➢ 与VAE的相似性

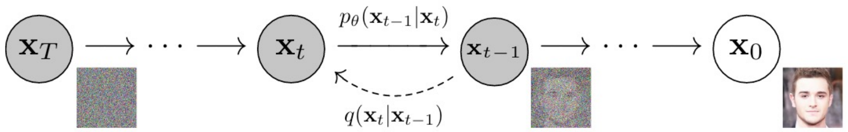
- 条件概率$q(x_t|x_{t-1})$与$p(x_{t-1}|x_t)$的组合可以看作一个VAE
- 可以类比VAE中的Variational lower bound (ELBO) [1]，计算两个概率之间的KL散度来构造目标函数

➢ ELBO目标函数

- 对于整个前向、反向扩散过程，相当于每一步损失的总和
- $L = L_0 + L_1 + \cdots + L_T$，其中$L_t$等价于$q(x_t|x_{t-1})$和$p(x_{t-1}|x_t)$均值的l2损失



VAE的编码与解码过程



Diffusion Model的加噪与降噪过程

1. Kingma D P, Welling M. Auto-encoding variational bayes[J]. arXiv preprint arXiv:1312.6114, 2013.

目标函数构造

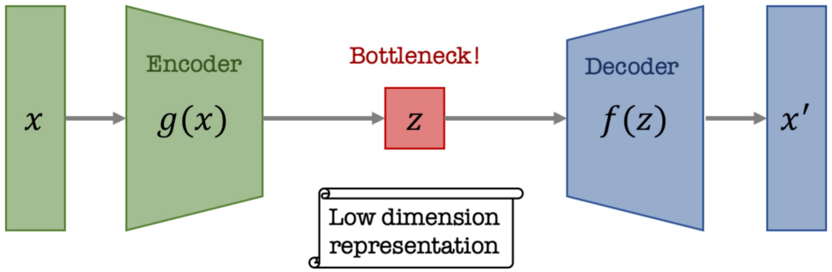
➢ 简化的计算方式

- 对$p(x_{t-1}|x_t)$的均值$\mu_\theta(x_t,t)$进行重参数化
- 让神经网络仅需预测添加的噪音，而非去噪后图像分布的均值

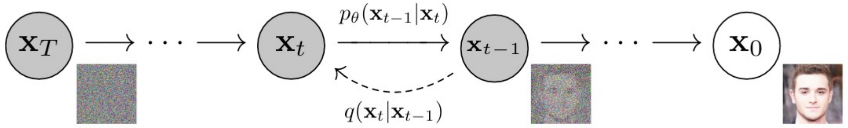
➢ 均值的重参数化

- 去噪后分布的均值$\mu_\theta(x_t,t)=\frac{1}{\sqrt{\alpha_t}}\left(x_t-\frac{\beta_t}{\sqrt{1-\overline{\alpha_t}}}\epsilon_\theta(x_t,t)\right)$，其中

  $\epsilon_\theta(x_t,t)$是预测的噪音
- 简化后的t步目标函数$L_t=\|\epsilon-\epsilon_\theta(x_t,t)\|^2$，其中$\epsilon\sim\mathcal{N}(0,\mathbf{I})$



VAE的编码与解码过程



Diffusion Model的加噪与降噪过程

1. Kingma D P, Welling M. Auto-encoding variational bayes[J]. arXiv preprint arXiv:1312.6114, 2013.

训练过程

➢ 训练步骤

- 从真实数据集中采集x_0
- 随机选取一个马尔科夫过程步骤下标t
- 计算噪音分布，向x_0中添加噪音，得到t步的噪音样本x_t
- 训练神经网络，根据x_t预测添加的噪音

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \mathrm{Uniform}(\{1, \dots, T\})$
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

Diffusion Model 训练算法 [1]

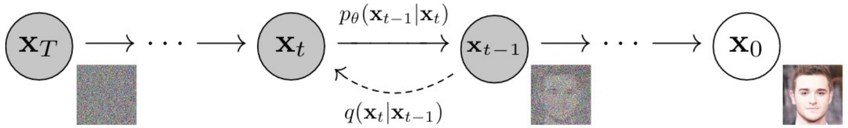
1. Niels Rogge, Kashif Rasul. The Annotated Diffusion Model. https://huggingface.co/blog/annotated-diffusion

➢ Diffusion Model中神经网络的功能

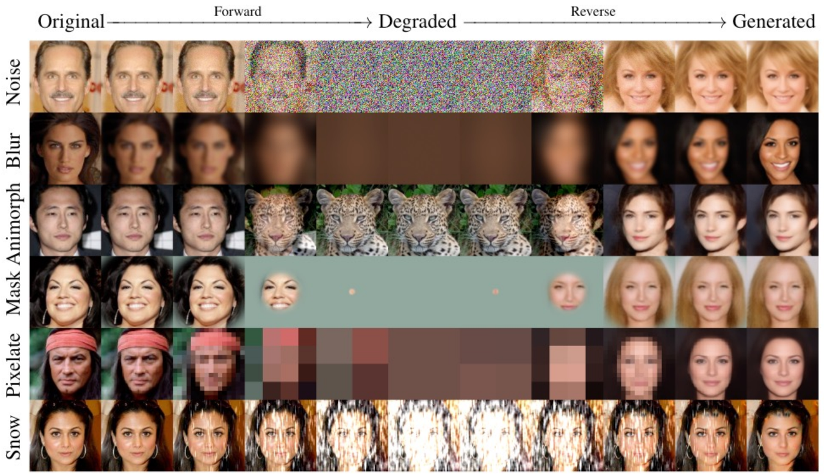
- 给定马尔科夫过程步骤下标t，和此步的噪音图片x_t，预测添加的噪音

➢ 模型选择的原则

- 噪音图片x_t和预测的噪音$\epsilon_\theta(x_t,t)$，两个张量大小一致
- 神经网络模型需要拥有相同形状的输入和输出



Diffusion Model的加噪与降噪过程[1]



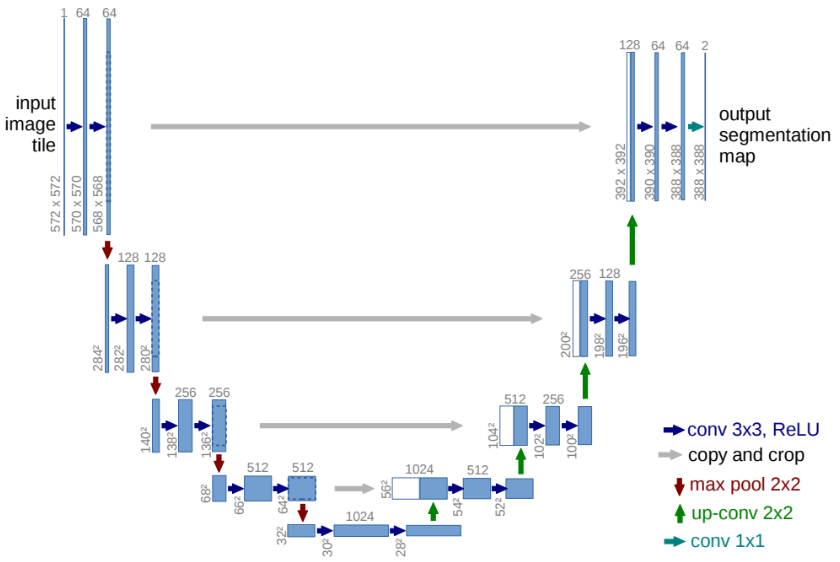
Diffusion Model的加噪与降噪步骤可视化[2]

1. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in Neural Information Processing Systems, 2020.
2. Bansal A, Borgnia E, Chu H M, et al. Cold diffusion: Inverting arbitrary image transforms without noise[J]. arXiv preprint arXiv:2208.09392, 2022.

模型选择

➢ 常见模型选择

- 自编码结构的神经网络满足条件，同时其瓶颈式结构能够减少模型参数量
- 处理图像数据时，常用U-Net [1]，一种基于CNN的自编码器模型
  - U-Net特色在于Encoder和Decoder之间存在残差连接，能够改善梯度传播效率



U-Net 结构

1. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.

模型组件

➢ 位置编码

- 受Transformer中的Positional Encoding [1] 启发
- 将马尔科夫过程步骤的下标t编码，使得模型能够知晓当前所在的步骤

➢ 带残差的卷积层

- 使用卷积层对输入的图像进行建模
- 在多层卷积之间添加残差连接

➢ 注意力层

- 添加在卷积层之间，增强模型容量

```python
class SinusoidalPosEmb(nn.Module):
    def __init__(self, dim):
        super().__init__()
        self.dim = dim

    def forward(self, x):
        device = x.device
        half_dim = self.dim // 2
        emb = math.log(10000) / (half_dim - 1)
        emb = torch.exp(torch.arange(half_dim, device=device) * -emb)
        emb = x[:, None] * emb[None, :]
        emb = torch.cat((emb.sin(), emb.cos()), dim=-1)
        return emb
```

Positional Encoding 编码马尔科夫过程步长

```python
class ResnetBlock(nn.Module):
    def __init__(self, dim, dim_out, *, time_emb_dim=None, groups=8):
        super().__init__()
        self.mlp = nn.Sequential(
            nn.SiLU(),
            nn.Linear(time_emb_dim, dim_out * 2)
        ) if exists(time_emb_dim) else None

        self.block1 = Block(dim, dim_out, groups=groups)
        self.block2 = Block(dim_out, dim_out, groups=groups)
        self.res_conv = nn.Conv2d(dim, dim_out, 1) if dim != dim_out else nn.Identity()

    def forward(self, x, time_emb=None):
        scale_shift = None
        if exists(self.mlp) and exists(time_emb):
            time_emb = self.mlp(time_emb)
            time_emb = rearrange(time_emb, 'b c -> b c 1 1')
            scale_shift = time_emb.chunk(2, dim=1)

        h = self.block1(x, scale_shift=scale_shift)

        h = self.block2(h)

        return h + self.res_conv(x)
```

带残差的卷积层

1. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

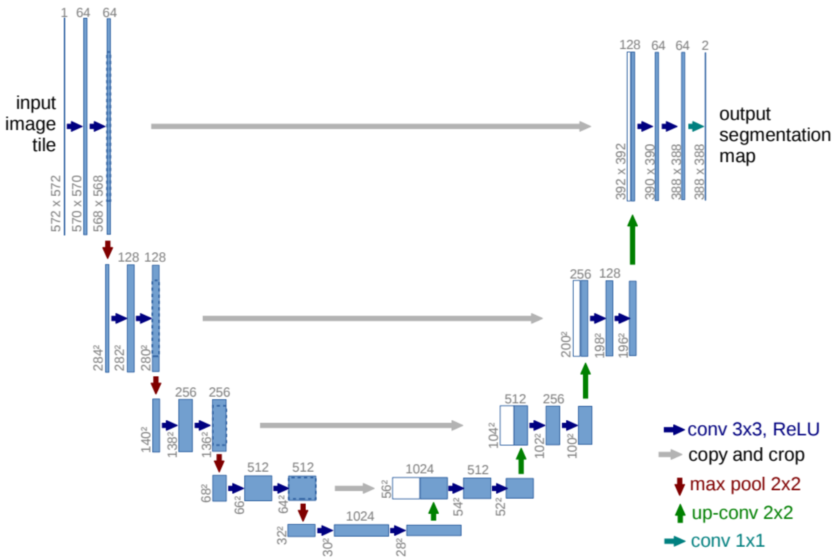
U-Net构建

➢ 预期输入与输出

- 输入添加噪音后的图像x(batch_size, num_channels, height, width)，以及噪音水平标识t(batch_size, 1)
- 输出对噪音的预测$\hat{\epsilon}$(batch_size, num_channels, height, width)

➢ 计算顺序

- 噪音图像x输入卷积层，噪音水平t进行位置编码
- 使用残差卷积层对数据进行多层降采样，同时添加正则化、注意力层
- 得到中间层隐藏状态
- 使用残差卷积层对隐藏状态进行多层升采样
- 输出对噪音的预测



U-Net 结构

1. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.

图片生成

花朵图片生成

人脸图片生成

1. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in Neural Information Processing Systems, 2020, 33: 6840-6851.

文本生成

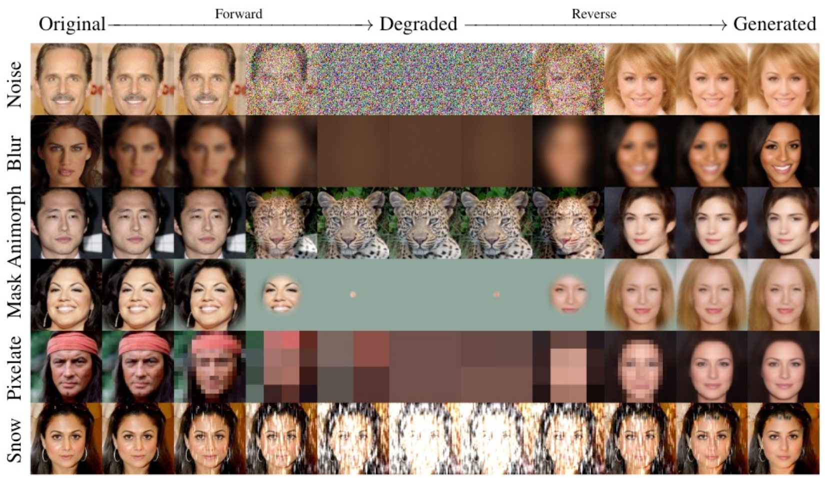
input (Semantic Content) output text	food : Japanese Browns Cambridge is good for Japanese food and also children friendly near The Sorrento .
input (Parts-of-speech) output text	PROPN AUX DET ADJ NOUN NOUN VERB ADP DET NOUN ADP DET NOUN PUNCT Zizzi is a local coffee shop located on the outskirts of the city .
input (Syntax Tree) output text	(TOP (S (NP (*) (*) (*)) (VP (*) (NP (NP (*) (*))))) The Twenty Two has great food
input (Syntax Spans) output text	(7, 10, VP) Wildwood pub serves multicultural dishes and is ranked 3 stars
input (Length) output text	14 Browns Cambridge offers Japanese food located near The Sorrento in the city centre .
input (left context) input (right context) output text	My dog loved tennis balls. My dog had stolen every one and put it under there. One day, I found all of my lost tennis balls underneath the bed.

可控制的文本生成

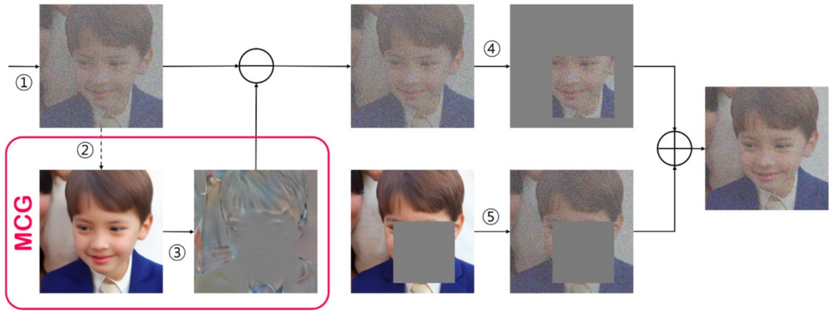
1. Li X L, Thickstun J, Gulrajani I, et al. Diffusion-LM Improves Controllable Text Generation[J]. arXiv preprint arXiv:2205.14217, 2022.

扩展阅读

- 前向扩散过程并不一定要通过添加高斯噪音实现
  - Bansal A, Borgnia E, Chu H M, et al. Cold diffusion: Inverting arbitrary image transforms without noise[J]. arXiv:2208.09392, 2022.
- 通过添加额外限制达成更佳的效果
  - Chung H, Sim B, Ryu D, et al. Improving Diffusion Models for Inverse Problems using Manifold Constraints[J]. arXiv:2206.00941, 2022.
- 在低维隐空间而非原始空间上进行扩散
  - Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 10684-10695.
- 应用于序列预测、序列插补等
  - Rasul K, Seward C, Schuster I, et al. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting[C]//International Conference on Machine Learning. PMLR, 2021: 8857-8868.
  - Alcaraz J M L, Strodthoff N. Diffusion-based Time Series Imputation and Forecasting with Structured State Space Models[J]. arXiv:2208.09399, 2022.



Cold diffusion: Inverting arbitrary image transforms without noise



Improving Diffusion Models for Inverse Problems using Manifold Constraints