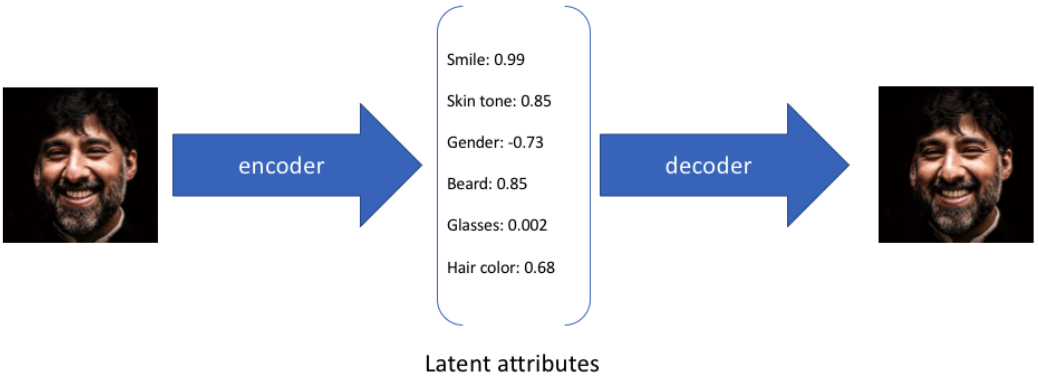
VAE

参考1(上半部分)

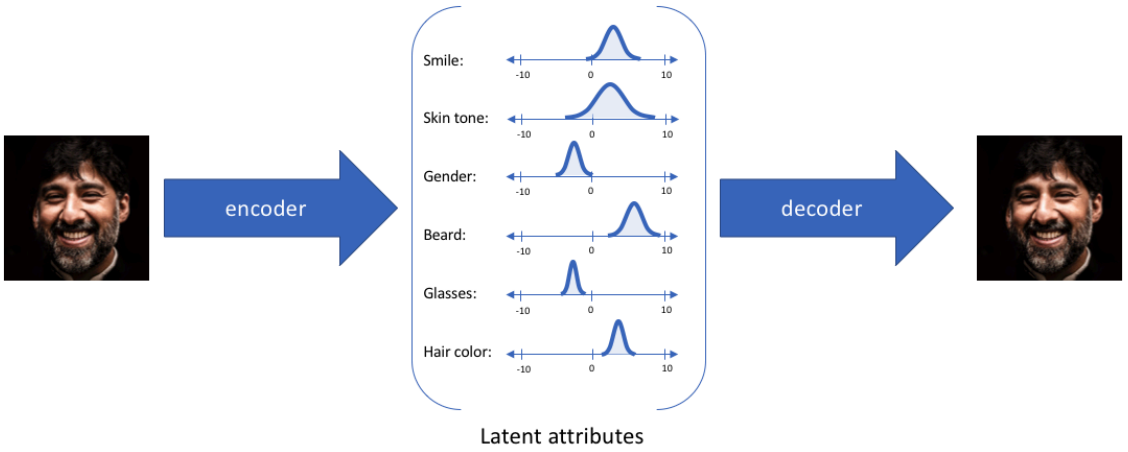
参考2

Insight

AE将输入编码为隐空间(几个特征值)中的向量z，然后解码，其假定一张图片可由几个特征对应的取值唯一确定，如下图



VAE则更进一步，不再假定图片对应的特征z是单一值，而是一个取值范围，也就是一个概率分布，如下图



令特征z为高斯分布$\mathcal{N}(\mu_x, \sigma_x)$，则编码器输出这个分布的参数，也就是均值$\mu_x$和方差$\sigma_x$

我们强迫$\mathcal{N}(\mu_x, \sigma_x)$向标准正态分布$\mathcal{N}(0, \mathbf{I})$靠拢，原因如下

1. 实现隐空间的规整性与连续性

- **目的**：确保潜在空间 z 的结构具有良好数学性质（平滑、连续、可插值）。
- **效果**：
  - 任意采样 $z \sim \mathcal{N}(0, I)$ 都能解码为有意义的图像（避免"空洞"区域）。
  - 隐空间的线性插值（如 $z_1 \to z_2$）能生成语义连续的过渡图像。
- **对比**：普通自编码器（AE）的隐空间可能不连续，随机采样 z 可能解码出无意义结果。

---

2. 正则化约束（KL散度项）

- VAE 的损失函数包含两部分：

$$\mathcal{L} = \text{重建损失} + \beta \cdot \mathrm{KL}(q(z|x) \| p(z))$$

  其中 $p(z) = \mathcal{N}(0, I)$ 是标准正态先验，$q(z|x)$ 是编码器输出的分布。
- **KL散度的作用**：
  - 强迫所有 $q(z|x)$ 向 $\mathcal{N}(0, I)$ 靠近，避免过拟合。
  - 控制隐变量的分布范围，防止编码器将不同图像映射到彼此远离的 z（隐空间坍塌）。

也就是希望不同图像对应的隐向量z相互靠近、聚集以保证隐空间的连续性，防止出现部分区域无法采样生成真实图像的情况

Principle Derivation

目标是最大化对数似然$\log p_\theta(x)$，利用变分推断，有

$$
\begin{aligned}
\log p(x) &\geq \mathbb{E}_{q(z|x)} \log \frac{p(x,z)}{q(z|x)} \\
&= \mathbb{E}_{q(z|x)}[\log p(x|z) + \log p(z) - \log q(z|x)] \\
&= \underbrace{-\mathrm{KL}(q(z|x)||p(z))}_{\text{prior matching}} + \underbrace{\mathbb{E}_{q(z|x)} \log p(x|z)}_{\text{reconstruction loss}}
\end{aligned}
$$

其中$q(z|x), p(x|z)$分别由θ, ϕ参数化，分别表示Encoder,Decoder；$p(z)$为z的先验分布，设定为标准正态分布

为了保证采样z的过程保持梯度可传导，采用重参数化方法，Encoder输出隐空间分布的均值μ_x和方差σ_x，z通过如下方式采样

$$
z = \mu_x + \sigma_x \epsilon, \epsilon \sim \mathcal{N}(0, \mathbf{I})
$$

$p(x|z)$也建模为高斯分布，因此Decoder输出该高斯分布的均值μ_z与方差σ_z，x通过重采样获取

$$
x = \mu_z + \sigma_z \epsilon, \epsilon \sim \mathcal{N}(0, \mathbf{I})
$$