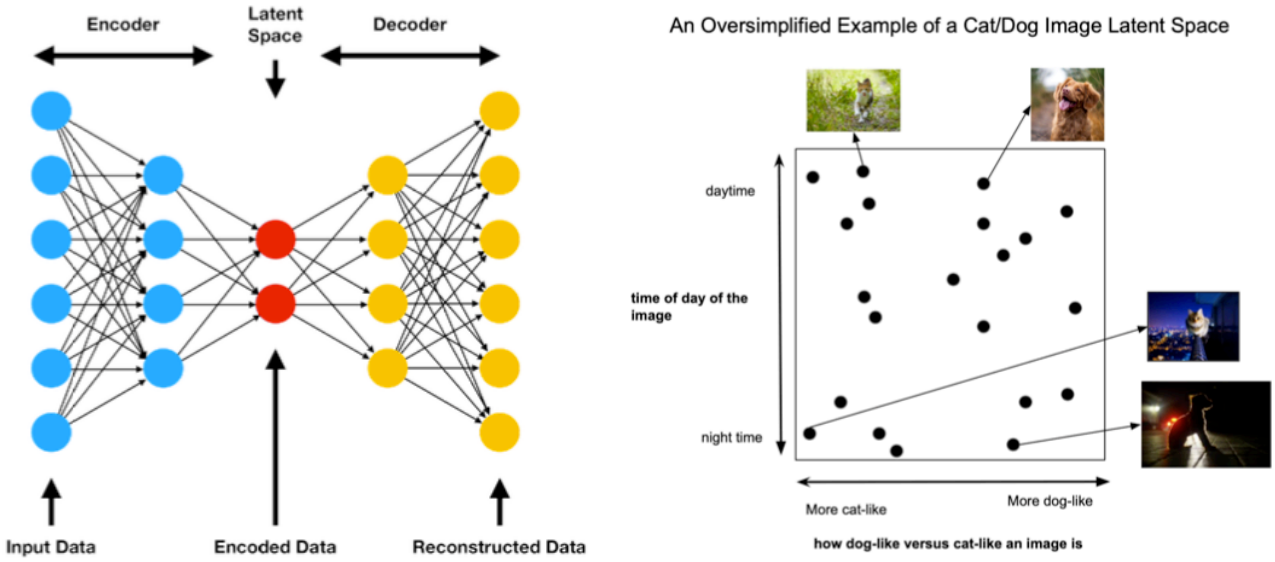
1 VAE 模型理解与思考

1.1 AutoEncoders 模型

我对 VAE 模型的理解从其所基于的 AutoEncoders 模型开始。AutoEncoders 模型是一种无监督学习模型，其目的是将输入的数据进行编码，并通过 Encoder 映射到隐层空间；然后再解码重建原始图像，使得解码后的数据与原始数据尽可能相似，即最小化重建误差。AutoEncoders 模型的结构如图1(a)所示。



(a) AE 模型图示

(b) 可能的 latent space

图 1: AE 模型及其隐层空间

对于输入数据 $\boldsymbol{x}$，先通过 Encoder 映射到隐层空间，得到隐层表示 $\boldsymbol{z}$；再通过 Decoder 将隐层表示映射到输出空间，得到重建数据 $\boldsymbol{x'}$。损失函数则可以使用均方误差，即 $||\boldsymbol{x}-\boldsymbol{x'}||_2^2$，或者根据特定的任务使用其他损失函数。AutoEncoders 模型并没有指定其训练的数据类型，所以其使用的 Encoder 和 Decoder 也可以相应地根据任务进行选取。

1.2 隐层空间的思考

神经网络强大的拟合能力使得其可以得到输入数据的一种特征表示，也就体现为隐层空间的向量。我们可以自由地选取隐层向量的维度，一般而言，维度越高，意味着能表达的语义越丰富。如图1(b)所示是二维隐层空间的可能情况，以输入数据集是猫/狗在白天/夜晚的图片为例，如果我们将图片映射到二维的隐层空间，其在不同维度上的隐层数值应该表示不同的特征，即反应图片像猫/狗的程度和拍摄的时间。理想情况下，潜在空间应该使语义相似的数据点距离彼此更近，语义不同的数据点距离彼此更远，而普通的 AutoEncoders 模型并不能保证这一点，VAE 模型则是在此基础上进行改进，以得到更好的隐层空间表示。

1.3 VAE 模型对 AE 的改进

在没有对隐层空间做出任何假设，即隐层空间是整个 $\mathcal{R}^n$，数据分布是任意的情况下，AutoEncoders 模型所得到的隐层表达完全可以自由地分布在整个空间中，但实际上我们可能希望数据分布在隐空间中所占据的区域体积较小而比较紧凑，而不是自由地分布到无穷远的地方。如下图2所示，大多数情况下，AE 只是通

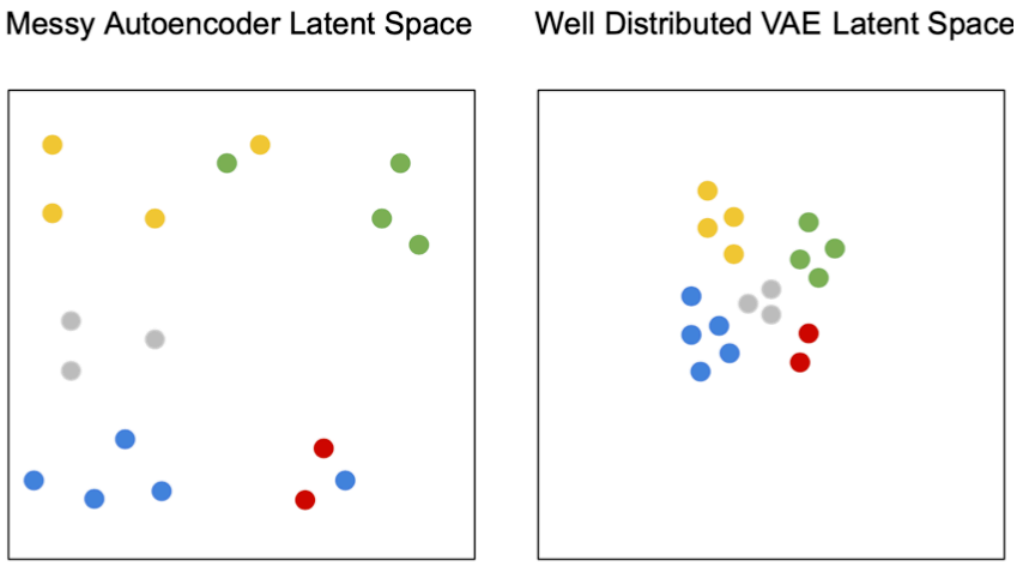


图 2: VAE 模型隐层空间的改进

过强大的编解码器记住了样本点在隐层空间的位置，这种强制性的要求就使得同类数据不一定能在隐层分布到邻近的区域。VAE 模型通过假设隐层空间的分布是高斯分布，即 $\boldsymbol{z} \sim N(\mu, \sigma)$ 来解决这个问题。如果隐层的先验分布用 $\Pr(\boldsymbol{z})$ 表示，输入数据仍以 $\boldsymbol{x}$ 表示，通过贝叶斯公式

$$\Pr(\boldsymbol{z}|\boldsymbol{x}) = \frac{\Pr(\boldsymbol{x}|\boldsymbol{z})\Pr(\boldsymbol{z})}{\Pr(\boldsymbol{x})} \tag{1}$$

我们就可以得到一个基本的 VAE 模型中 Encoder 和 Decoder 的数学形式，从而得到优化目标。VAE 模型的结构图如图3所示：Encoder 从输入数据 $\boldsymbol{x}$ 中得到隐层表示 $\boldsymbol{z}$，其服从的条件分布表示为 $q_\phi(\boldsymbol{z}|\boldsymbol{x})$；Decoder

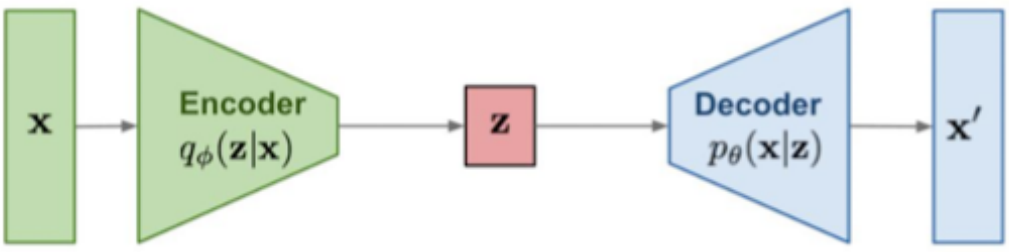


图 3: VAE 模型图示

从隐层表示 $\boldsymbol{z}$ 中重建出 $\boldsymbol{x}'$，其服从的条件分布表示为 $p_\theta(\boldsymbol{x}|\boldsymbol{z})$。$\phi$ 和 θ 则是 Encoder 和 Decoder 的参数。在 AE 最小化重建误差的基础上，VAE 模型还要最小化后验分布 $q(\boldsymbol{z}|\boldsymbol{x})$ 与先验分布 $p(\boldsymbol{z})$ 的差异，即最小化 KL

散度, 那么 VAE 模型的损失函数可以表示为:

$$\mathcal{L}(\boldsymbol{x};\theta,\phi) = -\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] + \mathcal{D}_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})) \tag{2}$$

至此，对最基本的 VAE 模型的理解就完成了，其推理过程如图4所示。后续部分将根据 VAE 的原理进行实现，并完成相关的实验。

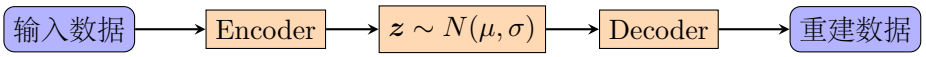


图 4: VAE 模型的推理过程

2 实验过程和分析

2.1 VAE 模型搭建和训练

2.2 隐层空间 $\mathcal{R}^1$ 的生成图片效果

2.3 隐层空间 $\mathcal{R}^2, [-5, 5]^2$ 的生成图片效果