1 提交文件简要说明

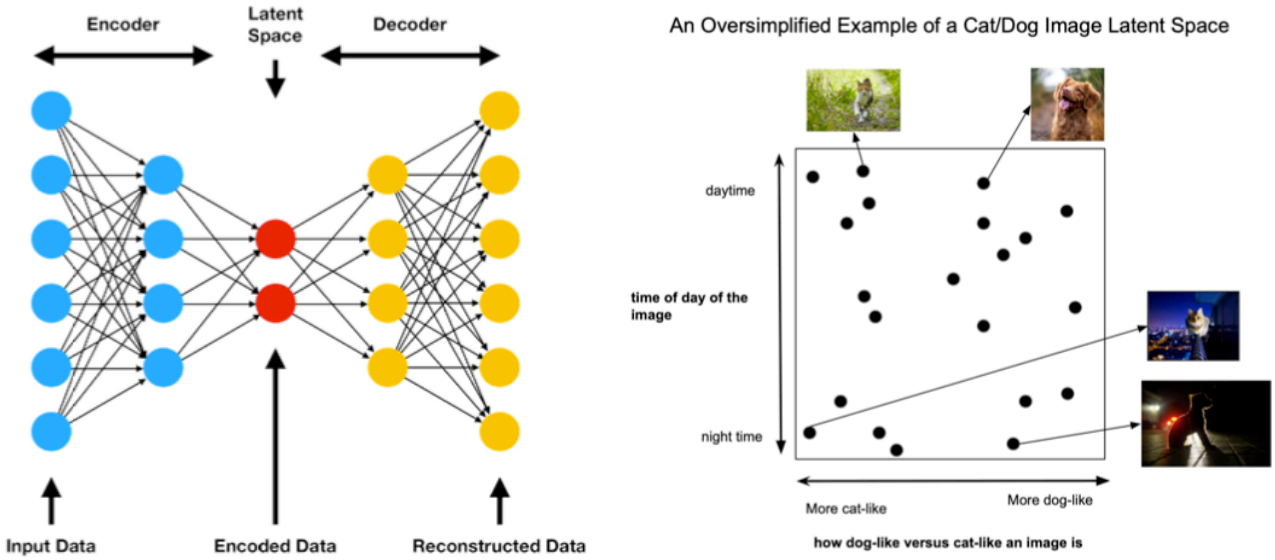
本实验最终提交的文件夹 `main.py` 即包含了数据集下载、训练、图片生成、保存的全部流程。重现最优结果的脚本配置则为 argparser 的默认配置。实际训练过程中 `z_dim` 是可以自由调整的。

本报告中所包含的生成图片是手动选取的较好结果，其他更多实验结果详见 `img` 文件夹。

2 VAE 模型理解与思考

2.1 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。

2.2 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。

2.3 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

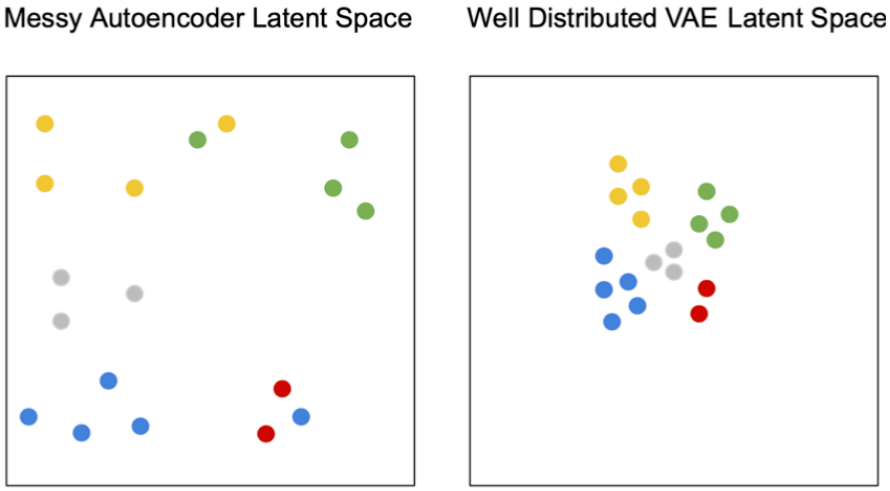


图 2: VAE 模型隐层空间的改进

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 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。

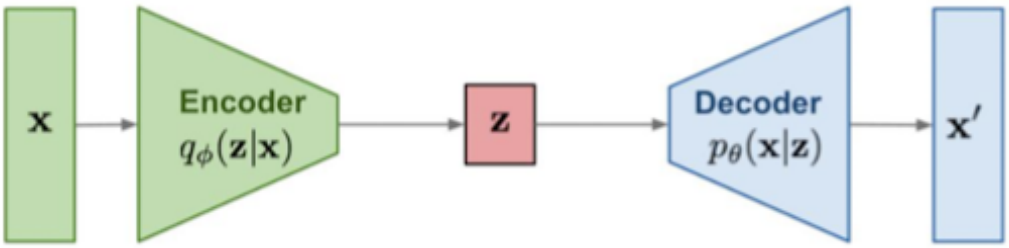


图 3: VAE 模型图示

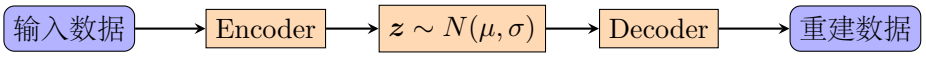


图 4: VAE 模型的推理过程

3 实验过程和分析

3.1 VAE 模型搭建和训练

对于我们的任务而言，Encoder 和 Decoder 使用全连接神经网络，即可收获不错的效果。对于正态分布的均值和标准差，Encoder 需要通过维度相同的全连接神经网络映射到隐空间，每次 forward 的过程中，首先通过 Encoder 得到隐层服从分布的均值和标准差，之后通过再参数化的方式，即 $\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \cdot \boldsymbol{\epsilon}$，其中 $\boldsymbol{\epsilon} \sim N(0, 1)$，来得到基于当前均值标准差下的隐层向量，然后再通过全连接的 Decoder 得到重建数据。

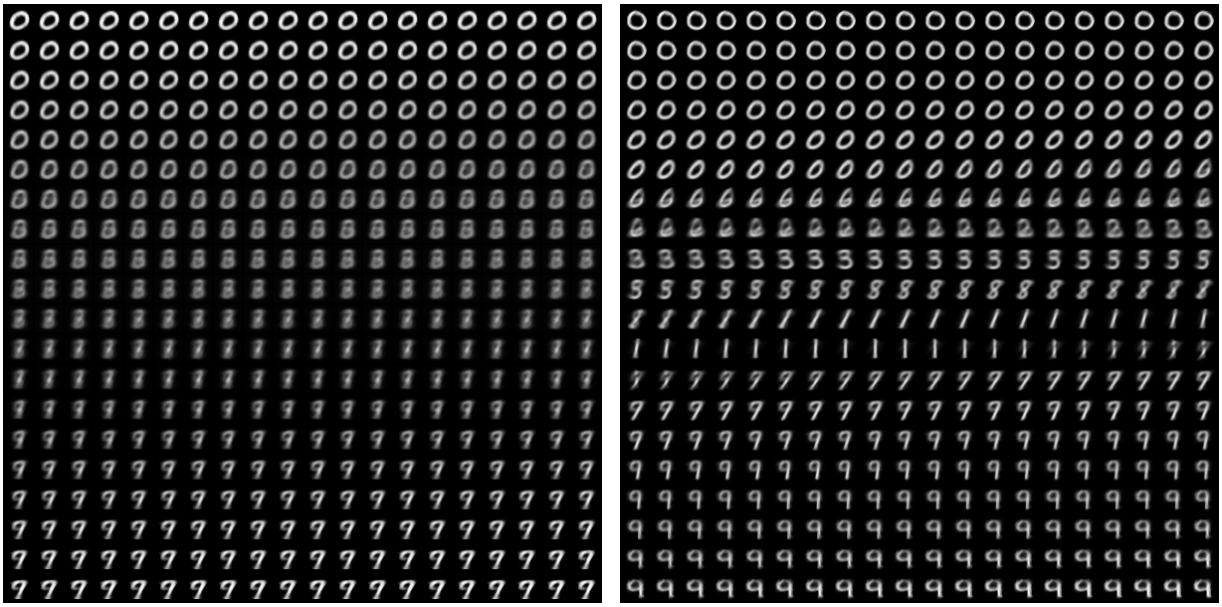
在实现上需要注意的是，Encoder 返回的标准差应该是取对数之后的结果，这样可以使数据分布范围更合理，便于更好地发挥神经网络的性能。损失函数则根据公式2来计算并内置于 VAE 类内。

3.2 隐层空间 $\mathcal{R}^1$ 的生成图片效果

既然 VAE 模型假定隐层空间服从高斯分布，那么理论上来说，在给定区间范围内的一组隐层向量 Decode 得到的图片应该服从一定规律。以下是我从区间 $[-3, 3]$ 中生成均匀分布且递增的实数，然后通过 Decoder 得到的图片效果，如图5所示。为了方便观察，将其 reshape 成为二维形式。之后，分别在第一个 Epoch 和最后一个 Epoch 时，将隐层空间的分布可视化，如图5(a)和图5(b)所示。可见，通过训练，模型的质量确实有所提高，而生成的图像分布确实服从某种规律，在水平方向上相邻的图片之间存在某种相关性。

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

同样地，我们可以将隐层空间的维度增加到 2 维，然后在 $[-5, 5]^2$ 的区间内生成均匀分布的隐层向量，然后通过 Decoder 得到的图片效果，如图6所示。通过查阅资料得知，VAE 的 Encoder 的效果可以被认为是找到了高维数据在低维上的一个 manifold(流形)，或者说高维数据在低维上的投影。与隐层空间是一维的情况对比明显可以发现，隐层空间是二维的情况下，生成的图片在两个维度上都存在着相关性。可以发现，生成图片的质量在接近原点的时候较为模糊，而在接近边缘的地方模式较为单一，这是因为隐层空间符合高斯分布，而高斯分布的概率密度函数在接近均值的地方较大，而在边缘的地方较小，所以在接近原点的地方，隐层空间的分布较为稠密，而在接近边缘的地方，隐层空间的分布较为稀疏，实验结果是符合理论的。当然，为了取得



(a) 一个 Epoch 后效果　　(b) Final Epoch 后效果

图 5: 隐层空间 $\mathcal{R}^1$ 的生成图片效果

更高的生成图片质量，我们可以增加模型参数，不过受限于计算资源和学期时间，目前的结果已经能够充分帮助我们理解 VAE 模型的理解和提高性能相关的因素了。

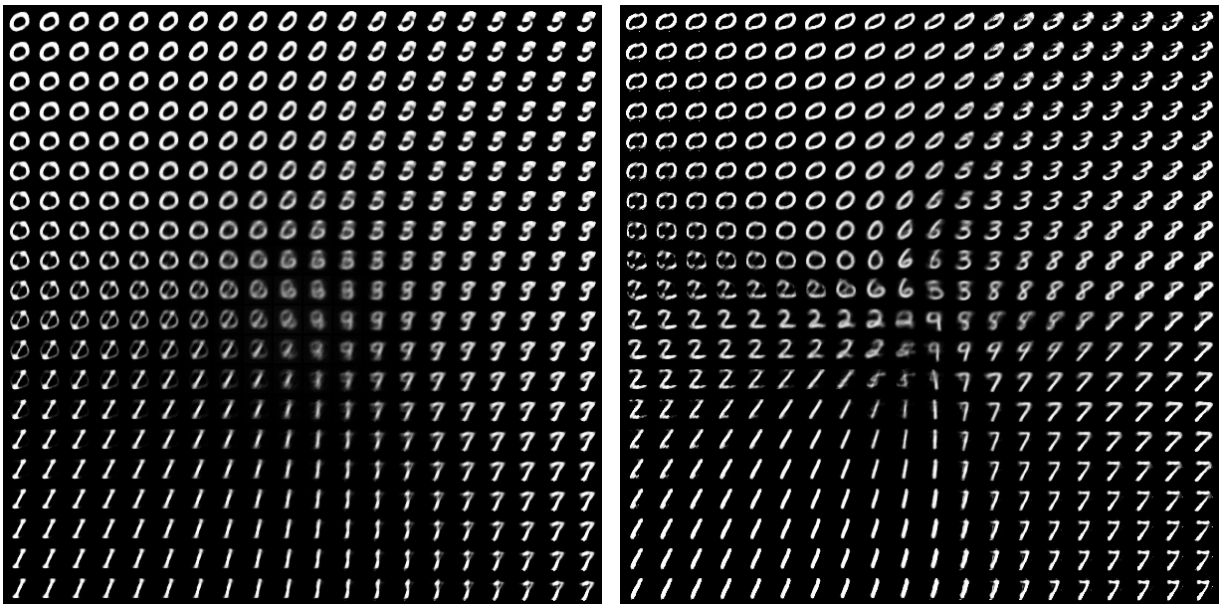
3.4 附：其他图片的生成效果选取

对于 VAE 而言，隐层空间越大，生成图片的质量也就越高。以下是我实验过程中一些具有代表性的结果选取，如图7所示。

4 实验总结与思考

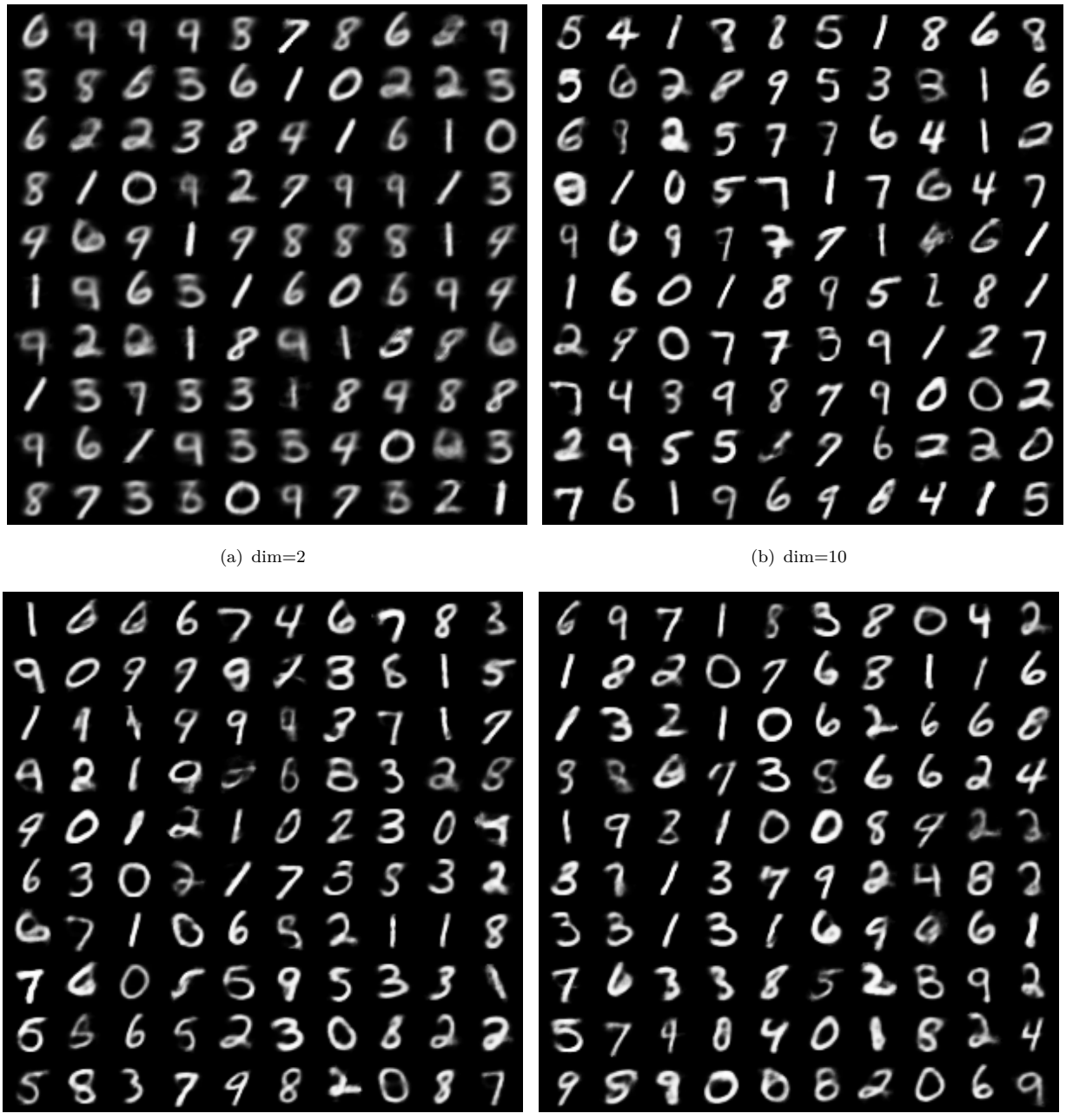
在本次课程接触到并实践 VAE 之前，我还曾尝试过 GAN 的实践；相比之下，VAE 更加深了我对“隐层空间”的理解。回顾并与 GAN 的训练流程相比，在我们训练好并得到一个 Generator 之后，我们也得到了一个潜在的“隐层空间”，但是我们很难知道这样一个隐层空间所服从的概率分布，也就是说，我们没有办法控制自己给定输入所生成的图片，因为我们不知道隐层向量在隐层空间下的“语义”。VAE 则不然，通过预先假定隐层空间服从高斯分布，Decoder 对隐层向量的解码则更有“目的性”。

本次实验启示我，神经网络有时并不是进行了语义匮乏的“非线性映射”，在良好数学形式的约束下，神经网络所能得到的特征表达是有一定意义的，而这种意义也是我们可以通过一定的方式来控制的，这也是深度学习科学性的体现。



(a) 一个 Epoch 后效果　　(b) Final Epoch 后效果

图 6: 隐层空间 $\mathcal{R}^2$ 的生成图片效果



(a) dim=2

(b) dim=10

(c) dim=20

(d) dim=30

图 7: 部分实验结果节选