

机器学习与数据挖掘 期末课程报告

林宇浩 21311274

一、报告主题

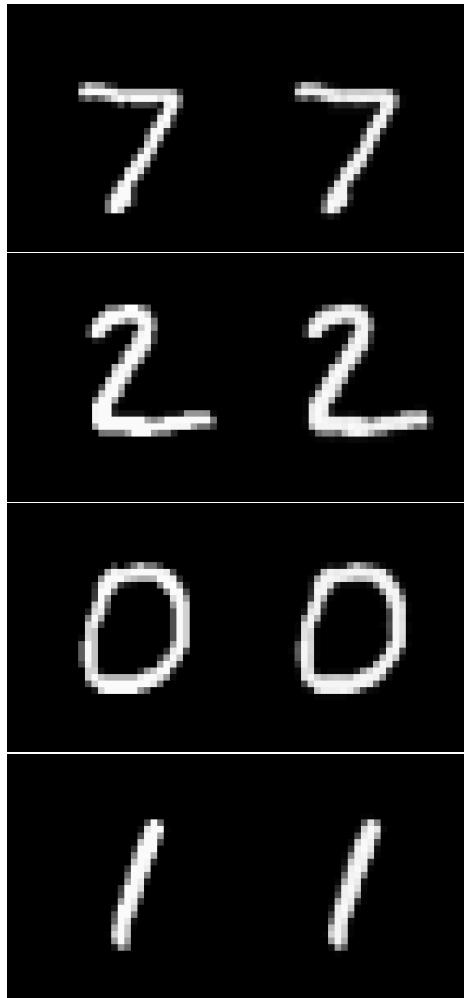
本次报告的主题为生成模型，主要关注于课程第 16 章节中介绍的自编码器。实验中将对目前已有的自编码器模型进行实验和分析，通过 MNIST 手写数字数据集比较不同模型之间的效果。

二、已有方法

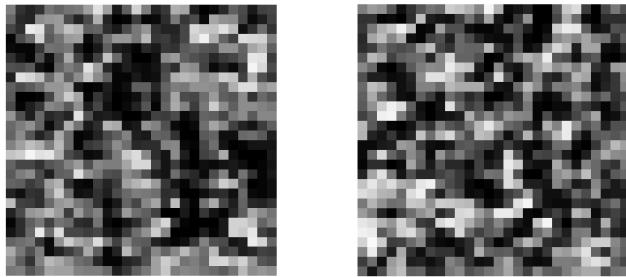
1、AE (autoencoder, 自编码器)

自编码器 AE 是一种无监督学习算法，属于神经网络的一种结构。它的目标是通过学习数据的压缩表示来重建输入数据，从而实现数据的降维或特征提取。自编码器包含两个主要部分：编码器（Encoder）和解码器（Decoder）。整个结构的目标是通过训练将输入数据映射到一个低维表示，然后再将该表示映射回原始数据。这个过程迫使网络学会捕捉输入数据的关键特征，所以也可以用于降维、特征提取、去噪等任务。

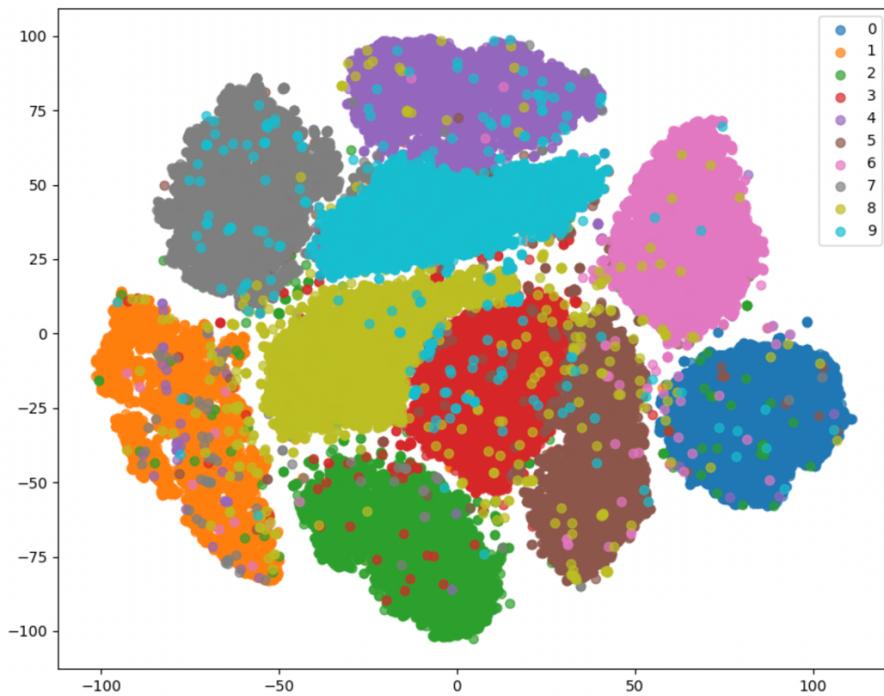
仅用较短的时间训练 5 轮后，自编码器就能达到较好的恢复效果，下面左侧是原图，右侧是根据原图恢复出的图像：



通过将随机生成特征向量并输入到解码器可以生成图片，但是使用随机生成的特征向量，并无法获得随机的某个数字的图片，而是得到一些乱码，如下所示：

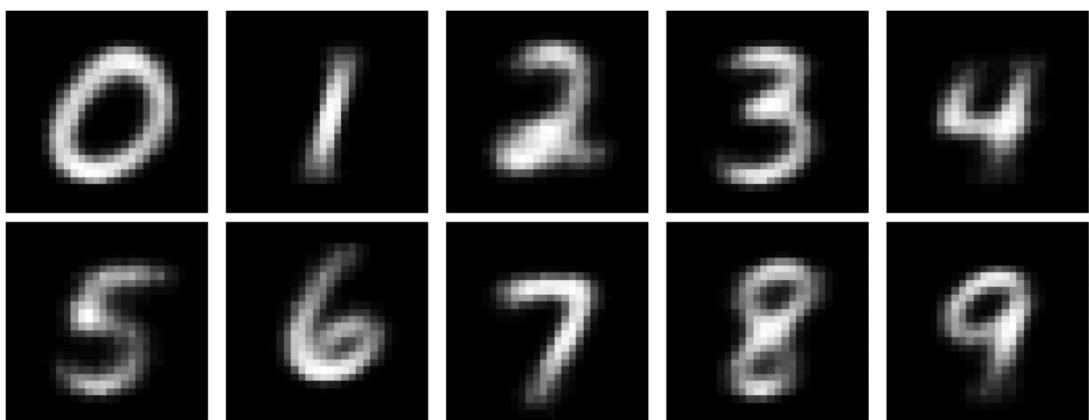


为了能够生成数字，我们对训练集中各个样本通过编码器生成的特征向量进行了统计处理，通过计算平均值得到了各个类别的聚类中心，下图为将各类别的特征向量映射到二维空间后得到的可视化效果。

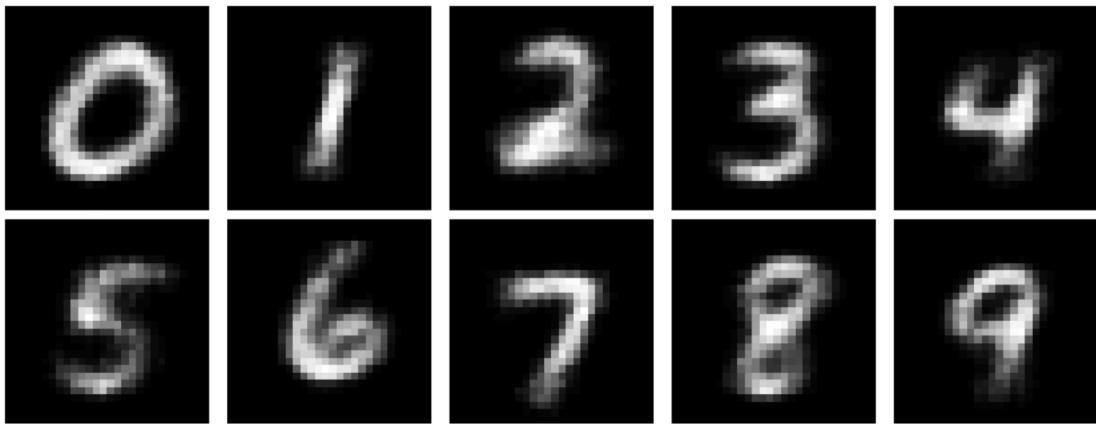


可以看到各个类别能在特征空间中被区分开，说明 AE 对不同数字的特征能够有较好的学习。

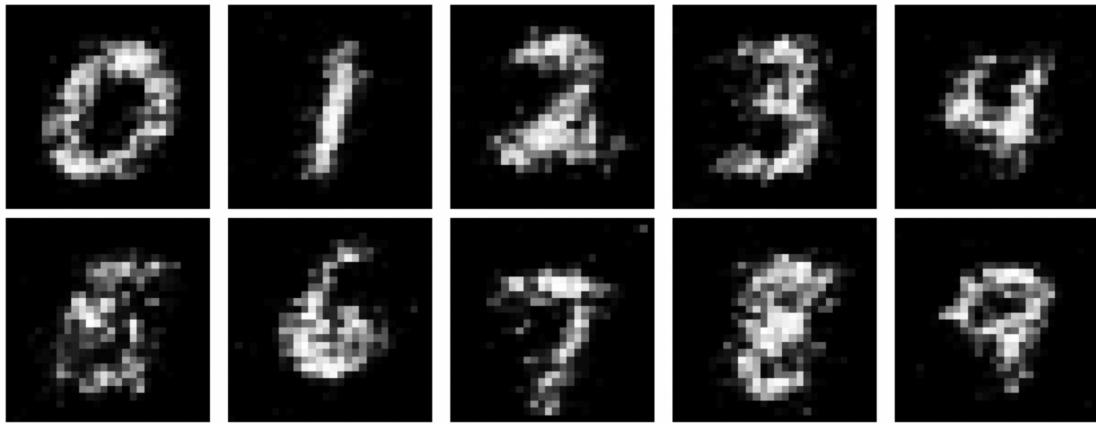
通过将各个类别特征向量的聚类中心输入近解码器，可以生成出不同的数字图像：



通过对各个类别特征向量的聚类中心增加一些噪声，可以对生成的效果有一些微小的改变：



但是随着噪声值的增加，数字形态并没有出现改变，生成出的图像的效果是形态变得模糊，如下图所示：



这说明 AE 的隐变量空间是比较割裂的。从训练结果我们可以看到，每张图像都复原得很好，说明解码器是有能力生成图像的。但是从随机生成的结果可以看到，隐空间中大部分的编码都是没有图像与之对应的。从上面的加噪情况可以看出，隐空间中一些数据点是能够生成图像的，但是在这些数据点周围采样出的特征并无法生成图像，各个能够生成图像的数据点之间是出于一个孤立状态，想要找出这些能生成图像的数据点不容易。虽然可以通过将训练数据输入到编码器得到能生成图像的数据点，但是在这些数据点上加噪并不能很好地生成类似的数据，而是像上面这样产生模糊的效果，这样即使在隐空间中找到了能生成图像的数据点，我们也无法生成新的数据，产生的数据都会和已有数据类似，所以主要的困难在于从隐空间中找到可以产生新图片的数据点。

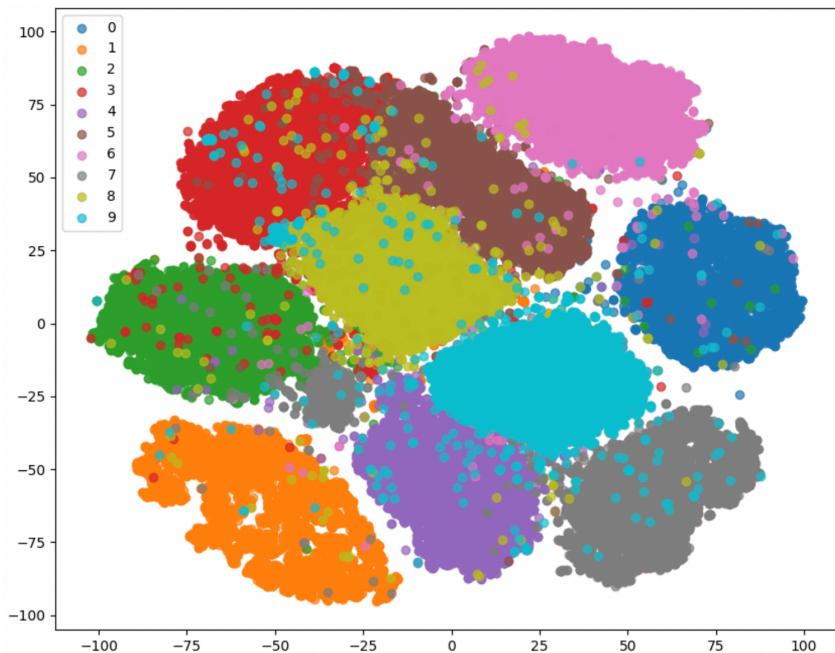
2、VAE (Variational Autoencoder, 变分自编码器)

变分自编码器是自编码器的一种变体，引入了概率潜在空间的概念。与传统自编码器不同，VAE 通过学习数据的概率分布，使得潜在表示更具有连续性和结构性。VAE 不仅学习将输入映射到潜在表示的函数，还学习潜在表示的概率分布。这意味着潜在表示不再是确定性的点，而是一个概率分布，通常假设为正态分布。

在 VAE 中，编码器的映射不仅包括均值也包括方差。编码器的输出参数表示潜在表示的均值和方差。在训练和生成时，从潜在表示的分布中进行采样，以获得实际的潜在表示。通常使用重参数化技巧来确保反向传播的可行性。解码器接收来自潜在表示的采样样本，并将其解码为生成的数据。VAE 的训练目标不仅包括重构误差，还包括潜在表示的分布与预定义的正态分布等先验分布之间的差异，通过 KL 散度来衡量。总体损失函数包括重构误差项和 KL 散度项。

VAE 的引入使得自编码器能够生成更具有连续性的潜在表示，使得在潜在空间中的插值和采样更有意义。

下面我们通过同样的方法，训练完 VAE 之后，通过编码器得到各个类别的聚类中心，映射到二维空间后得到的可视化效果如下所示：

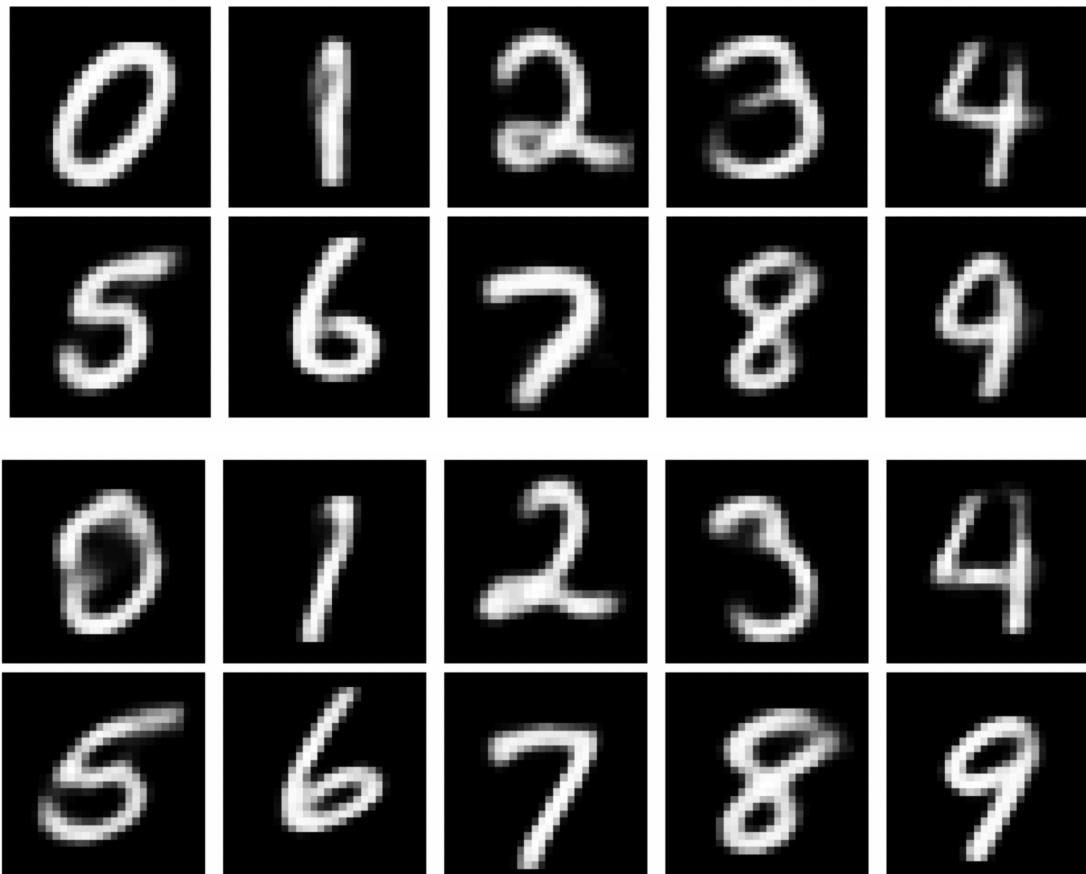


我们将各个数字的平均特征输入解码器，生成出的数字图像如下所示：



可以看到，相比 AE，VAE 生成出的图片效果好了非常多，和原始数据集的图像已经十分相像，没有出现 AE 中边缘非常不清晰的情况。

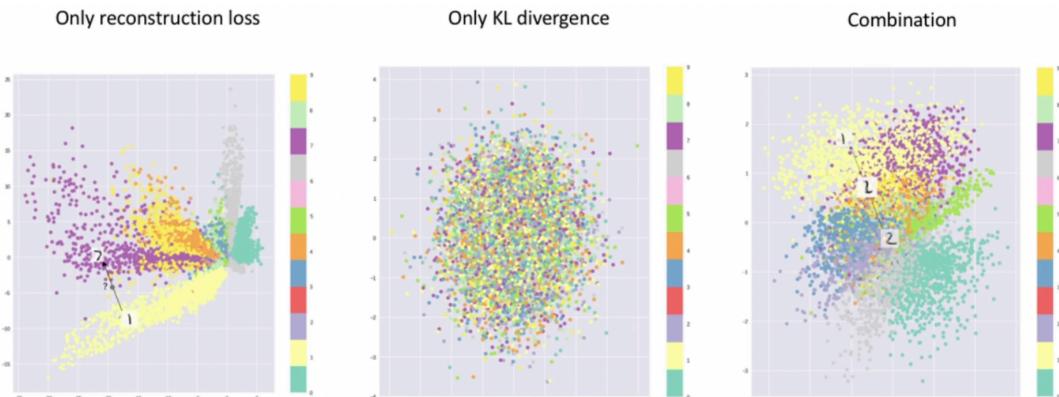
然后我们添加一些噪声后再次输入到解码器中，得到的效果如下所示：



可以看到，添加噪声之后，图像没有变得模糊，而是形态发生了改变，说明了隐空间有较好的连续性和结构性。查阅资料后进一步了解到可以通过控制损失函数的系数来调整隐空间的形态，我们知道 VAE 的两项损失如下所示：

$$\frac{E_{q_\phi(z|x)}[\log(p_\theta(x|z))]}{\text{Reconstruction Loss}} - \frac{D_{KL}(q_\phi(z|x)||p_\theta(z))}{\text{Regularization Loss}}$$

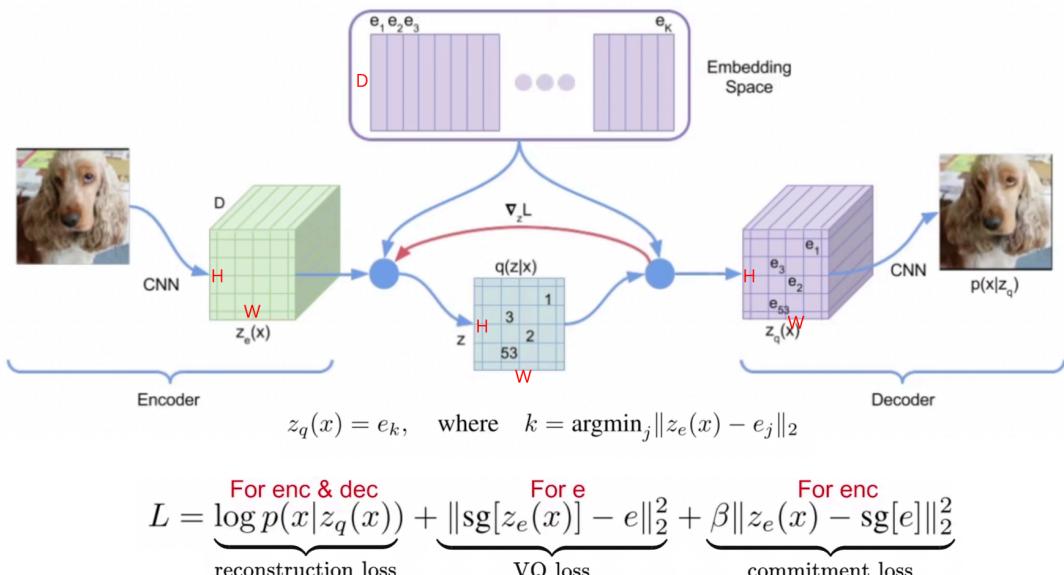
对于重构损失，其直观含义是希望生成出的图片和原始数据越接近越好，在隐空间中其效果如下方左图所示，不同类别的数据会被分开。对于 KL 散度，其直观含义是希望后验分布和先验分布越接近越好，在隐空间中其效果如中图所示，会将数据点映射成正态分布。两个损失的综合效果如右图所示，隐空间中数据点整体上符合正态分布有较好的连续性，同时各个类别的数据能够被分开，保持了较好的结构性。



3、VQ-VAE (Vector Quantised Variational AutoEncoder, 向量量化变分自动编码器)

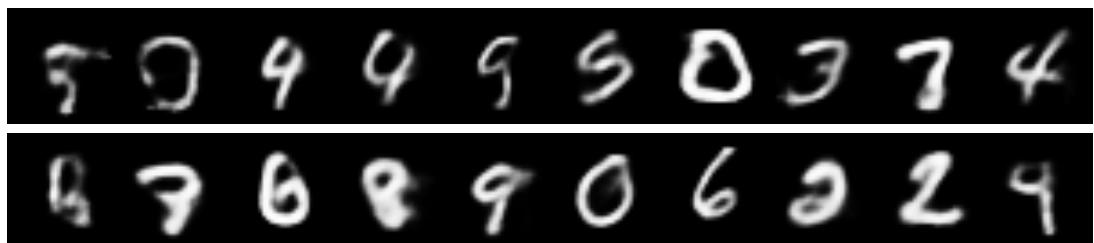
VQ-VAE 是一种结合了向量量化和变分自编码器的神经网络模型。该模型结合了两者的优势，实现了更强大的表示学习和生成模型的性能。VQ-VAE 引入了向量量化的概念，其中潜在特征表示被映射到称为码本的一组离散的向量上，而不是连续的值。在训练过程中，将潜在表示分配给离散的码本向量，使得模型学会一种数据的离散表示方式。

与传统 VAE 相似，VQ-VAE 的编码器将输入数据映射到潜在表示，但输出的是最接近的码本向量的索引而不是连续的值。编码器的输出是一个代表码本索引的离散值。VQ-VAE 包含一个离散码本，其中存储了一组离散的向量，这些向量被用于量化潜在表示。潜在表示通过比较与码本中向量的距离来进行向量量化，选择最接近的码本向量作为最终的离散表示。VQ-VAE 的损失函数包括来自于变分自编码器的重构误差项和来自于向量量化的码本匹配项。码本匹配项通常使用均方差来衡量潜在表示与最近码本向量的差异。解码器接收离散的潜在表示，并将其解码为生成的数据。与传统 VAE 一样，解码器的结构与编码器相对应。



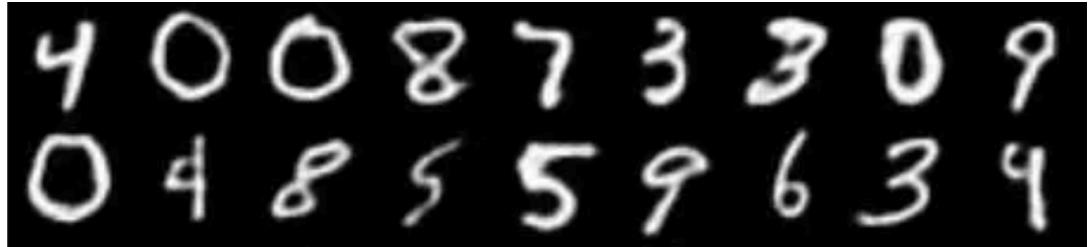
VQ-VAE 通过引入向量量化的机制，使得模型能够学习具有离散结构的表示，从而更有效地捕捉数据中的结构信息。

作为效果对比，下面是使用 VAE 在正态分布中采样生成的数据：



可以看到，在 VAE 随机采样生成的图片中产生了很多不是数字的图形，比如第一行第 1 张和第 6 张图片，第二行第 1 张、第 2 张、第 8 张图片，平均每 10 张图片里面至少有两三张图片完全不是数字的形态。

下面是使用 VQ-VAE 在随机采样后生成的数据：

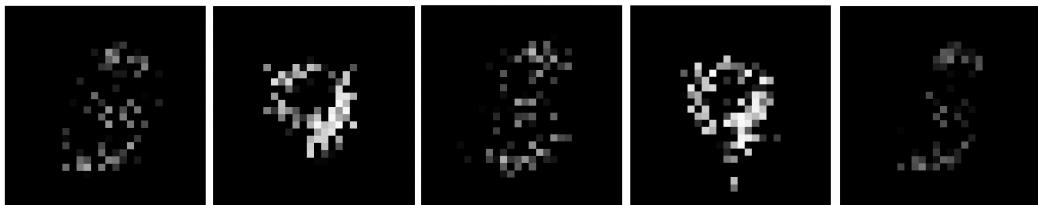


可以看到，相比 VAE，VQ-VAE 生成非数字形态的图片更少，大部分数字的形状都更规则，这应该是得益于 VQ-VAE 将隐空间离散化后有了更好的结构性。

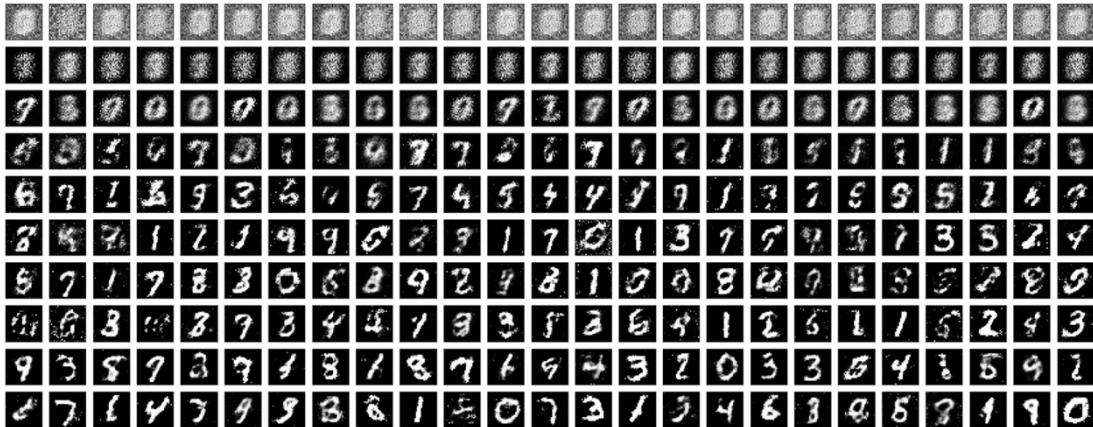
4、GAN (Generative Adversarial Nets，生成对抗网络)

生成对抗网络 (GAN) 是一种深度学习模型，由生成器和判别器组成，通过对抗训练的方式共同学习。生成器负责从随机噪声生成逼真的数据，而判别器则致力于区分生成器生成的假数据和真实数据。在训练中，生成器迭代改进生成样本的逼真程度，同时判别器努力提高对真伪数据的分类准确度。GAN 的损失函数通常由两部分组成：生成器的损失和判别器的损失。生成器的损失旨在最小化生成的数据与真实数据之间的差异，使生成器生成逼真的样本。判别器的损失旨在最大化对真实数据的正确分类以及对生成数据的正确分类，使其更好地区分真伪。这种对抗过程使生成器不断提高生成样本的逼真度，判别器也不断提高识别真伪的能力，最终达到生成逼真样本的目标。训练结束后，生成器可以使用随机噪声生成逼真的样本，这些样本将与训练数据相似。

下面是一些 GAN 的训练后的采样结果：



可以看到 GAN 生成的效果并不好，生成的图像只能勉强大致分辨出其内容，在通过多次改动模型调整参数后也仍然无法取得较好效果，通过查阅资料发现主要的原因是训练轮数不够，下图为网络上的一个 GAN 实验：



想要获得最后一行的效果需要 300 轮以上的迭代，由于硬件条件的限制以及 GAN 训练

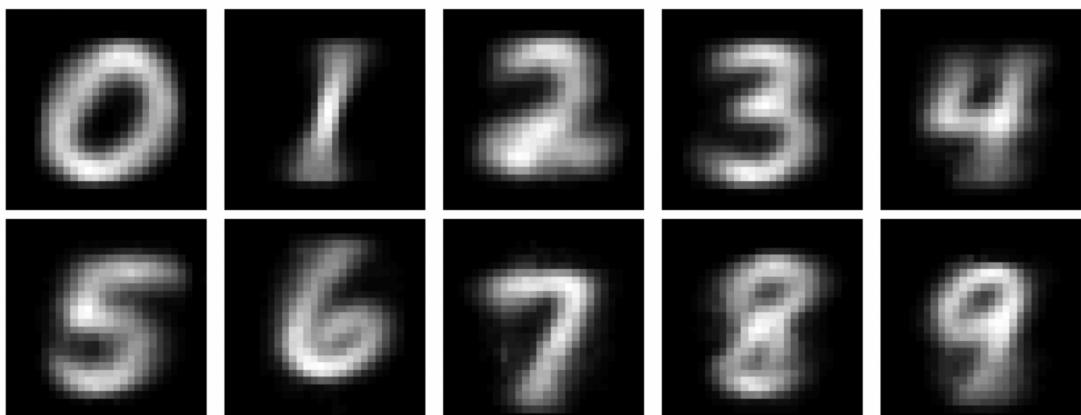
时间花费较长，我们这里没有进行重复。同时我们也可以看到，即使在经过了大量训练后，GAN 的生成效果仍然不是很好

三、新尝试

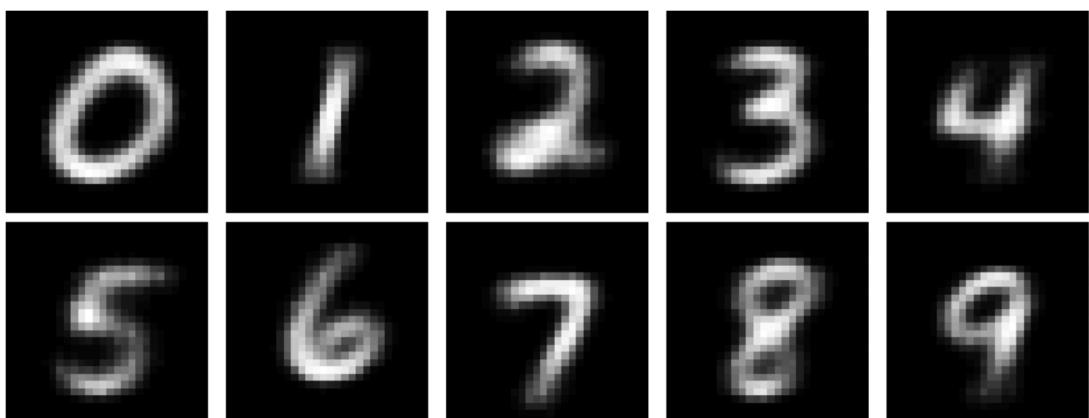
尝试 1

我们知道在 AE 模型中，在隐空间中随机采样容易生成乱码，如果要生成特定数字我们可以通过数据的标签得到各个类别在隐空间中的特征中心，在对应的特征中心采样能够获得对应的数字，或者通过将图片输入编码器得到特征值，再在特征值附近采样，但是这个方法图片没有什么变化，采样的意义不大。

这里的改进是，既然为了采样特定类别的数字我们通过标签进行了一个聚类，或许我们可以直接在训练时就让某一类的图片的特征都集中在某个空间内。于是我们的模型直接去掉了编码器，通过从人工设计的编码空间中采样来对解码器进行训练，比如对于某张数字 1 的图片，我们会直接对 $[1,1,1, \dots, 1,1,1]$ 这个向量添加一个噪声，然后将这个向量作为特征值输入到解码器，通过损失函数和梯度下降，使得在 $[1,1,1, \dots, 1,1,1]$ 附近的特征值生成出来的图片其形状越接近数字 1 的形状越好。同理对于数字 2 我们就在 $[2,2,2, \dots, 2,2,2]$ 附近采样，以此类推。最后训练出的效果如下所示，我们需要生成某个数字的图片只要在对应的向量附近进行采样即可，不需要进行聚类来寻找：



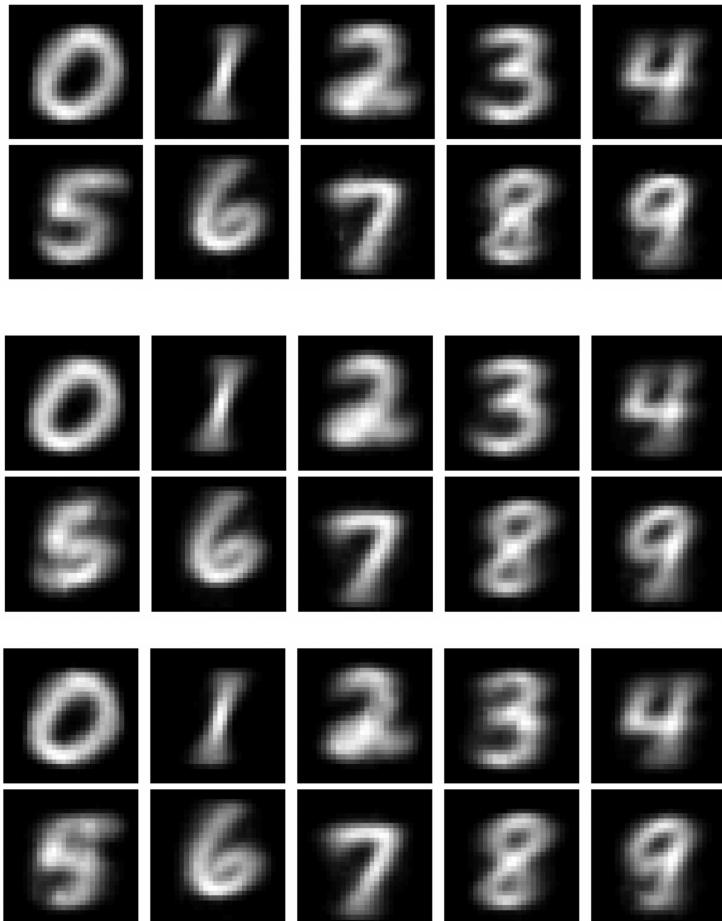
作为对比，下面是原始 AE 训练出的结果：



可以看到，这种方法生成的图片效果在某种程度上比原来的 AE 的生成效果还好一点，不过跟 VAE 的效果还是相差很大的，VAE 的先验远更丰富，而且 VAE 所需的训练时长

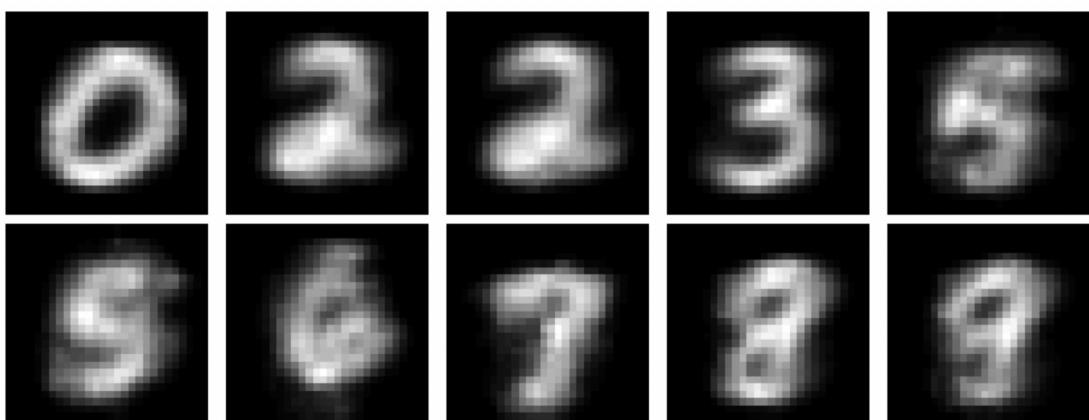
也是 AE 的几倍。

我们对输入的向量增加一些噪声，得到的效果如下：



可以看到，随着噪声的增加能有一些微弱的改变，比如 5 的形态变化稍微明显一些，当然这远比不上 VAE 的变化能力，但是新方法不会再有原来 AE 中加噪后图片会变得模糊破碎的情况。

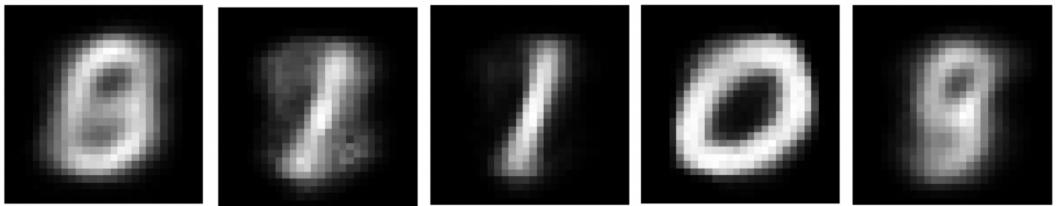
如果我们继续加噪的话，数字可能会变成相邻的数字，如下图所示，因为我们隐空间的设计是特征值分布在形为的向量 $[x,x,x,\dots,x,x,x]$ 的附近，所以形状是一条直线。通过这个实验也说明了我们是可以根据需要人为地设计隐空间的分布情况的。



尝试 2

第二个改进是将模型的编码器用 PCA 代替。我们看到对于生成模型而言，可以分成两类，一类是像 GAN 和 Diffusion Model 这类从完全的噪声中生成图像，另一类是像 AE 和 VAE 这样从一个特征空间中生成图像。对于 GAN 和 DM 这类模型，我们是比较难通过输入噪声的不同来控制生成相关结果的，想要控制生成结果的不同可能需要 transformer 等方法的辅助，比如文生图模型。而 AE 和 VAE 这种模型其本身也可以作为特征提取模型和聚类模型使用，在生成时可以通过对隐空间的调整来生成不同的结果。我们的想法是能否通过已有的可靠特征提取方式，直接将数据映射到特征空间，然后直接从一个教好的隐空间中学习解码。

对于 AE，我们使用 PCA 将数据映射到 3 为空间，然后从这个 3 维空间中解码进行训练，训练完成后在 3 维空间随机采样的效果如下所示：

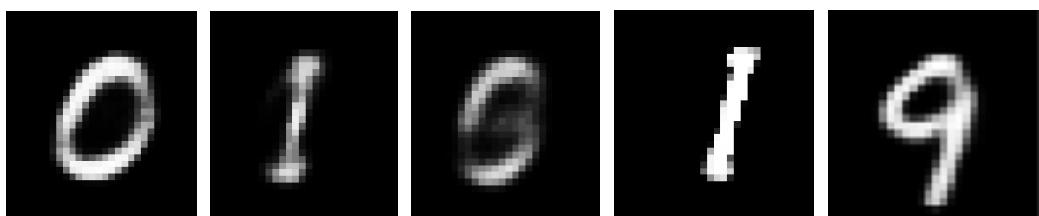


可以看到，原本在 AE 中是无法进行随机采样的，随机采样会生成之前我们展示的乱码，但是通过 PCA 映射之后，我们现在能够进行随机采样生成数字的图片，当然也会有一些生成的不好的结果，比如下面这些：



但是现在这些生成的图片都是有规则的图案，不再是乱码了，这使得我们有了一个易于采样的隐空间。

对于 VAE，从整体分布中随机采样生成的图片如下所示：

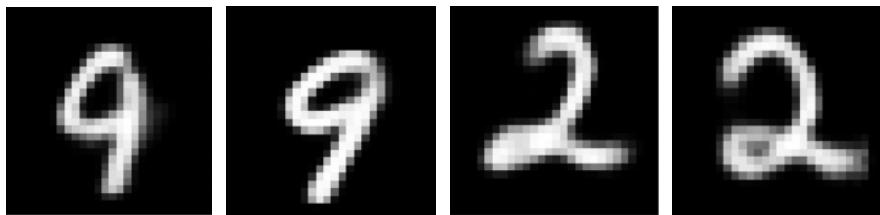


可以看到 VAE 生成出的图像的效果比 AE 更加清晰一些，和之前不使用 PCA 时的随机效果相比，使用 PCA 后生成的数字图像更加规则，而不使用 PCA 时经常会生成一些和数字无关的图像，但是使用 PCA 的隐空间更加单调，数字形态比较单一，多次随机采样生成得到的图像经常差不多，虽然生成的图像更加一致，但是丧失了多样性。

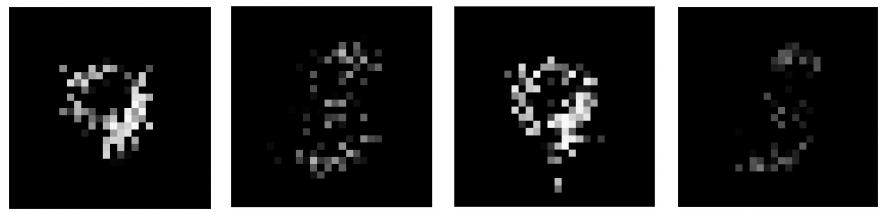
四、总结

通过对比，几个模型中最好的生成方法是 VAE 加上标签辅助获得聚类中心，将聚类中

心输入解码器即可生成出非常标准的数字图案，添加噪声后即可对数字形态进行改变。



效果最差的是 GAN 模型，在同样的训练资源下无法通过 GAN 模型训练出较好的数字图片，其在此任务上耗时最长且效果最差。



通过分析，认为可能的原因是手写数字图像的图形中具有较强的逻辑性，且 10 个数字类别差异很大。通过网上的一些已有实验发现，GAN 模型对于人脸、风景等这些相同性高、图形没有较强的对错之分、可改动范围较大的数据类型有着较好的效果。而对于数字图像，训练过程中判别器损失很快就降得比较低，生成器较难从噪声中尝试出数字图形。我们在实验中也尝试了给生成器添加重构损失来帮助生成器的生成，实验还尝试了通过人为设定不同类型的噪声来尝试通过控制噪声从而生成不同的数据，从而实现 GAN 的可控生成，但是由于 GAN 的训练消耗资源较大，硬件水平有限，只进行了几轮测试，由于没有太多的改进机会，最后结果不是很理想，所以没有写进报告里，将在假期进一步实验。

在几种模型中，训练耗时最短的是 AE 模型，实践中只需较短的时间通过几轮训练就已经能有较小的损失达到一定的效果。我们看到，手写数字数据集由于其图形的逻辑性强，有非常明确的结构和较为鲜明的对错差异，从隐空间中采样的模型更有可能比从噪声中采样的模型有着更好的效果，因为隐空间中带有数据的特征信息。通过改进 AE 模型的隐空间我们使得其生成有了更好的效果，与此同时改进后的 AE 仍然训练耗时非常低，训练中每次都是较短时间几轮训练即可完成，所以如果能够通过先验信息构建出较好的生成结构，AE 是能够同时做到较少的训练消耗的同时又有较好的生成效果的，这或许会在某些特定的高消耗任务中有所帮助。

如今主流的生成模型已经是 CLIP、LDM 等新架构，但是在这些新模型中仍然会使用到自编码器这些结构，这些经典模型已常被当作关键部件进行使用，所以能够通过这次实验对这些模型进行实操还是非常有意义的，由于时间和硬件条件的限制还有很多模型没有进行实验，将会在后面继续完成。