队伍：刺客五六七　Rank 6

- 赛题回顾

利用已知的小规模无线信道数据集，通过基于 AI 的数据生成方案，构建大规模无线信道数据集，同时，综合考察生成的大规模数据集与真实数据的相似程度、离散程度。

- 设计思路

针对赛题，我们有以下几个思考：

训练数据集是小规模的，因此模型、算法的设计时需要能适用小规模数据，这里主要有两个思路，一个是扩充训练数据集，比如利用一些数据增强的方式 ；第二个思路是控制模型的容量，数据集不多的情况下用大模型可能只是在过拟合，不能一味怼大模型。

考察生成的大规模数据集与真实数据的相似程度、离散程度，要做到既相似又离散往往是比较难的，提高一个指标的同时大概率会导致另一个指标下降，应该合理权衡两个指标。

- 模型架构

在前期的调研中我们了解到，无线通信领域经常需要在用户端对信道信息进行压缩，发送压缩比特流到基站端，基站通过对压缩比特流进行解码，恢复出原始信道。也就是说，从压缩比特流来生成信号其实是工业界中比较认可且广泛使用的一种方式。鉴于此，我们采用自编码器的结构来训练模型，然后仅需要随机生成比特流送入解码器就可以生成信号了。问题转化为自编码器结构的设计，包含编码器和解码器。

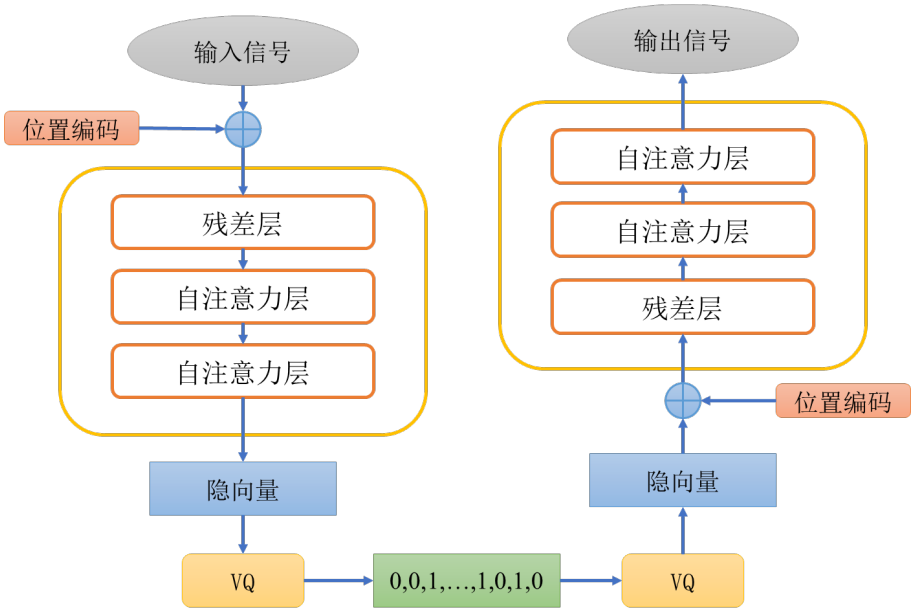
编码器和解码器模型设计的时候我们经过了以下历程：

初期使用官方 baseline 采用的卷积模型，卷积更关注局部，且具有“平滑”作用，生成出来的信号更加“平滑”且缺乏一些细节，这也导致训练出来的模型去生成样本的结果相似度较低、离散程度较高。

中期模型迭代时，为了提高生成样本的相似度，我们抛弃了卷积模型，转向 MLP、Transformer，训练时明显发现生成样本的相似度有了较大的提高，但是离散程度缺较差，我们的理解是 MLP、Transformer 本质上是对信号的每一个维度进行运算，在数据量较少的情况下容易过拟合，也就是把细节学得太好了，所以相似度就上去了，但是过拟合导致离散程度也变差了。

后期模型其实是上面两种思想的结合，我们既考虑了局部视野模型（残差网络），又考

虑了全局视野模型（注意力网络），最终的模型结构如下图所示



详见代码 qae_main.py 中的 Encoder 类和 Decoder 类。

- 数据处理流程
  - ✓ 数据预处理

    在设计思路时有需要考虑到训练数据的小规模问题，观察数据时我们有以下发现

    1. 数据集 1 每 5 个数据为一组，数据集 2 每 20 个数据为一组，组内相似度较高
    2. 由于信号是复数，对复数的实部虚部进行交换不会改变相似度

因此我们设计了如下两种数据增强模型

1. 对于每一组数据，任意选择其中两个信号，插值得到一个新的信号，可以理解二维平面上两个点连线的中点作为一个新样本
2. 对于复数信号进行取相反数，实部虚部交换等操作

详见代码 dataset.py 中的 load_data 函数

- ✓ 数据后处理

针对相似程度、离散程度的权衡问题，我们发现真实样本通常具有较高的方差，而模型生成的样本包含许多低方差样本，因此通过对生成信号进行后处理，以方差进行筛选，可以有效提高生成数据的相似度，当然这也会导致离散程度变差，但是权衡起来整体效果还是更好的。我们团队在数据集 2 也取得了 sim 为 0.38765134 的成绩（全部队伍中最高）。

详见 generatorFun.py 中我们对数据集 2 的生成样本的后处理。（数据集 1 我们也进行过同样操作，但是由于数据集 1 本身模型生成的结果相似度就比较高，离散程度也较好，采用后

处理后离散程度变得太差，得不偿失）

- 训练超参数

训练超参数这里主要有网络的层数，如设计思路所说，数据量有限时不能一味怼模型容量，我们也尝试过将残差网络、注意力网络加深，然而没有取得更好的结果。

最终方案为一层残差网络，两层注意力网络。

在模型以及优化器等其他超参数上我们并没有进行搜索调参，而是遵循一些常规的习惯。