

1. 环境

```
numpy==2.4.0
pandas==2.3.3
scikit_learn==1.8.0
scipy==1.16.3
torch==2.9.1+cu126
```

2. 重构

- 设置config配置文件：便于后期包装成函数或者修改模型参数
- 包装全流程：将函数全流程包装成`DeepETD_Train`和`DeepETD_Predict`两个函数
- 修改了一些源代码中的报错并增加部分函数的注释

3. 结果

最终得到的模型本模型在17个epoch后因早停机制终止训练，最佳验证集AUC为0.8200（第7个epoch）。

阶段	最佳AUC	对应Loss	准确率
训练集	0.9530	0.3651	0.8726
验证集	0.8200	0.7342	0.7471

注：由于早停机制，最佳验证集性能出现在epoch 7，而训练集性能在最终epoch(17)达到最佳。

训练过程分析过程参数发现以下问题：

- **验证集表现先升后降**：验证AUC在前7个epoch持续上升至峰值0.8200，随后进入平台期并略有下降，显示出典型的过拟合迹象。
- **内存使用稳定**：GPU内存占用始终保持在42.6MB，模型规模适中，可考虑添加其他维度参数，扩充多模态数据

4. 思考

- **数据维度过小**：最开始考虑用GPU加速，后来使用后发现每个Epoch里面，平均GPU使用只有0.00几s，而且在测试过程中发现了过拟合的现象，推测是模型参数量太小并且数据维度比较低，考虑添加中药网络药理学的特征数据进去，也就是纳入辩证-方剂-中药-化合物-靶点的信息进去，一起丢进去学习，可以增加维度并且大大提升数据量，只是有这个想法，暂时还没想清楚应该如何解决

- 关于text文档：最开始对text文档理解错误，随便尝试了一个化合物，做了一个其相关信息的json文件进去预测，但是预测效果很不好，考虑还是过拟合问题。现在对text文档的作用理解是：“输入化合物及其靶点的名称及其相关信息，可用预训练过的模型进行预测，得到其相互作用分数”