数据科学导论实验 7：ESI 深度学习排名预测 V2

姓名：杨云天

学号：10245501405

1. 数据说明

学科总数：22 个
数据特征：
- Web of Science Documents
- Cites
- Top Papers
- 机构在该学科的全球排名

- 忽略 Cites/Paper 原因：篇均引用可以通过 Cites 与 Web of Science Documents 计算获得，相关性极高，无需选取。
数据形状：(row_size, 4)

2. 数据集划分

训练集：60%
验证集：20%
测试集：20%
注: 每个学科的数据在划分前进行随机打乱

3. 数据预处理

3.1. 数据归一化
- 排名：反向归一化到 [0,1]，排名越靠前（数字越小），归一化值越接近 1
- 特征：正向归一化到 [0,1]
- 使用 sklearn 的 MinMaxScaler
- 归一化原因：量纲差异会导致数值大的特征在计算中占据主导地位，例如引用量（万级）会掩盖顶级论文数（百级）。对于回归模型，量纲差异会影响距离计算或系数估计，导致结果偏向数值范围大的特征，无法客观反映各特征的真实重要性。

3.2. 反归一化:
预先保存所有学科的排名范围，便于后续将预测的归一化值转换回实际排名。

4. 深度学习模型说明

4.1. 模型架构

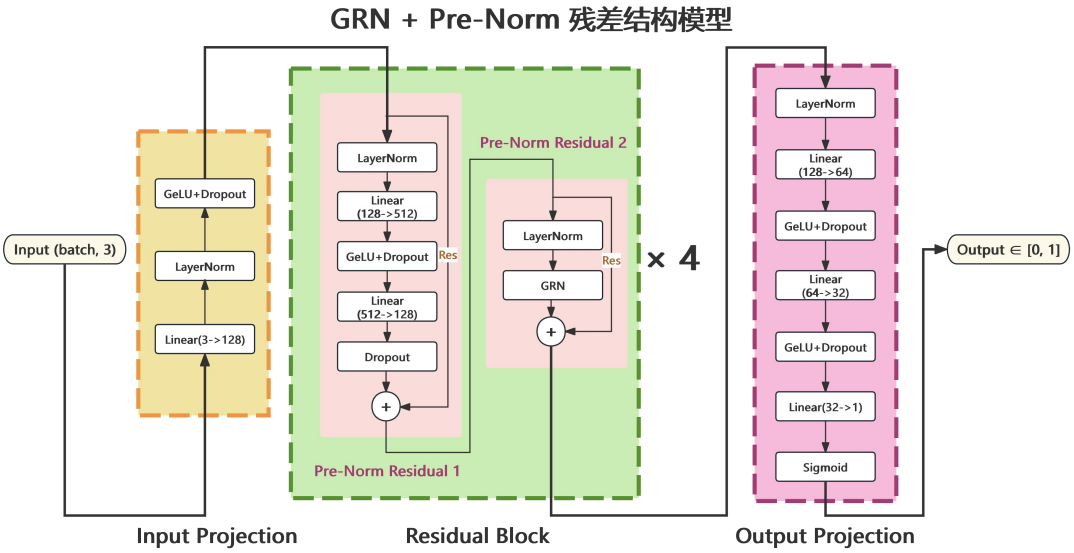
（1） 主干网络结构表：GRN + Pre-Norm 残差模型

操作	输入形状	输出形状	说明
线性层 1（输入投影）	(batch, 3)	(batch, 128)	将输入特征（3 个）映射到隐藏空间
LayerNorm	(batch, 128)	(batch, 128)	对输入特征标准化，稳定训练过程
GELU 激活	(batch, 128)	(batch, 128)	平滑非线性激活，梯度连续
Dropout (0.2)	(batch, 128)	(batch, 128)	随机丢弃特征，防止过拟合
Residual Block*4	(batch, 128)	(batch, 128)	四层堆叠的残差结构模块（每层内部结构见下表）
LayerNorm（输出前）	(batch, 128)	(batch, 128)	输出前再次标准化，防止分布漂移
Linear (128 → 64)	(batch, 128)	(batch, 64)	降维提取核心特征
GELU 激活	(batch, 64)	(batch, 64)	非线性转换
Dropout (0.2)	(batch, 64)	(batch, 64)	防止过拟合
Linear (64 → 32)	(batch, 64)	(batch, 32)	进一步降维
GELU 激活	(batch, 32)	(batch, 32)	稳定激活函数
Dropout (0.1)	(batch, 32)	(batch, 32)	轻微正则化
输出层 Linear (32 → 1)	(batch, 32)	(batch, 1)	输出单个预测值
Sigmoid **激活**	(batch, 1)	(batch, 1)	将结果映射到[0,1]区间（归一化预测）

（2） 残差块结构与连接表

操作	输入形状	输出形状	说明
LayerNorm(Pre-Norm 1)	(batch, 128)	(batch, 128)	对输入进行归一化，避免梯度爆炸/消失
Linear (128 → 512)	(batch, 128)	(batch, 512)	扩展维度 *4，用于特征增强
GELU 激活	(batch, 512)	(batch, 512)	平滑非线性特征提取
Dropout (0.2)	(batch, 512)	(batch, 512)	防止过拟合
Linear (512 → 128)	(batch, 512)	(batch, 128)	将特征映射回原维度
Dropout (0.2)	(batch, 128)	(batch, 128)	保持输出稳定性
+ 残差连接 1：x = x + residual	(batch, 128) +(batch, 128)	(batch, 128)	第一层残差路径，增强梯度流动性
LayerNorm(Pre-Norm 2)	(batch, 128)	(batch, 128)	第二次归一化，为 GRN 做准备
Global Response Normalization (GRN)	(batch, 128)	(batch, 128)	动态缩放特征响应，增强特征多样性
+ 残差连接 2：x = x + residual	(batch, 128) +(batch, 128)	(batch, 128)	第二层残差路径，输出稳定且信息完整

4.2. 模型架构示意图



5. 模型训练

5.1. 训练目标

针对每个学科单独训练一个 RankingPredictor 模型，该模型基于深度学习的全连接网络架构，结合 Pre-Norm 残差结构（Residual Block）与全局响应归一化（Global Response Normalization, GRN），以更好地适应表格型特征数据的分布差异。每个学科对应一个独立模型，以实现跨学科的定制化排名预测。

5.2. 数据处理

各学科数据独立划分训练集 (X_train, y_train)、验证集 (X_val, y_val) 和测试集 (X_test, y_test)。

所有特征在输入模型前进行了归一化处理，保证各输入维度的数值分布稳定。目标变量（排名）采用反向归一化策略，使得“排名越靠前”对应数值越接近 1。

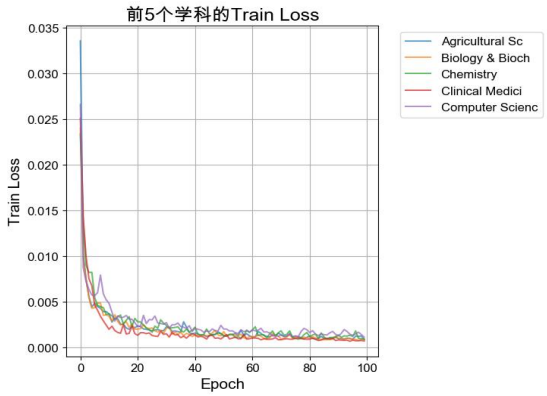
所有特征与标签均转换为 PyTorch 张量。

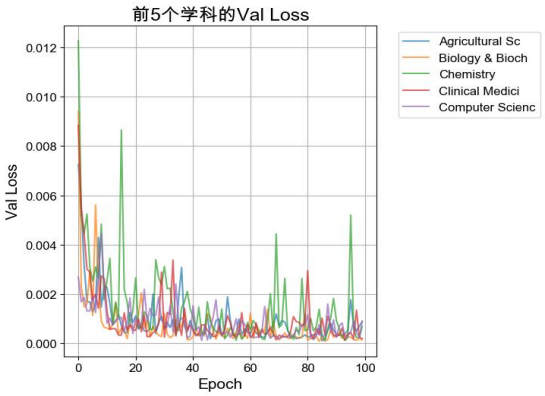
5.3. 训练过程

- 损失函数：均方误差损失（nn.MSELoss），适用于回归任务。
- 优化器：Adam 优化器，学习率 lr=0.001，自适应调整参数更新步长。
- 训练轮次（epochs）：100 轮，每 20 轮输出一次训练与验证损失。
- 批处理大小（batch_size）：32
- 训练流程

① 清空优化器梯度（optimizer.zero_grad()）
② 前向传播计算预测值
③ 计算损失并反向传播（loss.backward()）
④ 优化器更新参数（optimizer.step()）

- 验证阶段：关闭梯度计算（torch.no_grad()），计算验证集损失以监控过拟合风险。
- 记录数据：记录 train loss 和 val loss（下图）





5.4. 模型对比分析

对比项	模型 V1 （MLP 多层感知器）	模型 V2 （GRN + 残差结构）	改进效果
输入层结构	Linear(3→64) + ReLU + Dropout	Linear(3→128) + LayerNorm + GELU + Dropout	提升输入特征稳定性与非线性表达能力
隐藏层设计	3 层全连接，逐步缩小维度 (64→32→16)	4 个残差块，每个包含双残差路径与 GRN 模块	提高深层特征传递与梯度流动性，残差结构防止信息丢失，适合多维相关性强的表格特征学习
归一化方法	无归一化层	LayerNorm + GRN	GRN 动态调节特征响应，增强模型在不同特征尺度下的稳定性
激活函数	ReLU（硬激活）	GELU（平滑激活）	在表格数据中可捕获连续特征间的细微变化，训练更稳定
参数规模	约 6K 参数	约 40K 参数	增强模型容量，捕获复杂非线性关系，更大的参数空间可建模复杂非线性关系，提升对特征交互的表达能力
泛化能力（测试集表现）	易过拟合	泛化良好	在不同学科数据上均保持稳健性能

5.5. 模型保存

模型保存为每个学科保存训练完成的模型，model_V2/subject_name.csv

6. 评价指标

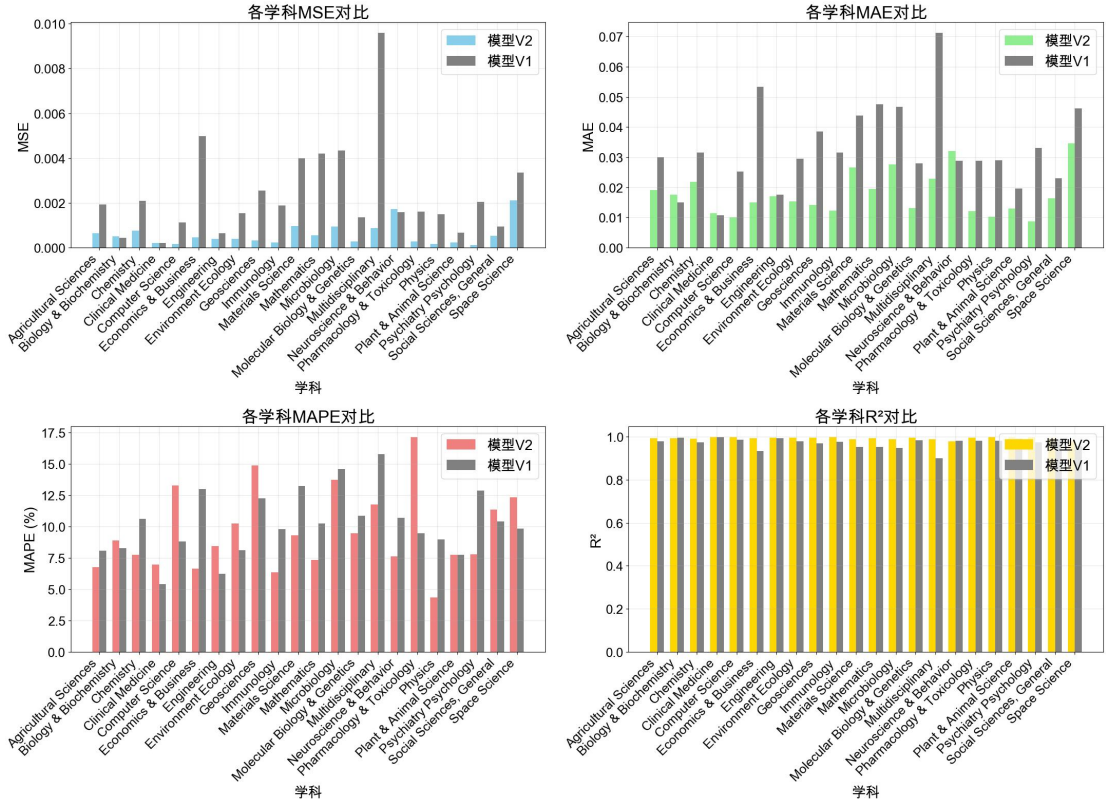
6.1. 各项评价指标含义

指标名称	含义	数值
MAE：平均绝对误差	预测排名与实际排名的平均差距	越小越好
MAE：平均绝对误差	预测值与实际值差值的绝对值的平均值，反映预测值与实际值的平均差距	越小越好
MAPE(%)：平均绝对百分比误差	预测值与实际值（反归一化后）差值的绝对值除以实际值后取平均值	为百分数形式，越小越好，超过 50%表示误差较大
R^2：决定系数	反映模型对数据变异的解释能力	取值范围为(-∞,1]，1 表示模型完全拟合数据，0 表示模型效果等同于均值预测，越接近 1 越好

6.2. 训练结果：各项指标分析

	学科名称	MSE	MAE	MAPE(%)	R^2
0	Agricultural Sciences	0.000639	0.019033	6.74	0.992066
1	Biology & Biochemistry	0.000495	0.017521	8.87	0.994204
2	Chemistry	0.000756	0.021873	7.75	0.990944
3	Clinical Medicine	0.000212	0.011463	6.94	0.997460
4	Computer Science	0.000170	0.010119	13.29	0.998047
5	Economics & Business	0.000458	0.014929	6.61	0.994207
6	Engineering	0.000399	0.017011	8.44	0.994869
7	Environment Ecology	0.000397	0.015357	10.23	0.995410
8	Geosciences	0.000325	0.014088	14.86	0.996306
9	Immunology	0.000227	0.012251	6.33	0.997167
10	Materials Science	0.000970	0.026650	9.28	0.988654
11	Mathematics	0.000543	0.019407	7.33	0.993389
12	Microbiology	0.000947	0.027592	13.70	0.988830
13	Molecular Biology & Genetics	0.000275	0.013180	9.46	0.996690
14	Multidisciplinary	0.000870	0.022868	11.73	0.988926
15	Neuroscience & Behavior	0.001720	0.031997	7.61	0.979989
16	Pharmacology & Toxicology	0.000268	0.012071	17.13	0.996772
17	Physics	0.000162	0.010271	4.34	0.998300
18	Plant & Animal Science	0.000227	0.012941	7.73	0.997293
19	Psychiatry Psychology	0.000108	0.008639	7.76	0.998682
20	Social Sciences, General	0.000517	0.016379	11.33	0.993701
21	Space Science	0.002106	0.034630	12.32	0.977728
22	平均值	0.000581	0.017739	9.54	0.993165

本模型（模型 V2）与上一版模型（模型 V1）进行对比：



分析：

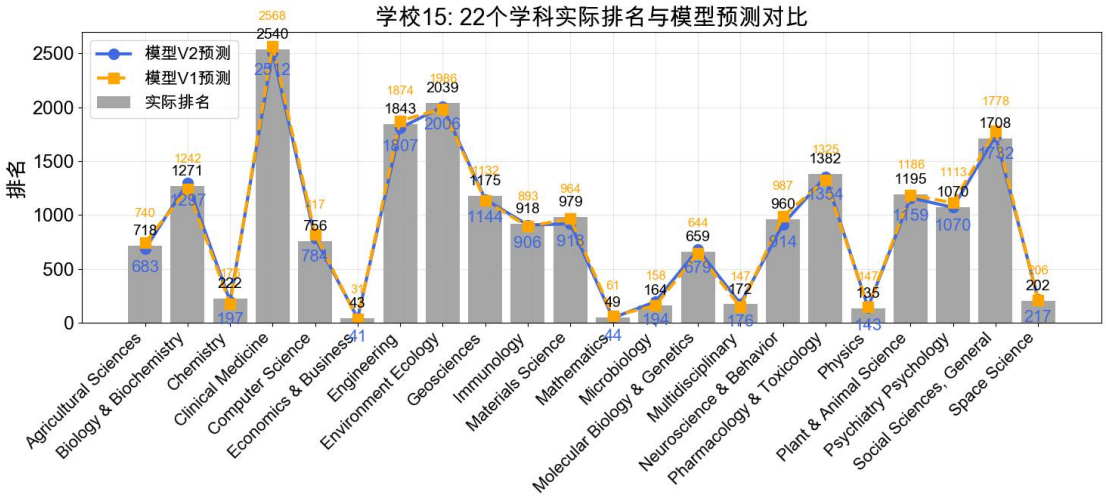
- MSE：模型 V2 在所有学科上显著低于模型 V1，说明预测误差平方项更小，整体拟合更精确。
- MAE：所有学科模型 V2 显著低于模型 V1，表示平均绝对误差更低，预测偏差更小。
- MAPE：在大多数学科中，模型 V2 均低于模型 V1，说明模型在不同尺度数据下保持了稳定的相对误差表现，提升泛化一致性。存在部分学科模型 V2 劣于模型 V1，原因可能为：模型 V2 中引入了 GRN 和多层残差结构，提升了整体表达能力，但也可能在样本量较小或特征噪声较高的学科中引入过拟合，导致局部预测偏差。
- R^2：两个模型普遍接近 1，说明两个模型拥有解释力强，对各学科排名预测的相关性高的特点，模型 V2 略高于模型 V1。
- 综上，模型 V2 在精度、稳健性与泛化性能上均优于模型 V1。

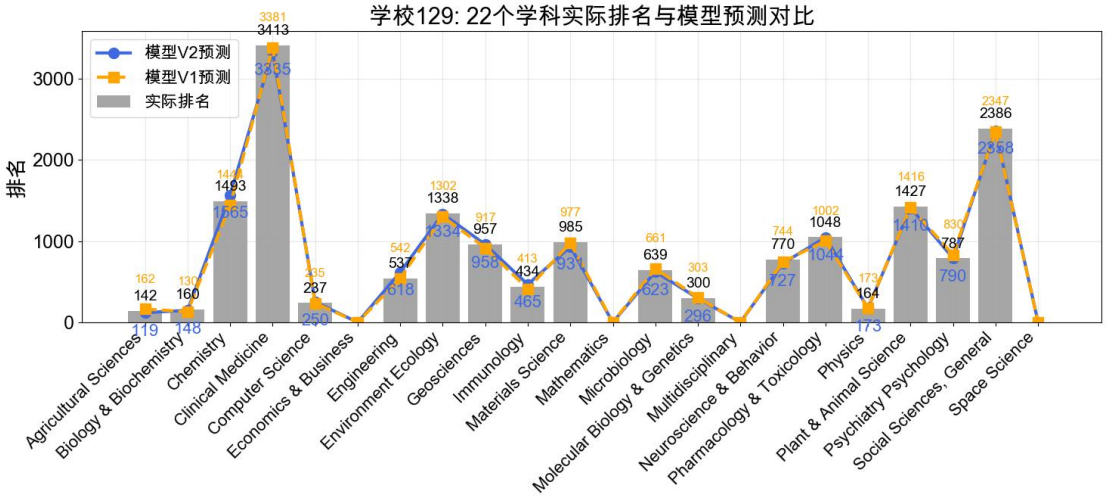
- 总结

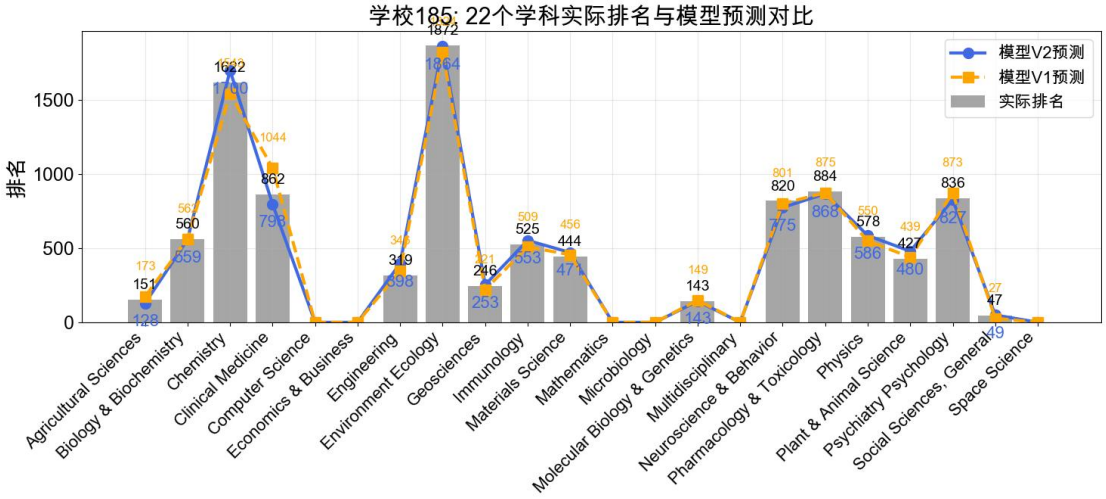
该排名预测模型（模型 V2）在四个核心评估指标上均展现出优于模型 V1（MLP 多层感知器）较好性能。

7. 预测

从测试集中随机选择三个样本，分别通过本模型（模型 V2）与上一版模型（模型 V1）对 22 个学科的进行排名预测







22 个学科所有学校 V2 模型累计排名误差: 1468

22 个学科所有学校 V1 模型累计排名误差: 1631