

# 一斤鸭梨队（现第九名）答辩PPT分享

潼关君

© 2022-09-23 16:37:59

0

22

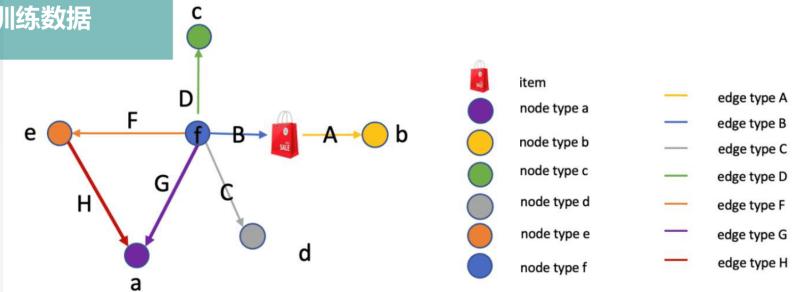
## 赛题分析

**比赛要求：**在大规模的异构图数据集中检测出异常商品。

数据集特点：

1. 噪声很大

2. 测试数据的正负样本比例远低于训练数据

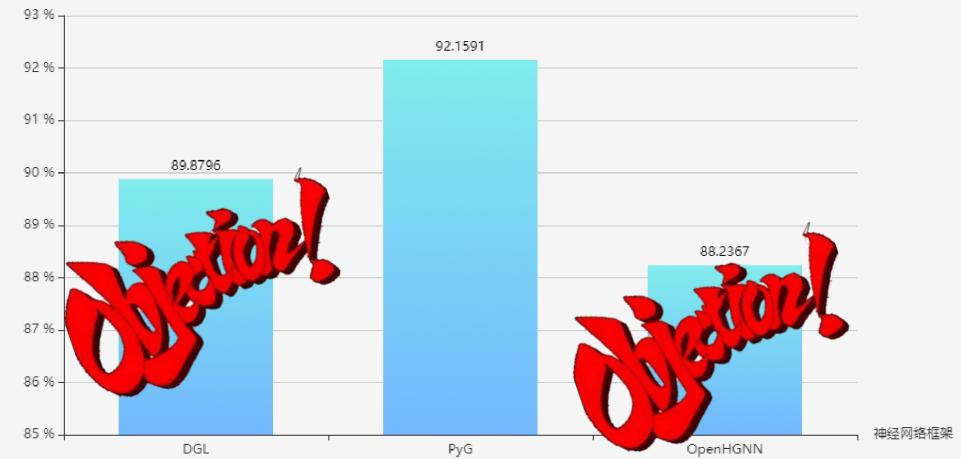


enter image description here

## 框架选择

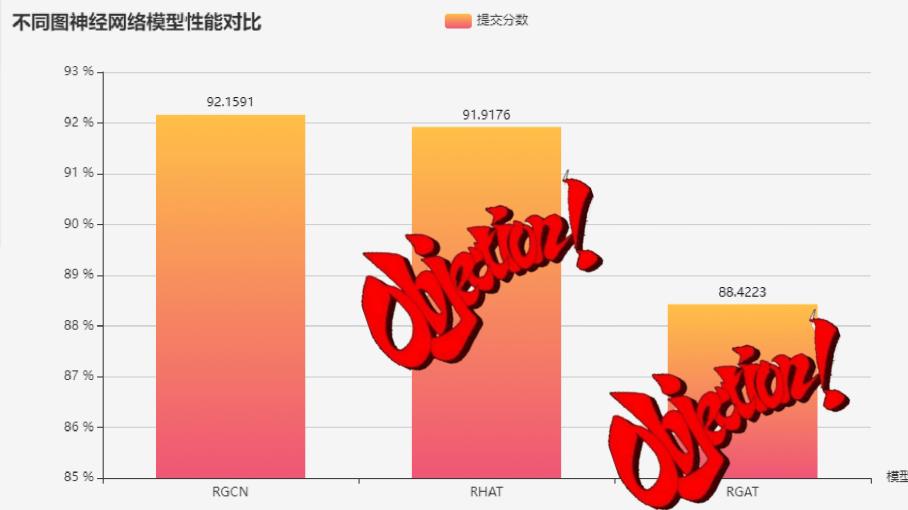
不同框架baseline提交分数对比

提交分数



对比不同框架的性能，我们最终选择了pyg图神经网络框架

## 模型选择



复杂的图结构反而使成绩下降了。

我们认为可能是数据噪声太大，导致过于复杂的模型不能发挥其在论文中的效果；而且，复杂的图结构更容易对噪声过拟合，从而在测试集上效果不佳。

## 策略改进

baseline

k折交叉验证

根据以往机器学习的经验，  
我们尝试了k折交叉验证，然  
而效果不理想

2层RGCN编码器+1层  
RGCN解码器+下游分  
类

预训练了121轮后接下游任务，  
用线性层做简单的二分类

自监督

使用给定的数据集节点训练  
自编码器，使编码器能够完  
全学习到网络特征；  
构建下游分类任务，精调预  
训练出来的模型。

4层RGCN编码器+1层  
RGCN解码器+下游分  
类

将编码器的层数逐步加到4层，  
效果有所提升

## 模型策略

PART 1

## 自编码器

PART 2

## 下游分类

训练自编码器

使用给定的数据集节点训练自编码器（编码器-解码器），使编码器能够完全学习到网络特征

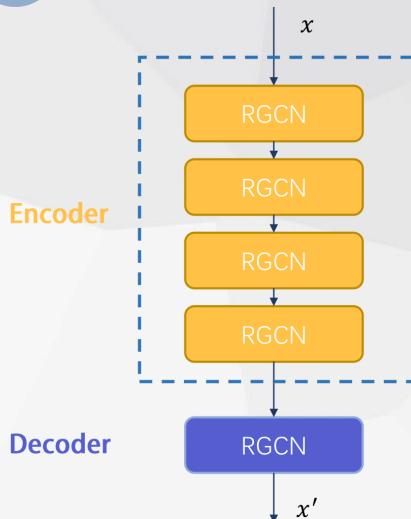
构建下游分类任务

使用训练得到的编码器构建下游分类任务，精调预训练出来的模型

预测

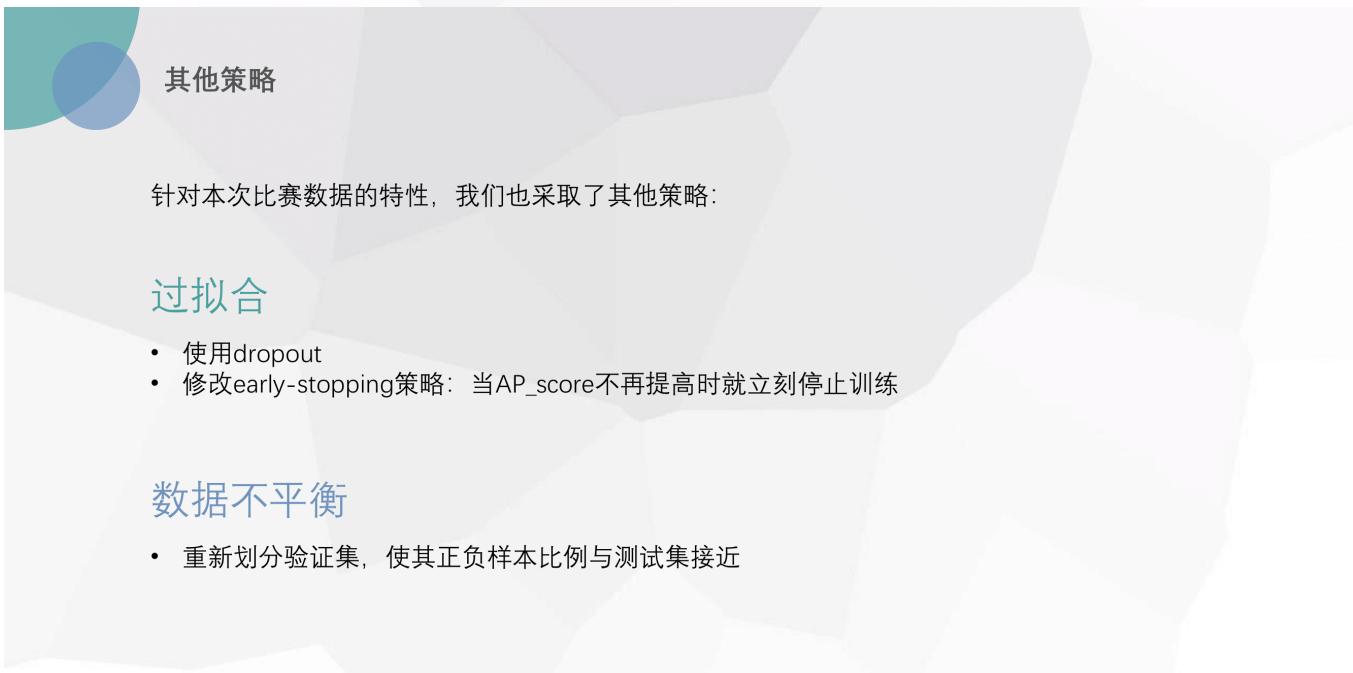
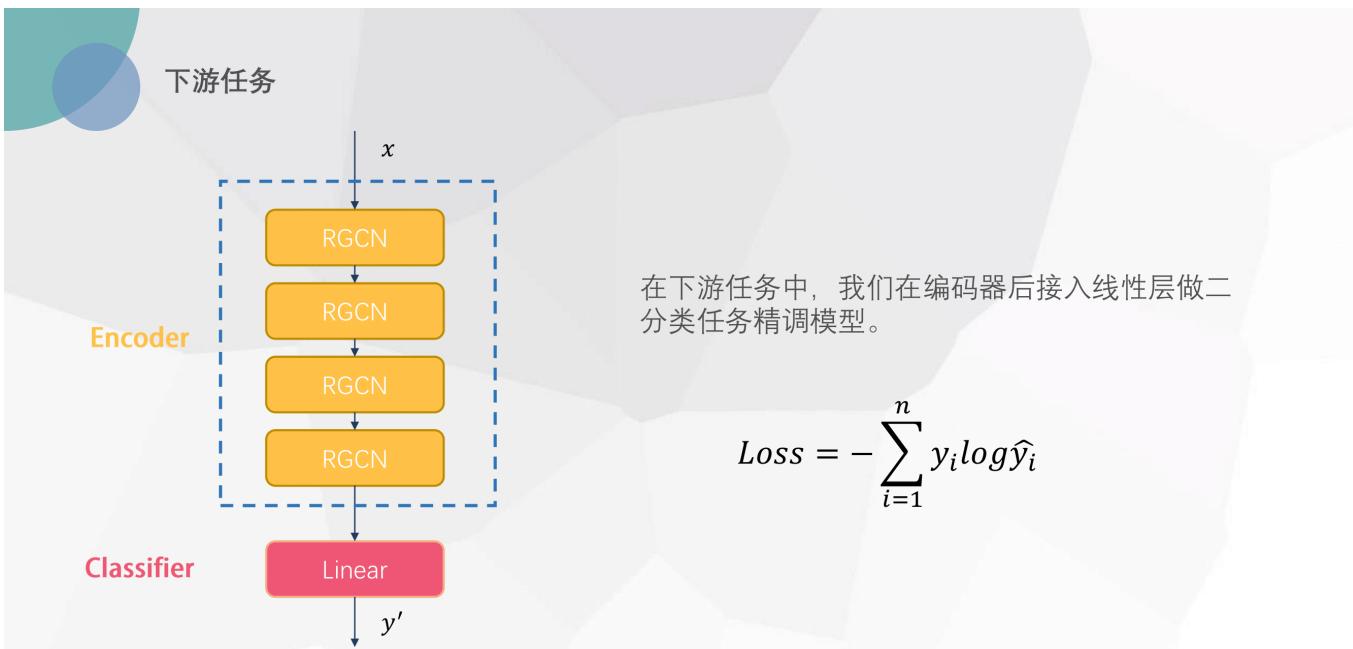
使用下游任务训练出来的模型进行预测

## 自编码器



最后我们选择4层RGCN编码器和1层RGCN解码器作为最终的模型结构。

$$MSE(x, x') = \frac{\sum_{i=1}^n (x_i - x'_i)^2}{n}$$

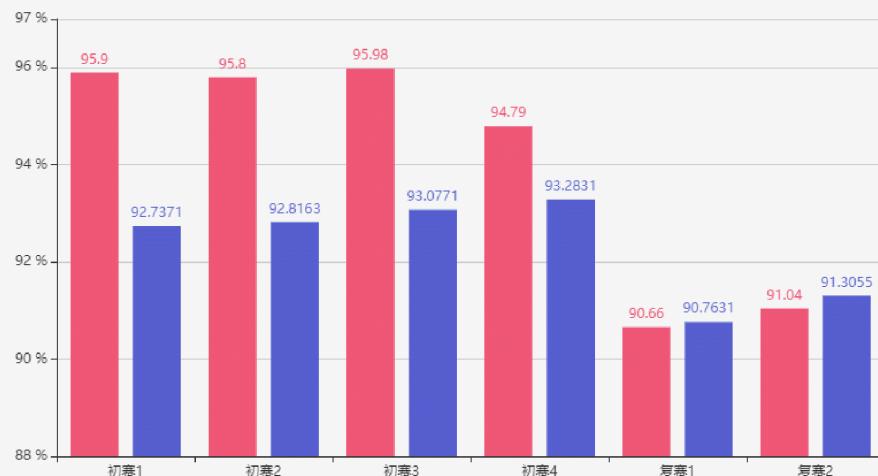


## 其他策略

重新划分验证集

线下验证与线上分数对比

■ 验证集AP ■ 提交结果



## 赛事总结

### ✓ 感谢

- 感谢阿里安全官方耐心的答疑
- 比赛提供的baseline非常清晰易懂，对萌新非常友好
- 感谢OpenI平台提供的算力支持

### ✓ 收获

- 作为第一次参加图神经网络的比赛，我们取得了不错的成绩
- 比赛过程中学到了很多图神经网络和机器学习的知识

### ✓ 遗憾

- 自监督学习这一策略虽然有很好的效果，但是需要消耗大量的时间在自编码器的预训练上，导致我们很早就放弃了对其他图模型的探索



关联比赛: ICDM 2022 : 大规模电商图上的风险商品检测

**版权声明:** 本文内容由阿里云天池用户自发贡献，版权归原作者所有，天池社区不拥有其著作权，亦不承担相应法律责任。如果您发现本社区中有涉嫌抄袭的内容，填写[侵权投诉表单](#)进行举报，一经查实，本社区将立刻删除涉嫌侵权内容。