

第5名队伍方案PPT分享

孤帆明灭 2022-09-22 06:22:13 0 55



大规模电商图上的风险商品检测竞赛 答辩报告

队名：提桶跑路

成员：牛津

哈尔滨工业大学 交通科学与工程学院

2022年9月22日

报告提纲

➤ 1 赛题简述

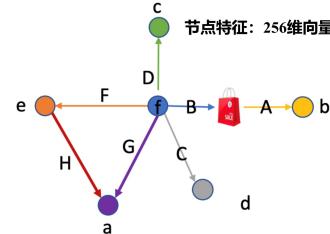
➤ 2 模型方案

➤ 3 实验与分析



1 赛题简述

竞赛题目——正负样本不均衡的异构图节点分类问题



训练集数据

# node type	# edge type	# node	# edge
7	7	13,806,619	157,814,864

正样本: 8364 (9.78%)

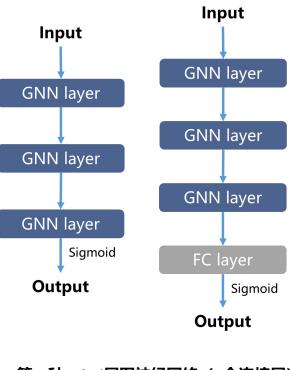
负样本: 77198 (90.2%)

2 模型方案

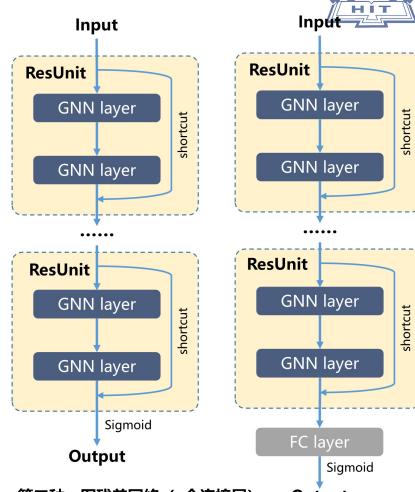


网络结构比选

异构图结构数据的二分类逻辑回归模型



第一种: 3~4层图神经网络 (+全连接层)

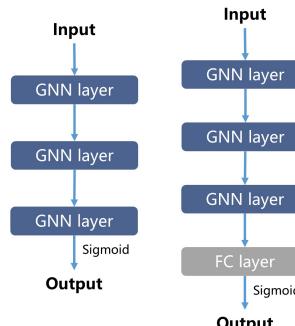


第二种: 图残差网络 (+全连接层)

2 模型方案



GNN层的选取



初赛GNN层选取试验(训练集: 验证集=8: 2)

1. R-GCN
验证集分数: 0.9615 初赛分数: 0.9295
2. R-GAT
验证集分数: 0.9310
3. GNN-FiLM
需要结合1~2层全连接层或R-GCN使用
验证集分数: 0.9643 初赛分数: 0.9335
4. SAGE
验证集分数略低于R-GCN

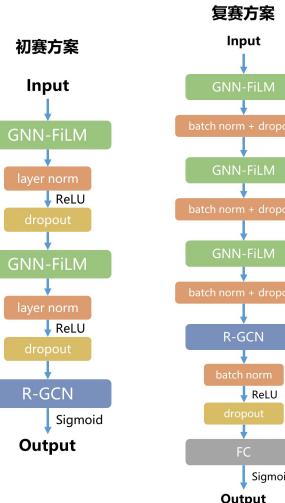
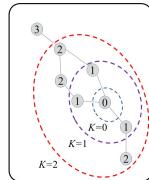


2 模型方案

初赛及复赛方案

实际使用中，考虑计算成本，子图采样范围较小（仅2阶），深层图残差网络并没有给计算结果带来提升。

采用方案：**3~4层图神经网络（+全连接层）**



3 实验与分析

训练方案

训练集/验证集划分

1. 正负样本**分层抽样**
2. 训练集取84%，验证集取16%

损失函数

训练采用的损失函数为二值交叉熵损失函数 (Binary Cross Entropy Loss, **BCELoss**)

子图采样

取**两阶**相邻顶点 num_neighbors=[256, 256]

学习率及其变化曲线

初始学习率取0.001，学习率的scheduler采用余弦下降曲线(Cosine Annealing)，经过24个epoch 学习率趋近于0，停止训练。

最终提交方案：

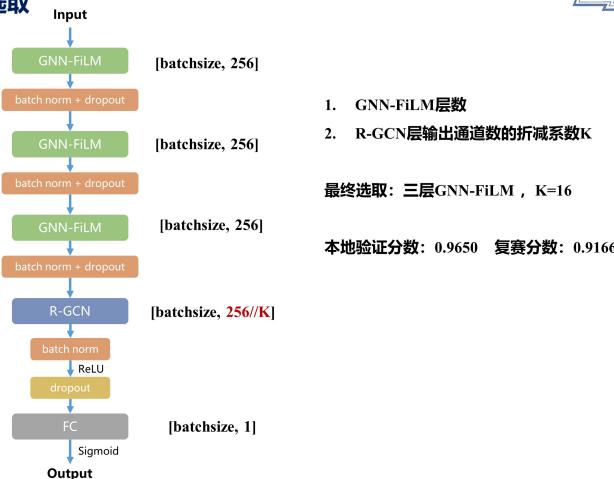
在保证训练稳定的前提下，**在全体数据上训练模型，不设置验证集**。

注：为保证在全体数据集上训练结果的稳定性，以**训练迭代批次相同**为原则，因此共训练 $24 \times 0.84 \approx 20 (个) epoch。$

3 实验与分析



超参数选取





3 实验与分析

正负样本不平衡问题

解决类别不平衡问题的尝试

1. BCE Loss + 正负样本权重调整

2. BCE Focal Loss

3. 对正样本过采样

可能原因：

1. 竞赛的评价指标为AP score，本身更关注正负样本的相对大小。

2. 虽然正样本较少，但由于正负样本差别尚处于可接受范围，在训练数据较大的情况下，该模型仍可以有效完成正样本的学习。

上述几种方法只能将正负样本的划分阈值向

0.5靠近，对AP score没有明显改善

$$AP_D(f) = \frac{1}{n} \sum_{i=1}^n \frac{1}{|Y_i|} \sum_{y \in Y_i} \frac{|\{rank_f(x_i, y) \leq rank_f(x_i, y'), y' \in Y_i\}|}{rank_f(x_i, y)}$$

一些失败/未充分尝试的方案

1. 通过对顶点mask进行预训练

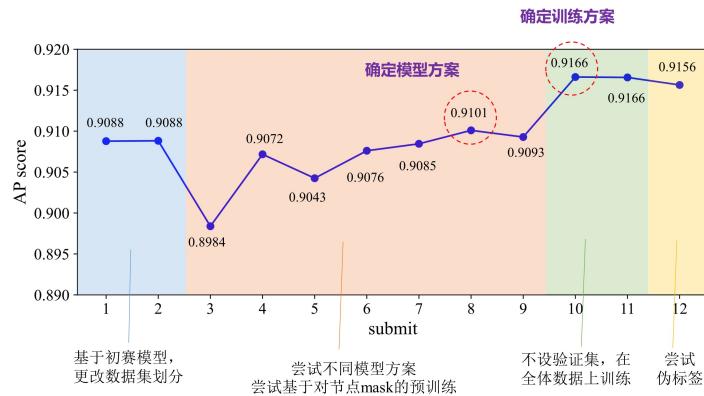
2. 伪标签

3. 对训练数据引入随机噪声

参考其他组实验结果，本组开展的为数不多的预训练尝试存在问题

3 实验与分析

复赛提交



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。

全部评论(0)