浙江大學 ZHEJIANG UNIVERSITY

EAGLE-Lab

ICDM 2022：大规模电商图上的风险商品检测
复赛答辩

队名：happy_fish

目录

CONTENTS

01

团队介绍

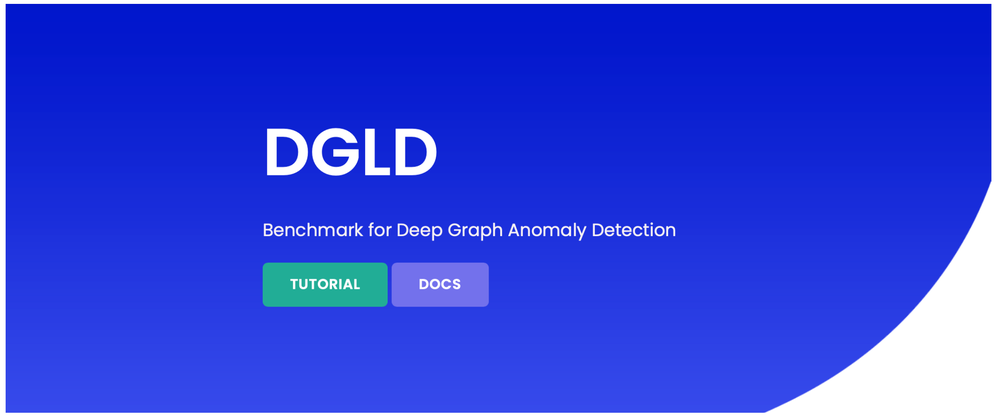
01 团队介绍

Eagle-Lab 实验室

方梦成

杨高明

周晟



DGLD

DGLD is an open-source library for Deep Graph Anomaly Detection based on pytorch and DGL. It provides unified interface of popular graph anomaly detection methods, including the data loader, data augmentation, model training and evaluation. Also, the widely used modules are well organized so that developers and researchers can quickly implement their own designed models.

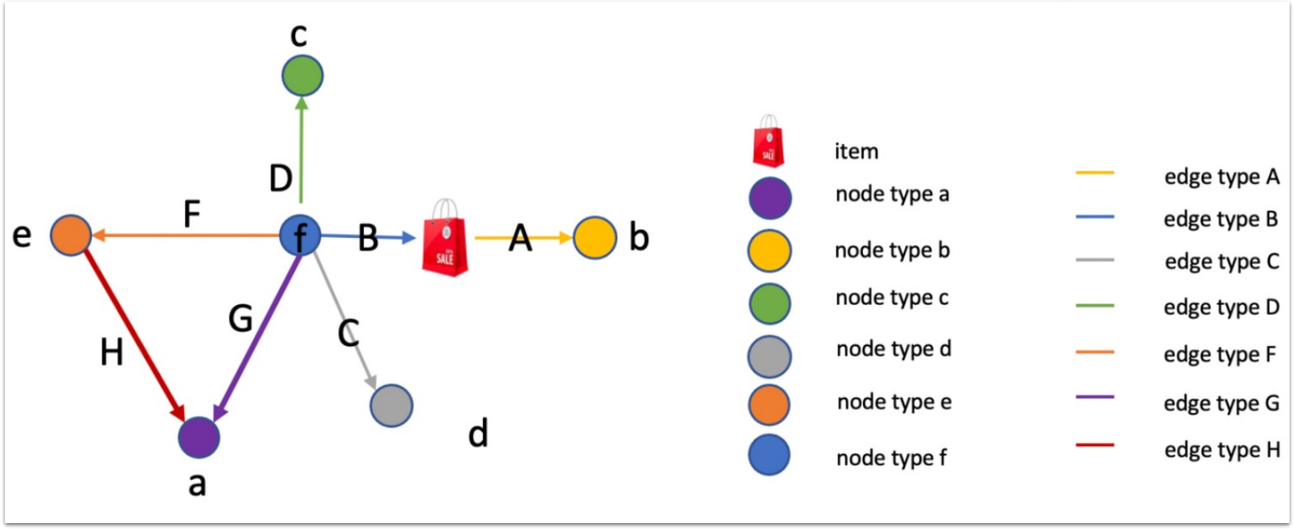
实验室开源项目——DGLD，目前包括17种主流的图上无监督异常检测算法

GitHub：https://github.com/EagleLab-ZJU/DGLD

02

赛题理解

图结构



统计信息

阶段	节点类型	边类型	节点总数	边总数
初赛	7	7	13,806,619	157,814,864
复赛	7	7	10,284,026	120,691,444

评价指标

$$\mathrm{AP} = \sum_{n} (R_n - R_{n-1}) P_n$$

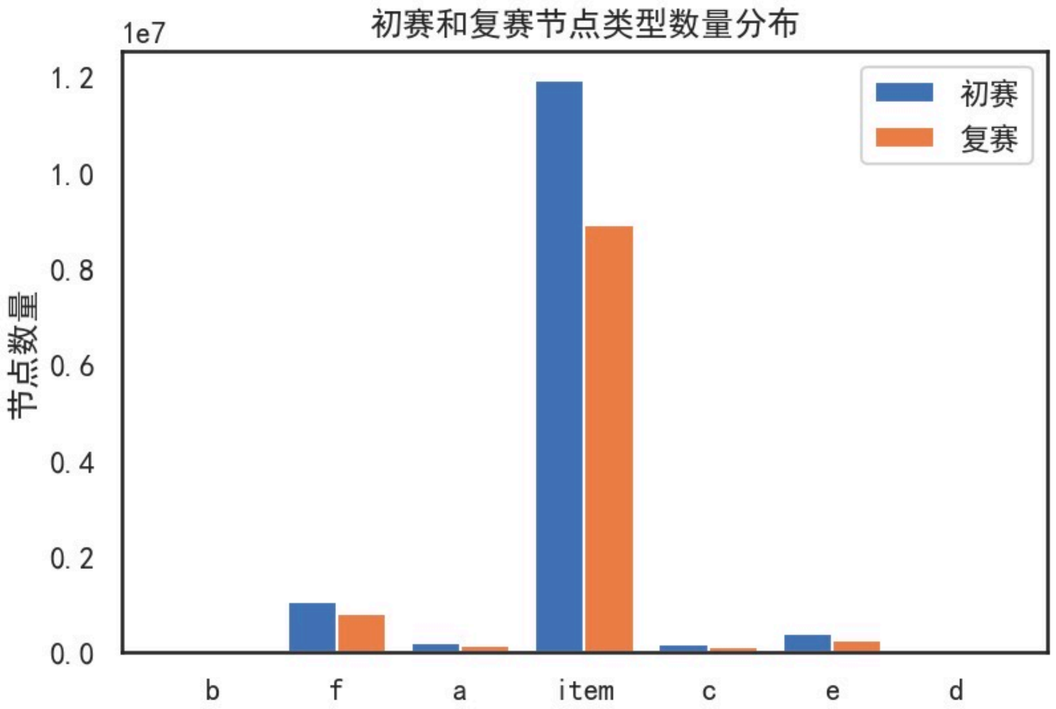
黑白样本比例

阶段	白样本数	黑样本数	比例
初赛	77,198	8,364	约 9:1
复赛	?	?	?

面临挑战

- 大规模异构图。节点数，边数上亿级。
- 黑白样本分布不均衡。
- 数据存在噪声。
- 复赛数据无标签。

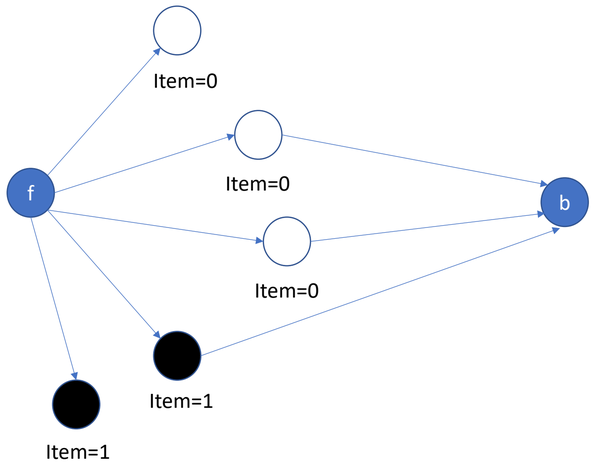
数据分析



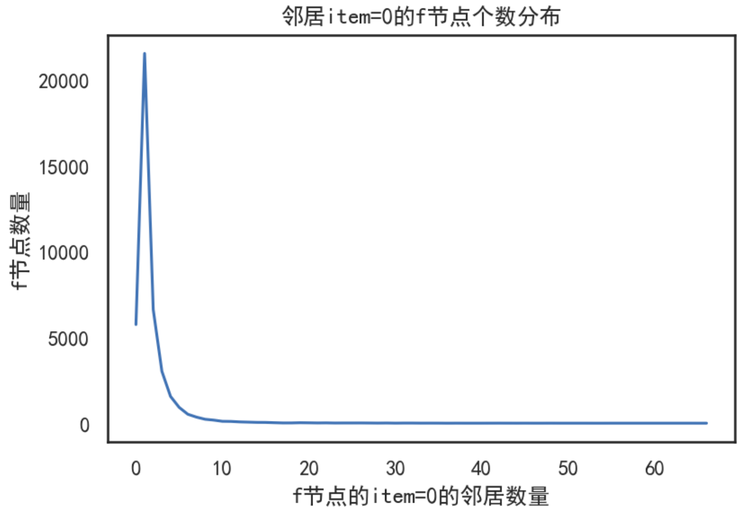
item节点标签数量

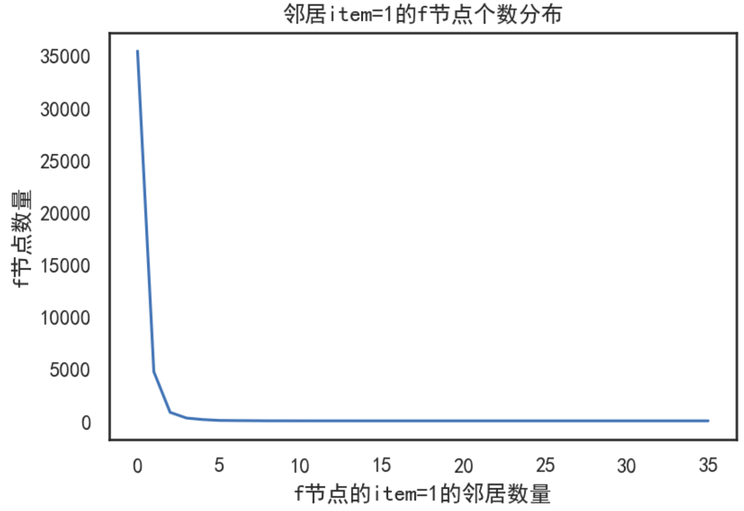
阶段	白样本数	黑样本数	不明确样本数
初赛	77,198	8,364	11,847,804

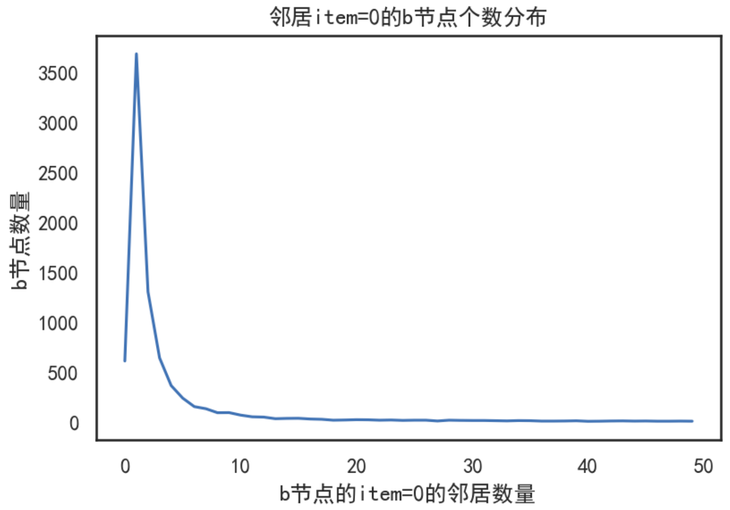
数据分析

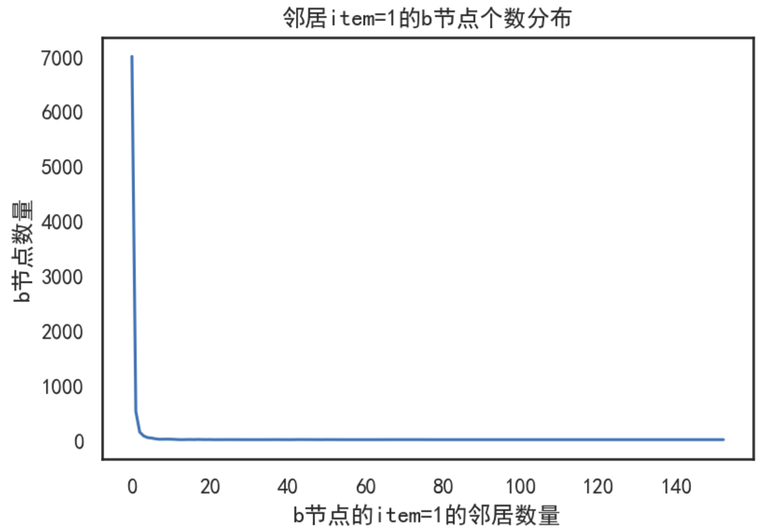


Item异常，其邻居节点b和f节点是否异常呢？





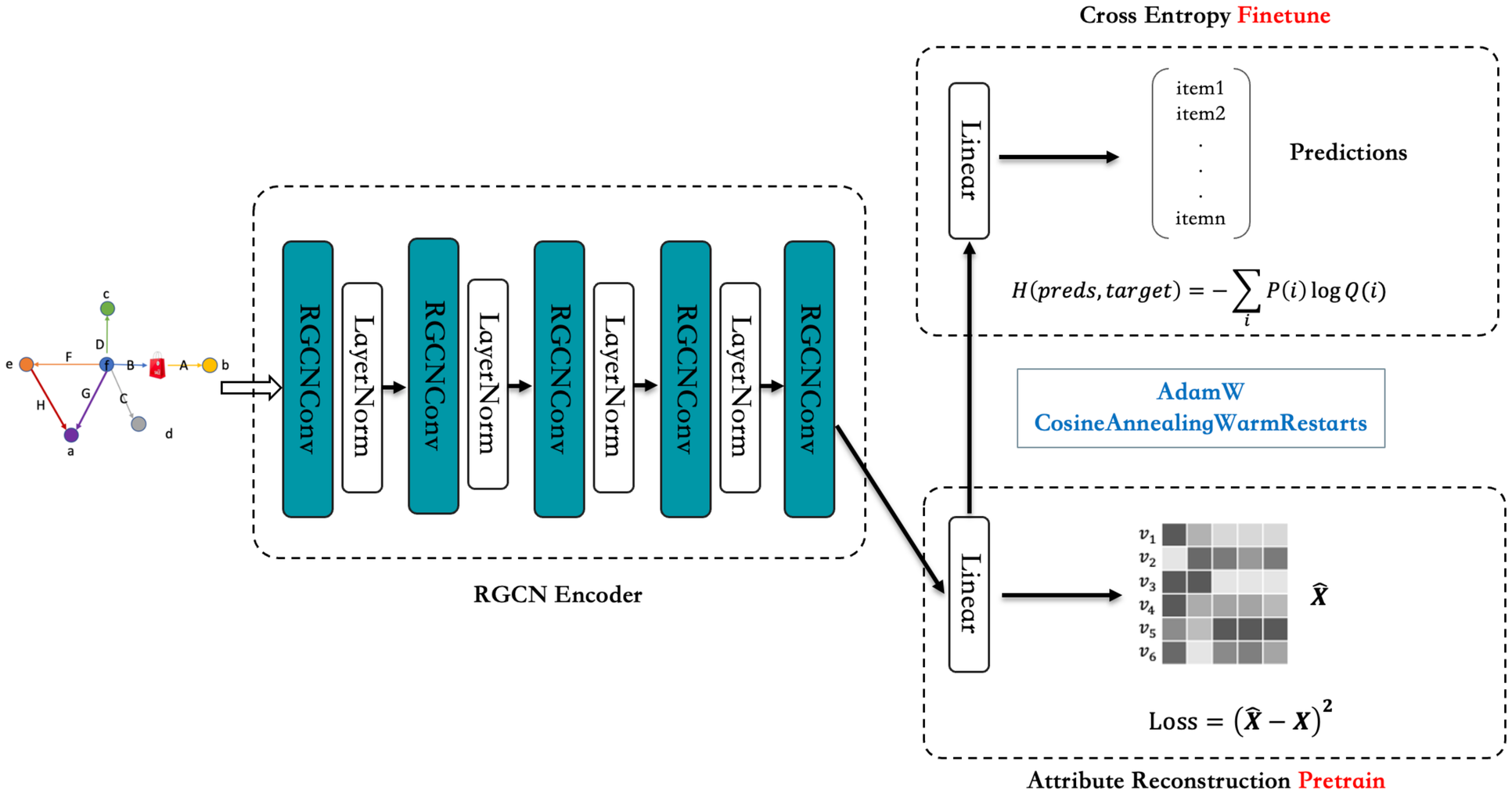




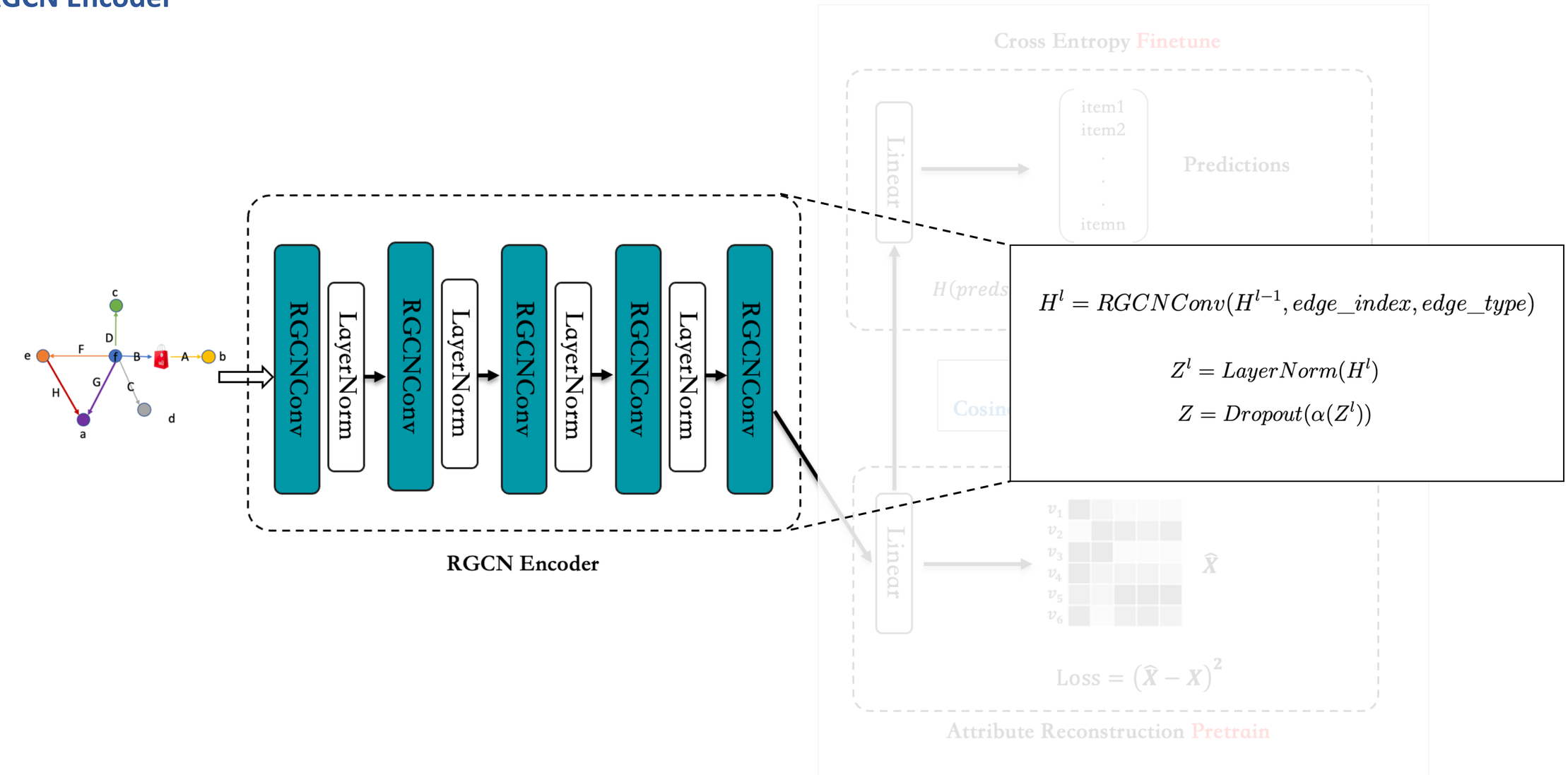
03

模型方案

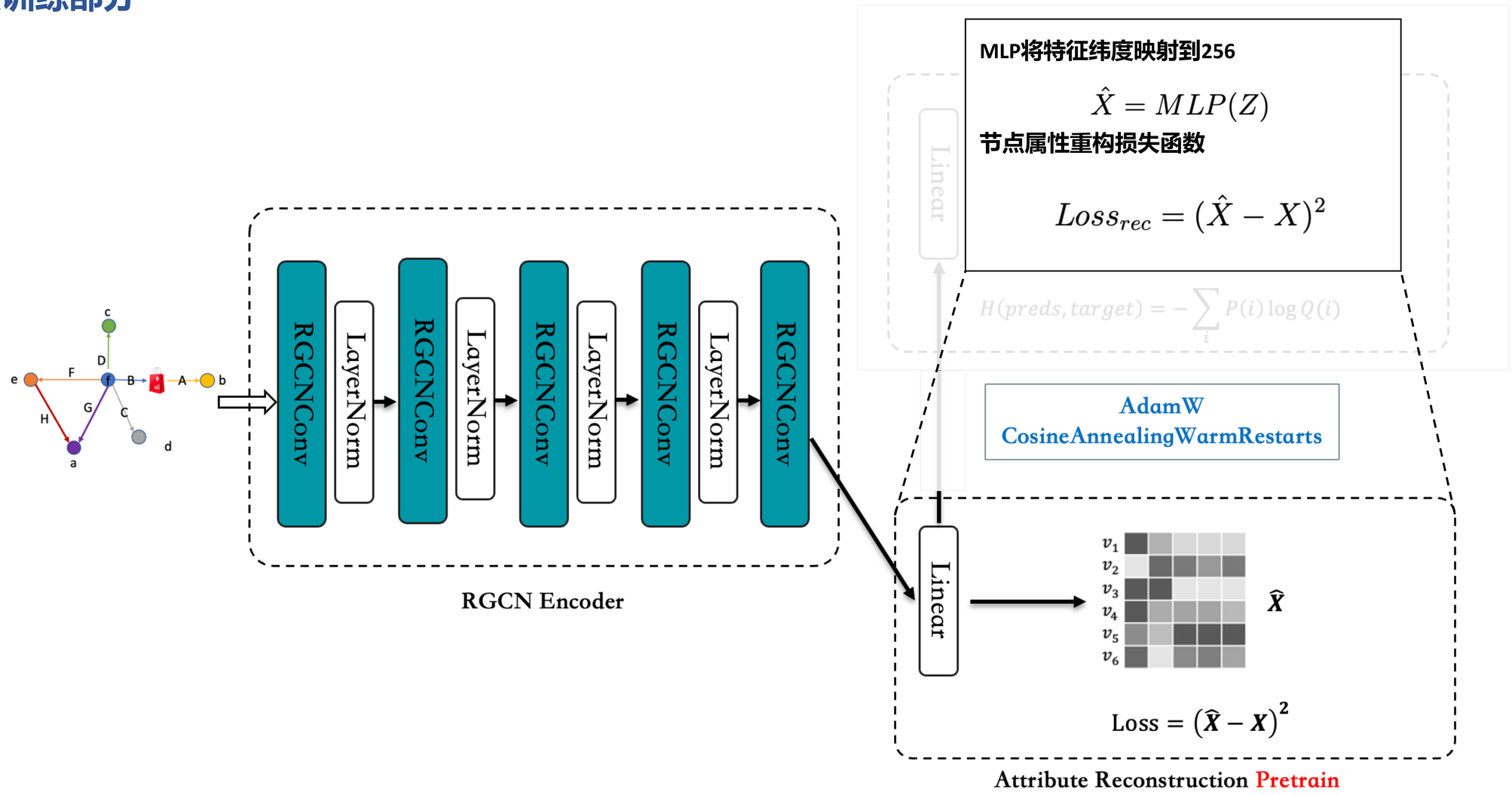
模型设计


Cross Entropy Finetune
Linear
item1
item2
itemn
Predictions
$H(preds, target) = -\sum_i P(i) \log Q(i)$
AdamW
CosineAnnealingWarmRestarts
RGCNConv
LayerNorm
RGCN Encoder
Linear
v_1
v_2
v_3
v_4
v_5
v_6
$\hat{X}$
Loss $= (\hat{X} - X)^2$
Attribute Reconstruction Pretrain

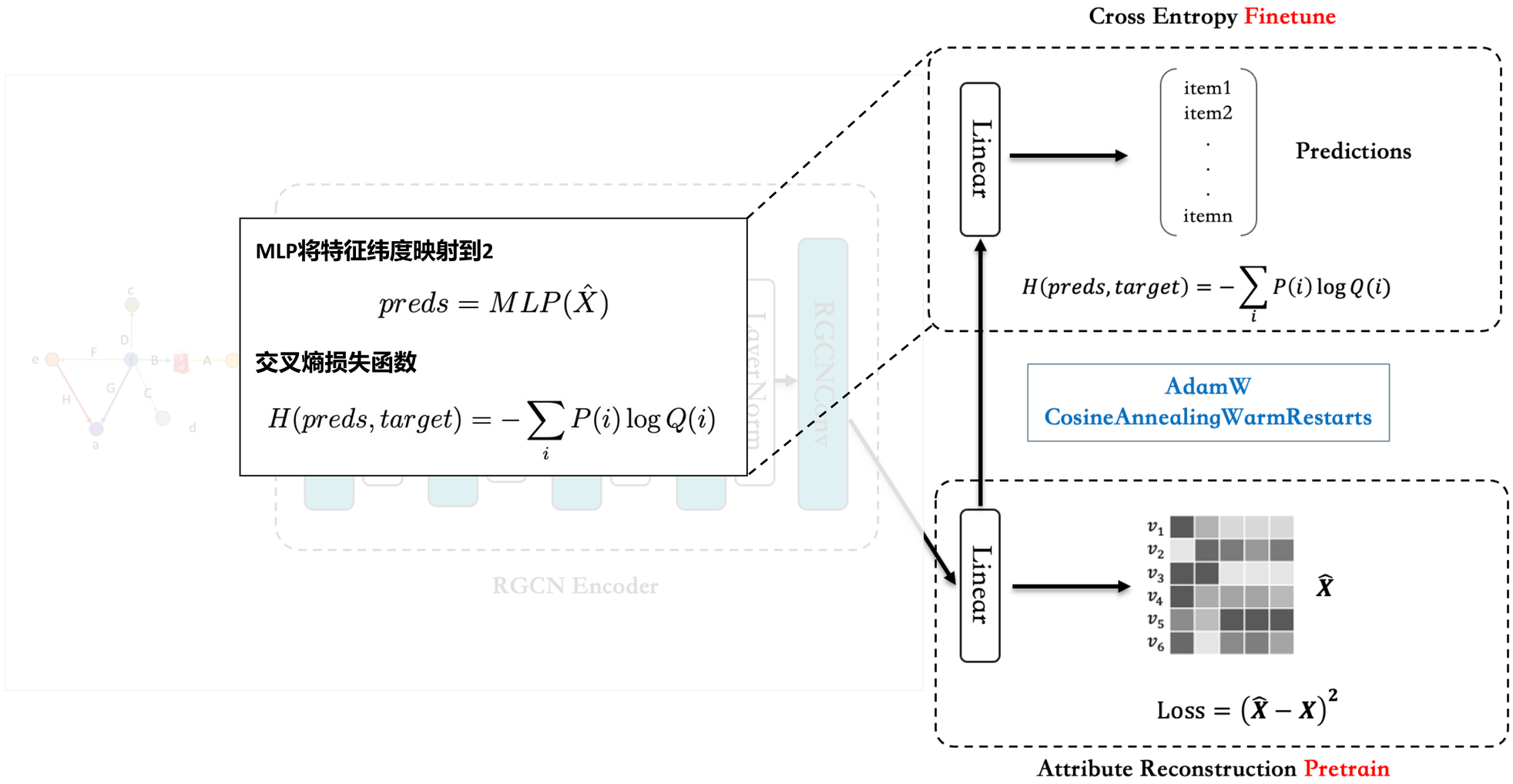
RGCN Encoder



预训练部分


MLP将特征纬度映射到256
$\hat{X} = MLP(Z)$
节点属性重构损失函数
$Loss_{rec} = (\hat{X} - X)^2$
Linear
$H(preds, target) = -\sum_i P(i) \log Q(i)$
AdamW
CosineAnnealingWarmRestarts
RGCNConv
LayerNorm
RGCNConv
LayerNorm
RGCNConv
LayerNorm
RGCNConv
LayerNorm
RGCNConv
RGCN Encoder
Linear
v_1
v_2
v_3
v_4
v_5
v_6
$\hat{X}$
$\text{Loss} = (\hat{X} - X)^2$
Attribute Reconstruction Pretrain

微调部分


Cross Entropy Finetune
Linear
item1
item2
.
.
.
itemn
Predictions
$H(preds, target) = -\sum_i P(i) \log Q(i)$
AdamW
CosineAnnealingWarmRestarts
Linear
v_1
v_2
v_3
v_4
v_5
v_6
$\hat{X}$
Loss $= (\hat{X} - X)^2$
Attribute Reconstruction Pretrain
MLP将特征纬度映射到2
$preds = MLP(\hat{X})$
交叉熵损失函数
$H(preds, target) = -\sum_i P(i) \log Q(i)$
LayerNorm
RGCNConv
RGCN Encoder

04

实验部分

04 实验部分——实验设置

模型参数设置

- 使用RGCNConv层数： 5
- 每层神经元数目： 768
- 预训练模型的 MLP 输出纬度： 256
- Dropout 的概率： 0.4

训练参数设置

- 采样邻居节点层数： 2
- 每层采样邻居节点数目： 300
- batch-size为： 256

简单模型实验

表 3: 简单模型实验结果

模型	每层采样数目	隐层大小	初赛验证集	初赛测试集
RGCN	300	768	0.9352	0.9253
RGAT	64	128	0.9273	0.9151
HGT	300	768	0.8991	0.8703
GAT	300	768	0.9293	0.9182
TransformerConv	300	768	0.9155	0.8812

表 4: 基于 dgl 的模型实验结果

方案	初赛验证集	初赛测试集
使用 2 层 RelGraphConv	0.928	0.92263
使用 2 层 RelGraphConv，增加神经元个数	0.93281	0.923319
使用 3 层 RelGraphConv，增加神经元个数	0.938824	0.925188
2 到 3 层之间加类似残差连接，并添加 BatchNorm	0.94065	0.929349

预训练模型实验

表 5: 预训练模型下游 finetune 结果

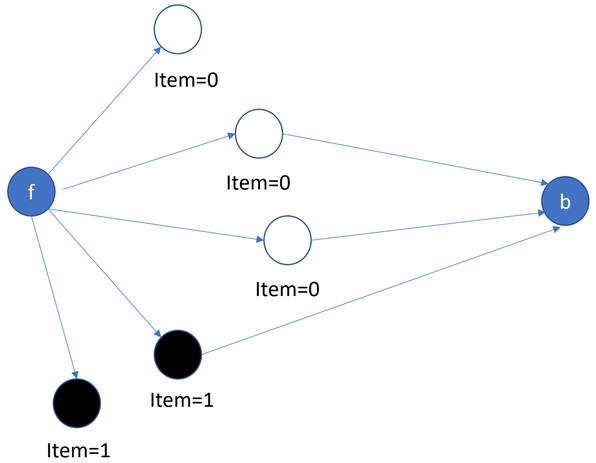
预训练轮数	微调最优轮数	初赛验证集	初赛测试集
200	9	0.9432	0.9357
300	8	0.9443	0.9398
500	8	0.9450	0.9402

表 6: 全量训练轮数与初赛测试集分数对比

实验次数	全量训练 6 轮	全量训练 7 轮	全量训练 8 轮
1	0.9367	0.9436	0.9412
2	0.9353	0.9444	0.9399
3	0.9372	0.9438	0.9420
4	0.9361	0.9431	0.9421

04 实验部分——复赛消融实验

引入b和f节点扩充训练集



Item异常，其邻居节点b和f直接打标签为1，否则为0！

表 7: 伪标签数据增强实验结果

实验设置	初赛验证集	全量训练 7 轮复赛测试集
无增强	0.9450	0.9189
item 伪标签增强	0.9421	0.9170
b 和 f 伪标签增强	0.9462	0.9212

表 8: 引入复赛预训练实验结果

继续预训练轮数	全量训练 7 轮复赛测试集
50	0.9222
85	0.9243(最终结果)
100	0.9221
150	0.9191

初赛item节点500轮预训练

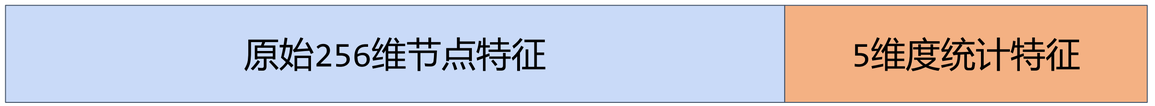
引入与初赛带标签item节点相连的b和f节点进一步预训练300轮

引入复赛的item节点和与其相连的b和f节点进一步预训练

04 实验部分——其他尝试

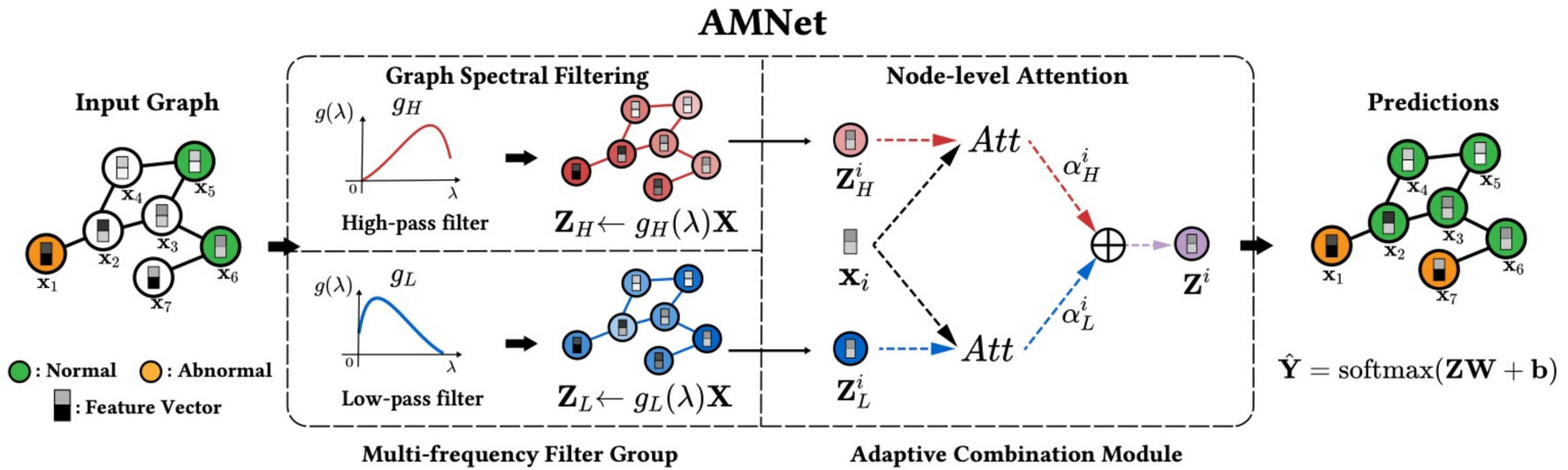
节点特征扩充策略

- 节点特征扩充策略，使用node2vec给item节点扩充128维特征，由于总特征变多，模型训练非常慢，而且服务器资源使用过大，最终效果没有明显提升，最后放弃。
- 在原始256维节点特征基础上添加，给item节点添加节点度特征，以及周围邻居节点度特征的统计值(max，min，std，mean)，效果也不好。



04 实验部分——其他尝试

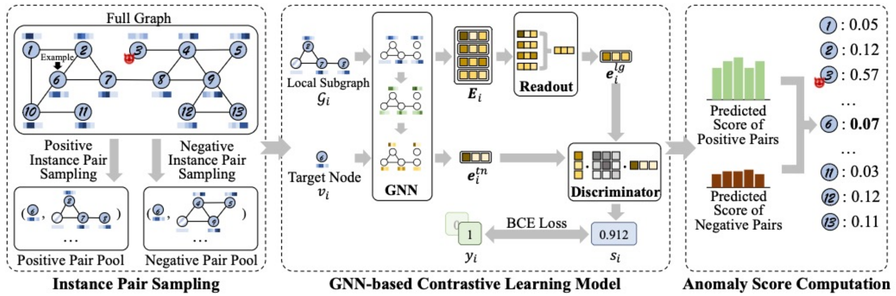
其他图异常检测模型尝试



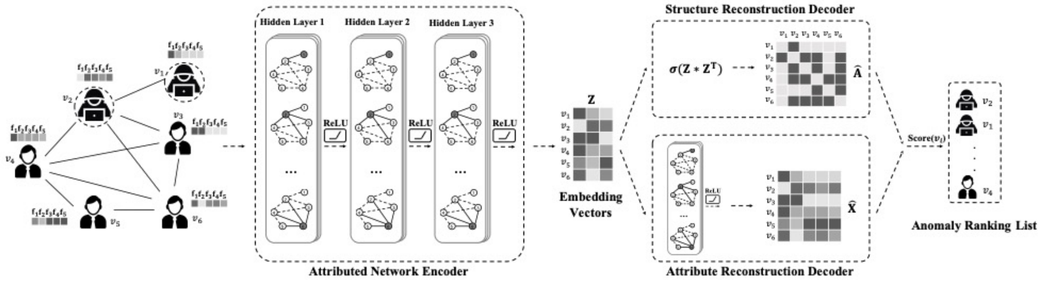
- AMNet[6]模型尝试。考虑异常节点和正常节点偏好不同的频段，设计自适应多频图神经网络，但是这个模型在这个数据集上效果无提升。
- DGL 模型尝试。用基于DGL的RGCN搭建相同的预训练模型，达不到pyg的效果，因此最终选择pyg框架。

[1] Chai, Ziwei, Siqi You, Yang Yang, Shiliang Pu, Jiarong Xu, Haoyang Cai, and Weihao Jiang. Can Abnormality be Detected by Graph Neural Networks?.

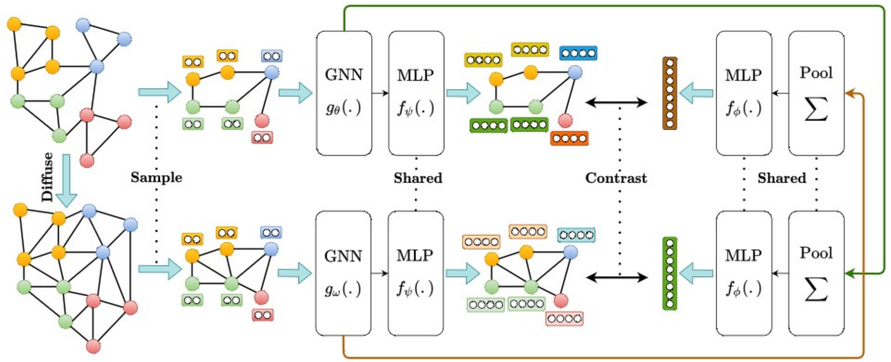
预训练模型尝试



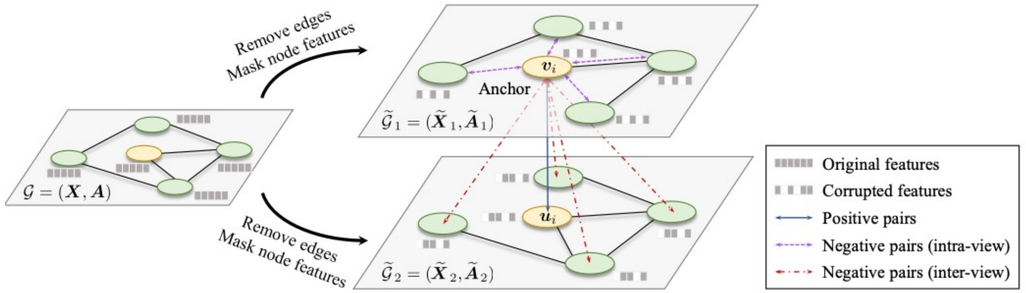
CoLA



Dominant



MVGRL



GRACE

- 使用DGLD[1]开源库中纯无监督模型进行预训练，比如CoLA[2]和Dominant[3]，效果无提升。
- 设计对比学习模型，参考MVGRL[4]和GRACE[5]模型，引入数据增强策略进行对比学习自监督预训练，最后效果也没提升。

[1] Sheng Zhou. https://github.com/EagleLab-ZJU/DGLD, 2022.
[2] Yixin Liu, Zhao Li, Shirui Pan, Chen Gong, Chuan Zhou, and George Karypis.Anomaly detection on attributed networks via contrastive self-supervised learning.CoRR,abs/2103.00113, 2021.
[3] Kaize Ding, Jundong Li, Rohit Bhanushali, and Huan Liu. Deep anomaly detection on attributed networks. In Proceedings of the 2019 SIAM International Conference on Data Mining, pages 594–602. SIAM, 2019.
[4] Kaveh Hassani and Amir Hosein Khas Ahmadi. Contrastive multi-view representation learning on graphs. CoRR, abs/2006.05582, 2020.
[5] Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Deep graph contrastive representation learning. CoRR, abs/2006.04131, 2020.11

05

总结与思考

05 总结与思考

总结

历时一个半月的大规模电商图上风险商品检测算法赛已经结束了，在这个比赛中我们学到了很多知识，锻炼了实践动手能力，提高了团队合作能力。非常感谢主办方提供的真实业务场景数据，让我们对图上的异常检测算法有更深的认识！

思考

- 如何设计更有效的预训练模型？(千万级item节点尝试)
- 论文中的深度图嵌入算法为什么效果不好？
- b 和 f 节点打标签策略?

浙江大學 ZHEJIANG UNIVERSITY

EAGLE-LAB

Thanks !