ICDM 2022：大规模电商图上的风险商品检测

队伍名：躺平不如起立
队员：杨思超

A.问题分析

B.技术思路

C.总结与收获

1. 问题分析

比赛类型：图分类问题

特点：

- 大规模异构图：大于国内同时期举办的其他图相关比赛，数据前处理需要的单机内存~46GB
- 黑白样本分布不均：正负样本~1：10
- 噪声大：图的homophily非常低，导致黑白样本的结构区分度不大
- Inductive：复赛需要在新图上直接做推理，label-propagation类型的方法无法使用
- 要求单模型：不能做单模型cv，多模型集成

graph	edge	node	edge_feat	node_feat	edge_type	node_type	directed	hetero	label	source
DGraph-Fin2022	4300999	3700550	1	17	-	-	y	n	4	https://dgraph.xinye.com/introduction
icdm2022comp	157814864	13806619	-	256	7	7	y	y	2	https://tianchi.aliyun.com/competition/entrance/531976/information
	36.7	3.7								

2. 技术思路

本方案里采用的主要方法可以归纳为：

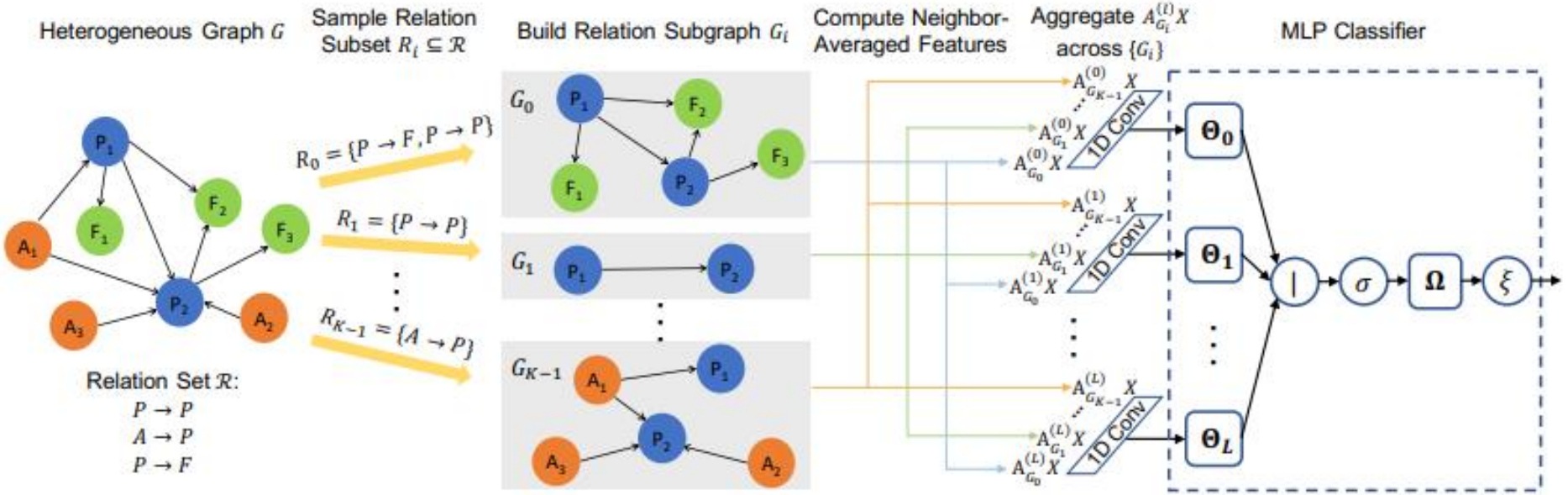
* NARS: Neighbor Averaging over Relation Subgraphs
* Semi-supervised learning: mutual-mean teaching, fixmatch
* rebalanced loss weights & hard example learning

Scalable Graph Neural Networks for Heterogeneous Graphs
FixMatch: Simplifying Semi-Supervised Learningwith Consistency and Confidence
Mutual Mean-Teaching: Pseudo Label Refinery for Unsupervised Domain Adaptation on Person Re-identification
Focal Loss for Dense Object Detection

2. 技术思路

* NARS: 把simplify gcn在heterogeneous graph上进行了拓展，simplify gcn的核心思路是之前的大规模图上sample的方法（sage，saint）由于限制采样邻居数且只在subset上采样会引入bias，且sampling上做的多层non-linear transformation没有明显性能提升，所以直接去掉了message passing里的NL transform。NARS在异构图上sample不同的metapath作为subsets聚合feature，供后续分类器学习。

在这次竞赛的实验中，nars表现优于rgcn和hgt，分析原因是这次的图噪声很大，所以与其在包含所有类型节点全图上采样不如在只包含部分类型节点的metapath上采样再做ensemble更鲁棒，采样深度也可以做得更深。另外NARS可以把特征生成和标签分类两部分分开，两部分单独调优，整体学习更高效。

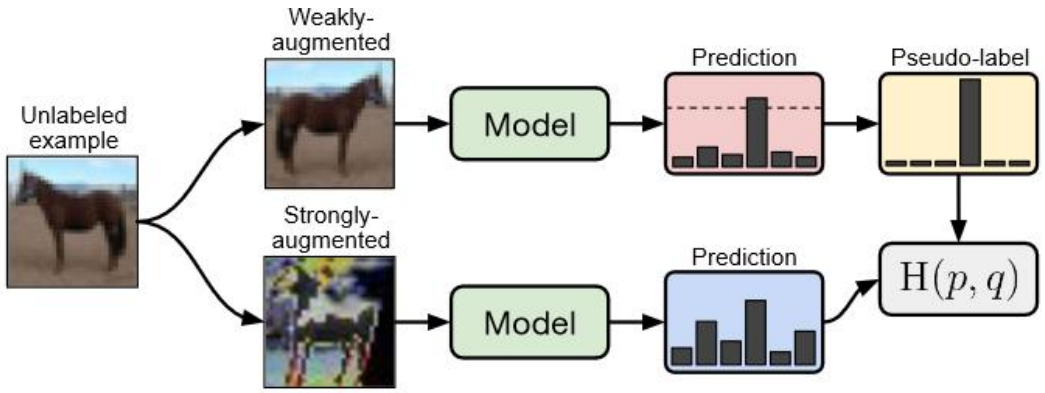


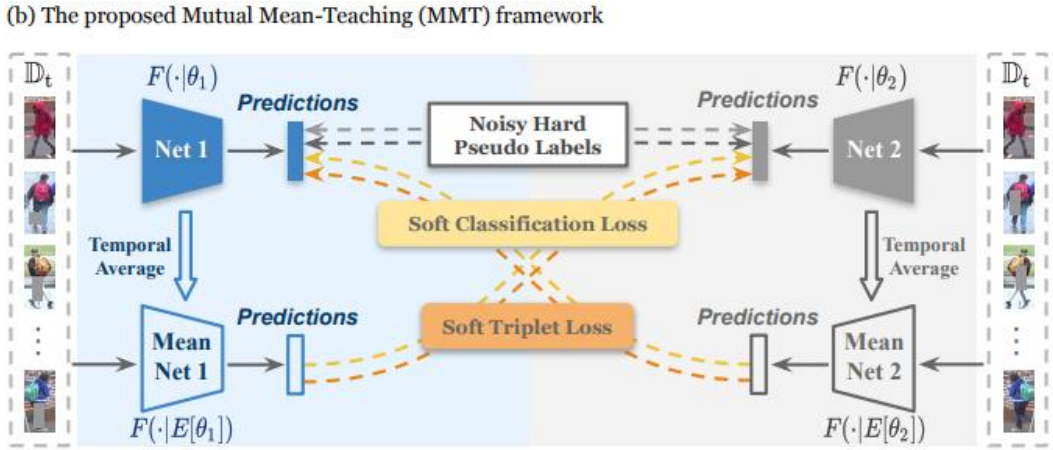
Scalable Graph Neural Networks for Heterogeneous Graphs

2. 技术思路

* Semi-supervised learning: mutual-mean teaching, fixmatch，比赛有标签的数据只有全量数据0.6%，为了更好利用没有标签的数据，方案里借鉴了两个流行的自监督学习方法：Fixmatch和mutual-mean teaching。第一个方法做的是图像任务，利用噪声数据来做consistency regularization。对于图任务，我随机的dropout输入的节点feature，然后要求预测分布和没有regularized的分布距离接近。第二个方法为了解决自监督学习常见的bias confirmation问题，设计了两个网络互相在线和离线学习。两个网络的参数设置和数据可以相互独立。针对本次比赛，我利用cv里当前模型以外的其他模型来做unlabelled data的pseudo labelling，并利用confidence_masking来控制引入无标签样本的比例和label_smoothing来控制loss分布。

实验发现结合regularization，dev set上的精度能够小幅提升。



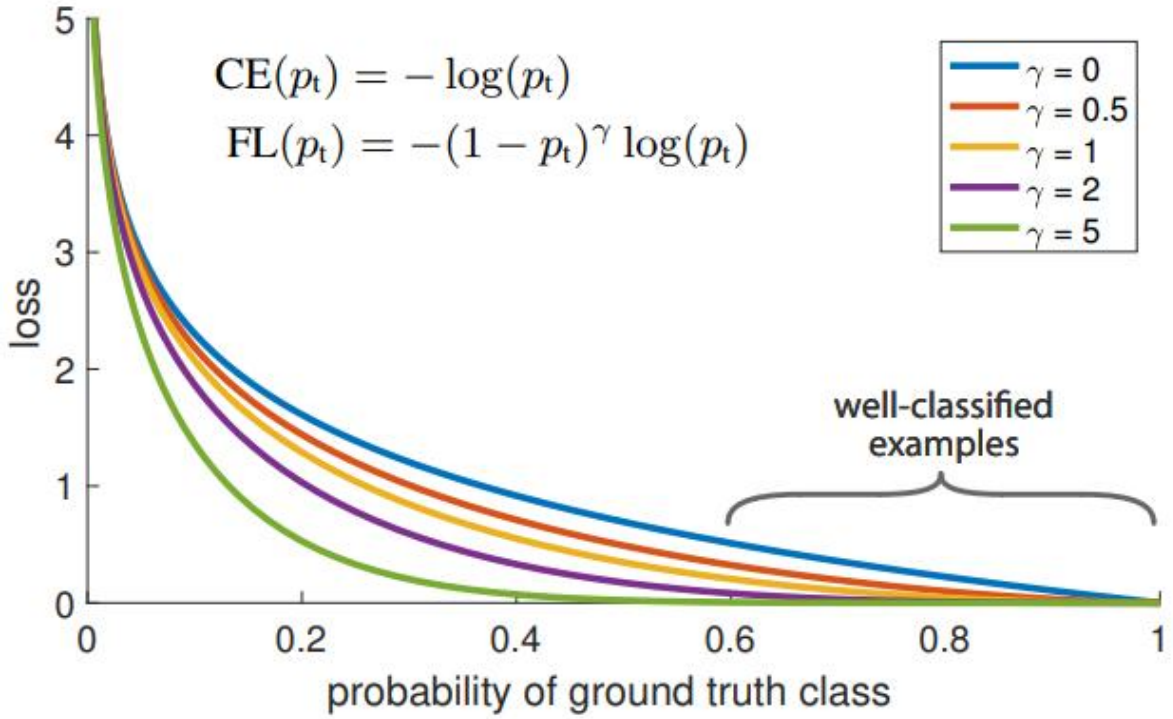


Scalable Graph Neural Networks for Heterogeneous Graphs

Mutual Mean-Teaching: Pseudo Label Refinery for Unsupervised Domain Adaptation on Person Re-identification

2. 技术思路

* rebalanced loss weights & hard example learning，为了解决正负样本分布不均衡的问题，且基于比赛测试集正负比例低于训练集的公开信息，这里使用了focal loss和设置负权重略高于正权重的策略。这两项技术也帮助模型精度的线上成绩小幅提升。



Focal Loss for Dense Object Detection

2. 技术思路

下面分别是对模型结构，半监督学习策略和损失函数的消融实验结果。

	train_ap	val_ap	param
rgcn	9467	9350	-
hgt	9476	9321	-
SIGNV1	9608	9523	1511245
SIGNV2	9875	9498	1510863
SIGNV3	9795	9507	5002289
SIGNV4	9407	9379	5150513
SIGNV5	9868	9504	1215961
NARS_R_GAMLP	9357	9369	1672337
NARS_JK_GAMLP	9333	9308	928720

	train_ap	val_ap
w/o cr	9757	9540
scr	9807	9545
dcr	9792	9548
scr+dcr	9806	9555

	train_ap	val_ap
cross_e	9842	9538
focal_1	9864	9566
focal_2	9806	9555
focal_3	9719	9511

3. 收获与总结

- 比赛过程中，实验的有效记录和针对性的误差分析非常重要。初赛排名第5，但复赛一度跌到20名开外，其中就包括了模型结果无法复现，优化策略遇到瓶颈的问题。
- 比赛结果很重要，但享受学习的过程更重要，否则难免会陷入调参刷分的焦虑中。
- 其他选手的优胜方案里很多都提出了更贴近于图结构的无监督学习方法来更有效的利用海量的无标签数据，深受启发。
- 图比赛很有趣，第一次接触，希望这只是一个开始。
- 最后，特别感谢阿里安全高效的组织耐心地解答，以及OpenI给力的算力支持，畅享A100的感受实在难忘。