软件学报 ISSN 1000-9825, CODEN RUXUEW
2025,36(3):1218−1239 [doi: 10.13328/j.cnki.jos.007170] [CSTR: 32375.14.jos.007170]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向电子商务社交知识图谱高效增量预训练的双向模仿蒸馏*

朱渝珊[1], 张 文[2], 王晓珂[3], 李志宇[3], 陈名杨[1], 姚 祯[2], 陈 辉[3], 陈华钧[1]

[1](浙江大学 计算机科学与技术学院, 浙江 杭州 310012)

[2](浙江大学 软件学院, 浙江 宁波 315048)

[3](阿里巴巴集团, 浙江 杭州 311121)

通信作者: 张文, E-mail: zhang.wen@zju.edu.cn

摘 要: 知识图谱 (knowledge graph, KG) 预训练模型有助于电子商务应用中各种下游任务, 然而, 对于具有高动态性的大规模电商社交知识图谱来说, 预训练模型需要及时更新以感知由用户交互引起的节点特征变化. 提出一种针对电商社交知识图谱预训练模型的高效增量学习方法, 该方法通过基于双向模仿蒸馏的训练策略充分挖掘不同样本对模型更新的作用, 并通过基于样本常规性和反常性的采样策略来减少训练数据规模, 提升模型更新效率. 此外, 还提出一种逆重放机制, 为社交知识图谱预训练模型的增量训练生成高质量的负样本. 在真实的电子商务数据集和相关下游任务上的实验结果表明, 相较于最先进的方法, 所提方法可以更有效且高效地增量更新社交知识图谱预训练模型.

关键词: 知识图谱; 知识图谱预训练; 增量学习; 知识蒸馏

中图法分类号: TP181

中文引用格式: 朱渝珊, 张文, 王晓珂, 李志宇, 陈名杨, 姚祯, 陈辉, 陈华钧. 面向电子商务社交知识图谱高效增量预训练的双向模仿蒸馏. 软件学报, 2025, 36(3): 1218–1239. http://www.jos.org.cn/1000-9825/7170.htm

英文引用格式: Zhu YS, Zhang W, Wang XK, Li ZY, Chen MY, Yao Z, Chen H, Chen HJ. Bidirectional Imitation Distillation for Efficient Incremental Pre-training of E-commerce Social Knowledge Graph. Ruan Jian Xue Bao/Journal of Software, 2025, 36(3): 1218–1239 (in Chinese). http://www.jos.org.cn/1000-9825/7170.htm

Bidirectional Imitation Distillation for Efficient Incremental Pre-training of E-commerce Social Knowledge Graph

ZHU Yu-Shan[1], ZHANG Wen[2], WANG Xiao-Ke[3], LI Zhi-Yu[3], CHEN Ming-Yang[1], YAO Zhen[2], CHEN Hui[3], CHEN Hua-Jun[1]

[1](College of Computer Science and Technology, Zhejiang University, Hangzhou 310012, China)

[2](School of Software Technology, Zhejiang University, Ningbo 315048, China)

[3](Alibaba Group, Hangzhou 311121, China)

Abstract: Pre-training knowledge graph (KG) models facilitate various downstream tasks in e-commerce applications. However, large-scale social KGs are highly dynamic, and the pre-training models need to be updated regularly to reflect the changes in node features caused by user interactions. This study proposes an efficient incremental update framework for the pre-training KG models. The framework mainly includes a bidirectional imitation distillation method to fully use the different types of facts in new data, and a sampling strategy based on samples' normality and abnormality is proposed to sample the most valuable facts from all new facts to reduce the training data size, and a reverse replay mechanism is proposed to generate high-quality negative facts that are more suitable for the incremental training of social KGs in e-commerce. Experimental results on real-world e-commerce datasets and related downstream tasks demonstrate that the proposed

* 基金项目: 国家自然科学基金 (62306276, U23B2055, U19B2027, 91846204); 浙江省自然科学基金 (LQ23F020017); 宁波市自然科学基金 (2023J291)
收稿时间: 2023-06-25; 修改时间: 2023-10-10, 2023-12-29; 采用时间: 2024-02-28; jos 在线出版时间: 2024-06-14
CNKI 网络首发时间: 2024-06-19

framework can incrementally update the pre-training KG models more effectively and efficiently compared to state-of-the-art methods.
Key words: knowledge graph (KG); knowledge graph pre-training; incremental learning; knowledge distillation

知识图谱 (KG) 在电子商务场景如商品分类[1,2]、商品对齐[3,4]和推荐系统[2,5]等任务中有着广泛应用. 知识图谱预训练模型, 如 PKGM[6]和 K-DCN[7], 在大规模商品知识图谱上进行预训练, 并为各种下游任务提供向量形式的商品知识. 然而, 尽管这些知识图谱预训练模型是有效的, 它们并不适用于电子商务场景中一些具有高动态性的知识图谱, 例如社交知识图谱.

电子商务场景中社交知识图谱记录了用户之间如邀请、回复、分享商品等交互行为. 它是一个典型的时序知识图谱[8,9], 由若干形如 (头实体、关系、尾实体、时间戳) 的四元组组成, 简记为 (s,r,o,t) . 例如四元组 (Mike, share_product, David, 4 March 2021) 表示用户 Mike 在 2021 年 3 月 4 日向用户 David 分享了一个商品的链接. 随着新用户不断加入社交网络, 以及平台提供的新型互动, 社交知识图谱变得更加丰富和稠密. 每个月甚至每个星期, 多达数亿的新事实被增加到社交知识图谱中. 此外, 社交知识图谱中用户节点的特征也会根据用户的交互行为不断更新. 尽管使用传统的知识图谱预训练模型能为不同的下游任务提供用户特征向量服务, 但这类只经过一次训练并长期不更新的模型可能无法感知社交知识图谱的动态变化, 不能满足当前的业务需求. 为了使预训练模型捕获知识图谱的动态性, 一个直接的解决方案是定期从头训练整个模型, 然而这种方式不仅浪费计算资源且效率低下.

因此, 基于社交知识图谱中不断新增的用户交互数据, 以一种高效的方式增量地更新预训练模型是亟待解决的问题. 这将有助于确保预训练模型及时感知用户的最新交互动态, 并为当前的下游业务场景提供更准确的用户信息, 从而长期维持这类时序知识图谱预训练模型的可用性.

根据新增数据更新已有模型属于增量学习[10−12]的研究范围, 其重点关注更新模型过程中的灾难性遗忘问题. 总的来说, 增量学习方法有两大类: 1) 基于正则化的方法[13−15]通过对新任务的损失函数施加约束, 以保护旧知识不被新知识覆盖; 2) 基于数据重放的方法[16−18]通过重用部分旧数据来使模型回顾其所学的旧知识. 增量学习在图像处理[19]、数据可视化[20]和自然语言处理[21,22]等领域已经得到了广泛的探索. 然而, 当前对社交知识图谱这类时序知识图谱的增量学习的研究尚不充分, 目前唯一的工作 TIE[9]提出了一个将正则化和数据重放结合的时序知识图谱表示模型的增量更新框架. 尽管 TIE 能较好地保护模型已学到的旧知识, 但它忽略了区分不同训练样本对模型更新的作用且数据重放模块导致训练数据非常庞大, 难以应用在真实的大规模社交知识图谱预训练模型上. 本文提出了一种新颖的基于双向模仿蒸馏的高效增量学习方法 BID-NAS, 用于有效且高效地更新电子商务场景中的社交知识图谱预训练模型.

在我们的方法中, 当前时间的样本根据其在模型更新的不同作用被分为两种类型: 1) 常规样本, 它们几乎不影响旧模型中相关用户表示, 这类样本有助于保护旧模型已学习到的知识; 2) 反常样本, 它们会对旧模型中相关用户表示产生较大影响, 这类样本有助于模型在更新过程中捕获时序知识图谱当前的变化. 为充分利用常规样本和反常样本, 我们提出基于双向模仿蒸馏的训练策略 BID (bidirectional imitation distillation), 对不同类型的样本采用不同的学习策略, 使模型在增量训练过程中既能保留已学习到的知识, 避免灾难性遗忘, 也能捕捉到当前知识图谱中的变化. 具体来说, 对于常规样本, 我们通过正向模仿蒸馏缩小当前模型和旧模型的输出差异, 从而能够保留旧模型已学到的知识. 对于反常样本, 我们通过反向模仿蒸馏扩大当前模型和旧模型的输出差异, 使模型能够建模当前知识图谱中节点特征的变化. 为提升大规模知识图谱预训练模型的增量更新效率, 我们还提出了一种有效的基于样本常规性和反常性的采样策略 NAS (normality and abnormality-based sampling) 以减少训练数据的规模. 该策略根据旧模型对当前样本的掌握程度和样本自身对时间的敏感程度两个因素, 计算样本的常规性和反常性并从当前样本中采样部分具有高常规性或反常性的样本作为最终的训练数据, 从而在保障模型性能的前提下有效提升了模型的更新效率. 此外, 我们还提出了一种逆重放机制为社交知识图谱预训练模型的增量训练生成高质量的负样本. 具体来说, 我们将过去一段时间内被删除的事实 (即过去出现但当前时间未出现的事实) 作为负样本候选集, 并计算不同负样本的质量得分作为其优化权重, 进一步提升模型性能.

我们在一个真实电子商务场景中大规模 (千万级) 社交知识图谱上进行了实验, 对预训练模型进行多次增量更新, 并在用户分类、好友推荐和商品推荐 3 个下游任务上对更新前后的预训练模型进行测试. 实验结果表明, 我们的方法在性能和效率上均优于其他基线方法, 包括当前最先进的方法. 综上所述, 我们的主要贡献如下.

• 提出了一种高效的增量学习方法 BID-NAS, 用于在电子商务场景中快速更新大规模社交知识图谱预训练模型.

• 在 BID-NAS 中提出了一种基于双向模仿蒸馏的训练策略 BID 以避免灾难性遗忘并捕获知识图谱的变化, 一种基于样本常规性和反常性的采样策略 NAS 筛选高价值训练样本以提升训练效率. 此外, 我们还提出一种逆重放机制生成更适合社交知识图谱场景的高质量负样本.

• 实验证明, 本文提出的增量学习方法在电子商务社交知识图谱的实际应用场景中是有效且高效的.

本文第 1 节介绍知识图谱结合预训练技术和增量学习的相关方法和研究现状. 第 2 节介绍本文所需的基础知识, 包括时序知识图谱增量学习的问题定义、知识图谱预训练模型以及时序知识图谱的预训练模型. 第 3 节介绍我们针对电子商务社交知识图谱预训练模型提出的基于双向模仿蒸馏的高效增量学习方法. 第 4 节通过对比实验验证了所提方法的有效性. 最后总结全文.

1 相关工作

1.1 知识图谱结合预训练

预训练在计算机视觉领域, 如 VGG[23]、Google Inception[24]和 ResNet[25], 以及自然语言处理领域, 如 BERT[26], XLNet[27]和 GPT-3[28]的成功应用, 启发了知识图谱和预训练相结合的发展. 目前将预训练与知识图谱相结合的工作主要包括两大类: 知识图谱增强的预训练和知识图谱自身的预训练.

第 1 种类型的工作主要在预训练语言模型中注入知识使其达到更好的性能. K-BERT[29]将三元组注入到句子中, 以生成统一的知识丰富的语言表示. ERNIE[30]将知识模块中的实体表示集成到语义模块中, 以融合单词和实体的异构信息. KnowBert[31]使用一个集成的实体链接器, 通过一种词到实体注意力的形式来生成知识增强的实体跨度表示. K-Adapter[32]通过针对每种知识的神经适配器, 将事实和语言知识注入预训练语言. K3M[3]向多模态预训练模型注入知识, 解决多模态场景中的模态噪声和模态缺失问题. 更多探索在不同应用任务中融合知识发现和预训练模型的研究包括 ProQA[33]、PLM-KDA[34]等.

第 2 种工作直接对知识图谱本身进行预训练. 最具代表性的研究是 PKGM[6]和 K-DCN[7]. 他们在没有语言模型的情况下, 利用知识图谱的结构和拓扑信息训练模型, 并将预训练模型应用于基于知识图谱的下游任务. 然后, 知识图谱预训练模型可以服务于各种下游任务, 为提高基于知识图谱的电商业务的性能做出不可或缺的贡献, 如商品分类, 产品对齐和推荐系统. 这类知识图谱预训练模型均是经过一次性离线训练得到, 并上传到在线业务供其查询相关实体表示. 然而在电子商务场景社交知识图谱具有高度动态性, 源源不断的用户交互行为使新事实不断加入知识图谱中, 不更新的知识图谱预训练模型无法及时反映用户节点特征的变化. 但是, 从头开始训练模型浪费计算资源且效率很低, 难以投入实际应用. 因此, 根据新增事实高效更新预训练模型是大规模知识图谱在实际应用中亟待解决的问题.

1.2 增量学习

增量学习[10], 又称终身学习[11]或持续学习[12], 是指通过不断到达的数据更新已有模型的一种学习形式. 在更新模型时, 模型可能会因为遗忘已学到的知识性能大幅降低, 这一问题被称为灾难性遗忘[13,35]. 研究者提出了多种方法来解决这一问题, 其中基于数据重放和基于正则化的增量学习方法最为常见.

基于正则化的增量学习方法的主要思想是通过对当前模型的损失函数施加约束来保护旧模型已学到的知识. LwF[13]是最具代表性的工作之一, 它通过知识蒸馏技术[36]使当前模型模仿旧模型的输出来避免灾难性遗忘. Rannen 等人[14]提出在增量训练模型同时为每个旧任务学习一个自动编码器, 以保护此任务的最重要的特性. EWC[37]是一种基于贝叶斯的参数约束方法, 它鼓励当前模型的参数尽可能接近旧模型的参数. Liu 等人[15]通过对

模型参数的 Fisher 信息矩阵进行近似对角化显著提升了 EWC 在序列任务学习场景中的表现. EEIL[35]设计了一个具有交叉蒸馏损失函数的端到端增量学习框架来学习新类别. LwM[38]提出了注意力蒸馏损失, 以避免图像分类模型在增量训练过程中注意力区域转移, 达到在添加新类时保留基类信息的目的.

基于数据重放的增量学习方法主要通过将一部分旧训练数据重放到当前模型中以维持模型对旧知识的掌握能力. iCaRL[16]在模型学习新类别的分类任务时, 为旧的类别选择了部分接近该类特征平均值的旧样本用于模型回忆旧知识. GEM[17]为了避免模型对旧数据过度拟合, 通过只更新模型中新任务相关的参数而不改变旧任务相关的参数来改进 iCaRL. 对 GEM 的改进工作还包括 A-GEM[18]和 GSS-Greedy[39]. GR[40]利用生成对抗网络[41]来生成用于重放的数据, 避免了直接将旧训练数据重放潜在的数据隐私泄露问题. CLEAR[42]提出在多任务强化学习中动态调整使用旧数据的数量. BiC[43]应用知识蒸馏解决了大数据集的不平衡类问题. 还有一些针对图神经网络(GNN)[44, 45]提出的基于数据重放的方法, 包括 GraphSAIL[46]、ER-GNN[47]和 TWP[48]. Zhang 等人[49]结合基于正则化和数据重放的方法进行图神经网络的增量学习. 然而, 与基于正则化的方法相比, 基于数据重放的方法需要模型对旧数据重复学习以回忆已学到的知识, 增加了增量训练的计算量和时间消耗.

目前, 对知识图谱的增量学习研究尚不充分. Song 等人[50]在知识图谱嵌入模型利用基于正则化的增量学习方法来增强实体表示. TIE[9]是与本文更相关的一项工作, 提出了针对时序知识图谱的增量训练框架, 该框架结合了正则化和数据重放方法避免灾难性遗忘. 然而, 它忽略了区分不同训练样本对模型更新的作用, 且数据重放模块虽然有效避免了灾难性遗忘, 却导致训练数据非常庞大, 这并不适用于电子商务场景大规模社交知识图谱预训练模型. 本文提出了一种新颖的基于双向模仿蒸馏的高效增量学习方法, 用于有效且高效地完成模型大规模社交知识图谱预训练模型的增量更新.

2 基础知识

本文所提方法主要针对社交 (时序) 知识图谱预训练模型的增量学习, 下面就相关概念和基本知识予以介绍. 我们首先在第 2.1 节中介绍时序知识图谱增量学习的相关概念和问题定义, 然后在第 2.2 节中介绍知识图谱预训练模型 PKGM[6]以及它服务于下游任务的方式, 最后在第 2.3 节中描述我们如何将预训练模型 PKGM 应用在时序知识图谱上, 作为我们的基础预训练模型.

2.1 时序知识图谱增量学习

时序知识图谱 $\mathcal{G}=\{G^1,G^2,\ldots,G^T\}$ 由一个知识图谱在 T 个时间步下的快照序列组成, G^t $(t=1,2,\ldots,T)$ 表示时间步 t 时的知识图谱快照. 每个时间步的知识图谱记为 $G^t=(E^t,R^t,F^T_{\text{all}})$, 其中 E^t 和 R^t 是在时间 t 下的实体和关系的集合, F^T_{all} 是在时间 t 下的所有事实 (s,r,o,t) 的集合, 其中 $s\in E^t, r\in R^t, o\in E^t$. 在增量训练过程中, F^T_{all} 通常被划分用于训练的部分 F^t 和用于测试的部分 $F^t_{\text{test}}=F^T_{\text{all}}\backslash F^t$.

时序知识图谱预训练模型的增量学习任务定义如下: 对于每个时间步 t, 在上一时间步模型的基础上, 用数据 F^t 对其进行增量训练, 更新后的新模型作为时间 t 的当前模型. 由于在真实的电子商务应用场景中, 模型最需要的是预测用户当下或者未来的行为而不是用户过去的行为能力, 因此, 不同于之前的工作[9]在已出现过的所有测试数据 $\cup^t_{i=1}F^t_{\text{test}}$ 上评估当前模型, 本文只在当前最新的测试数据 F^t_{test} 上评估当前模型. 在每个时间步, 模型可以访问的实体是当前和更早时间出现的所有实体, 即 $E^t_{\text{known}}=\cup^t_{i=1}E^i$.

2.2 知识图谱预训练模型

我们以先前的工作 PKGM[6]为预训练模型的基础框架. PKGM 提出了由三元组 (头实体, 关系, 尾实体)、缩写为 (s,r,o) 组成的非时序知识图谱的预训练模型, 它通过一个三元组查询模块编码三元组的真实性, 并通过一个关系查询模块编码关系的存在性, 以向量形式为下游任务提供服务. 知识图谱预训练模型与常规知识图谱嵌入模型的区别在于, 常规知识图谱嵌入模型关注对知识图谱中实体和关系编码的方法, 针对不同场景和任务通常需要训练特定的知识图谱嵌入模型, 在应用过程中, 实体自身的嵌入表示将直接作为对应的实体输入; 知识图谱预训练模型更强调增强大规模知识图谱嵌入的通用性, 经过一次预训练后, 在应用过程中, 融合了实体相关上下文信息的向

量可服务于多种下游任务.

2.2.1 知识图谱预训练过程

知识图谱预训练模型可基于任何知识图谱嵌入方法构建[6], 例如 TransE[51]、ComplEx[52]等, 记预训练模型为 M, 三元组 (s,r,o) 在该模型中最终得分可表示为:

$$M(s,r,o)=f_{\mathrm{KGE}}(s,r,o)+f_{\mathrm{rel}}(s,r) \tag{1}$$

其中, $f_{\mathrm{KGE}}(s,r,o)$ 是三元组查询模块 (原知识图谱嵌入方法的评分函数) 给出的三元组真实性得分, 得分越高表示知识图谱嵌入方法判断三元组 (s,r,o) 为真的可能性越大. 不同的知识图谱嵌入方法的 $f_{\mathrm{KGE}}(s,r,o)$ 的计算方式不同, 例如在基于 TransE[51]的预训练模型中 $f_{\mathrm{KGE}}(s,r,o)=-\|z_s+z_r-z_o\|$, 其中 z_s,z_r 和 z_o 表示 s,r 和 o 的嵌入向量, $\|x\|$ 表示 x 的 L1 范式, 在基于 ComplEx[52]的预训练模型中 $f_{\mathrm{KGE}}(s,r,o)=\mathrm{Re}(z_s^{\top}\mathrm{diag}(z_r)\bar{z}_o)$, 其中 $\mathrm{Re}(x)$ 和 $\bar{x}$ 分别表示 x 的实部和虚部. $f_{\mathrm{rel}}(s,r)=-\|M_r z_s-z_r\|$ 是关系查询模块给出的实体-关系对得分, 得分越高表示实体 s 具有关系 r 的可能性越大, M_r 为关系 r 的转换矩阵. 三元组最终得分 $M(s,r,o)$ 越高, 则模型将三元组 (s,r,o) 判断为正样本的概率越大. 预训练的损失函数为:

$$\mathcal{L}=-\sum_{(s,r,o)} y\log\sigma(M(s,r,o))+(1-y)\log(1-\sigma(M(s,r,o))) \tag{2}$$

其中, σ 是 Sigmoid 函数. y 是三元组的真实标签, 对于正三元组, $y=1$, 对于负三元组, $y=0$.

2.2.2 为下游任务提供向量服务

知识图谱预训练模型 PKGM 以提供实体表征向量的方式服务于下游任务. 对于给定实体 e, 除了提供模型中实体自身的嵌入表示 z_e, PKGM 还提供包含更多信息的上下文表示 c_e 供下游模型使用.

具体来说, 首先将实体 e 分别与每个关系 $r_i\in R$ 组合成实体-关系对 (e,r_i), $i=1,2,\ldots,|R|$. 然后基于预训练模型获得 (e,r_i) 的实体上下文向量 ec_i, 它编码了与 (e,r_i) 相关的目标实体信息. ec_i 的计算方式与知识图谱嵌入方法评分函数相关, 如基于 TransE[51]的 PKGM 中 $ec_i=z_e+z_{r_i}$, 基于 ComplEx[52]的 PKGM 中 $ec_i=z_e^{\top}\mathrm{diag}(z_{r_i})$, 其中 z_e 和 z_{r_i} 均来自预训练后的 PKGM. 再基于预训练模型获得 (e,r_i) 的关系上下文向量 rc_i, 它编码了实体 e 与关系 r_i 关联性信息. rc_i 的计算方式为 $rc_i=M_{r_i}z_e-z_{r_i}$, 其中 M_{r_i}, z_e 和 z_{r_i} 均来自预训练后的 PKGM. 所有实体-关系对 (e,r_i) 的实体上下文向量 ec_i 和关系上下文向量 rc_i 被拼接并聚合为实体 e 最终的上下文表示 c_e:

$$c_e=\frac{1}{|R|}\sum_{i=1}^{|R|}[ec_i;rc_i] \tag{3}$$

其中, $[x;y]$ 表示将向量 x 和 y 拼接.

2.3 时序知识图谱的预训练模型

为了将 PKGM 应用于时序知识图谱, 我们根据 Wu 等人[9]对实体和关系时间感知嵌入的定义, 定义实体对时间 t 的感知嵌入 z_e^t 为:

$$z_e^t[n]=\begin{cases} z_e[n]\sin(w_e[n]t+b_e[n]), & 1\leqslant n\leqslant \gamma d \\ z_e[n], & \gamma d<n\leqslant d \end{cases} \tag{4}$$

其中, w_e 和 b_e 是特定于实体的可学习向量参数, d 是嵌入总维度, γ 是一个 0–1 之间的超参数用于嵌入分段, n 代表嵌入的第 n 维, 向量 z_e 的前 γd 维用于捕获实体的时序特性, z_e^t 的第 $\gamma d+1$ 到第 d 维捕获实体的静态特性, sin 是 sine 激活函数. 同样的定义也被应用于时间感知的关系嵌入 z_r^t.

具体来说, 在时序知识图谱预训练过程中, 对于给定四元组 (s,r,o,t), 我们用实体和关系对时间 t 的感知嵌入 (z_s^t, z_o^t 和 z_r^t) 替换公式 (1)–(3) 中所有实体和关系嵌入 (z_s, z_o 和 z_r), 其他部分均与第 2.2 节保持一致. 以预训练模型对四元组 (s,r,o,t) 的评分为例, 根据公式 (1) 和公式 (4), 我们有:

$$M(s,r,o,t)=f_{\mathrm{KGE}}(s,r,o,t)+f_{\mathrm{rel}}(s,r,t) \tag{5}$$

其中, $f_{\mathrm{rel}}(s,r,t)=-\|M_r z_s^t-z_r^t\|$, 对基于 TransE 的预训练模型 $f_{\mathrm{KGE}}(s,r,o,t)=-\|z_s^t+z_r^t-z_o^t\|$, 对基于 ComplEx 的预训练模型 $f_{\mathrm{KGE}}(s,r,o,t)=\mathrm{Re}(z_s^{t\top}\mathrm{diag}(z_r^t)\bar{z}_o^t)$.

同样的, 在为下游任务提供向量服务过程中, 对于给定实体 e 和时间 t, 时序知识图谱预训练模型提供实体自身的嵌入表示 z_e^t 和该实体的上下文表示 c_e^t.

3 基于双向模仿蒸馏的高效增量学习方法

本节介绍我们提出的基于双向模仿蒸馏的高效增量学习方法 BID-NAS. 第 3.1 节介绍基于双向模仿蒸馏的训练策略 BID, 它旨在捕捉新数据中的变化同时保留旧模型已学到的知识. 第 3.2 节介绍基于样本常规性和反常性的采样策略 NAS, 它旨在减少训练数据规模以提升训练效率. 第 3.3 节介绍逆重放机制, 该机制有助于提升负样本质量. 第 3.4 节介绍最终增量训练的损失函数. 图 1 显示了基于双向模仿蒸馏的高效增量学习方法的整体框架.

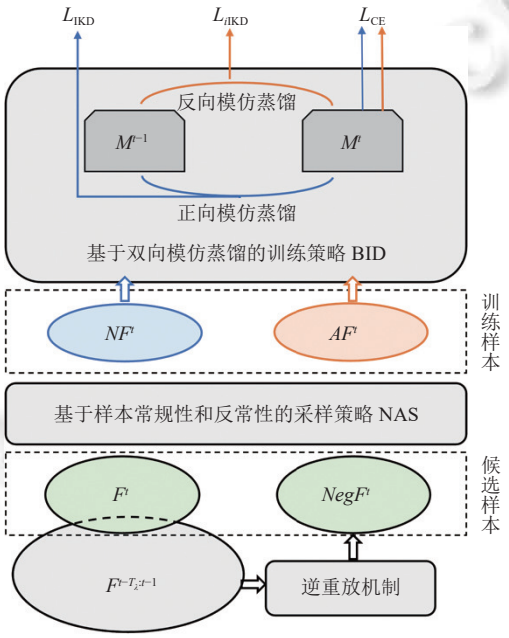

图 1 基于双向模仿蒸馏的高效增量学习方法 BID-NAS 框架图

3.1 基于双向模仿蒸馏的训练策略 (BID)

知识蒸馏技术[36]是增量学习中常用的技术, 核心思想是将当前数据同时输入当前模型 M^t 和旧模型 M^{t-1} 中, 使 M^t 模仿 M^{t-1} 输出的以保留旧模型已学到的知识, 避免灾难性遗忘. 然而, 对于具有高动态性的社交知识图谱, 传统基于知识蒸馏的增量学习方法往往忽略了区分不同样本 (四元组) 对于模型更新的作用. 我们根据对模型更新的作用不同将当前样本划分为两类: 常规样本 (normal sample, NS^t) 和反常样本 (abnormal sample, AS^t), 并提出基于双向模仿蒸馏的训练策略 BID 以充分发挥不同类型样本对模型更新的作用, 包含针对常规样本的正向模仿蒸馏和针对反常样本的反向模仿蒸馏.

3.1.1 常规样本和反常样本

1) 常规样本不会使当前模型中相关用户的表示产生较大变动, 主要起到保护旧模型已经学习到的知识的作用. 这类样本可能是旧模型已经能很好掌握的样本, 即旧模型能直接给出较高分数的正样本或者较低分数的负样本, 或者对当前时间敏感程度低的样本, 即样本标签一段时间内未发生变化 (始终作为正样本或负样本). 它们在旧模型 M^{t-1} 中的四元组分数对于当前模型 M^t 具有较高的模仿价值.

2) 反常样本主要起到帮助当前模型捕获知识图谱变化的作用, 会显著改变当前模型中相关用户的表示. 这类样本可能是旧模型还不能掌握的样本, 即旧模型会给出较低 (高) 分数的正 (负) 样本, 或者对当前时间敏感程度高

的样本, 即样本在过去和当前时间的标签相反, 过去为正 (负) 样本但现在为负 (正) 样本. 它们在旧模型 M^{t-1} 中的四元组分数对于当前模型 M^t 的模仿价值较低, 若使 M^t 强行模仿 M^{t-1} 对反常样本的评分不利于 M^t 学习用户的特征变化, 降低模型性能.

3.1.2 针对常规样本的正向模仿蒸馏

正向模仿蒸馏旨在使当前模型 M^t 模仿旧模型 M^{t-1} 对常规样本的评分, 使当前模型保留旧模型已经学习的知识. 具体来说, 给定一个常规样本 $(s,r,o,t)\in NS^t$, 正向模仿蒸馏的目的是希望当前模型给出的四元组分数 $M^t(s,r,o,t)$ 与旧模型给出的分数 $M^{t-1}(s,r,o,t)$ 近似, 且常规性越高的样本会被分配更高的正向模仿优化权重. 正向模仿蒸馏损失被定义为 $M^{t-1}(s,r,o,t)$ 与 $M^t(s,r,o,t)$ 之差的加权和:

$$\mathcal{L}_{IKD}=\sum_{(s,r,o,t)\in NS^t} nor_{srot}\cdot d_\delta(M^{t-1}(s,r,o,t),M^t(s,r,o,t)) \tag{6}$$

其中, nor_{srot} 是样本常规性大小 (具体计算方法见公式 (9)), d_δ 是 $\delta=1$ 的 Huber 损失:

$$d_\delta(a,b)=\begin{cases}\dfrac{1}{2}(a-b)^2, & |a-b|\leqslant 1\\ |a-b|-\dfrac{1}{2}, & |a-b|>1\end{cases} \tag{7}$$

3.1.3 针对反常样本的反向模仿蒸馏

反向模仿蒸馏旨在拉开当前模型 M^t 和旧模型 M^{t-1} 对反常样本的评分的差距, 促使模型学习知识图谱的变化. 给定一个反常样本 $(s,r,o,t)\in AS^t$, 反向模仿蒸馏的目的是希望当前模型给出的四元组分数 $M^t(s,r,o,t)$ 与旧模型给出的分数 $M^{t-1}(s,r,o,t)$ 的差值尽可能大, 且反常性越高的样本会被分配更高的反向模仿优化权重. 反向模仿蒸馏损失被定义为 $M^{t-1}(s,r,o,t)$ 与 $M^t(s,r,o,t)$ 的加权边际损失:

$$\mathcal{L}_{iIKD}=\sum_{(s,r,o,t)\in AS^t} abn_{srot}\cdot\max\{\eta+M^{t-1}(s,r,o,t)-M^t(s,r,o,t),0\} \tag{8}$$

其中, abn_{srot} 是样本反常性大小 (具体计算方法见公式 (10)), η 是代表四元组分数差距阈值的超参数.

3.2 基于样本常规性和反常性的采样策略 (NAS)

在大规模社交知识图谱中, 由于用户数量庞大且用户间交互频繁, 每周都会新增数千万的事实. 将所有新增样本作为模型更新的训练数据是不现实的, 其中大量的冗余数据会降低模型的更新效率. 为了在保障模型性能的前提下尽可能减少训练数据集规模, 我们提出一种基于样本常规性和反常性的采样策略 (NAS).

3.2.1 样本常规性和反常性计算

我们根据以下两个因素评估四元组 (s,r,o,t) 的常规性和反常性: 旧模型 M^{t-1} 对该样本的掌握程度和样本对当前时间 t 的敏感程度, 然后将具有高常规性 (反常性) 的样本采样为常规 (反常) 样本.

对于给定的四元组 (s,r,o,t), 旧模型 M^{t-1} 对它的掌握程度取决于旧模型对该四元组的评分 $M^{t-1}(s,r,o,t)$ (具体计算方法参考第 2.3 节), 对于正 (负) 样本, 旧模型给出的评分越高 (低) 则该样本常规性越高; 反之, 该样本常规性越低.

四元组 (s,r,o,t) 对当前时间的敏感程度取决于事实 $(s,r,o,*)$ 在过去的一段时间内出现的情况. 一个直觉是, 如果事实 $(s,r,o,*)$ 在时间 t 之前刚发生过或者在过去一段时间反复出现, 我们可以认为时间变化对这个事实是否发生的影响不大, 即 (s,r,o,t) 是时间不敏感的; 相反, 如果事实 $(s,r,o,*)$ 在时间 t 之前已经很久没有出现过或者在过去一段时间几乎没有出现, 则可以认为时间变化对这个事实是否发生可能有较大影响, 即 (s,r,o,t) 是时间敏感的. 具体来说, 对于正样本 $(s,r,o,t)\in F^t$, 在过去的事实集 $F^{t-T_\lambda:t-1}$ 中事实 $(s,r,o,*)$ 出现的频率越高或出现时间越接近当前时间 t, 它对当前时间的敏感程度越低, 样本常规性越高; 反之, 其对当前时间的敏感程度越高, 样本常规性越低. 对于负样本 $(s,r,o,t)\notin F^t$ 则相反, 在过去的事实集 $F^{t-T_\lambda:t-1}$ 中事实 $(s,r,o,*)$ 出现的频率越高或出现时间越接近当前时间 t, 它对当前时间的敏感程度越高, 样本常规性越低; 反之, 其对当前时间的敏感程度越低, 样本常规性越高. 因此, 我们定义四元组 (s,r,o,t) 的常规性为:

$$nor_{srot}=\begin{cases}\sigma(M^{t-1}(s,r,o,t))\cdot\sigma\left(w_0\dfrac{1+n_{sro}^{t-T_\lambda:t-1}}{\min\{t-t',T_\lambda\}}\right), y=1\\ (1-\sigma(M^{t-1}(s,r,o,t)))\cdot\sigma\left(w_1\dfrac{\min\{t-t',T_\lambda\}}{1+n_{sro}^{t-T_\lambda:t-1}}\right), y=0\end{cases} \tag{9}$$

其中, T_λ 是有效的旧数据时间窗口, 它是一个人为设置的超参数. t' 是事实 $(s,r,o,*)$ 在过去时间窗口 $t-T_\lambda:t-1$ 内最近一次出现的时间, 如果事实 $(s,r,o,*)$ 未在时间窗口 $t-T_\lambda:t-1$ 内出现, 则 $t'=0$. $n_{sro}^{t-T_\lambda:t-1}$ 是事实 $(s,r,o,*)$ 在过去的事实集 $F^{t-T_\lambda:t-1}$ 中出现的总次数. w_0 和 w_1 是大于 0 的可学习放缩系数. y 是四元组的真实标签, 对于正样本 $y=1$, 对于负样本 $y=0$. 与常规性负相关, 四元组 (s,r,o,t) 的反常性被简单定义为:

$$abn_{srot}=1-nor_{srot} \tag{10}$$

注意, 虽然公式 (9) 计算时间敏感程度 (事实在过去出现的次数 $n_{sro}^{t-T_\lambda:t-1}$ 和当前最短时间间隔 $\min\{t-t',T_\lambda\}$) 需要扫描所有历史数据, 但本质上 $n_{sro}^{t-T_\lambda:t-1}$ 和 $\min\{t-t',T_\lambda\}$ 只与数据本身有关而与模型无关, 因此它们可以在数据预处理阶段离线并行计算得到, 供训练阶段直接查询使用, 这不会增加模型增量更新的训练开销.

3.2.2 样本采样过程

为了在保证模型性能的前提下缩减训练数据规模, 我们从当前正样本候选集 F^t 和负样本候选集 Neg^t (Neg^t 的具体生成方法见第 3.3 节) 中采样具有高常规性的样本 NS^t 和具有高反常性样本 AS^t 作为最终的训练数据 $NS^t\cup AS^t$. 我们用 r_n 代表常规事实采样率, 用 r_a 代表反常事实采样率. 记正负样本采样比为 $1:n_{\text{neg}}$.

具体的, 对于常规样本采样, 我们以公式 (9) 的样本常规性 nor_{srot} 作为样本被采样的概率, 将被采样的样本从候选集 $F^t\cup Neg^t$ 中删除并加入常规样本集 NS^t, 直到 NS^t 中正样本数达到 $r_n\cdot|F^t|$, 负样本数达到 $n_{\text{neg}}\cdot r_n\cdot|F^t|$; 对于反常样本采样, 我们以公式 (10) 的样本反常性 abn_{srot} 作为样本被采样的概率, 将被采样的样本从候选集 $F^t\cup Neg^t$ 中删除并加入反常样本集 AS^t, 直到 AS^t 中正样本数达到 $r_a\cdot|F^t|$, 负样本数达到 $n_{\text{neg}}\cdot r_a\cdot|F^t|$. 一个四元组不能被同时采样为常规样本和反常样本, $NS^t\cap AS^t=\varnothing$. 记总的事实采样率 $r=r_n+r_a(0<r\leqslant 1)$, 则最终的训练数据中正样本数量为 $r\cdot|F^t|$, 总样本量为 $|NS^t\cup AS^t|=(1+n_{\text{neg}})\cdot r\cdot|F^t|$.

3.3 逆重放机制

知识图谱表示学习中传统的负样本生成机制是通过随机破坏正样本, 即对于给定的事实 $(s,r,o,t)\in F^t$, 将实体 s 或实体 o 随机替换为 E_{known}^t 中任意其他实体, 得到负样本 $(s,r,e,t)\notin F^t$ 或 $(e,r,o,t)\notin F^t$. 然而, 由于知识图谱通常是稀疏的, 这种方法的负样本候选集 $Neg^t=\{s,r,o,t|s\in E_{\text{known}}^t\wedge r\in R^t\wedge o\in E_{\text{known}}^t\}\backslash F^t$ 非常庞大, 很容易采样到过于简单的低质量负样本.

TIE[9]提出从最近被删除的事实中采样负样本, 即在过去时间 $t-1$ 发生但当前时间 t 未发生的事实被视为当前的负样本, 这种方式的负样本候选集 $Neg^t=\{(s,r,o,t)|(s,r,o,t-1)\in F^{t-1}\wedge(s,r,o,t)\notin F^t\}$ 又过于限制, 对社交知识图谱来说很容易采样到不合理的负样本. 不同于常识知识图谱, 社交知识图谱中同样的事件可以在不同时间多次出现, 这本质上意味着两个用户亲密程度的积累和增加. 例如对于两个近期有大量交互历史如 (Mike, share_product, David, *) 的用户 Mike 和 David, 尽管当前暂时未观察到他们新的交互行为, 但这不足以表明他们亲密度降低, 他们仍然被认为是亲密的且有很高的概率会再次交互, 因此 (Mike, share_product, David, t) 并不适合作为当前的负样本.

3.3.1 逆重放机制提升负样本质量

为了生成更适合社交知识图谱表示学习的高质量负样本, 我们提出了逆重放机制. 该机制将负样本候选空间限制在过去时间窗口 $t-t_\lambda:t-1$ 发生的事实中, 将过去的正样本部分逆向重放为当前的负样本, 并计算负样本的质量得分作为其优化权重. 具体来说, 我们首先从过去的事实集 $F^{t-t_\lambda:t-1}$ 中删除那些也出现在当前事实集 F^t 中的事实, 即如果满足 $(s,r,o,*)\in F^{t-t_\lambda:t-1}\wedge(s,r,o,t)\in F^t$, 则将 $(s,r,o,*)$ 从 $F^{t-t_\lambda:t-1}$ 中删除. 基于过去事实集的剩余部分, 我们修改四元组 $(s,r,o,*)$ 的时间为 t, 并将其标签逆转 (由过去的 $y=1$ 改为 $y=0$), 获得最终的负样本候选集 $Neg^t=\{(s,r,o,t)|(s,r,o,*)\in F^{t-T_\lambda:t-1}\wedge(s,r,o,t)\notin F^t\}$.

负样本 (s,r,o,t) 的质量得分与当前模型 M^t 对该事实的掌握程度有关[53], 若分数 $M^t(s,r,o,t)$ 较高, 说明模型倾向于将其判断为正样本, 该样本有较大的可能是高质量负样本. 在此基础上, 为了保证负样本的合理性, 降低与两个亲密 (有大量交互历史或近期交互记录) 用户相关的负样本质量得分, 我们统计在时间窗口 $t-T_\lambda:t-1$ 中用户 s 和用户 o 的历史交互情况, 即在过去的事实集 $F^{t-T_\lambda:t-1}$ 中事实 $(s,*,o,*)$ 出现的次数 $n_{so}^{t-T_\lambda:t-1}$ 以及最近一次出现的时间 t'. 若事实 $(s,*,o,*)$ 在过去出现次数较少, 出现时间距离当前时间较远, 则负样本 (s,r,o,t) 的质量较高; 反之, 负样本质量较低. 我们定义负样本质量得分为:

$$q_{srot}=\sigma(M^t(s,r,o,t))\cdot\sigma\left(w_2\frac{\min\{t-t',T_\lambda\}}{1+n_{so}^{t-T_\lambda:t-1}}\right) \tag{11}$$

其中, σ 是 Sigmoid 激活函数, w_2 是大于 0 的可学习放缩系数. 与第 3.2.1 节所述相同, 公式 (11) 中 $n_{so}^{t-T_\lambda:t-1}$ 和 $\min\{t-t',T_\lambda\}$ 只与数据本身有关而与模型无关, 因此它们可以在数据预处理阶段离线并行计算得到, 供训练阶段直接查询使用, 这不会增加模型增量更新的训练开销. 在计算以真实标签为监督信号的交叉熵损失时, 我们将负样本质量得分作为负样本的优化权重, 使模型更关注高质量负样本:

$$\mathcal{L}_{CE}=-\sum_{(s,r,o,t)\in NS^t\cup AS^t}y\log\sigma(M^t(s,r,o,t))+q_{srot}\cdot(1-y)\log(1-\sigma(M^t(s,r,o,t))) \tag{12}$$

其中, y 是四元组的真实标签, 对于正样本, $y=1$; 对于负样本, $y=0$.

3.4 增量训练损失函数

由于本文所提出的训练框架涉及多个学习任务, 手动调整每个任务的损失权重对于大规模知识图谱是昂贵和难以实现的, 为减少增量训练期间超参数调整的工作量, 我们使用多任务自适应损失[54]自动为不同任务分配学习权重. 最终的增量训练损失函数为:

$$\mathcal{L}=\frac{1}{\gamma_0^2}\mathcal{L}_{CE}+\frac{1}{\gamma_1^2}\mathcal{L}_{iIKD}+\frac{1}{\gamma_2^2}\mathcal{L}_{IKD}+\sum_{i=0}^{2}2\cdot\log\gamma_i^2 \tag{13}$$

其中, $\gamma_i\ (i=0,1,2)$ 是可学习的参数.

4 实验分析

本节首先在第 4.1 节介绍增量训练相关的实验设置, 包含数据集和基础模型介绍、对比的基线增量学习方法和增量训练设置. 然后, 第 4.2 节在模型自身的训练任务上评估了不同增量学习方法的性能和效率, 并在第 4.3 节中对比了使用不同增量学习方法更新的基础模型以及未更新的基础模型在 3 个常见的下游任务中提供服务的能力.

4.1 实验设置

4.1.1 数据集和基础模型介绍

我们的 10 亿规模的电商社交知识图谱 (SKG) 包含了淘宝平台上 2021 全年 (2021.1.1–2021.12.30) 淘宝用户的社交数据, 涉及 4 亿用户共计 90.4 亿条记录. 我们在社交知识图谱的子集 SKG-sub 上进行了实验, 该数据集包含 50 775 620 个事实 (四元组), 涉及 6 974 959 个用户和 15 种关系, 我们以周为单位重置四元组的时间戳 t, 整个数据集 SKG-sub 时间跨度为 52 周, $t=[0,1,2,\ldots,51]$. 我们用前 28 周的事实 $F^{0:27}$ 预训练基础模型 M^{base}. 与 TIE[9] 相同, 我们的知识图谱预训练模型基于知识图谱嵌入方法 ComplEx[52]构建, ComplEx 在多关系知识图谱上能够较好地平衡模型的参数量与表达能力, 适合大规模社交知识图谱场景. 具体的, 我们设置公式 (4) 中模型维度 $d=64$ 和 $\gamma=0.5$, 最终模型 M^{base} 大小为 13.3 GB.

4.1.2 增量学习对比基线

为了证明我们提出的基于双向模仿蒸馏的高效增量学习方法, 包括基于双向模仿蒸馏的训练策略 BID (第 3.1 节), 基于样本常规性和反常性的采样策略 NAS (第 3.2 节) 和逆重放机制 (第 3.3 节) 的有效性和高效性, 我们总共在 3 种增量学习的训练策略、5 种采样策略和 3 种负样本生成方法上进行了实验.

4.1.2.1 3 种增量学习的训练策略

(1) 微调 (fine tuning, FT), 直接在旧模型的基础上用当前样本微调模型, 最终训练损失函数为公式 (12).

(2) 当前唯一的时序知识图谱增量学习方法 (TIE[9]), 除了对当前样本计算交叉熵损失 (公式 (12)), 该方法还从历史事实 $F^{t-T_\lambda:t-1}$ 中筛选一个子集 $H^{t-T_\lambda:t-1}$ 用于重放, 使当前模型模仿旧模型对重放事实的输出, 并对新旧模型参数增加正则化避免灾难性遗忘. 在我们的实验中, 历史重放事实 $H^{t-T_\lambda:t-1}$ 的筛选方式, 重放损失和参数正则化损失的计算方法均与 TIE[9]原文相同.

(3) 我们的基于双向模仿蒸馏的训练策略 (BID), 参考第 3.1 节.

为了公平对比不同方法的性能和效率, 我们控制 3 种增量学习训练策略在训练过程使用的采样样本 (包括负样本) 完全相同, 即 $NS^t \cup AS^t$. 具体的, BID 和 FT 的训练样本集为 $NS^t \cup AS^t$. TIE 的训练样本集为 $NS^t \cup AS^t \cup H^{t-T_\lambda:t-1}$.

4.1.2.2 5 种获得训练样本 $NS^t \cup AS^t$ 的采样策略

(1) 随机采样 (randomly sampling, 简记为 RND), 所有样本的常规性 nor_{srot} 和反常性 abn_{srot} 均为 1/2.

(2) 基于模式频数的采样 (pattern count sampling, 简记为 PAT), 这种方法被 TIE[9]用来筛选用于重放的历史事实 H^t, 其中拥有较高模式频数的事实被认为可以保护旧模型已学到的知识. 我们通过模式频数计算样本常规性. 对于给定样本 (s,r,o,t), 其模式集为 $P=\{(s,r,o),(s,*,o),(s,r,*),(*,r,o),(s,*,o),(s,*,*),(*,*,o)\}$, 首先统计每个模式的历史出现频数, 即过去事实集 $F^{t-T_\lambda:t-1}$ 中满足该模式的事实总数 $n_p^{t-T_\lambda:t-1}$, 并将对数尺度下各模式频数的加权和作为该样本的模式频数:

$$pat_{srot}=\sum_{p\in P}\lambda_p \log(n_p^{t-T_\lambda:t-1}+1) \tag{14}$$

其中, λ_p 是各模式频数与所有模式总频数的比值. 对于正样本, 模式频数越大, 样本常规性越高; 对于负样本, 模式频数越大, 样本常规性越低. 公式 (9) 被重写为:

$$nor_{srot}=\begin{cases}\sigma(w_0 pat_{srot}), & y=1\\ \sigma\left(\dfrac{w_1}{pat_{srot}}\right), & y=0\end{cases} \tag{15}$$

(3) 基于样本时间敏感程度的采样 (简记为 NASt), 这是我们采样策略的第一种变体, 即只通过样本对当前时间敏感程度评估样本常规性, 公式 (9) 被重写为:

$$nor_{srot}=\begin{cases}\sigma\left(w_0\dfrac{1+n_{sro}^{t-T_\lambda:t-1}}{\min\{t-t',T_\lambda\}}\right), & y=1\\ \sigma\left(w_1\dfrac{\min\{t-t',T_\lambda\}}{1+n_{sro}^{t-T_\lambda:t-1}}\right), & y=0\end{cases} \tag{16}$$

(4) 基于样本旧模型得分的采样 (简记为 NASs), 这是我们采样策略的第二种变体, 即只通过旧模型对该样本的掌握程度评估样本常规性, 公式 (9) 被重写为:

$$nor_{srot}=\begin{cases}\sigma(M^{t-1}(s,r,o,t)), & y=1\\ 1-\sigma(M^{t-1}(s,r,o,t)), & y=0\end{cases} \tag{17}$$

(5) 我们的基于样本常规性和反常性的采样策略 (NAS), 参考第 3.2 节.

4.1.2.3 3 种负样本生成机制

(1) 传统负样本生成机制 (简记为 trad), 即通过随机破坏正样本来生成负样本, 负样本候选集为 $Neg^t=\{(s,r,o,t)|s\in E^t_{\text{known}}\wedge r\in R^t\wedge o\in E^t_{\text{known}}\}\backslash F^t$, 公式 (12) 中的所有负样本质量得分 q_{srot} 均为 1.

(2) 最近删除机制 (简记为 del)[9], 将时间 $t-1$ 发生但时间 t 未发生的事实作为负样本, 负样本候选集为 $Neg^t=\{(s,r,o,t)|(s,r,o,t-1)\in F^{t-1}\wedge(s,r,o,t)\notin F^t\}$, 公式 (12) 中的所有负样本质量得分 q_{srot} 均为 1.

(3) 我们的逆重放机制, 参考第 3.3 节.

具体来说, 我们将 3 种增量学习训练策略 (FT、TIE 和 BID) 和 5 种采样策略 (RND、PAT、NASt、NASs 和 NAS) 组合成 15 种不同的增量学习方法. 首先, 我们在统一应用逆重放机制的设定下, 对比了我们的基于双向模仿蒸馏的高效增量学习方法 BID-NAS 和其他 14 种基线增量学习方法的性能和效率, 证明我们的基于双向模仿蒸馏

的训练策略和基于样本常规性和反常性的采样策略的有效性和高效性. 然后, 我们将负样本生成方式由逆重放机制替换为其他两种负样本生成机制 (trad、del) 并进行对比, 证明我们的逆重放机制在提升负样本质量上的优越性.

4.1.3 增量训练设置

我们用第 28–51 周 ($t=[28,29,\ldots,51]$) 的数据进行增量学习实验. 在每个增量训练时间步 t, 先用上一时间步的模型 M^{t-1} 初始化当前模型, 然后用当前数据对其进行增量训练, 更新训练后的结果记为模型 M^t. 具体来说, 当 $t=28$, 在 M^{base} 的基础上, 用第 28 周的数据对其增量训练, 记更新后的模型为 M^{28}. 当 $t=29$, 在 M^{28} 基础上, 用第 29 周的数据对其增量训练, 记更新后的模型为 M^{29}, 以此类推, 最后一次更新得到的模型为 M^{51}. 我们的实验在 PyTorch-lightning 上实现. 训练过程中, 我们设置正负样本比例为 $1:n_{neg}=1:200$, 采用 Adam[55]优化器, 设置初始学习率为 1E–4, 公式 (8) 中的四元组分数差距阈值 $\eta=1$. 相较于更早期的用户行为, 近 3 个月内的用户行为更能影响当前用户表征[56], 我们设置旧数据时间窗口 $T_\lambda=12$. 总的事实采样率 $r=[0.2,0.4,0.6,0.8,1]$ (第 3.2.2 节), 其中常规事实和反常事实采样比例 $r_n:r_a=7:3$. 为公平对比不同增量学习方法的性能和效率, 所有实验都在单个 Tesla-V100 GPU (32 GB) 上进行, 以批量大小 32 训练 10 轮.

4.2 模型自身训练任务的评估

4.2.1 评估指标

对于增量学习方法的性能, 评估指标为模型自身的训练任务, 链接预测的标准指标 MRR 和 Hit@k (k=1, 3, 10). 具体来说, 对于每个时间步 t 更新后的模型 M^t, 给定一个测试四元组 $(s,r,o,t)\in F^t_{test}$, 我们用每一个实体 $e\in E^t_{known}$ 替换实体 s 生成候选四元组 (e,r,o,t), 然后计算所有候选四元组的得分并降序排列, 将四元组 (s,r,o,t) 自身得分的排名作为其头实体预测排名 $rank_s$. 类似的, 替换尾实体 o 并计算其尾实体预测排名 $rank_o$. 将 $rank_s$ 和 $rank_o$ 的均值作为 (s,r,o,t) 的最终排名. 通过所有测试四元组的最终排名计算 MRR 和 Hit@k, MRR 是它们的平均倒数排名, Hit@k 衡量最终排名在前 k 的测试四元组的百分比.

对于增量学习方法的效率, 评估指标为每个时间步 t 完成包含样本采样过程以及模型训练更新的总时间消耗, 以小时为单位. 最后, 我们将每种增量学习方法在 24 个增量更新时间步的平均 MRR、平均 Hit@k 和平均时间消耗作为该方法的最终指标结果.

4.2.2 增量学习方法性能对比

表 1 显示了在事实采样率为从 $r=1$ 到 $r=0.2$ 的情况下, 基础模型 M^{base} 通过由 3 种训练策略和 5 种采样策略组合的 15 种增量学习方法 (第 4.1.2 节) 在 24 个时间步增量训练后的 MRR 和 Hit@k ($k=1, 3, 10$).

表 1 不同事实采样率 (r) 下, 15 种增量学习方法在 24 个增量训练步的平均 MRR 和 Hit@k ($k=1, 3, 10$)

增量学习方法	MRR					Hit@10					Hit@3					Hit@1				
	$r=1$	0.8	0.6	0.4	0.2	$r=1$	0.8	0.6	0.4	0.2	$r=1$	0.8	0.6	0.4	0.2	$r=1$	0.8	0.6	0.4	0.2
FT-RND	0.253	0.238	0.210	0.168	0.122	0.370	0.351	0.317	0.262	0.207	0.300	0.283	0.252	0.206	0.157	0.188	0.170	0.148	0.112	0.072
FT-PAT	0.253	0.240	0.212	0.170	0.124	0.370	0.354	0.320	0.266	0.211	0.300	0.287	0.255	0.210	0.159	0.188	0.173	0.150	0.115	0.076
FT-NASt	0.253	0.243	0.215	0.176	0.131	0.370	0.356	0.324	0.272	0.219	0.300	0.289	0.259	0.215	0.167	0.188	0.176	0.154	0.121	0.083
FT-NASs	0.253	0.245	0.218	0.181	0.138	0.370	0.359	0.328	0.279	0.227	0.300	0.292	0.263	0.221	0.174	0.188	0.180	0.157	0.125	0.089
FT-NAS	0.253	0.248	0.221	0.184	0.144	0.370	0.364	0.330	0.284	0.234	0.300	0.294	0.266	0.225	0.181	0.188	0.183	0.161	0.129	0.095
TIE-RND	0.262	0.251	0.235	0.204	0.168	0.381	0.368	0.347	0.308	0.262	0.310	0.298	0.279	0.245	0.207	0.197	0.185	0.169	0.144	0.114
TIE-PAT	0.262	0.260	0.243	0.216	0.185	0.381	0.379	0.358	0.321	0.282	0.310	0.308	0.289	0.260	0.226	0.197	0.195	0.181	0.157	0.131
TIE-NASt	0.262	0.253	0.236	0.204	0.169	0.381	0.371	0.348	0.308	0.262	0.310	0.300	0.280	0.246	0.207	0.197	0.188	0.171	0.145	0.114
TIE-NASs	0.262	0.256	0.239	0.210	0.177	0.381	0.373	0.353	0.315	0.272	0.310	0.303	0.284	0.252	0.216	0.197	0.190	0.175	0.151	0.122
TIE-NAS	0.262	0.257	0.241	0.213	0.180	0.381	0.376	0.355	0.318	0.276	0.310	0.305	0.286	0.255	0.220	0.197	0.193	0.178	0.153	0.126
BID-RND	0.273	0.263	0.247	0.222	0.192	0.396	0.382	0.363	0.328	0.292	0.322	0.311	0.294	0.264	0.234	0.207	0.195	0.183	0.161	0.135
BID-PAT	0.273	0.265	0.250	0.224	0.196	0.396	0.384	0.365	0.332	0.298	0.322	0.313	0.297	0.268	0.237	0.207	0.198	0.185	0.164	0.138
BID-NASt	0.273	0.267	0.254	0.231	0.207	0.396	0.387	0.370	0.341	0.312	0.322	0.315	0.301	0.276	0.249	0.207	0.201	0.190	0.171	0.148
BID-NASs	0.273	0.269	0.256	0.235	0.212	0.396	0.390	0.374	0.347	0.318	0.322	0.318	0.304	0.280	0.255	0.207	0.203	0.193	0.174	0.154
BID-NAS	0.273	0.271	0.263	0.240	0.220	0.396	0.393	0.378	0.353	0.327	0.322	0.320	0.312	0.286	0.264	0.207	0.206	0.196	0.179	0.161

首先, 我们对比不同的采样策略, 图 2 可视化了表 1 中每种采样策略的平均 MRR 和 Hit@k, 其中一条曲线对应一种采样策略, 例如 x-RND 代表基于随机采样策略 RND 的增量学习方法 (FT-RND, TIE-RND, BID-RND) 的平均结果.

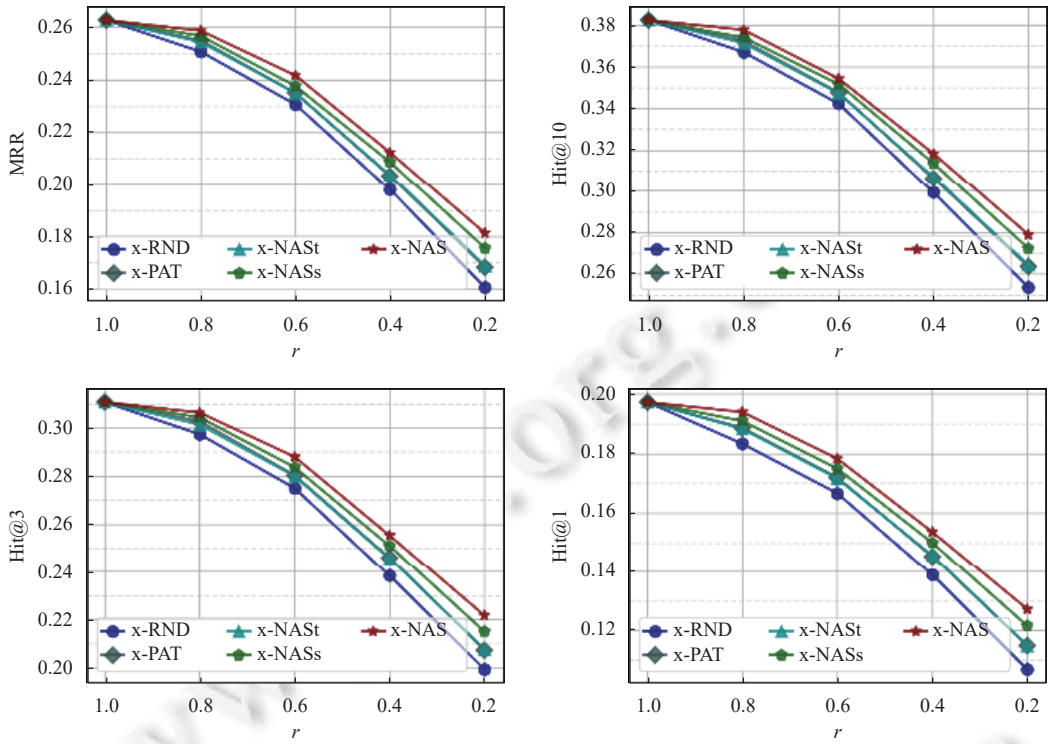

图 2 不同事实采样率 (r) 下, 基于各种采样策略的增量学习方法的平均性能

从图 2 可知, 我们的采样策略 NAS 取得了优于所有基线采样策略的性能. 整体上, 随着训练数据的减少 (事实采样率降低), 所有方法都表现出不同程度的性能下降, 其中基于 NAS 的增量学习方法 (x-NAS) 性能下降最缓, 特别是在使用 20% 的训练数据 ($r = 0.2$) 时, 基于 NAS 的增量学习方法 (x-NAS) 相较于基于随机采样策略 RND 的增量学习方法 (x-RND), 平均 MRR、Hit@3 和 Hit@1 都提升了超过 10%, 这是因为 NAS 同时采样了高质量的常规样本和反常样本, 有效保证了训练数据的质量. 在任何采样率下, 随机采样策略 RND 的性能始终是最差的, 这说明随机采样很容易遗漏掉高价值的样本. 基于模式频数的采样策略 PAT 并没有表现出明显优势, 这是因为这种采样策略只能筛选出高质量的常规样本, 忽略了反常样本. 此外, NAS 的两个变体, 即基于样本旧模型得分的采样策略 NASs 和基于样本时间敏感程度的采样策略 NASt, 相较于 NAS 均表现出一定程度的性能退化, 说明旧模型对样本的掌握情况和样本自身对时间的敏感程度均会影响对样本常规性和反常性的判断, NASs 略优于 NASt 反映出旧模型对样本的掌握程度的影响更大.

然后, 我们对比不同的训练策略, 图 3 可视化了表 1 中每种训练策略的平均 MRR 和 Hit@k, 其中一条曲线对应一种训练策略, 例如 FT-x 代表基于微调训练策略 FT 的增量学习方法 (FT-RND, FT-PAT, FT-NASt, FT-NASs, FT-NAS) 的平均结果.

从图 3 可知, 我们的基于双向模仿蒸馏的训练策略 BID 明显优于基于微调的训练策略 FT 和当前最先进的训练策略 TIE. 在使用全量训练数据 ($r = 1$)、60% 的训练数据 ($r = 0.6$) 和 20% 的训练数据 ($r = 0.2$) 的情况下, 基于 BID 的增量学习方法 (BID-x) 比基于 TIE 的增量学习方法 (TIE-x) 的平均 MRR 分别提升了 4.2%、6.4% 和 16.8%, 比基于 FT 的增量学习方法 (FT-x) 的平均 MRR 分别提升了 7.9%、18.0% 和 55.8%. FT 因为在更新模型的

过程中不能保护模型已学到知识, 受灾难性遗忘的影响表现出最差的性能. TIE 虽然一定程度上缓解了灾难性遗忘问题, 但忽略了对知识图谱中反常样本的利用. 不同于 FT 和 TIE, 对于常规样本, BID 通过正向模仿蒸馏使当前模型的输出拟合旧模型的输出以防止灾难性遗忘, 对于反常样本, BID 通过反向模仿蒸馏使当前模型的输出远离旧模型的输出以捕获知识图谱中的变化, 因此基于 BID 的增量学习方法表现出最佳且稳定的性能.

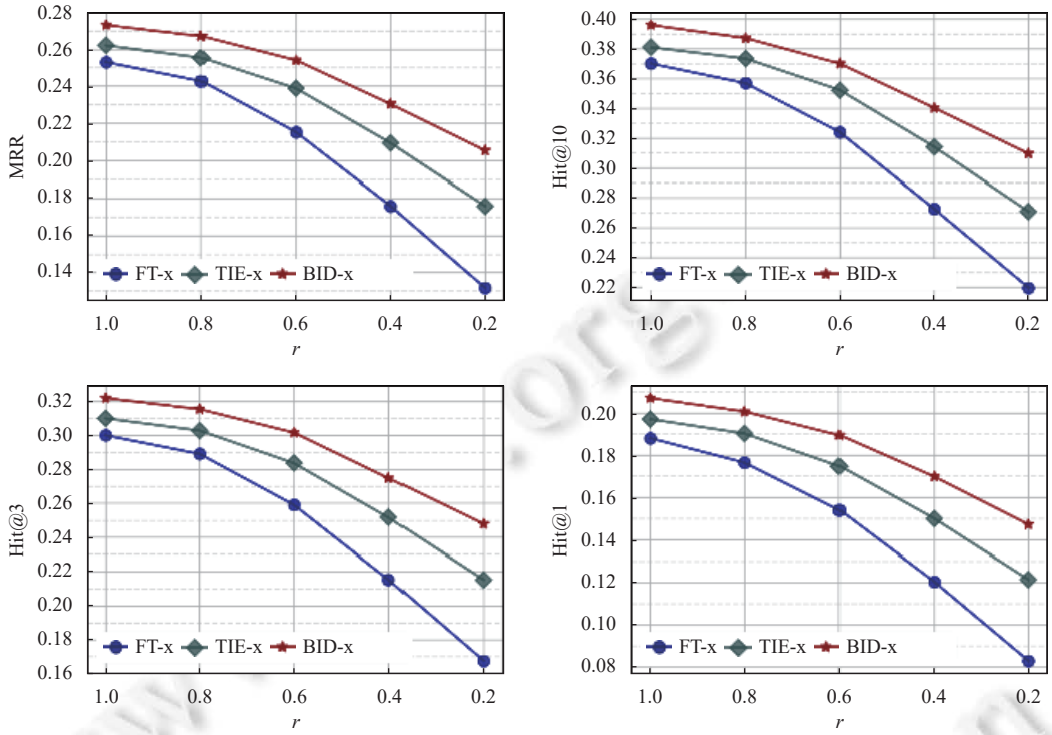

图 3 不同事实采样率 (r) 下, 基于各种训练策略的增量学习方法的平均性能

与当前最先进的时序知识图谱增量学习方法 TIE-PAT 相比, 我们的 BID-NAS 能在训练数据减少 40% 的情况下与 TIE-PAT 使用全部训练数据的表现相当: BID-NAS 在 $r=0.6$ 时的 MRR、Hit@10、Hit@3 和 Hit@1 分别达到 TIE-PAT 在 $r=1$ 时对应指标结果的 100.4%、99.2%、100.7% 和 99.5%. 此外, BID-NAS 还表现出最优的稳定性, 即随着训练数据减少, 模型性能的下降幅度也最小: 在使用 80% 的训练数据 ($r=0.8$) 时, BID-NAS 几乎保持了使用全量数据时的性能, MRR 和 Hit@10 达到了使用全量数据 ($r=1$) 时的 99.3% 和 99.2%, 在使用 60% 的训练数据 ($r=0.6$) 时, BID-NAS 的 MRR 和 Hit@10 达到了使用全量数据时的 96.3% 和 95.5%. 总的来说, BID-NAS 在不同训练样本采样率下均达到了最佳性能, 证明了我们提出的增量学习方法的有效性.

4.2.3 增量学习方法效率对比

图 4 显示了在事实采样率为从 $r=1$ 到 $r=0.2$ 的情况下, 通过 15 种增量学习方法对基础模型完成 24 个时间步的增量训练消耗的平均时间 (以 h 为单位). 从中可以看出训练策略对训练速度有较大的影响, 总体上基于 BID 的方法 (BID-x) 的效率远高于基于 TIE 的方法 (TIE-x), 比基于微调策略的方法 (FT-x) 稍差. 虽然 FT-x 具有最快的训练速度, 但由表 1 结果可知, 由于忽略了保护旧模型已学到知识, 很容易发生灾难性遗忘而导致其性能很差, 难以投入实际的工业应用. TIE-x 在全量训练数据上的训练速度则过于缓慢, 减少训练数据也不能显著提升其训练速度, TIE-x 在 $r=0.6$ 时的训练耗时仍接近于 FT-x 在 $r=1$ 时的训练耗时, 这是因为训练策略 TIE 需要为所有样本的模式筛选部分旧样本并进行数据重放以避免灾难性遗忘, 增加了计算开销. 而 BID-x 的训练速度随着事实采样率降低有显著的提高, 与 $r=1$ 时相比, BID-x 的平均训练时间在 $r=0.8$、0.6、0.4 和 0.2 时分别减少 18.3%、

34.1%、54.1% 和 74.5%. 而采样策略对训练速度仅有轻微影响, 其中基于我们采样策略 NAS 和其变体 NASs 的增量学习方法耗时略多, 这是因为它们在采样过程中需要计算旧模型对样本的评分.

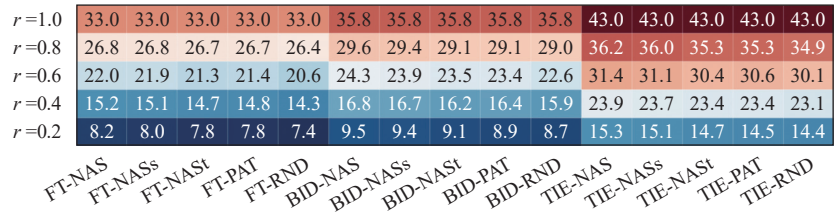

图 4 不同事实采样率 (r) 下, 15 种增量学习方法在 24 个增量训练步的平均消耗时间 (h)

结合第 4.2.2 节对增量学习方法性能的分析, 我们的方法 BID-NAS 能在 $r = 0.6$ 时达到与当前最先进方法 TIE-PAT 在 $r = 1$ 时相当的性能, 而前者完成一次模型更新消耗的时间 (24.3 h) 仅为后者 (43.0 h) 的 56.5%, 即训练速度提升约 77.0%. 这说明当我们的增量学习方法 BID-NAS 是有效且高效的.

4.2.4 对不同模块的有效性分析

本节对我们提出的增量学习方法中每个模块进行有效性分析, 包括基于双向模仿蒸馏的训练策略 BID, 基于样本常规性和反常性的采样策略 NAS 和逆重放机制.

4.2.4.1 对训练策略 BID 的有效性分析

对比 3 种训练方法, 我们的基于双向模仿蒸馏的训练策略 (BID) 总体上表现更好. 例如, 在使用全量训练数据 ($r = 1$)、60% 的训练数据 ($r = 0.6$) 和 20% 的训练数据 ($r = 0.2$) 的情况下, BID 与 5 种采样策略组合 (记为 BID-x) 的平均 MRR 比当前最先进的 TIE 与 5 种采样策略组合 (记为 TIE-x) 的平均 MRR 分别提高 4.2%、6.4% 和 16.8%, 比直接微调的训练策略 (FT-x) 的平均分别提高 7.9%、18.0% 和 55.8%. 另外, 随着训练样本的减少, BID 表现出更好的稳定性, 基于不同训练策略的增量学习方法 (BID-x、TIE-x 和 FT-x) 都表现出不同程度的性能下降, 而 BID-x 的性能下降幅度最小: 从 $r = 1$ 降低到 $r = 0.2$, BID-x、TIE-x 和 FT-x 的平均 MRR 分别降低 24.8%、32.9% 和 47.9%.

另外, 我们对基于不同训练策略在全量数据 ($r = 1$) 下更新的模型进行了具体的案例分析, 结果显示在表 2 中, 具体来说, 我们选择用户 A*为头实体, 用户 Z*为尾实体的交互事件为分析对象, 共包含 5 个测试事实: 在时间 $t = 34$ 时, 用户 A*向用户 Z*发送了一条消息 (A*, send_message, Z*, 29 August 2021), 分享给用户 Z*一件商品 (A, share_product, B, 28 August 2021) 以及回复了用户 Z*的评论 (A*, reply_comment, Z*, 29 August 2021), 在时间 $t = 39$ 时, 用户 A*向用户 Z*发送了一条消息 (A*, send_message, Z*, 3 October 2021) 并帮用户 Z*支付了一次交易 (A*, pay_for, Z*, 3 October 2021). 我们令 $t = 39$ 作为当前时间, 更新后的模型记为 M^{39}, 并分别让基于 FT、TIE 和 BID 训练的 M^{39} 预测这 5 个事实.

表 2 对使用全量数据 ($r = 1$) 及不同训练策略更新的模型案例分析

训练策略	M^{39} 对 $t = 34$ 的过去事实的预测			M^{39} 对 $t = 39$ 的当前事实的预测	
	常规样本	反常样本		常规样本	反常样本
	(A*, send_message, Z*, 29 Aug. 2021)	(A*, share_product, Z*, 28 Aug. 2021)	(A*, reply_comment, Z*, 29 Aug. 2021)	(A*, send_message, Z*, 3 Oct. 2021)	(A*, pay_for, Z*, 3 Oct. 2021)
FT	×	×	×	√	×
TIE	√	×	√	√	×
BID	√	√	√	√	√

由表 2 结果可知, 基于 FT 更新的模型发生了严重灾难性遗忘, 对过去 ($t = 34$) 的 3 个事实都预测失败, 因为 FT 在更当前模型时没有保护模型已学到知识的措施, 此外它不能正确预测当前 ($t = 39$) 的反常样本 (A*, pay_for, Z*, 3 October 2021) 说明 FT 也很难捕获例如用户之间出现新的交互关系这类变化. TIE 虽然能一定程度上克服灾难性遗忘问题, 但由于同样忽略了关注知识图谱中发生的变化, 它对部分反常样本的预测仍然不理想. 其他两种训

练策略相比, BID 一方面可以通过正向模仿蒸馏使当前模型模仿旧模型对常规样本的输出来防止灾难性遗忘, 另一方面, 可以通过反向模仿蒸馏使当前模型对反常样本的输出远离旧模型对反常样本的输出, 来捕获知识图谱中的变化, 因此我们的 BID 表现出更好、更稳定的性能.

4.2.4.2 对采样策略 NAS 的有效性分析

在训练策略 BID 和 FT 下, 我们的基于样本常规性和反常性的采样策略 (NAS) 相较于其他 4 种采样策略均表现出明显优势. 在训练策略 TIE 下, 我们的采样策略 (TIE-NAS) 的表现不如基于模式频数的采样策略 (TIE-PAT), 这是由于相较于 PAT, 我们的 NAS 会采样出更多具有高反常性的样本, 但是 TIE 并没有针对反常样本的特有学习方式, 仍然使当前模型模仿旧模型的输出, 这种做法不利于模型捕获知识图谱中的变化, 这也再次印证了训练策略 TIE 并不能充分利用常规样本和反常样本. 而无论在哪种训练策略下, 随机采样策略 (记为 x-RND) 的性能都是最差的, 这说明随机采样很容易遗漏掉高价值的样本.

此外, 在同一训练策略下, 我们采样策略的两个变体 (基于样本旧模型得分的采样策略 NASs 和基于样本时间敏感程度的采样策略 NASt) 相较于 NAS 都有一定程度的下降, 证明了旧模型对该样本的掌握情况和样本自身对时间的敏感程度均会影响样本常规性和反常性, 其中 NASs 优于 NASt, 表明样本常规性和反常性受旧模型对样本的掌握程度影响更大. 总体上, 与其他的采样策略相比, 我们的采样策略能 NAS 有利于采样出具有更高价值的样本, 在保障方法性能的前提下缩减训练数据规模, 提升训练效率.

4.2.4.3 对逆重放机制的有效性分析

为了验证我们提出的逆重放机制对提升负样本质量的有效性, 我们在训练策略 BID 和 5 种采样策略组合的不同增量学习方法上进行实验, 将原本的逆重放机制 (orig) 分别替换为第 4.1.2 节中介绍的传统的随机负样本生成机制 (trad) 和最近删除机制 (del). 注意, 使用不同的负样本生成机制只影响训练数据中的负样本, 不改变正样本.

表 3 对比了替换不同负样本生成机制后的增量学习方法的性能. 应用逆重放机制的方法 (BID-x orig) 始终优于应用其他两种负样本生成机制的方法 (BID-x trad 和 BID-x del), 并且可以看出随着事实采样率降低, 逆重放机制的优势更加明显, 说明在训练数据有限的情况下更需要高质量的负样本来保障模型性能. 此外, 我们还发现基于最近删除机制的方法 (BID-x del) 在社交图谱上的表现甚至差于传统的随机负样本生成方法 (BID-x trad), 说明将最近被删除 (上一时间步发生但当前未发生) 的事实作为负样本在学习用户社交表征场景并不适用. 我们的逆重放机制考虑了用户在过去一段时间内的交互情况和当前模型对负样本的掌握程度, 有助于生成更适合社交知识图谱的高质量负样本.

表 3 使用不同负样本生成机制在 24 个增量训练步的平均 MRR 和 Hit@k (k = 1, 3, 10)

负样本生成机制		MRR					Hit@10					Hit@3					Hit@1				
		r = 1	0.8	0.6	0.4	0.2	r = 1	0.8	0.6	0.4	0.2	r = 1	0.8	0.6	0.4	0.2	r = 1	0.8	0.6	0.4	0.2
BID-RND	trad	0.270	0.259	0.243	0.215	0.184	0.391	0.377	0.357	0.320	0.282	0.319	0.307	0.289	0.257	0.225	0.205	0.192	0.179	0.156	0.128
	del	0.267	0.255	0.237	0.208	0.174	0.388	0.372	0.350	0.310	0.268	0.316	0.302	0.283	0.248	0.214	0.202	0.189	0.174	0.149	0.119
	orig	0.273	0.263	0.247	0.222	0.192	0.396	0.382	0.363	0.328	0.292	0.322	0.311	0.294	0.264	0.234	0.207	0.195	0.183	0.161	0.135
BID-PAT	trad	0.270	0.261	0.245	0.218	0.188	0.391	0.379	0.360	0.324	0.288	0.319	0.309	0.292	0.262	0.229	0.205	0.195	0.182	0.159	0.132
	del	0.267	0.257	0.240	0.210	0.177	0.388	0.374	0.353	0.313	0.273	0.316	0.305	0.286	0.252	0.216	0.202	0.191	0.177	0.152	0.122
	orig	0.273	0.265	0.250	0.224	0.196	0.396	0.384	0.365	0.332	0.298	0.322	0.313	0.297	0.268	0.237	0.207	0.198	0.185	0.164	0.138
BID-NASt	trad	0.270	0.263	0.249	0.224	0.198	0.391	0.383	0.364	0.333	0.300	0.319	0.312	0.296	0.269	0.239	0.205	0.198	0.186	0.165	0.140
	del	0.267	0.260	0.244	0.216	0.187	0.388	0.378	0.357	0.322	0.286	0.316	0.308	0.290	0.260	0.226	0.202	0.194	0.181	0.158	0.130
	orig	0.273	0.267	0.254	0.231	0.207	0.396	0.387	0.370	0.341	0.312	0.322	0.315	0.301	0.276	0.249	0.207	0.201	0.190	0.171	0.148
BID-NASs	trad	0.270	0.265	0.252	0.229	0.204	0.391	0.385	0.368	0.338	0.307	0.319	0.314	0.299	0.273	0.246	0.205	0.201	0.189	0.169	0.147
	del	0.267	0.262	0.247	0.221	0.193	0.388	0.380	0.361	0.328	0.293	0.316	0.310	0.293	0.264	0.233	0.202	0.198	0.185	0.162	0.136
	orig	0.273	0.269	0.256	0.235	0.212	0.396	0.390	0.374	0.347	0.318	0.322	0.318	0.304	0.280	0.255	0.207	0.203	0.193	0.174	0.154
BID-NAS	trad	0.270	0.268	0.255	0.234	0.212	0.391	0.388	0.372	0.345	0.316	0.319	0.317	0.303	0.279	0.255	0.205	0.203	0.192	0.174	0.154
	del	0.267	0.265	0.250	0.226	0.201	0.388	0.384	0.366	0.335	0.302	0.316	0.313	0.298	0.270	0.243	0.202	0.200	0.188	0.167	0.145
	orig	0.273	0.271	0.263	0.240	0.220	0.396	0.393	0.378	0.353	0.327	0.322	0.320	0.312	0.286	0.264	0.207	0.206	0.196	0.179	0.161

4.3 模型服务下游任务的能力评估

为了进一步说明及时更新预训练模型以维护用户最新社交特征表示的必要性和本文所提方法的有效性, 我们将预训练模型提供的用户表示应用在 3 个以用户为中心的下游任务上, 包括用户分类、好友推荐和商品推荐. 具体的, 我们在每个增量更新时间步 $t=[28,29,\ldots,51]$, 评估了未更新的基础模型 M^{base} 提供的用户表示和更新后的当前最新模型 M^t 提供的用户表示在下游任务中的性能.

4.3.1 下游任务数据集

用户分类任务的数据集包含 24 个更新时间步下的共计 2 384 745 条用户标签记录, 涉及 120 590 个用户, 28 个用户类别, 用户的类别可随时间发生变化. 样本形式是 (user_A, cls_B, t), 表示用户 user_A 在时间 t 被分到类别 cls_B 中. 我们设置每个时间步的训练集、验证集和测试集中样本数比例为 7:1:2.

好友推荐任务的数据集包含 24 个更新时间步下的共计 1 642 383 条好友添加记录, 共涉及 86 712 个用户. 样本形式是 (user_A, user_B, t), 表示用户 user_A 在时间 t 添加用户 user_B 为平台好友. 我们设置每个时间步的训练集、验证集和测试集中样本数比例为 7:1.5:1.5.

商品推荐任务的数据集包含 24 个更新时间步下的共计 1 667 247 条商品浏览 (点击) 记录, 共涉及 13 440 个用户和 30 683 种商品. 样本形式是 (user_A, product_B, t), 表示 user_A 在时间 t 浏览了商品 product_B. 我们将用户在每个时间步下最后两条浏览记录分别作为该时间步的验证和测试样本, 其余的作为该时间步的训练样本.

3 个下游任务的数据详细统计情况如表 4 所示.

表 4 下游任务数据统计情况

任务名称	用户数目	标签类别数目	商品数目	24个更新时间步下累计样本数目		
				训练集	验证集	测试集
用户分类	120 590	28	—	1 669 322	238 475	476 948
好友推荐	86 712	—	—	1 149 668	246 357	246 358
商品推荐	13 440	—	30 683	1 167 073	250 087	250 087

4.3.2 任务定义和训练目标

如第 2.3 节所述, 在服务下游任务时, 对于时间 t 的给定用户实体 e, 时序知识图谱预训练模型提供用户实体自身的嵌入表示 z_e^t 和该用户实体的上下文表示 c_e^t. 具体地, 3 个下游任务对预训练模型提供的用户表示的利用方式和训练目标如下.

用户分类是指将用户划分到当前时间 t 的对应类别中, 是电子商务平台与用户管理相关的一个多标签分类任务. 在用户分类任务中, 我们将用户时间感知的自身嵌入表示 z_e^t 和上下文表示 c_e^t 拼接起来, 并将其输入到一个 3 层 MLP 中训练多分类器, 优化目标为最小化交叉熵损失:

$$\mathcal{L}=-\sum_{e\in E}\sum_{l\in CLS}y_{el}\log(MLP([z_e^t;c_e^t])) \tag{18}$$

其中, y_{el} 是用户标签, 如果用户 e 在时间步 t 时属于类别 l, 则 $y_{el}=1$, 否则 $y_{el}=0$.

好友推荐是指向目标用户推荐他们当前时间 t 可能感兴趣的其他用户, 他们可能发展为好友, 这有助于增加电商平台社交网络的活力. 该任务被定义为给定目标用户 e_i 一组候选推荐用户 e_k 的排序问题. 我们将用户时间感知的自身嵌入表示 z_e^t 和上下文表示 c_e^t 拼接并通过一个 3 层 MLP 学习用户特征向量, 最终推荐得分计算为两个用户特征向量的点积:

$$sim_{ij}=\left\langle MLP([z_{e_i}^t;c_{e_i}^t]),MLP([z_{e_k}^t;c_{e_k}^t])\right\rangle \tag{19}$$

其中, $\langle x,y\rangle$ 表示向量 x 和 y 的点积, 优化目标为最小化交叉熵损失:

$$\mathcal{L}=-\sum_{(e_i,e_k)}y_{ik}\log sim_{ik}+(1-y_{ik})\log(1-sim_{ik}) \tag{20}$$

其中, y_{ik} 是好友添加标签, 如果用户 e_i 和用户 e_k 在时间 t 时添加为好友, 则 y_{ik} =1, 否则 y_{ik} =0.

商品推荐是指向目标用户推荐他们当前时间 t 感兴趣的商品, 该任务被定义为给定目标用户和一组候选商品的推荐排序问题. 在该任务中, 社交知识图谱预训练模型提供的用户表示向量被注入并增强基于协同过滤的推荐模型 NCF[57]. NCF 包含通过线性核建模用户和商品潜在特征交互的 GMF 层, 以及通过非线性核从数据中学习交互的 MLP 层. 我们遵循 PKGM[6]通过预训练模型提供的服务向量增强 NCF 的方法, 对于给定的用户 e 和候选商品 p, 将预训练模型提供的用户自身嵌入表示 z_e^t 和上下文表示 c_e^t 集成到 NCF 的 MLP 交互层中:

$$\phi_1^{MLP}(x_e, x_p, [z_e^t; c_e^t]) = \begin{bmatrix} x_e \\ x_p \\ [z_e^t; c_e^t] \end{bmatrix} \tag{21}$$

其中, x_e 和 x_p 是 NCF 中用户和商品的潜在向量, 推荐模型 NCF 的其他部分均与 PKGM[6]中描述一致.

4.3.3 实验设置和评估指标

用户分类任务中, 我们设置 MLP 隐藏层大小为 32, 使用 ReLU 激活函数和 Adam 优化器, 设置学习率为 5E–4, 正负样本比例为 1:3. 将所有类别的分类准确率和 F1-score 的平均值作为评估指标.

好友分类任务中, 我们设置 MLP 隐藏层大小为 32, 使用 ReLU 激活函数和 Adam 优化器, 设置学习率为 1E–3, 正负样本比例为 1:10. 我们为每条测试样本的目标用户随机采样 100 个候选好友 (负样本), 计算真实好友 (正样本) 与所有候选好友对目标用户的推荐得分并降序排列. 将正样本排序结果的 MRR 和 Hit@k (k=1, 3, 10) 作为最终指标.

商品推荐任务中, 我们设置正负样本比例为 1:3, 使用 Adam 优化器, 设置学习率为 5E–4, NCF 框架的所有设置 (包括 GMF、MLP 和预测层) 与 PKGM[6]相同. 我们为每条测试样本的用户随机采样 50 个候选商品 (负样本), 计算真实浏览商品 (正样本) 和所有候选商品对给定用户的推荐得分并降序排列. 将正样本排序结果的命中率 HR@k (k = 5, 10) 和归一化折损累计增益 NDCG@k 作为最终指标.

4.3.4 实验结果

在每个下游任务中, 我们分别使用了由未更新的基础模型 M^{base} 和更新的模型 M^t 提供的服务向量 (用户实体嵌入表示 z_e^t 和上下文表示 c_e^t) 并对比它们在下游任务的性能. 根据第 4.2 节的结论, 我们分别通过 3 种训练策略各自最佳的增量学习方法 (BID-NAS、TIE-FS 和 FT-NAS) 训练更新得到的 M^t .

图 5 显示了不同预模型, 即未更新的基础模型 M^{base} 和通过不同增量学习方法 (事实采样率 r 从 0.2 到 1) 更新的模型 M^t , 提供的服务向量在 3 个下游任务上 24 个时间步的平均指标结果. 与第 4.2 节的结论类似, 我们的增量学习方法 BID-NAS 显著优于其他基线方法, 且训练样本数量减少基于 BID-NAS 更新的模型提供的服务质量下降趋势也越缓, 这显示出我们方法具有更强的鲁棒性. 此外, 我们发现在用户分类和商品推荐任务中, 在训练样本量较少 (事实采样率 r = 0.2) 的情况下, 通过 FT-NAS 更新的模型比基础模型 M^{base} 提供的用户表示性能更差. 我们认为, 这是由于直接微调的训练策略 FT 自身容易引发灾难性遗忘, 而对少量训练样本的过度拟合进一步损害了原有预训练模型的性能.

图 6 显示了不同预模型 (基础模型 M^{base} 和通过不同增量学习方法在 r = 0.6 时更新的模型 M^t , 第 4.2.3 节已证明 r = 0.6 是有效且高效的) 提供的服务向量在 3 个下游任务上每个时间步的详细指标. 可以发现, 随着时间的推移, 基础模型 M^{base} 提供的服务向量的质量越来越差, 在最后一个更新时间步 (t = 51), 与保持更新的模型 M^t 相比, M^{base} 提供的用户表示使下游任务的性能下降超过 10%.

在下游任务上的实验结果表明, 对于具有高动态性的大规模社交知识图谱, 及时更新预训练模型是非常必要的, 且通过我们的增量学习方法 BID-NAS 更新的预训练模型可以在捕获知识图谱的数据变化的同时避免灾难性遗忘, 帮助预训练模型长期维持为当前下游任务提供高质量用户特征表示的能力.

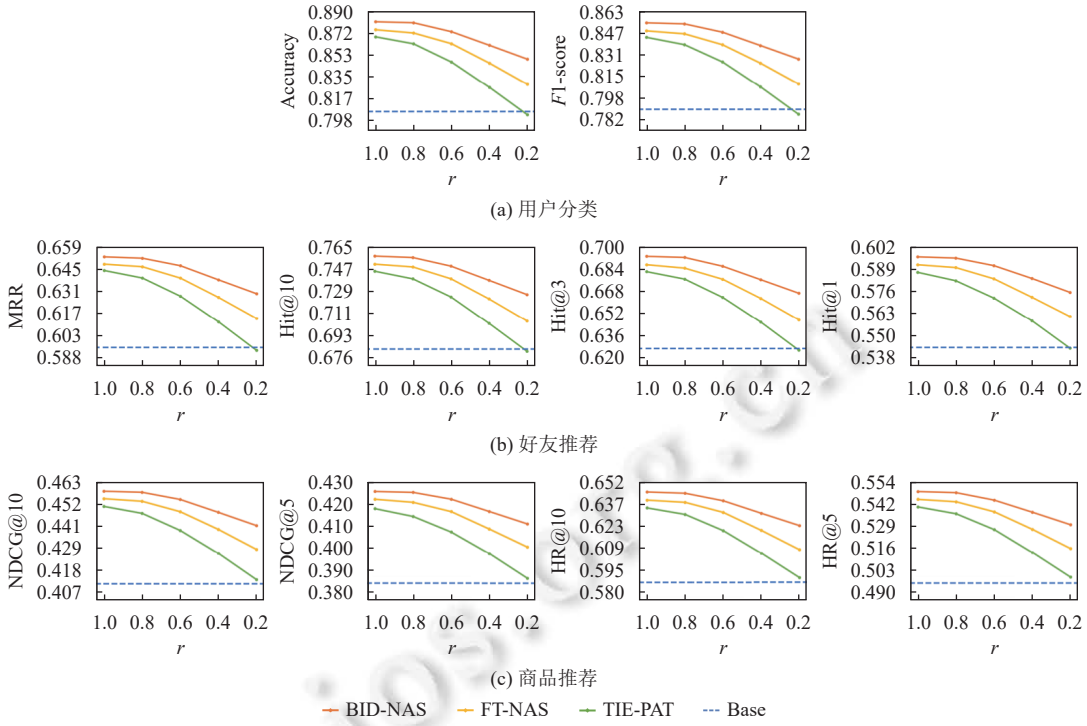

图 5 由不同增量学习方法 (r 从 0.2 到 1) 更新的模型提供的用户表示在下游任务中 24 个时间步的平均指标

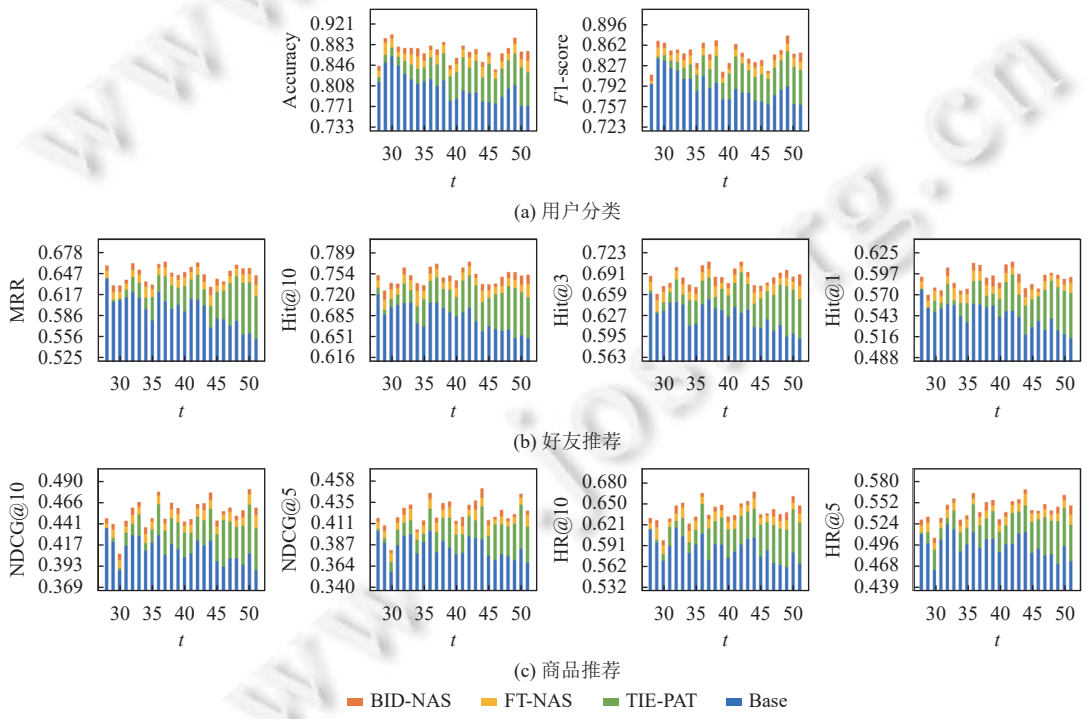

图 6 由不同增量学习方法更新的模型提供的用户表示在下游任务中 24 个时间步 (t) 的详细指标

5 总 结

本文提出了一种电子商务场景下针对社交知识图谱预训练模型的高效增量学习方法 BID-NAS, 该方法通过基于双向模仿蒸馏的训练策略 BID 使模型在更新过程中避免灾难性遗忘并能捕获知识图谱的变化. 在训练过程中, 我们通过基于事实常规性和反常性的事实采用策略 NAS 对新数据中具有高价值的事实进行采样, 以减小训练数据规模. 最后, 我们还提出更适合社交知识图谱的逆重放机制提升负样本质量. 我们在一个真实的大规模电子商务社交知识图谱数据集和一系列下游任务上进行了实验, 实验结果验证了我们方法的高效性和有效性. 未来, 我们希望将社交知识图谱预训练模型的增量学习方法扩展到多模态知识图谱.

References:

[1] Wakil K, Alyari F, Ghasvari M, Lesani Z, Rajabion L. A new model for assessing the role of customer behavior history, product classification, and prices on the success of the recommender systems in e-commerce. Kybernetes, 2019, 49(5): 1325–1346. [doi: 10.1108/K-03-2019-0199]

[2] Zhao Q, Chen JL, Chen MM, Jain S, Beutel A, Belletti F, Chi EH. Categorical-attributes-based item classification for recommender systems. In: Proc. of the 12th ACM Conf. on Recommender Systems. New York: Association for Computing Machinery, 2018. 320–328. [doi: 10.1145/3240323.3240367]

[3] Zhu YS, Zhao HY, Zhang W, Ye GQ, Chen H, Zhang NY, Chen HJ. Knowledge perceived multi-modal pretraining in E-commerce. In: Proc. of the 29th ACM Int'l Conf. on Multimedia. New York: Association for Computing Machinery, 2021. 2744–2752. [doi: 10.1145/3474085.3475648]

[4] Stein D, Shterionov D, Way A. Towards language-agnostic alignment of product titles and descriptions: A neural approach. In: Companion Proc. of the 2019 World Wide Web Conf. (WWW2019). New York: Association for Computing Machinery, 2019. 387–392. [doi: 10.1145/3308560.3316602]

[5] Neve J, Palomares I. Hybrid reciprocal recommender systems: Integrating item-to-user principles in reciprocal recommendation. In: Companion Proc. of the 2020 Web Conf. (WWW2020). New York: Association for Computing Machinery, 2020. 848–853. [doi: 10.1145/3366424.3383295]

[6] Zhang W, Wong CM, Ye GQ, Wen B, Zhang W, Chen HJ. Billion-scale pre-trained e-commerce product knowledge graph model. In: Proc. of the 37th Int'l Conf. on Data Engineering (ICDE). Chania: IEEE, 2021. 2476–2487. [doi: 10.1109/ICDE51399.2021.00280]

[7] Wong CM, Feng F, Zhang W, Vong CM, Chen H, Zhang YC, He P, Chen H, Zhao K, Chen HJ. Improving conversational recommender system by pretraining billion-scale knowledge graph. In: Proc. of the 37th Int'l Conf. on Data Engineering (ICDE). Chania: IEEE, 2021. 2607–2612. [doi: 10.1109/ICDE51399.2021.00291]

[8] Li ZX, Jin XL, Li W, Guan SP, Guo JF, Shen HW, Wang YZ, Cheng XQ. Temporal knowledge graph reasoning based on evolutional representation learning. In: Proc. of the 44th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: Association for Computing Machinery, 2021. 408–417. [doi: 10.1145/3404835.3462963]

[9] Wu JP, Xu YS, Zhang YX, Ma C, Coates M, Cheung JCK. TIE: A framework for embedding-based incremental temporal knowledge graph completion. In: Proc. of the 44th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: Association for Computing Machinery, 2021. 428–437. [doi: 10.1145/3404835.3462961]

[10] Losing V, Hammer B, Wersing H. Incremental on-line learning: A review and comparison of state of the art algorithms. Neurocomputing, 2018, 275: 1261–1274. [doi: 10.1016/j.neucom.2017.06.084]

[11] Yoon J, Yang E, Lee J, Hwang SJ. Lifelong learning with dynamically expandable networks. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018.

[12] Nguyen CV, Li YZ, Bui TD, Turner RE. Variational continual learning. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018.

[13] Li ZZ, Hoiem D. Learning without forgetting. In: Leibe B, Matas J, Sebe N, Welling M, eds. Computer Vision—ECCV 2016. Cham: Springer, 2016. 614–629. [doi: 10.1007/978-3-319-46493-0_37]

[14] Rannen A, Aljundi R, Blaschko MB, Tuytelaars T. Encoder based lifelong learning. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision (ICCV). Venice: IEEE, 2017. 1329–1337. [doi: 10.1109/ICCV.2017.148]

[15] Liu XL, Masana M, Herranz L, Van De Weijer J, López AM, Bagdanov AD. Rotate your networks: Better weight consolidation and less

catastrophic forgetting. In: Proc. of the 24th Int'l Conf. on Pattern Recognition (ICPR). Beijing: IEEE, 2018. 2262–2268. [doi: 10.1109/ICPR.2018.8545895]

[16] Rebuffi SA, Kolesnikov A, Sperl G, Lampert CH. iCaRL: Incremental classifier and representation learning. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 5533–5542. [doi: 10.1109/CVPR.2017.587]

[17] Lopez-Paz D, Ranzato MA. Gradient episodic memory for continual learning. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6470–6479.

[18] Chaudhry A, Ranzato M, Rohrbach M, Elhoseiny M. Efficient lifelong learning with A-GEM. In: Proc. of the 2019 Int'l Conf. on Learning Representations. New Orleans: OpenReview.net, 2019.

[19] Liu L, Bai X, Zhang HG, Zhou J, Tang WZ. Describing and learning of related parts based on latent structural model in big data. Neurocomputing, 2016, 173: 355–363. [doi: 10.1016/j.neucom.2014.12.120]

[20] Malik ZK, Hussain A, Wu J. An online generalized eigenvalue version of Laplacian eigenmaps for visual big data. Neurocomputing, 2016, 173: 127–136. [doi: 10.1016/j.neucom.2014.12.119]

[21] Liang Z, Wang HZ, Dai JJ, Shao XY, Ding XO, Mu TY. Interpretability of entity matching based on pre-trained language model. Ruan Jian Xue Bao/Journal of Software, 2023, 34(3): 1087–1108 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6794.htm [doi: 10.13328/j.cnki.jos.006794]

[22] Sun KL, Luo XD, Luo YR. Survey of applications of pretrained language models. Computer Science, 2023, 50(1): 176–184 (in Chinese with English abstract). [doi: 10.11896/jsjkx.220800223]

[23] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego: ICLR, 2015.

[24] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 2818–2826. [doi: 10.1109/CVPR.2016.308]

[25] Xie SN, Girshick R, Dollár P, Tu ZW, He KM. Aggregated residual transformations for deep neural networks. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 5987–5995. [doi: 10.1109/CVPR.2017.634]

[26] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics, 2019. 4171–4186. [doi: 10.18653/v1/N19-1423]

[27] Yang ZL, Dai ZH, Yang YM, Carbonell J, Salakhutdinov RR, Le QV. XLNet: Generalized autoregressive pretraining for language understanding. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 517.

[28] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D. Language models are few-shot learners. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 159. [doi: 10.5555/3495724.3495883]

[29] Liu WJ, Zhou P, Zhao Z, Wang ZR, Ju Q, Deng HT, Wang P. K-bert: Enabling language representation with knowledge graph. Proc. of the AAAI Conf. on Artificial Intelligence, 2020, 34(3): 2901–2908. [doi: 10.1609/aaai.v34i03.5681]

[30] Zhang ZY, Han X, Liu ZY, Jiang X, Sun MS, Liu Q. ERNIE: Enhanced language representation with informative entities. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 1441–1451. [doi: 10.18653/v1/P19-1139]

[31] Peters ME, Neumann M, Logan R, Schwartz R, Joshi V, Singh S, Smith NA. Knowledge enhanced contextual word representations. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics, 2019. 43–54. [doi: 10.18653/v1/D19-1005]

[32] Wang RZ, Tang DY, Duan N, Wei ZY, Huang XJ, Ji JS, Cao GH, Jiang DX, Zhou M. K-adapter: Infusing knowledge into pre-trained models with adapters. In: Proc. of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021. 1405–1418. [doi: 10.18653/v1/2021.findings-acl.121]

[33] Xiong WH, Wang H, Wang WY. Progressively pretrained dense corpus index for open-domain question answering. In: Proc. of the 16th Conf. of the European Chapter of the Association for Computational Linguistics: Main Volume. Association for Computational Linguistics, 2021. 2803–2815. [doi: 10.18653/v1/2021.eacl-main.244]

[34] Xiong K, Du L, Ding X, Liu T, Qin B, Fu B. Knowledge enhanced pre-trained language model for textual inference. Journal of Chinese Information Processing, 2022, 36(12): 27–35 (in Chinese with English abstract). [doi: 10.3969/j.issn.1003-0077.2022.12.003]

[35] Castro FM, Marín-Jiménez MJ, Guil N, Schmid C, Alahari K. End-to-end incremental learning. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, eds. Computer Vision—ECCV 2018. Cham: Springer, 2018. 241–257. [doi: 10.1007/978-3-030-01258-8_15]

[36] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv:1503.02531, 2015.

[37] Kirkpatrick J, Pascanu R, Rabinowitz N, Veness J, Desjardins G, Rusu AA, Milan K, Quan J, Ramalho T, Grabska-Barwinska A, Hassabis D, Clopath C, Kumaran D, Hadsell R. Overcoming catastrophic forgetting in neural networks. Proc. of the National Academy of Sciences of the United States of America, 2017, 114(13): 3521–3526. [doi: 10.1073/pnas.1611835114]

[38] Dhar P, Singh RV, Peng KC, Wu ZY, Chellappa R. Learning without memorizing. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 5133–5141. [doi: 10.1109/CVPR.2019.00528]

[39] Aljundi R, Lin M, Goujaud B, Bengio Y. Gradient based sample selection for online continual learning. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 1058.

[40] Shin H, Lee JK, Kim J, Kim J. Continual learning with deep generative replay. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 2994–3003.

[41] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2014. 2672–2680.

[42] Rolnick D, Ahuja A, Schwarz J, Lillicrap T, Wayne G. Experience replay for continual learning. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 32.

[43] Wu Y, Chen YP, Wang LJ, Ye YC, Liu ZC, Guo YD, Fu Y. Large scale incremental learning. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 374–382. [doi: 10.1109/CVPR.2019.00046]

[44] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: OpenReview.net, 2017.

[45] Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph attention networks. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018.

[46] Xu YS, Zhang YX, Guo W, Guo HF, Tang RM, Coates M. GraphSAIL: graph structure aware incremental learning for recommender systems. In: Proc. of the 29th ACM Int'l Conf. on Information & Knowledge Management. New York: Association for Computing Machinery, 2020. 2861–2868. [doi: 10.1145/3340531.3412754]

[47] Zhou F, Cao CT. Overcoming catastrophic forgetting in graph neural networks with experience replay. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. AAAI Press, 2021. 4714–4722. [doi: 10.1609/AAAI.V35I5.16602]

[48] Liu HH, Yang YD, Wang XC. Overcoming catastrophic forgetting in graph neural networks. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. AAAI Press, 2021. 8653–8661. [doi: 10.1609/AAAI.V35I10.17049]

[49] Zhang Y, Zhu YL. Incremental partial discharge recognition method combining knowledge distillation with graph neural network. Trans. of China Electrotechnical Society, 2023, 38(5): 1390–1400 (in Chinese with English abstract). [doi: 10.19595/j.cnki.1000-6753.tces.220285]

[50] Song HJ, Park SB. Enriching translation-based knowledge graph embeddings through continual learning. IEEE Access, 2018, 6: 60489–60497. [doi: 10.1109/ACCESS.2018.2874656]

[51] Bordes A, Usunier N, Garcia-Durán A, Weston J, Yakhnenko O. Translating embeddings for modeling multi-relational data. In: Proc. of the 26th Int'l Conf. on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2013. 2787–2795.

[52] Trouillon T, Welbl J, Riedel S, Gaussier É, Bouchard G. Complex embeddings for simple link prediction. In: Proc. of the 33rd Int'l Conf. on Machine Learning. New York: JMLR.org, 2016. 2071–2080.

[53] Sun ZQ, Deng ZH, Nie JY, Tang J. RotatE: Knowledge graph embedding by relational rotation in complex space. In: Proc. of the 2019 Int'l Conf. on Learning Representations. New Orleans: OpenReview.net, 2019.

[54] Cipolla R, Gal Y, Kendall A. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. 2018. 7482–7491. [doi: 10.1109/CVPR.2018.00781]

[55] Kingma DP, Ba J. Adam: A method for stochastic optimization. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego: ICLR, 2015.

[56] Zhu YS, Tong X, Fan D, Wang X. Identifying privacy leakage from user-generated content in an online health community. In: Chen HC, Zeng D, Yan XB, Xing CX, eds. Smart Health. Cham: Springer, 2019. 257–268. [doi: 10.1007/978-3-030-34482-5_23]

[57] He XN, Liao LZ, Zhang HW, Nie LQ, Hu X, Chua TS. Neural collaborative filtering. In: Proc. of the 26th Int'l Conf. on World Wide Web. Perth: Int'l World Wide Web Conf. Steering Committee, 2017. 173–182. [doi: 10.1145/3038912.3052569]

附中文参考文献:

[21] 梁峥, 王宏志, 戴加佳, 邵心玥, 丁小欧, 穆添愉. 预训练语言模型实体匹配的可解释性. 软件学报, 2023, 34(3): 1087–1108. http://www.jos.org.cn/1000-9825/6794.htm [doi: 10.13328/j.cnki.jos.006794]

[22] 孙凯丽, 罗旭东, 罗有容. 预训练语言模型的应用综述. 计算机科学, 2023, 50(1): 176–184. [doi: 10.11896/jsjkx.220800223]

[34] 熊凯, 杜理, 丁效, 刘挺, 秦兵, 付博. 面向文本推理的知识增强预训练语言模型. 中文信息学报, 2022, 36(12): 27–35. [doi: 10.3969/j.issn.1003-0077.2022.12.003]

[49] 张翼, 朱永利. 结合知识蒸馏和图神经网络的局部放电增量识别方法. 电工技术学报, 2023, 38(5): 1390–1400. [doi: 10.19595/j.cnki.1000-6753.tces.220285]

朱渝珊(1998－), 女, 博士生, 主要研究领域为知识图谱, 知识图谱表示学习和推理.

陈名杨(1997－), 男, 博士生, 主要研究领域为知识图谱, 知识图谱表示学习和推理.

张文(1992－), 女, 博士, 特聘研究员, 博士生导师, CCF 专业会员, 主要研究领域为知识图谱, 图数据处理, 大数据系统.

姚祯(1999－), 男, 硕士, 主要研究领域为知识图谱, 知识图谱表示学习.

王晓珂(1995－), 女, 硕士, 主要研究领域为电子商务平台的链接预测, 图表示学习.

陈辉(1991－), 男, 博士, 主要研究领域为自然语言处理, 知识图谱, 表示学习.

李志宇(1992－), 男, 博士, 主要研究领域为电子商务平台的社会计算, 数据挖掘.

陈华钧(1978－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为知识图谱, 大数据系统, 自然语言处理.