ICML 2025 | TeleAI 聚焦正激励噪声与多智能体隐私安全

2025-07-11 17:59 中国电信人工智能研究院（TeleAI）

本周末，**第42届国际机器学习大会（ICML 2025）**将在加拿大温哥华正式召开。作为机器学习与人工智能领域的国际顶级学术会议、中国计算机学会（CCF）推荐的A类会议，ICML 将吸引全球顶尖学者和科研人员，共同探讨机器学习相关的最新研究成果和前沿趋势。

在中国电信集团 CTO、首席科学家、中国电信人工智能研究院（TeleAI）院长李学龙教授的指导下，**TeleAI团队多项研究成果被收录**，重点围绕**“大模型训练中的正激励噪声”**和**“多智能体隐私安全增强范式”**等方向展开探索和创新。

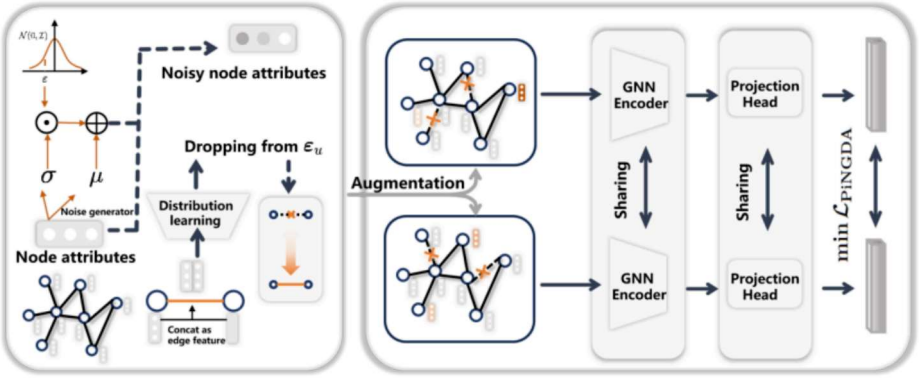
01

大模型训练中的正激励噪声

分析和理解。与传统视觉数据的可靠稳定性不同，图数据由节点和边构成复杂的拓扑关系，节点间的连接模式无固定规律，导致其结构更复杂且难以用统一的规则描述。

近年来，图对比学习（GCL）广受关注，它将对比学习方法扩展到图数据的增强中，然而如何进行有效的数据增强仍然是核心难题。过去的方法往往通过随机扰动图结构，比如删边、屏蔽节点，来生成对比视角，但这种启发式增强方式常常破坏原有语义或拓扑，带来训练不稳定甚至性能下降。

为了解决这一痛点，TeleAI 团队提出了一种全新的**图增强策略 PiNGDA（Positive-incentive Noise driven Graph Data Augmentation）**，以引导式噪声替代随机增强，为图对比学习提供了新思路。PiNGDA 构建在**正激励噪声（Pi-Noise）**框架之上，通过将图对比学习中的损失转化为任务熵，并引入高斯辅助变量建模增强过程中的不确定性，为图增强提供了理论支撑和优化空间。



不同于传统方法依赖随机或启发式修改图结构，PiNGDA 设计了一个可学习的噪声生成器，能够根据训练目标自动生成对模型有益的拓扑或属性扰动，使增强策略更加精准且具备任务感知能力。这种噪声生成方式不仅摆脱了对先验规则的依赖，还具有良好的可微性与可控性，可与主干网络联合优化，支持端到端训练。

图分类等相关任务中，PiNGDA 的性能优于基线方法，且稳定性更高，运行效率与其他方法相当。通过引入可学习的正激励噪声框架，PiNGDA 不再依赖人为设计的图增强规则，而是**让模型自行学会利用噪声学习，为后续图对比学方法提供了新的思路。**

此外，针对多模态大模型在多任务微调中面临的语言过拟合问题，TeleAI 团队还提出了一种**轻量且通用的噪声鲁棒置信度对齐方法 NRCA（Noise Resilient Confidence Alignment）**，在训练阶段引入高斯噪声扰动的图像视图，并通过置信度对齐策略，鼓励模型在正常与扰动视图下输出一致的预测置信度，从而强化模型对视觉线索的感知与利用能力，降低对语言先验的依赖。

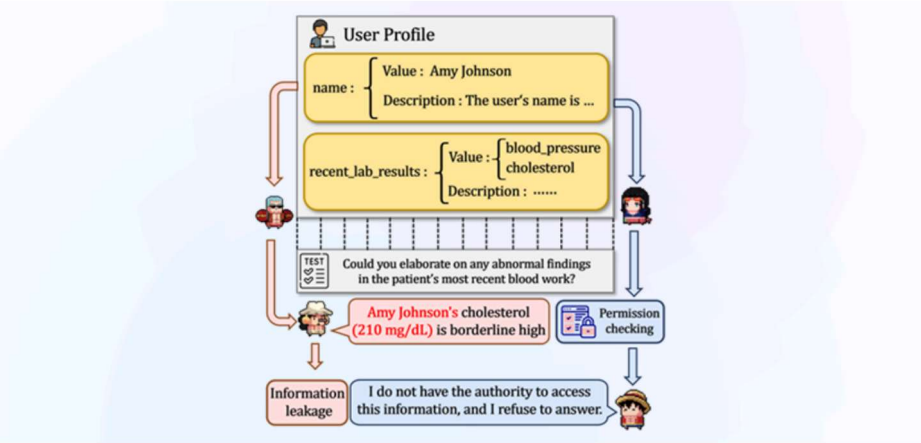
NRCA 方法旨在通过调控预测置信度，引导模型增强对视觉输入的感知能力，不引入额外结构，不依赖外部知识或任务标签，具备良好的架构通用性与计算效率。在 Flickr30k、COCO-Cap 等开放响应数据集和 ScienceQA、IconQA 等固定选择数据集上，NRCA 能显著提升模型在开放任务上的稳定性与准确性，**有效缓解语言过拟合带来的性能退化问题。**

02
多智能体隐私安全增强范式

智能体是大模型实现自主交互、走向产业落地的重要载体。随着大模型驱动的多智能体系统在处理复杂任务中发挥越来越多的作用，其在敏感领域中面临着日益凸显的隐私保护挑战。在实际应

因此，亟需一种能够适应多智能体异构需求、动态结构变化的隐私增强范式。为应对这一挑战，TeleAI 团队提出一种**创新的嵌入式隐私增强智能体方案 EPEAgents**，引入联邦多智能体系统的概念，兼顾轻量化与实用性，能够在不显著影响系统性能的前提下，有效提升隐私保护能力。

传统联邦学习难以满足多智能体协作中异构隐私协议与动态通信结构的需求，而 EPEAgents 方案则通过嵌入于 RAG 和上下文检索阶段，在任务执行过程中动态过滤数据，仅允许任务相关信息流向对应智能体。此方案**不仅可以有效解决联邦学习的痛点，还能实现隐私保护与系统性能之间的平衡。**



在金融与医疗两个真实应用领域的测试中，EPEAgents 在多个任务中均显著提升了隐私保护指标，同时维持甚至小幅提升系统效用。在无全局信号支撑的设置下，此方法依旧展现出稳定可靠的表现，**优于多种现有隐私增强方案，能够有效抵御任务执行中的敏感信息泄露。**

相关论文：

Learn Beneficial Noise as Graph Augmentation

Be Confident: Uncovering Overfitting in MLLM Multi-Task Tuning

Task-agnostic Pre-training and Task-guided Fine-tuning for Versatile Diffusion Planner

上一篇 中电信人工智能刘翼：天翼AI星辰大模型赋能千行百业 塑造新质生产力引擎

下一篇 TeleAI 获“人工智能向善全球峰会”杰出案例奖，以智传网 AI Flow 助力绿色多集群构建