许艺怀

人工智能安全性 · 最优化理论

19357384223 | yihuai1024@outlook.com | yihuai-xu.github.io | Yihuai-Xu

教育经历

浙江工商大学-统计与数学学院-计算科学（硕士） *浙江杭州*

一作发表/投稿机器学习三大顶会论文两篇（AIGC鉴伪领域），其中一篇ICLR25已接收，一篇ICML25审稿中。
二作（导师一作）投稿中科院二区期刊论文一篇（数值最优化理论与方法领域）。
获“华为杯”中国研究生数学建模竞赛全国二等奖（2次），其他国家级/省级学科竞赛奖3项。
获硕士研究生国家奖学金，一等奖学金，综合测评全系第一（1/9）各1次。 *2022.09-2025.01*

杭州师范大学-数学学院-数据科学与大数据技术（本科） *浙江杭州*

获全国大学生数学竞赛（数学类）等国家级/省级学科竞赛二等奖3项、三等奖5项。 *2018.09-2022.06*

论文成果

- **Y. Xu**, Y. Wang, Y. Bi, H. Cao, et al. Training-free LLM-generated Text Detection by Mining Token Probability Sequences. Proceedings of the Twelfth International Conference on Learning Representations (ICLR), 2025.
- **Y. Xu**, K. Lv, H. Cao, W. Fan, Y. Wang and F. Wu. Training-free Black-box LLM-Generated Text Detection By Token Probability Topology Modeling. (Under review), 2025.
- L. Peng, **Y. Xu**, G. Xiao, B. Wu and Y. Rui. A New Subspace Minimization Conjugate Gradient Algorithm Without Explicit Use of Gradient Terms. (Under review), 2025.
- K. Lv, H. Cao, K. Tu, **Y. Xu**, Z, Zhang, et al. Hyper Adversarial Tuning for Boosting Adversarial Robustness of Pretrained Large Vision Models. (Under review), 2025.

实习经历

浙江大学-上海高等研究院 *科研实习生*

AIGC合成内容鉴伪（LLM生成文本检测，元学习） *2024.04-2025.04*

- 创造性提出一种基于token概率序列动力学分析的免训练LLM生成文本检测算法。在全局统计量的基础上，通过挖掘概率序列的多尺度-多样化熵进一步开发出局部统计量，以更低或相当的成本实现了跨域、跨源、跨场景以及改写攻击前提下的LLM生成文本的实时检测以及同类型方法时下的最佳性能。相关文章发表于机器学习顶级会议ICLR-25。
- 创造性提出一种基于token概率序列的拓扑有向性重建的免训练LLM生成文本检测算法。通过对来自不同代理模型的概率序列进行B-样条插值以获得几何启发指标，提升了免训练方法的鲁棒性和可扩展性。相关文章投稿至机器学习顶级会议ICML-25。
- 参与基于超网络的视觉大模型对抗鲁棒性增强研究，相关文章投稿至CCF-B/中科院二区Top期刊“Pattern Recognition”。
- 参与构造多模态（可解释性）合成图像检测细粒度数据集。相关文章正在投稿过程中。

华院计算技术（上海）股份有限公司 *社会治理实习生*

NLP下游应用（多层级文本分类/摘要，监督微调） *2021年6月-2021年12月*

- 矛盾调度处置中心业务系统开发项目。负责对基层群众的信访留言等文本数据进行基于监督微调的多层级文本分类与内容摘要。
- 在真实数据集上对多种预训练语言模型进行测评，撰写测评报告。通过误差分析，细化数据集标注规则。
- 负责数据库的日常使用和维护以及其他移动端数据的格式化抽取。

项目经历

提升免训练LLM生成文本检测器的鲁棒性和可扩展性 *论文(投稿至ICML-25)*

第一作者 *2024.12-2025.04*

- 先前的免训练检测器在应对重写、模仿以及风格化等攻击时性能下降严重，体现出代理模型概率特征的数值不稳定性。
- 本文借助多个代理模型的推理信息，利用B-样条插值曲线重建离散token概率的拓扑有向性，导出曲率关联的几何启发检测器。
- 本文检测器发挥了多个代理模型推理结果的数值和几何优势，在性能和鲁棒性、可扩展性上均取得了优于同类检测器的表现。
- 文章“Training-free Black-box LLM-Generated Text Detection By Token Probability Topology Modeling”投稿至ICML-25。

黑盒情景下LLM生成文本的检测性能提升 *论文(已被ICLR-25接收)*

第一作者 *2024.04-2024.10*

- 针对免训练类检测器在进行黑盒、跨源检测时性能不佳的问题，提出了更强大的免训练通用检测器，该成果被百度Paddle采用。
- 先前的免训练检测器仅通过token概率序列的全局统计量计算得分，忽视了概率序列局部动态复杂度对检测性能的显著影响。
- 本文时间序列熵分析的角度出发，提出了全局与局部统计量联合评估指标Lastde（单样本评估）与Lastde++（采样分布评估）。
- 本文所提检测器在跨源、跨域、跨情景、跨语种以及改写攻击等广泛设置下的性能均优于同类型检测器，且成本相当甚至更低。
- 文章“Training-free LLM-generated Text Detection by Mining Token Probability Sequences”被ICLR-25接收。

扩展子空间最小化共轭梯度法在大规模无约束优化问题中的应用 *论文(投稿至中科院二区期刊)*

学生一作 *2024.04-2024.09*

- 提出了一种不显式包含梯度项的子空间最小化共轭梯度法，在规定成本内以更高的效率求解了更多的无约束优化问题。
- 将子空间最小化技术引入Nazareth型三项共轭梯度法中，扩展了共轭梯度法的种类，并证明了其数值可行性以及全局收敛性。
- 文章“A New Subspace Minimization Conjugate Gradient Algorithm Without Explicit Use of Gradient Terms”正在审稿中。

一种基于泰勒展开动量修正的神经网络优化方法 *专利(公示中)*

学生一作 *2023.09-2024.02*

- 本发明将带参softplus激活函数的二阶泰勒展开与Adam算法的二阶动量相结合，缓解了Adam泛化性常常不如SGDM算法的问题。
- 数值结果显示，本发明在多种数据类型和神经网络上的收敛速度和泛化性都优于Adam算法和SGDM算法。

神经网络的优化和收敛性分析 *导师项目*

作者 *2022.09-2023.04*

- 本项目提出了SGDM算法的一种新的推广形式——自适应动量系数的SGDM算法，提升了经典SGDM算法的加速效果。
- 数值结果显示，本文算法相比于此前多种SGDM算法具有更快的收敛速度和更小的误差。
- 本文在一定假设下证明了均方误差函数的强收敛和弱收敛结果。

浙江省新昌县专家推荐系统研究 *导师项目*

推荐算法负责人 *2022.05-2022.06*

- 基于新昌县众创共享科创云平台中的专家和企业难题数据库，设计相应的专家-企业双向推荐系统，实现二者的相互匹配与推荐。
- 信息压缩与专家画像。首先基于T5-Pegasus少样本监督微调对专家简历信息压缩（生成式摘要），而后利用CRF和keybert关键词提取算法分别对所得摘要进行知识三元组抽取。最后基于neo4j和所得三元组构建专家数据库的知识图谱，即专家画像。
- 匹配与推荐。首先抽取部分专家信息摘要和企业技术难题样本进行匹配标注，构造出正负样本对。随后使用SimCSE、CoSent依次进行有监督的对比学习。将测试集上Spearman系数最高（0.8405）的CoSent作为专家和企业的Top-K匹配模型实现相互匹配。

浙江省丽水市社会纠纷文本研究 *导师项目*

负责人 *2022.04-2022.05*

- 基于丽水市某数字治理数据库上的真实纠纷记录和信访文本数据进行挖掘，解决了内容分类和部门推荐、内容摘要生成等问题。
- 内容分类和部门推荐。对XLNet进行有监督微调。进一步结合数据增强、Focalloss、伪标签等策略将内容分类准确率提升至90%以上。最后根据纠纷文本的分类标签匹配出前K个最佳职责部门。
- 内容摘要。分别使用SentenceBERT和T5-Pegasus对纠纷文本（无标注）进行抽取式和生成式摘要。使用不确定性感知自训练思想在T5-Pegasus上进行增强，收敛后rouge-L提升0.2，并且生成的摘要格式工整，易于进行信息抽取等后续任务。

技能

技能 Python，PyTorch，Tensorflow，MATLAB，R，Neo-4j，Git，Linux，SQL，LaTeX
外语 CET4，CET6

荣誉奖励

2020.12 **浙江省大学生高等数学竞赛（工科类）** *省级·三等奖*
2020.12 **第十二届全国大学生数学竞赛（非数学类）** *省级·二等奖*
2020.12 **第九届浙江省大学生统计调查方案设计大赛** *省级·二等奖*
2021.05 **美国大学生数学建模竞赛** *国家级·H奖*
2021.06 **第九届“泰迪杯”数据挖掘挑战赛** *国家级·二等奖*
2021.12 **浙江省大学生高等数学竞赛（数学类）** *省级·三等奖*
2021.12 **第十三届全国大学生数学竞赛（数学B类）** *省级·二等奖*
2022.10 **第十九届“华为杯”中国研究生数学建模竞赛** *国家级·二等奖*
2022.11 **第三届“大湾区杯”金融数学建模大赛** *省级·三等奖*
2023.04 **第十一届“泰迪杯”数据挖掘挑战赛** *国家级·二等奖*
2023.09 **第二十届“华为杯”中国研究生数学建模竞赛** *国家级·二等奖*
2021.12 **杭州师范大学学业奖学金** *校级·获奖*
2023.06 **浙江工商大学研究生学业一等奖学金** *校级·数学系第一*
2024.09 **硕士研究生国家奖学金** *校级·获奖*