蔡怀广

130 6092 8548 caihuaiguang@gmail.com https://caihuaiguang.github.io/

教育背景

中国科学院自动化所 北京，2022.09 – 2025.06
多模态人工智能系统全国重点实验室 硕士（导师：张文生） 专业：模式识别与智能系统
研究方向：非合作博弈论，合作博弈论，可解释机器学习，大语言模型 GPA：3.84/4.00

中山大学 广州，2018.09 – 2022.06
先进网络与计算系统研究所 本科（导师：周知、王昌栋） 专业：计算机科学与技术
研究方向：边缘计算，推荐系统 GPA：3.9/4.0

实习经历

MiniMax - 技术 | 通用模型 | 大语言模型算法实习生 2024.11 – 2025.05

- **博弈论与推理模型**：应用合作博弈论量化DeepSeek R1在答案生成时query和think的贡献，验证“小模型筛选-大模型训练”可行性，定量分析R1适用场景、学科、难度、社区，部分阐明推理模型的原理和局限。[代码]
- **长文本推理能力提升**：提出推理步骤能力级别的 InsTag 方案来筛选高质量且多样化的 query，并对 response 三盲一致拒绝采样来构造长文本问答 SFT 数据集；通过剔除 unique token ratio 较低的数据大幅降低输出的重复率；提出基于思维链价值的数据选择算法；综合以上方案，通过 SFT 提升了模型的长文推理能力。
- **模型注意力分布优化**：提出定位回答问题所需原文片段的方案（prompt + fewshot + 最长公共子序列）来筛选长文本 SFT 数据，通过优化模型注意力分布的均衡性来提升模型信息检索能力。
- **长文本回复风格调研**：调研在模型回复同时生成对原文句子级别引用的方案（LongCite）以提高回复可信度。

字节跳动 - 巨量引擎 | 品牌广告中台 | 广告拍卖策略实习生 2023.11 – 2024.03

- 实现广告拍卖 A/B 实验的支撑技术（预算分桶机制）的代码，并修复复杂业务链路中潜藏四年的小Bug。

公开成果

Online Resource Allocation for Edge Intelligence with Colocated Model Retraining and Inference
Huaiguang Cai, Zhi Zhou, et al. INFOCOM 2024 (**CCF-A**, 计算机网络顶会)
简介：针对边缘智能场景中由于数据、模型、任务漂移导致推理精度下降的难题，首次理论建模**训练和推理混合部署**的计算范式。通过最优化理论解决了以下难点：处理未来不可预测信息、目标函数非凸、自变量离散。提出的具有性能理论保证（竞争比）的轻量级在线算法 ORRIC 能够自适应平衡训练与推理的资源分配，在保证用户体验的同时提升长期推理精度。为大模型时代训练和推理计算资源共享提供了启发。
[论文] [幻灯片] [代码]

CHG Shapley: Efficient Data Valuation and Selection towards Trustworthy Machine Learning
Huaiguang Cai Arxiv
简介：针对**深度神经网络训练可解释性**问题，基于合作博弈论推导出每个数据点对模型性能贡献的解析式 CHG Shapley，**首次**将衡量数据价值所需的模型训练次数从数据集大小的平方次数降为 1。此外在数据选择任务中，尤其是在数据存在噪声的情况下，基于 CHG Shapley 的数据选择方法显著优于现有方法。
[OpenReview] [代码1] [代码2]

Shapley value-based class activation mapping for improved explainability in neural networks
Huaiguang Cai, et al. The Visual Computer 2025 (SCI 期刊)
简介：针对**深度神经网络预测可解释性**问题，首次揭示了可解释性领域两大主流特征归因方法 CAM（Class Activation Mapping）与 SHAP（SHapley Additive exPlanations）之间的内在联系。通过将神经网络预测过程建模为一个合作博弈，从 Shapley value 角度阐明了2016年以来流行的启发式方法 GradCAM 的理论基础，提出的 ShapleyCAM 和 ReST 效用函数显著提升了神经网络预测可解释性。在 ImageNet 验证集上对12种流行网络的实验验证了有效性。ShapleyCAM 已合并到**最广泛使用**的可解释人工智能代码库 pytorch-grad-cam。
[论文] [代码] [pytorch-grad-cam, **11k stars**]

研究项目

PIDCFR：快速求解两人Nash均衡 [results]

- 在 8 种博弈中发现，Predicte CFR+ 算法在 last-iterate 的收敛速度比 average-iterate 快数倍。实验结果获得了 **MIT** 教授 Gabriele Farina（OpenAI 著名科学家 Noam Brown 的师弟）的赞赏。
- 目前正尝试建立控制理论中的 PID 算法（广泛应用于互联网广告出价）与非合作博弈理论中的 CFR 算法（广泛应用于求解不完美信息博弈均衡）之间的理论联系，以期推导出更快求解均衡且能大规模使用的算法。