孙一丁

+86 15522420639 emanual20.sun@gmail.com Github: Emanual20

教育经历

中国人民大学 2022.09 – 2025.06
人工智能 硕士 高瓴人工智能学院 GPA: 3.86/4.0 北京

南开大学 2018.09 – 2022.06
计算机科学与技术 本科 计算机学院 GPA: 3.84/4.0 专业排名: 6/122 天津

科研经历

Yulan大模型 2023.02 – 2024.02
基石模型数据组负责人 高瓴人工智能学院，中国人民大学

- 为了解决大模型社区缺乏统一预训练数据清洗框架的问题，构建通用数据处理框架 Yulan-GARDEN [Code] [Paper]。该框架集成了支持探测和评价的数据分析模块和由若干不同粒度算子构成的处理模块。用户可借助该工具对语料采样，并根据探测到的特征自由组合预定义的算子列表形成质量提升流水线。通过 ChatGPT 标注自动评测，Yulan-GARDEN 清洗后的数据有70.2%优于未清洗的数据；在端到端评价中，分别使用 Yulan-GARDEN 清洗前后的 CommonCrawl 从头预训练110M GPT-2-raw 和 GPT-2-ref. 对比语言建模能力，GPT-2-ref. 只需要 GPT-2-raw 约 65% 的训练步数就能达到与 GPT-2-raw 相近的表现，且全量训练的 GPT-2-ref. 语言建模能力比 GPT-2-raw 平均强14.77%。相关成果已以第一作者被 **SIGIR 2024** 的 **Demo Track** 录用。
- 作为预训练数据组负责人参与从头预训练 12B 的中英双语基石模型 Yulan-LLM。Yulan-LLM 使用 Yulan-GARDEN 清洗后的高质量中英文双语数据进行预训练，并利用检索器链接外部知识库知识增强的方式进行继续预训练。通过高质量双语指令微调，使模型在 MMLU/C-EVAL 等主流评测基准上与同规模模型达到可比水平。
- 担任指令收集和清洗职责参与 Yulan-Chat-2 项目 [Code]，以 LLaMA-2 为基座模型扩充中文词表与上下文长度，并使用高质量中英文双语指令进行指令微调，提升了 LLaMA-2 的中英文基础语义和理解能力。相较于同期基于LLaMA-2微调的模型，Yulan-Chat-2 具有显著性能优势。该项目在 Github 已获得 335 Stars。

实习经历

百川智能大语言模型部 - 大模型数据算法实习生 2024.03 – 至 今

- 验收外部采购数据，对百 GB 规模试题数据采样并进行质量探测和初步清洗、去重，撰写质量评测验收报告。

美团外卖推荐算法组 - 推荐算法实习生 2022.02 – 2023.12

- 为了解决可解释推荐领域对解释质量评价过度依赖用户实验的问题，提出了一种包括三个自动化指标有用性、保真度和泛化性的解释评价框架，通过公开数据、用户实验和在线日志实验评价，验证了提出指标的区分和诊断能力。初步成果已以第二作者身份于 **QUARE@SIGIR 2022** 上报告，扩充后拟投稿 **CCF-A** 类期刊。
- 为了缓解美团线上解释和推荐模型评分与物品特征对齐不显著的问题，提出了一种基于最大化解释和评分/物品特征互信息的强化学习框架改善了基石模型的性能。相关成果以第二作者在投于 **CCF-A** 类会议。
- 分析线上日志，开展用户实验，验证用户对线上推荐解释标签的感知能力，从多方面分析用户对解释标签的偏好，就改进标签线上展出策略提出建议。相关成果以第一作者在 **QUARE@SIGIR 2022** 上报告。

项目经历

NKUCS.ICU 经验交流平台 [Code] 2020.01 – 至 今

- 为了打通学校信息交流壁垒，作为网站创始人使用 Github Pages + Docsify 轻量解决方案搭建并长期维护平台。以南开大学师生为用户群体，累计访问量达 20k 人次。该项目在 Github 已获得 105 Stars，23 Forks。

NKU-TEDA ARM编译器 [Code] 2021.04 – 2021.08

- 在2021年全国大学生计算机系统能力大赛编译系统设计赛（华为毕昇杯）中作为主要参与者，在Raspberry Pi上使用 C++ 实现类 C 语言从词法分析、语法分析、中间代码生成及优化和目标代码生成的完整编译过程。
- 设计基于 llvm 的 SSA 中间代码形式的代码优化，实现数据流分析、循环展开、死代码删除、自动向量化等优化 Pass，在某些性能测试样例上达到可比 gcc-O2 的效率。该项目获得比赛全国二等奖。

获奖情况

中国人民大学研究生新生奖学金, 二等学业奖学金 2022, 2023
南开大学本科优秀毕业生 2022
中国大学生程序设计竞赛 (CCPC-2019) 秦皇岛站优胜奖 2020