张跃柯

| yueke.zhang@vanderbilt.edu | github.com/zyueek

教育背景

Vanderbilt University Nashville, US 美国纳什维尔
计算机科学博士生 *2023 年 8 月 – 至今*

西南大学 Chongqing, China 中国重庆
管理信息系统学士 *2019 年 9 月 – 2023 年 6 月*

论文发表

- Who' s Pushing the Code? An Exploration of GitHub Impersonation.
  Yueke Zhang, Anda Liang, Xiaohan Wang, Pamela Wisniewski, Fengwei Zhang, Kevin Leach, Yu Huang
  IEEE/ACM International Conference on Software Engineering (ICSE' 2025, CCF-A)
- An Empirical Study of Modeling Fault Localization as a Decision Making Process with Uncertainty and Ambiguity
  Yueke Zhang, Yu Huang, Kevin Leach
  ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM' 2023, CCF-B)
- Leveraging Fuzzy System to Reduce Uncertainty of Decision Making in Software Engineering Automation
  Yueke Zhang, Yu Huang
  Genetic Improvement (GI' 2022)
- Pre-Training Representations of Binary Code Using Contrastive Learning
  Yifan Zhang, Chen Huang, **Yueke Zhang**, Kevin Cao, Huajie Shao, Kevin Leach, Yu Huang
  IEEE Transactions on Software Engineering (TSE, major revision, CCF-A)
- CodeACT-R: A Cognitive Simulation Framework for Code Reading
  Yueke Zhang, Zihan Fang, Greg Trafton, Danniel Levin, Yu Huang
  IEEE/ACM International Conference on Software Engineering (ICSE' 2026, under review)

项目

通过模拟人类注意力提升代码模型 2024 年 8 月 – 至今

- 开发了 CodeACT-R，一种基于 ACT-R 架构的 cognitive simulation framework，用于模拟代码理解过程中类似人类的视觉注意信息 (visual attention patterns)，从而显著降低对昂贵眼动数据 (eye-tracking data) 的依赖。
- 相比于采用传统 numerical data augmentation techniques 的模型，在代码理解任务中模型性能提升了 26.24%，且模拟覆盖了多达 87% 的真实视觉注意信息。

开源社区提交作者身份辨别 2024 年 8 月 – 至今

- 进行了首次关于 GitHub commit 作者身份辨别中 impersonation threats 的实证研究，揭示了开发者对此问题认识不足，并凸显了 GitHub commit mechanisms 导致检测 impersonation 的挑战。
- 通过整合额外的 commit 信息，提升了 LLM 及其他 code models 的训练效果，在 GitHub 仓库中实现了最高 96.78% 的代码作者辨别准确率。

基于对比学习的二进制代码预训练表示 2023 年 8 月 – 至今

- 提出了 CONTRABIN，一种新颖的 contrastive learning framework，通过对源码、注释和二进制代码进行 simplex interpolation，为二进制代码分析构建了高效的 embeddings。
- 在 binary functionality classification 中相比于之前的模型（RoBERTa 和 CodeT5）最高提升 14.34%，在 binary function name recovery 中提升了 13.60%，实现了显著的性能改进。

静态分析工具定位软件漏洞能力评估 2023 年 11 月 – 2024 年 12 月

- 首次提出基于行距离、文件距离和逻辑距离的评估方法，以衡量静态分析工具在漏洞定位方面的实际有效性，并通过分析 243 个开源项目的 1320 个真实漏洞，揭示了不同静态分析工具在不同漏洞类型上的差异。
- 实验结果表明，Infer 在 87% 的情况下能够提供最接近真实补丁位置的文件级警告，而 Flawfinder 提供的警告行距离最小，但误报率较高；研究表明，静态分析工具在不同类别的漏洞检测中表现各异，可用于指导工具选择。

基于证据理论的故障定位 2022 年 8 月 – 2023 年 8 月

- 提出了一种基于证据理论的故障定位方法，通过融合多个 SBFL（Spectrum-Based Fault Localization）技术的可疑度评分，减少了单一 SBFL 方法的不确定性，从而提高了故障定位的准确性。
- 通过引入模糊窗口技术，使可疑度评分在代码的相邻行间平滑分布，在 Defects4J 数据集上的实验表明，该方法最高可减少 34.5% 的代码检查工作量，并将 Top-10% 可疑代码行的命中率提升 27.9%。

技术技能

语言：普通话（母语），英语（流利）
编程技能：Pytorch, Tensorflow, Java（使用静态分析工具：SOOT），Python, C（使用静态分析工具：Infer），Shell