Tong Ye
1998.08.20

Email : tongye@zju.edu.cn
Mobile : +86-18868102709

Education

- **ZheJiang University** — Hangzhou, China
  PhD candidate of Cyber Security — *Sep. 2020 – Now*
- **ZheJiang University** — Hangzhou, China
  Bachelor of Theoretical Physics; GPA: 3.86/4 — *Sep. 2016 – Jun. 2020*

Honorary award

- **Zhejiang Provincial Government Scholarship**
- **Outstanding Graduates of Zhejiang Province**

Research Papers

- **CP-BCS: Binary Code Summarization Guided by Control Flow Graph and Pseudo Code** 2023
  EMNLP 2023 Main
  Automatically generating function summaries for binaries is an extremely valuable but challenging task, since it involves translating the execution behavior and semantics of the low-level language (assembly code) into human-readable natural language. However, most current works on understanding assembly code are oriented towards generating function names, which involve numerous abbreviations that make them still confusing. To bridge this gap, we focus on generating complete summaries for binary functions, especially for stripped binary (no symbol table and debug information in reality). To fully exploit the semantics of assembly code, we present a control flow graph and pseudo code guided binary code summarization framework called CP-BCS. CP-BCS utilizes a bidirectional instruction-level control flow graph and pseudo code that incorporates expert knowledge to learn the comprehensive binary function execution behavior and logic semantics. CP-BCS can significantly improves the efficiency of reverse engineering.
- **Tram: A Token-level Retrieval-augmented Mechanism for Source Code Summarization** 2023
  NAACL 2024
  We delved into the mapping between code and natural language. Although Neural Language Models (CodeT5, StarCoder, GPT family) achieve significant performance in this field (Code $\longleftrightarrow$ Natural Language), an emerging trend is combining neural models with external knowledge. In this work, we explore a fine-grained token-level retrieval-augmented mechanism on the decoder side to help the vanilla neural model generate a better code summary.
- **State-of-the-Art Survey of Open-source Software Supply Chain Security** 2021
  Journal of Software
  Software development is changing. Since the Internet allows far-flung development teams to collaboratively create software, open-source software supply chains are becoming more complex and sophisticated. This work tries to define the new open-source software supply chain model and presents a detailed survey of the security issues in the new open-source software supply chain architecture. Various emerging technologies, such as blockchain, machine learning (ML), and continuous fuzzing as solutions to the vulnerabilities in the open-source software supply chain have also been discussed.

Projects

- **Code Clone Detection**: Code clone refers to more than two duplicate or similar code fragments existing in a software system. We extract the Structural information (e.g., AST, CFG) of code to enrich the representation ability of code encoder, to align in the semantic space, and achieve the effect of the alignment of similar code in semantic space.

Interests

- **NLP domain**: Text Generation, Code Generation, Code Summarization
- Math & Physics

叶童

tongye@zju.edu.cn · (+86) 18868102709

教育背景

浙江大学, 杭州 2020 – 至今
在读博士研究生 网络空间安全

浙江大学, 杭州 2016 – 2020
本科 理论物理学

获奖情况

浙江省政府奖学金
浙江大学优秀毕业生

顶会论文

CP-BCS: Binary Code Summarization Guided by Control Flow Graph and Pseudo Code 2023

EMNLP 2023

如何让 NLP 语言模型理解计算机底层代码（二进制/汇编代码）是一项极具价值但又充满挑战的任务，因为它涉及将底层代码的执行逻辑和语义信息翻译成人类可读的自然语言。然而，当前大多数理解汇编代码的工作是朝着生成汇编函数名称的方向发展，而且对于底层代码的执行逻辑和真实语义的并无深刻的认知，因此学术界对于语言模型是否能够理解计算机底层代码依旧是个未知的领域。为了弥合这一差距，我们专注于为二进制代码生成完整自然语言摘要，特别是对于剥离二进制文件（针对实际厂商软件发布场景）。为了充分挖掘汇编代码的语义，我们提出了一个基于控制流图和伪代码引导的二进制代码摘要框架，称为 CP-BCS。CP-BCS 利用双向指令级控制流图和伪代码，结合专家知识，以学习全面的二进制函数执行行为和逻辑语义。

Tram: A Token-level Retrieval-Augmented Mechanism for Source Code Summarization 2024

NAACL 2024

我们深入研究了基于检索增强下的代码与自然语言之间的映射关系。尽管语言模型（如 CodeT5、StarCoder、GPT-3.5/4）在这一领域（代码 $\longleftrightarrow$ 自然语言）取得了显著的性能，但一个新兴的趋势是如何将语言模型与外部知识相结合。在这项工作中，我们首次探索了在解码器端使用细粒度的 token 级别检索增强机制，以纠正原始语言模型的生成词概率分布，融合外部知识特征，以生成更加准确的代码注释，尤其是对于低频词的生成效果尤为显著。

State-of-the-Art Survey of Open-source Software Supply Chain Security 2021

Journal of Software 2021

软件开发正在发生变化。由于互联网允许分布在不同地方的开发团队协作创建软件，开源软件供应链变得更加复杂和精细。这项工作试图定义新的开源软件供应链模型，并详细调查了新开源软件供应链架构中的安全问题。此外，还讨论了各种新兴技术，如区块链、机器学习和持续模糊测试作为解决开源软件供应链中现有问题的解决方案。

项目

我们面像计算机软件安全、逆向工程领域构建了首套大规模计算机底层代码理解框架及系统，并集成进安全工程师、逆向工程师的日常工作流程中，显著提高工程师的日常逆向效率（平均函数级别效率提升达 9.7 倍)。该系统已获第五届华为杯人工智能创新大赛浙江大学一等奖，参赛项目排第二 (2/29)。

其他

- 研究方向: 文本生成、检索增强、代码理解、大模型代码生成
- 个人网站: https://tongye98.com