Saikat Chakraborty

saikatc.info

PHONE: +1 (434) 242-1306 E-MAIL : saikatc@microsoft.com

ABOUT ME

I am a senior researcher at the Research in Software Engineering (RiSE) group at Microsoft Research, USA. My research area is *Programming Language Processing (PLP)* – a coalescence between Programming Languages and Machine Learning. More specifically, my research interest is on bringing user trust in ML generated code. I have been working towards empowering AI models generate provably correct code.

WORK EXPERIENCE

SEPTEMBER 2022 -TILL DATE	**Senior Researcher** at **Microsoft Research**, USA Research in Software Engineering (RiSE) group.
JANUARY 2019 -SUMMER 2022	**Research Assistant** Arise Lab, Columbia University, New York, NY. Working in AI4SE (Artificial Intelligence for Software Engineering) sub-group.
SUMMER 2021	**Software Engineer Intern** at **Facebook Inc.**, Remote. Probability (bigcode) team. Worked in Bigcode team for designing and development of source code diff model. Such model initiates first step towards automating code review process and improved the performance of regression prediction and prediction of different code review metrics.
SUMMER 2019	**Software Engineer Intern** at **Google LLC.**, Sunnyvale, CA. BinEval team. Worked in designing ML based tool for analyzing security and privacy. Designed models for identifying embedded malicious code in cloud documents.
SUMMER 2017 & 2018	**Research Intern Fujitsu Laboratories of America, Sunnyvale, CA** Worked with AI based fault localization. Extracted subtle information from auxiliary sources to improve the performance of fault localization and program repair.

SELECTED PROJECTS

iRank [2]	In this work, we propose a *re-ranking* approach for the loop invariants generated by LLMs. We have designed a ranker that can distinguish between correct inductive invariants and incorrect attempts based on the problem definition.
CausalVul [1]	Wwe introduced causality into deep learning-based vulnerability detection, by discovering spurious features that the model may use to make predictions, subsequently applying the causal learning algorithms, specifically, do-calculus, to remove the use of spurious features and thus promote causality based prediction.
NatGen [6]	A new pre-training objective "Naturalizing" of source code, exploiting code's bimodal, dual-channel (formal & natural channels) nature. Learning to generate equivalent, but more natural code, at scale, over large corpora of open-source code, without explicit manual supervision, helps the model learn to both ingest & generate code.
Codit [11]	A tree based hierarchical NMT tool for learning frequent code change patterns. Tree based modeling technique guarantees syntactic correctness of the edited code. This is an industry collaboration with Microsoft Research, Cambridge.
PLBART [10]	A large scale pretrained model for multiple programming languages. PLBART is trained on several hundred millions source code in Java and Python and technical natural languages from stackoverflow.
Redcoder [9]	A framework combining code search and source code synthesis. Given a summary of programmer intention, REDCODER relevant source code and adapts those code based on developers' need.
Modit [8]	A multi-modal framework for source code editing. MODIT accounts for code edit context and developers' intention for editing to generate precise edited code.

DISCO [7]	A source code understanding pretrained model that learns to reason about the functional properties of the code. This is an industry collaboration with IBM Research.
ReVeal [15]	An empirical study for understanding the feasibility of Deep Learning Based Vulnerability Detection for detecting real world vulnerabilities. We identified major challenges in using DL-based systems for Vulnerability detection, and proposed prospective solution.

Selected Publications

[1]* [**ICSE'24**] Towards Causal Deep Learning for Vulnerability Detection, Md. Rahman, I. Ceka, C. Mao, S. Chakraborty, B. Ray, W. Le, accepted to be published at 46th International Conference on Software Engineering (ICSE) 2024.

[2] [**EMNLP'23 (findings**] Ranking LLM-Generated Loop Invariants for Program Verification, S. Chakraborty, S. K Lahiri, S. Fakhoury, M. Musuvathi, A. Lal, A. Rastogi, A. Senthilnathan, R. Sharma, N. Swamy, accepted to be published at the findings of The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP) 2023.

[3]* [**ESEC/FSE'23**] GrACE: Generation using Associated Code Edits, P. Gupta, A. Khare, Y. Bajpai, S. Chakraborty, S. Gulwani, A. Kanade, A. Radhakrishna, G. Soares, A. Tiwari, accepted to be published at The ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE) 2023.

[4]* [**ISSTA'23**] CONCORD: Clone-aware Contrastive Learning for Source Code, Y. Ding, S. Chakraborty, L. Buratti, S. Pujar, A. Morari, G. Kaiser, B. Ray,32nd ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA) 2023. [**Won ACM Sigsoft Distinguished Paper Award**].

[5] [**ICSE-NIER'23**] On ML-Based Program Translation: Perils and Promises, A. Malyala, K. Zhou, B. Ray, S. Chakraborty published at the IEEE/ACM International Conference on Software Engineering - New Ideas and Emerging Results (NIER) track (ICSE-NIER) 2023.

[6]* [**ESEC/FSE'22**] NatGen: Generative pre-training by "Naturalizing" source code, S. Chakraborty, T. Ahmed, Y. Ding, P. Devanbu, B. Ray, accepted to be published at The ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE) 2022 **(Acceptance rate : 22%)**.

[7]* [**ACL'22**] Towards Learning (Dis)-Similarity of Source Code from Program Contrasts, Y. Ding, L. Buratti, S. Pujar, A. Morari, B. Ray, S. Chakraborty, Published at 60th Annual Meeting of the Association for Computational Linguistics **(Acceptance rate : 20.75%)**.

[8]* [**ASE'21**] On Multi-Modal Learning of Editing Source Code, S. Chakraborty, B. Ray, Published in The 36th IEEE/ACM International Conference on Automated Software Engineering. **(Acceptance rate : 28%)**.

[9] [**EMNLP'21 (findings)**] Retrieval Augmented Code Generation and Summarization, MDR. Parvez, WU. Ahmad, S. Chakraborty, B. Ray, K. Chang, Findings of The 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP-findings), 2021. **(Acceptance rate : 38%)**.

[10]* [**NAACl'21**] A Unified Pre-training for Program Understanding and Generation, WU. Ahmad[§], S. Chakraborty[§], B. Ray, K. Chang, Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2021. [§] Co-first authors. **(Acceptance rate : 26%)**.

[11]* [**TSE'20**] CODIT: Code Edits with Tree Based Machine Translation, S. Chakraborty, Y. Ding, M. Allamanis, B. Ray, in IEEE Transactions on Software Engineering, 2020. **(Impact factor : 3.331)**.

[12]* [**ACL'20**] A Transformer-based Approach for Source Code Summarization (short paper), WU. Ahmad, S. Chakraborty, B. Ray, K. Chang, 58th Annual Meeting of the Association for Computational Linguistics (ACL) 2020. **(Acceptance rate : 17.6%)**.

[13] [**SCAM'19**] Toward Optimal Selection of Information Retrieval Models for Software Engineering Tasks, MM. Rahman, S Chakraborty, G. Kaiser, B. Ray, 19th International Working Conference on Source Code Analysis and Manipulation (SCAM) 2019. **(Acceptance rate : 39.6%)**.

[14]* **[ACL'18]** Building Language Models for Text with Named Entities, R. Parvez, S. Chakraborty, B. Ray, K. Chang, 56th Annual Meeting of the Association for Computational Linguistics (ACL) 2018. **(Acceptance rate : 24.9%)**.

[15]* **[TSE'21]** Deep Learning based Vulnerability Detection: Are We There Yet? S. Chakraborty, R. Krishna, Y. Ding, B. Ray, accepted to be published in IEEE Transaction of Software Engineering. **(Impact factor : 3.331%)**.

[16] **[ICSE'18 (poster)]** Which similarity metric to use for software documents?: a study on information retrieval based software engineering tasks. Md. Rahman, S. Chakraborty, and B. Ray, **Poster** at Proceedings of the 40th International Conference on Software Engineering: Companion Proceeedings. ACM, 2018.

Education

August 2022	**Ph.D. in Computer Science** **Columbia University, New York, NY, USA** **Area:** Artificial Intelligence for Software Engineering. **Expertise:** Source Code Analysis, Deep Learning, Natural Language Processing, Neural Machine Translation. **Thesis:** Learning to Edit Code. **Advisor:** Dr. Baishakhi Ray.
January 2019 February 2021	**Master of Science in Computer Science** **Columbia University, New York, NY, USA**
March 2009 - July 2014	**B.Sc. in Computer Science and Engineering** **Bangladesh University of Science and Technology** Dhaka, Bangladesh **Advisor:** Dr. Md. Monirul Islam **Thesis:** Diversity Guided Unified Evolutionary Framework for MDPCVRP.

Selected Talks

1. Detecting Vulnerabilities in Source Code at Vulnerability Detection and Security Research, National Security Agency, (NSA) (November 2021).
2. On Multi-Modal Learning of Editing Source Code at The 36th IEEE/ACM International Conference on Automated Software Engineering (ASE). (November 2021).
3. PLBART : Unified Pre-training for Program Understanding and Generation at Decal Lab - UC Davis (April 2021), NAACL 2021 (June 2021), IBM Research (June 2021), Facebook BigCode team (August 2021).
4. Programming Language Processing - Learning to Edit Code at Programming Systems Lab Research Seminar, Department of Computer Scicence, UC Berkeley (May 2021).
5. CODIT: Code Editing with Tree Based Neural Models at 43rd International Conference on Software Engineering (ICSE) (May 2021).
6. Machine Learning for Source Code Analysis at Open University, UK and Toshiba, UK (March 2021).
7. A transformer-based approach for source code summarization. at 58th Annual Meeting of the Association for Computational Linguistics (ACL) (April 2020).

Service Experience

Reviewer	IEEE Transaction of Software Engineering (TSE). ACM Transactions on Software Engineering and Methodology (TOSEM). IEEE Software. Conference and Workshop on Neural Information Processing Systems (NeurIPS) 2022. Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI), 2023. IEEE Transactions on Neural Networks and Learning Systems. Springer Journal of Automated Software Engineering. International Conference on Learning Representation (ICLR) 2023.

*Top tier publications.

PC member	ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (**ESEC/FSE**) - Tool Demonstration Track, 2022; Research Paper Track, 2023. International Conference on Software Engineering (**ICSE**) - Technical Track, 2024. MSR Mining Challenge, 2021. Workshop on Natural Language Processing for Programming (NLP4Prog), 2021.
Session Chair	**ESEC/FSE 2021.** Program Repair, Code Recommendation. **ESEC/FSE 2022.** Program Repair/Synthesis, Human Computer Interaction.
Leadership	**Secretary,** Association of Bangladeshi Students at UVa. (2017-2018). **Founding Vice President**, Engineering Students' Association of Bangladesh . **Organizer**, International Engineering Innovation Summit Bangladesh, 2015.

Teaching Experience

Fall 2020	**Lead Teaching Assistant** Programming Languages and Translators Columbia University (enrollment : 150). **Responsibilities** : Weekly office hour and recitation class, Designing course assignments and exam questions, Grading.
Fall 2016 - Spring 2017	**Teaching Assistant** Discrete Math, University of Virginia **Responsibilities** : Grading.
October 2014 -August 2016	**Lecturer** Ahsanullah University of Science and Technology, Dhaka, Bangladesh **Courses taught** : Compilers (both theory and lab), Digital system design (both theory and lab).