Xigao Li

Ph.D in Web Security

Email: lxgfrom2009@gmail.com
Web: https://xigaoli.com
Github: github.com/xigaoli
Phone: 631-305-1054
[LinkedIn Profile]

Summary

Software Engineer, result-oriented Ph.D proficient in C++, Python and Java. Experienced in machine learning, web security, familiar with industry-level codebase; my research build lightweight and pragmatic deep learning models for web security.

Education

Stony Brook University — Stony Brook, NY
Ph.D Candidate in Computer Science; GPA:4.0/4.0 — *Aug. 2018 – Aug 2023*
M.S. in Computer Science, GPA:4.0/4.0

State Grid Electric Power Research Institute (SGEPRI) — Nanjing, China
M.S. in Electrical Engineering; GPA:87/100 — *Sep. 2013 – July 2016*

North China University of Water Resources and Electric Power — Zhengzhou, China
Bachelor of Engineering in Computer Science and Technology; GPA:3.9/5 — *Sep. 2009 – July 2013*

Experience

Software Engineer — Aug 2023 – Present
Bloomberg LP — *New York, NY*

- Designed ticket throttling system for Bloomberg company-wide internal/external device monitoring, reduced the system maintenance burden

Software Engineer Intern — May 2022 – Aug 2022
Meta (Facebook) — *Menlo Park, CA*

- Led the generic score adjustor project, using C++ and Python to build expression based score adjustor for commerce indexer scoring stage;user rule update time reduced from 7 days to 30 minutes
- Optimized expression library, designed in-place optimization to expression AST, greatly reduced computational complexity of evaluation, designed benchmark suite, optimized expression gained 65.3% performance
- Migrated Facebook Cowatch client to new score adjustor; conducted client side A/A testing in production environment with real user traffic

Data Science Intern — May 2021 – Aug 2021
FireEye .Inc — *(Remote) Milpitas, CA*

- Led the malware emulation project, developed a malware emulation pipeline and extracted API call sequence, Memory Tag and access counter as features
- Trained a gradient boosting tree model over EMBER17 and EMBER18 dataset, achieved 0.99 AUROC
- Built a hybrid CNN model for malware family classification

Graduate Research Assistant — June 2019 – Present
Stony Brook University — *Stony Brook, NY*

- Created automatic systems that can deploy honeypot-like web servers to capture web bot activities
- Trained CNN and LSTM deep neural network to detect malicious URLs
- Clustered web bot by page traverse order through K-medoids and Hierarchical cluster algorithms

Skills

Programming Languages: Python (primary), C/C++, Java
Web server / database systems: MySQL, MSSQL, Apache, Nginx.
Developer Tools: Elasticsearch (ELK), RESTful API, Cloud Environments (AWS), web crawlers, kernel-level programming in C, Linux file systems
Libraries: PyTorch, Keras, Selenium, BeautifulSoup, Matplotlib,Sklearn
Area of expertise: Web security, Web privacy, Deep learning, Computer Vision

Selected Publications [Google Scholar]

- *Scan Me If You Can: Understanding and Detecting Unwanted Vulnerability Scanning*, **Xigao Li**, Babak Amin Azad, Amir Rahmati, Nick Nikiforakis, TheWebConf (WWW) 2023
- *Double and Nothing: Understanding and Detecting Cryptocurrency Giveaway Scams*, **Xigao Li**, Anurag Yepuri, Nick Nikiforakis, Network and Distributed Systems Security (NDSS) Symposium, 2023
- *Good Bot, Bad Bot: Characterizing Automated Browsing Activity*, **Xigao Li**, Babak Amin Azad, Amir Rahmati, Nick Nikiforakis, IEEE S&P (Oakland) 2021 [paper][poster]
- *Malware Classification with Deep Neural Network using Lightweight Emulation*, **Xigao Li**, David Krisiloff, CAMLIS 2021 (in submission)
- *A Hybrid Disaster-Tolerant Model with DDF Technology for MooseFS Open- Source Distributed File System*, **Xigao Li**, Lin Qian, Journal of Supercomputing, 2016 [paper] [github]
- *A Direct Data Fetch Technology Applied in Disaster-Tolerant Model of Distributed File System*, **Xigao Li**, Lin Qian, ICCSNT 2015 [paper]
- *Research on Testing Model of IT infrastructure*, Lin Qian, Jun Yu, Jia Wu, Guangxin Zhu, Hengmao Pang, **Xigao Li**, Xuran Wang, ICITMI 2015

Projects

Good Bot, Bad Bot: Characterizing Automated Browsing Activity | *Python, PHP, Crawlers, AWS*

- Created automatic systems that can deploy honeysites; crawled 7 month of data, including over 26.4M requests
- Developed behavioral fingerprinting techniques; found over 57% bots are malicious
- Utilized TLS fingerprinting to capture over 86% bots were lying identity
- Created visualization of captured bot dataset, finding 5 RCE exploits were quickly being abused
- **Best Video Editing Award** in IEEE S&P 2021 [award] [video]

Malware classification using lightweight emulation | *PyTorch, NLP* *Project at FireEye. Inc*

- Developed an automated malware emulation pipeline, emulated 1.1 Million malware with cost <10 hours (EMBER'17)
- Extracted malware API call sequence, memory access information and RWX counter
- Trained a lightGBM and a character level CNN model, achieved 0.99 AUROC / 0.98 accuracy
- Developed a hybrid CNN model classifying malware families, reached 0.96 accuracy

Malicious URL detection with deep neural network | *Python, Tensorflow, Keras*

- Crawled and collected various malicious and benign URLs to build large dataset.
- Trained and fine-tuned CNN and LSTM models to identify malicious URLs
- Implemented deep learning models into mobile browsers

Disaster-tolerant distributed file system [github] | *C/C++, Bash, Linux kernel*

- Developed hybrid disaster-tolerant model for open-sourced distributed file system
- Customized and recompiled CentOS kernel for performance optimization

Supervised Students

- *Karan Gada*, Web bot classification project, May 2021 - Present
- *Anurag Yepuri*, Crypto-currency scam detection project, May 2021 - Present

Certificates

Neural Networks and Deep Learning - Coursera [Cert] | *Python, NumPy, Tensorflow* Dec 2020

Python for Data Science & Machine Learning - Udemy [Cert] | *NumPy, Sklearn* Feb 2021

Engineering Virtual Program - Golden Sachs [Cert] | *C/C++, Bash, Crawlers* Dec 2020