Zitao Chen • zitaoc@ece.ubc.ca

Research interests

Machine learning security; intersection between ML and security

Education

University of British Columbia — Vancouver, Canada
Master of Applied Science in Electrical and Computer Engineering — *Sept. 2018 - May 2020*
Supervisor: Karthik Pattabiraman
Thesis: Understanding and Improving the Error Resilience of Machine Learning Systems
Thesis Committee: Karthik Pattabiraman, Sathish Gopalakrishnan, Prashant Nair

China University of Geosciences (Wuhan) — Wuhan, China
Bachelor degree in Information Security — *Sept. 2014 – Jun. 2018*

Research Experience

1. Security of Machine Learning (adversarial ML) — Ongoing Research

- *1.1. Research question:* How to **detect** *universal adversarial patch attack* that can trigger targeted misclassification on input embedded with input-independent adversarial patch (AP).
  - *Insight:* AP always exhibits the malicious behavior to cause targeted misclassification despite the image it locates in, which can be used as a symptom to detect AP.
  - (1) Locate potential AP by using attribution method to identify the high attribution features in the original input. (2) *Crop and transfer* the high attribution features (that contain potential AP) to a *new* image. (3) *Same* predicted label on the original and new image indicates the existence of AP.
  - *Result:* Detection accuracy of **97.64%** on AP input, with 18.4% false positive rate (FPR).
- *1.2. Research question:* How to **rectify** misclassification due to AP and reduce false detection on benign input.
  - *Insight:* AP is vulnerable to corruption due to its localized feature while benign features are more robust.
  - (1) Randomly *mask* the high attribution features and use *image inpainting* to reconstruct the masked contents. (2) This could force the AP to become non-adversarial and rectify misclassification. (3) Same predicted label on the input before and after inpainting indicates false detection on benign input.
  - *Result:* Robust accuracy of **75.51%** on *AP input*; FPR is reduced to **5.24%**.

2. Reliability of Machine Learning — Master thesis

2.1. Understanding the error resilience of ML

- *Research question:* How to efficiently **identify** the *critical faults* in ML systems due to hardware transient faults (i.e., bit-flips)? These are the faults that could corrupt the ML programs' output, e.g., misclassification.
- *Insight:* Common ML models consist of functions with *monotone* property and thus deviations by faults also propagate monotonically to the output layer. E.g., if a fault at one bit is identified as critical fault via fault injection (FI), any fault at *higher*-order bit will also be deemed as critical fault based on the monotone property, *without* FI.
- Design a **bin**ary-search-like **f**ault **i**njector (BinFI) to efficiently identify the critical faults in ML systems.
- *Result:* BinFI identifies 99%+ of critical faults with 99%+ precision, and significantly outperforms existing approach.
- Code: https://github.com/DependableSystemsLab/TensorFI-BinaryFI

2.2. Improving the error resilience of ML

- *Research question:* How to **protect** ML systems from hardware transient faults?
- *Insight:* ML inference is *inherently tolerant* to insignificant errors, which can be leveraged to mitigate critical faults.
- Propose to *selectively restrict the value ranges* in different layers of the models, which can dampen the large deviations due to critical faults. The reduced deviations can be *inherently tolerated* by the models to generate correct outputs, thus enabling fault correction *without* re-computation.
- Implement an *automated* transformation to convert unreliable ML models into the error-resilient ones.
- *Result:* The proposed technique significantly improves the ML reliability (e.g., reduce the chance of misclassification due to transient faults from $\sim 15\%$ to $\sim 0.4\%$) with negligible overhead ($\sim 0.5\%$).

Publications **(SC is a top venue in High Performance Computing indexed by csrankings.org)**

- **[IOLTS'20]** Karthik Pattabiraman, Guanpeng Li, Zitao Chen, "Error Resilient Machine Learning for Safety-Critical Systems: Position Paper" *IEEE 26th International Symposium on On-Line Testing and Robust System Design*, 4 pages, 2020. *Invited paper*
- **[ISSRE'20]** Zitao Chen*, Niranjhana Narayanan*, Bo Fang, Guanpeng Li, Karthik Pattabiraman, Nathan DeBardeleben, "TensorFI: A Flexible Fault Injection Framework for TensorFlow Applications" *The 31st International Symposium on Software Reliability Engineering*, 2020 **Acceptance rate: 25.7% (38/148)**

- **[SC'19]** Zitao Chen, Guanpeng Li, Karthik Pattabiraman, Nathan DeBardeleben "BinFI: An Efficient Fault Injector for Safety-Critical Machine Learning Systems", *In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2019 **Acceptance rate: 20.9% (72/344). Finalist for SC reproducibility challenge (one of 3 papers)**
- **[FGCS]** Zitao Chen, Wei Ren, Yi Ren and Kim-Kwang Raymond Choo, "LiReK: A Lightweight and Real-time Key Establishment Scheme for Wearable Embedded Devices by Gestures or Motions", *Future Generation Computer Systems* (2018) [Impact Factor: 6.125] *Undergrad research*

Preprint

- **[ArXiv]** Zitao Chen, Guanpeng Li, Karthik Pattabiraman"A Low-cost Fault Corrector for Deep Neural Networks through Range Restriction" **Submitted to DSN'21**

Others

Open-source project: https://github.com/DependableSystemsLab/TensorFI Sept. 2018 - present

Teaching Experience: CPEN400A 2019

Programming languages: Python, Java, C, C++

Award: UBC Faculty of Applied Science Graduate Award: $4000; $3000 2019,2020