Research Statement

Traveling through system security to AI security, I found my research interest around **robust, reliable, and explainable AI**.

In the first year, I joined the ZJU SuperComputing Team. Under guidance from Prof. *Jianhai Chen*, I participated as a member and earned the second prize in the ASC 20-21 Student Supercomputer Challenge. With the professional knowledge I garnered, I pursued my first research project on **secure heterogeneous computation**, advised by Prof. *Chen* again, at the ZJU INCAS Lab. The project aimed at protecting data privacy and performing secure computations (e.g., deep learning programs) on CPU-GPU architectures. Various concepts I did not have a background in, such as hardware and operating systems, confused me as a sophomore student. As the team leader, I decided to investigate a specific aspect – protecting the data flow during the communication period between host and device. Most of the existing work relied heavily on special secure hardware, while our focus only required system-level encryption and decryption. With the familiarity of CUDA programming and Trusted Execution Environment (TEE), we quickly planned a roadmap for *Enchecap*, our design on a secure computation protocol that extends the safe region beyond TEE and prevents malicious data-stealing attacks on both transmissions and host memory. We further implemented our design into a demo and an open-source library with Intel SGX and Nvidia CUDA toolkits. Our secure computing paradigm shelters sensitive data on runtime level, while importing only 19% overall overhead.

My first research experience motivated me to probe into data-driven security. I asked myself: What are the security demands in real ML/AI scenarios, and can we protect DNN models at algorithm level rather than system level? This May, soon after completing my last research project, I sought a chance to work on **AI security** as a remote research intern in Prof. *Ting Wang*'s ALPS group at the Pennsylvania State University and was co-advised by Prof. *Shouling Ji* at the ZJU NESA Lab. After diving into various aspects of AI security, I began my project on **verifying the (non-)existence of backdoors** in a pre-trained DNN model. DNN models with backdoors make targeted incorrect predictions on clean inputs with an attacker-specified trigger embedded. We aimed at providing certified robustness for DNN models against backdoor attacks, for example, to answer the question -- whether the model is free of backdoor under a certain amount of perturbation. This project explored more than just the widely discussed adversarial robustness verifications. I formulated the backdoor certification problem by adding an additional "minimax", and directly applied linear programming based on an existing NN verifier, CROWN, building up the initial verifier for backdoors. My algorithm worked as expected and experiment results demonstrated more than 15% improvement compared to certified adversarial robustness. To even advance our certification, I decided to apply a relaxation trick and exchange the "min" and "max". After that, I was able to optimize and achieve even tighter certified bounds, which allowed the experiment results to improve by at most 30% compared to the original. The project is still in progress.

During the experiments, I accidentally visualized the CROWN weight parameters and realized they could indicate the positions of the potential backdoor triggers. So, I pursued follow-up work, trying to utilize the parameters to reverse engineer triggers. Following the spirit of Gradient Descent, I proposed and implemented **a trigger restore algorithm**, searching for the potential trigger step by step. While testing our restored triggers on clean inputs, they achieved approximate attack success rates compared with the state-of-the-art methods. Furthermore, our algorithm requires no real input at all, and our results match the real injected triggers, making it an important first step towards faithful trigger restoration.

With the endless nights I have spent on the two projects, I consolidated my research abilities -- analyzing problems with an objective vision, actively experimenting with potential ideas, and more. Moreover, I was fascinated by existing work and unsolved problems about secure and robust AI and decided to dive even deeper.

My research continued when I was invited to work on a project about **deployment-stage backdoor attacks**, under the supervision of Principal Researcher *Jifeng Zhu* at the Tencent ZhuQue Security Lab and Professor *Kai Bu* at ZJU. This project focused on DNN's neglected vulnerabilities at the deployment stage concerning backdoor attacks. We proposed the first **gray-box and physically realizable weight attack algorithm for backdoor injection**, namely, **Subnet Replacement Attack (SRA)**. The idea of SRA is simple, replacing a narrow chain of a complete neural network by the attacker's backdoor subnet to inject a malicious backdoor. Based on previous experiences with neural network backdoors, I quickly implemented, tuned, and evaluated SRA on various models and datasets to manifest its universal compatibility. I also extended SRA to different trigger types and even physical-world triggers, making our proposed attack a lot more practical by triggering backdoor behaviors in the real world flexibly. On the other hand, my teammates showed feasible ways to apply SRA on system levels. Together, we combined insights from both AI security and system security and found SRA to be harmful, stealthy, and extremely practical. Attackers could secretly hack DNN models by tampering with user device memory through infectious trojan scripts. In addition to completing all of the simulation experiments, I wrote a majority of our paper and submitted it to CVPR'22 as a first co-author.

I journeyed through my undergraduate research like completing a puzzle. System-level secure computation provided me with the initial piece, after which AI security replenished even more intriguing pictures. Eventually, our work on SRA bridged the gap between my understanding of system security and AI security. Throughout my experiences and work, I found the robustness of machine learning being both a "dark cloud" and an attractive perspective to work on. I plan to conduct my research in two aspects. First, I intend to study and solve security concerns involving current non-robust deep learning models. Second, I would like to better understand AI's behaviors and make their predictions more human-like through explainable and causal methods. In summary, I hope to fully explore the breadth and depth of secure, robust, and reliable AI.