秦孟远

15035138797 | qinmengyuan@std.uestc.edu.cn

B.E. Software Engineering,UESTC,Class of 2021

RESEARCH INTERESTS:NLP,Large-scale Model,Federated Learning,Multimodal Learning,Adversarial Learning

TECHNOLOGY STACK

- Proficient in Deep Learning, especially in **Large-Scale Natural Language Processing** and **Federated Learning**, also skilled in **Adversarial Learning**, **session-based recommendation** and **image classification**
- Proficient in **Big Data Analysis and Calculation**, and familiar with **Hadoop**, **Spark**, **Hive**, **Trino**, **ClickHouse**, etc.
- Skilled in **parallel & distributed computing**, such as data parallel, model parallel, ZeRO and **MPI**, **NFS**
- Familiar with **Python**, **C/C++**, **Assembly Language**, skilled in **Java**; familiar with Linux programming, the principles of operating system
- Skilled in common protocols &security protocols of computer network and computer networks programming
- Proficient in **encrypted traffic analysis** and **APT**, familiar with **network&software security**, **reverse**, **cryptography**, **antivirus** and **social engineering attack**;familiar with Yara and Sigma
- Skilled in **flask** and **Django** frameworks, **Nginx**, **RESTful** API design specification; Skilled in **PgSQL**, **SQLite**, SQLAlchemy

AWARDS

Excellent Communist Party Member 2021

Consecutive **UESTC Pacesetter Scholarships** for the 2017-2018 and 2018-2019 academic years 2017-2019

The **national second prize** of 2019 (The 12th) China University Student Computer Design Competition 2019

School-level second prize of the 5th China International College Students' "Internet+" Innovation and Entrepreneurship Competition 2019

RESEARCH EXPERIENCE

Research on Large-Scale Language Model for detecting encrypted webshell communications Mar 2022 - Jul 2022

Responsible for the research on the encrypted webshell traffic detection algorithm, which had detected a large number of **uniquely reported attacks** in practical scenarios, reduced about 95% false positives contrasted to the rule-based detection(in a grey test with 10 billion open-environment data) and achieved the generic ability to distinguish different attack tools for the first time

- Investigated **the gap of webshell traffic detection using deep learning** and related fields in a short period, selected the direction and designed the technical plan
- Trained the **end-to-end transformer architecture** model with universal comprehension of traffic by pre-training it with **MLM** and **same-origin prediction** tasks on a large amount of actual traffic, and then **fine-tuned** the model for downstream tasks on the data including mainstream webshell tools' payloads to achieve the targeting ability
- To avoid skewing the model's comprehension of both encrypted and plain traffic caused by the more obvious morphology and syntax of the latter, the tokenizer was built with two vocabs trained separately for the designed task
- Planned to use **Prompt Tuning** or **FLAN** (especially their **Few-shot Learning** ability) by **injecting knowledge into the templates** to enable the model architecture to incorporate knowledge from other mature industry approaches whose targets are not part of the payloads(e.g. rule detection, behavioural patterns), as well as to be compatible with other tasks in business scenarios where the application determines the content of traffic payload (e.g. advanced persistent threats, backdoor software, botnets) and **evolve into a generic traffic processing model**
- For the first time, the model achieves **almost full coverage**(99.4%) to the detection of known attack tools, and still had a detection rate of ~50% for their variants & other unknown tools
- Deployed models on the k8s cluster with multi-vGPU parallel computing and used Spark to fetch data from pulsar

Research on federated learning algorithms |advised by 刘峤 Jan 2021 - Aug 2021

- Researched **improving** both **clients' local performances** and **the global generic performance** simultaneously **without defeating any original intention of FL**, rather than sacrificing either of them as previous works did, through learning a client-specific local bias
- Developed an image classification engine based on the above algorithms, supporting LR, CNN and ResNet on Non-IID datasets such as FEMNIST, CIFAR-10 and CIFAR-100 processed by the LEAF method
- Constructed a distributed computing cluster running the engine based on the gRPC specification, NFS file system and MPI specification

A national research project on situational awareness in cyber security |the Cyberspace Security Lab of UESTC Nov 2019 - Apr 2020

- Researched the clustering algorithm that is both **incremental** and **hierarchical**
- Constructed a BERT+AutoEncoder network architecture with Keras to embed mixed multilingual text into a vector representing the desired features and compressed it for clustering

PROFESSIONAL EXPERIENCE AND OTHER PROGRAMS

Sangfor Aug 2021 - Aug 2022

Machine Learning Engineer, Algorithm Team

- Led the webshell traffic detection research project mentioned above
- Analysed webshell-related attack chains, TTP (Tactics, Techniques and Procedures) and designed detection solutions
- Responsible for further reducing the WAF's false alarms (especially in command injection and SQL injection)

360 Mar 2020 - Aug 2020

Security AI Engineering Intern, HuntingZero Lab at the Core Security Technology Centre

- **APT attack clustering:** converted file samples of APT attacks and other file samples into vectors and clustered them through Word2Vector to train a model that can identify variants of known attack samples of each APT organisation for painting attackers' portraits and threat hunting
- Developed the backend of the intelligence operations management system and the knowledge base of detection rules with an architecture of **Nginx+flask/Djiango+SQLAlchemy+PgSQL/SQLite** to implement the requirements of operating threat intelligence, to manage the team's automated detection rules and to prepare for threat hunting

Member of the Lime cyber security team at the Zhichuang (Lime) Studio of UESTC

XinYi - Intelligent Emotion Management System based on physiological signals Aug 2018 - Aug 2019

responsible for designing the encryption algorithm and implementing the part of data security

The distributed crawler project with Scrapy+redis architecture Aug 2019 - Dec 2019

The port scanner wrote by C

VISITING PROGRAMS & LEADERSHIP EXPERIENCE

- **short-term visiting programs:** PolyU; National Research University of Electronic Technology (MIET)
- **Party branch committee** | **League branch secretary** | **Student union officer**