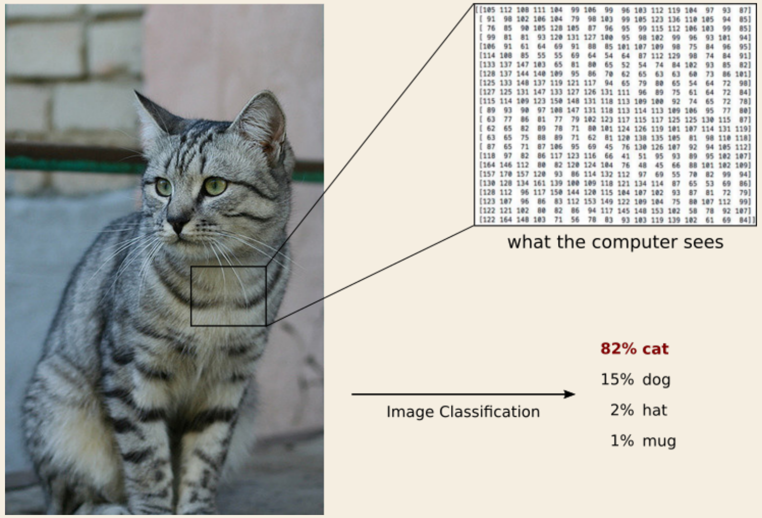
Securing Malware Cognitive Systems against Adversarial Attacks

Yuede Ji Benjamin Bowman **H. Howie Huang**

Graph Computing Lab
The George Washington University

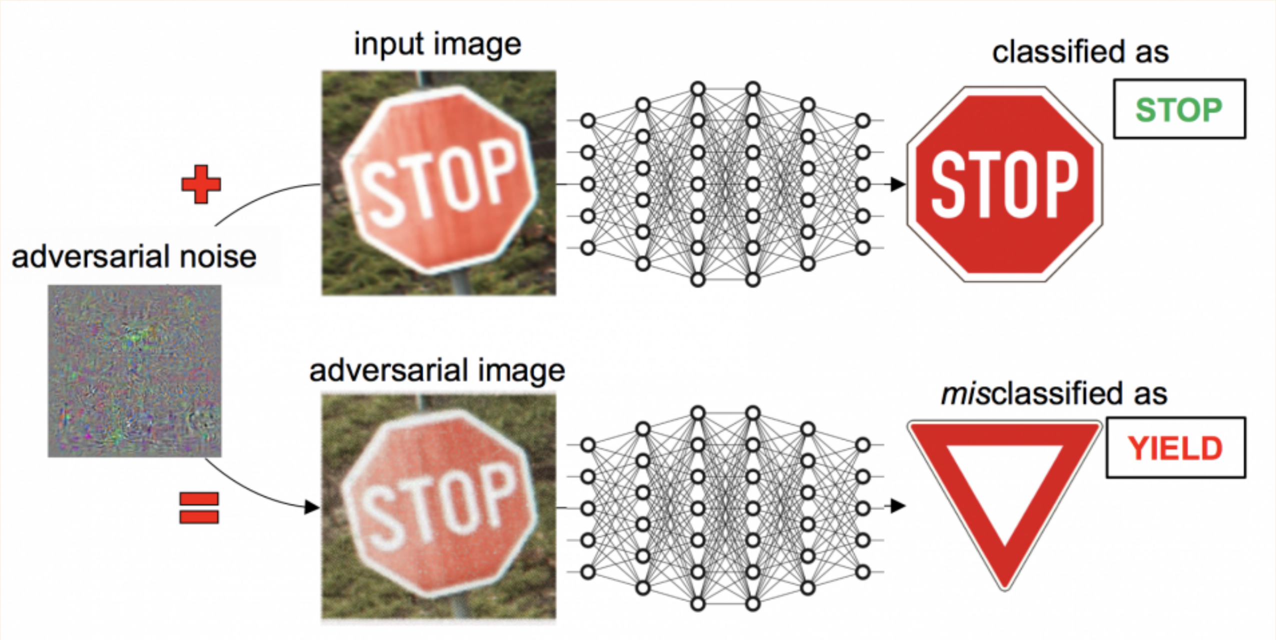
Cognitive System

- A self-learning system leverages a combination of intelligent techniques, such as machine learning (ML), and data mining.
- It has made breakthrough performance in many applications, such as image processing, self-driving vehicles, and cybersecurity.



Adversarial Attack

- Adversarial attacks try to cause the machine learning methods to misbehave or leak sensitive model information.
- The cognitive systems are vulnerable to adversarial attacks.



Picture credits to "Vaccinating machine learning against attacks"

Malware Cognitive Systems

- Applying cognitive intelligence to malware detection
  - Gained great popularity, which has been used in Sparkcognition, Cisco, IBM, Cybereason.
- Such systems are vulnerable to adversarial attacks.

Outline

- Background
- Problem Definition
- DeepArmour
- Experiment
- Conclusion

Background: Malware

Wikipedia
The Free Encyclopedia

Main page
Contents

2019 Baltimore ransomware attack

From Wikipedia

The Baltimore
ransomware c

Another Hack
a Ransom, Th

By Patricia Mazzei

June 27, 2019

MIAMI — Even the pho
City, Fla., after hackers
city's computer systems

Forensic science

Hacked forensic firm pays ransom after malware attack

Largest private provider Eurofins hands over undisclosed fee to regain control of systems

Hannah Devlin
Science correspondent

@hannahdev

Fri 5 Jul 2019 11.51 EDT

15

▲ Ransomware is a type of computer program that infiltrates IT systems and threatens to publish data or block access until money is paid. Photograph: Wilfredo Lee/AP

Ransomware Hits Georgia Courts as Municipal Attacks Spre

07:49 PM

HITS GEORGIA
NICIPAL ATTACKS

Background: Adversarial Attack

- Data poisoning attack
  - Training phase
  - Add "poisoned" training data to confuse the inference result.
- Evasion attack
  - Testing phase
  - Test multiple data to identify the network gradients, thus perform targeted attack.
- Exploratory attack
  - Testing phase
  - Aim to extract knowledge from a trained model instead of fooling it

Outline

- Background
- Problem Definition
- DeepArmour
- Experiment
- Conclusion

Problem Definition

- Task Definition
  - Aim to defend evasion attacks for malware classification
  - Five malware classes, no benign software

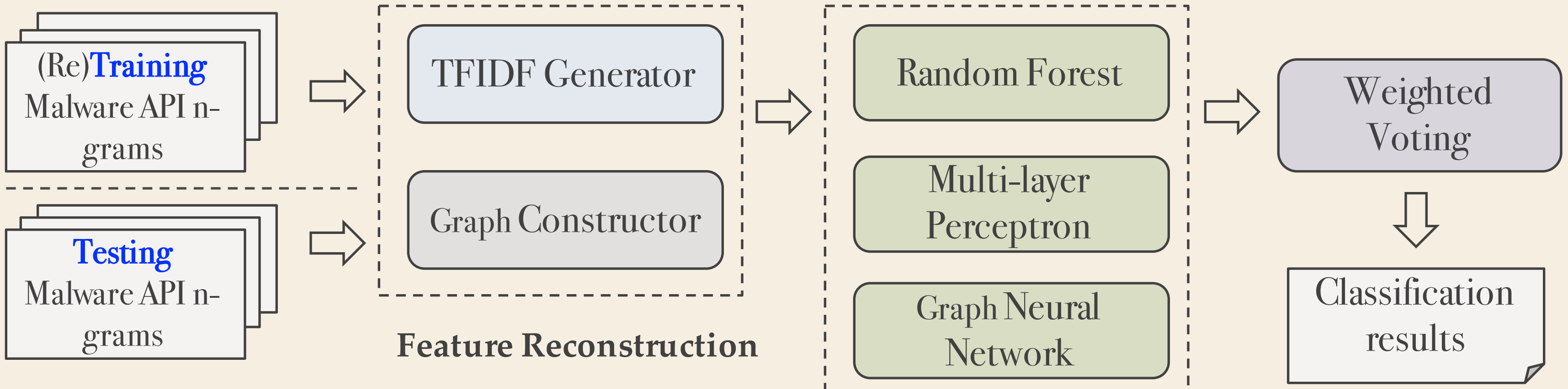
- Threat Model
  1. The adversarial attacks can only happen at the testing stage.
  2. The adversaries may have knowledge of the training dataset, but are not allowed to modify it.
  3. The adversaries have no knowledge of the trained model (architecture, parameters).
  4. The adversaries only aim at degrading the performance in terms of accuracy metrics and are not attacking any confidentiality or privacy issues.

Outline

- Background
- Problem Definition
- DeepArmour
- Experiment
- Conclusion

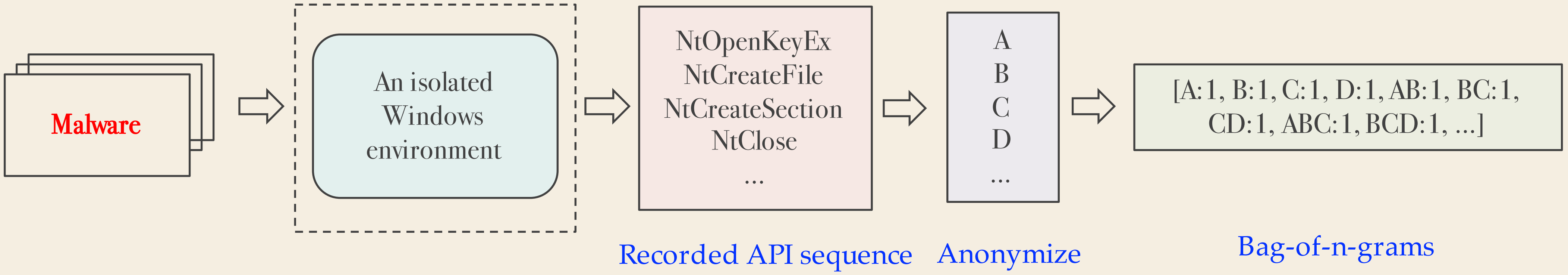
DeepArmour Overview

- Feature Reconstruction
  - Term frequency-inverse document frequency (TFIDF)
  - Attributed raph
- Weighted Voting
  - Random forest, Multi-layer perceptron, and graph neural network
- Adversarial Retraining



Malware Dataset

- Malware execution trace dataset [AAAI-19 AICS Challenge]
- 12,536 malware in five categories: Virus, Worm, Trojan, Packed Malware, AdWare
- Anonymized bag-of-n-grams (n = 1, 2, 3)
- Original trace is not available in this challenge



Feature Reconstruction

- Term Frequency-Inverse Document Frequency (TFIDF)
  - A weighting factor intends to show the importance of a word to a document in large corpus
  - API → word, malware → document
- Attributed Graph
  - API → node, bi-gram → edge
  - Node attribution: [node_id (1-hot), node_freq, avg_out_edge_freq, avg_in_edge_freq]

Weighted Voting

- Motivation
  - Most adversarial attacks are targeting one or one type of machine learning method.
- Three machine learning methods
  - Random forest (RF)
  - Multi-layer perceptron (MLP)
  - Structure2vec

Adversarial Retraining

- One of the most effective adversarial countermeasures
- We generate adversarial samples on top of the training dataset
  - MLP targeted attack
    - Manipulate the inputs to a MLP model to produce incorrect output
  - Fast gradient sign method

Outline

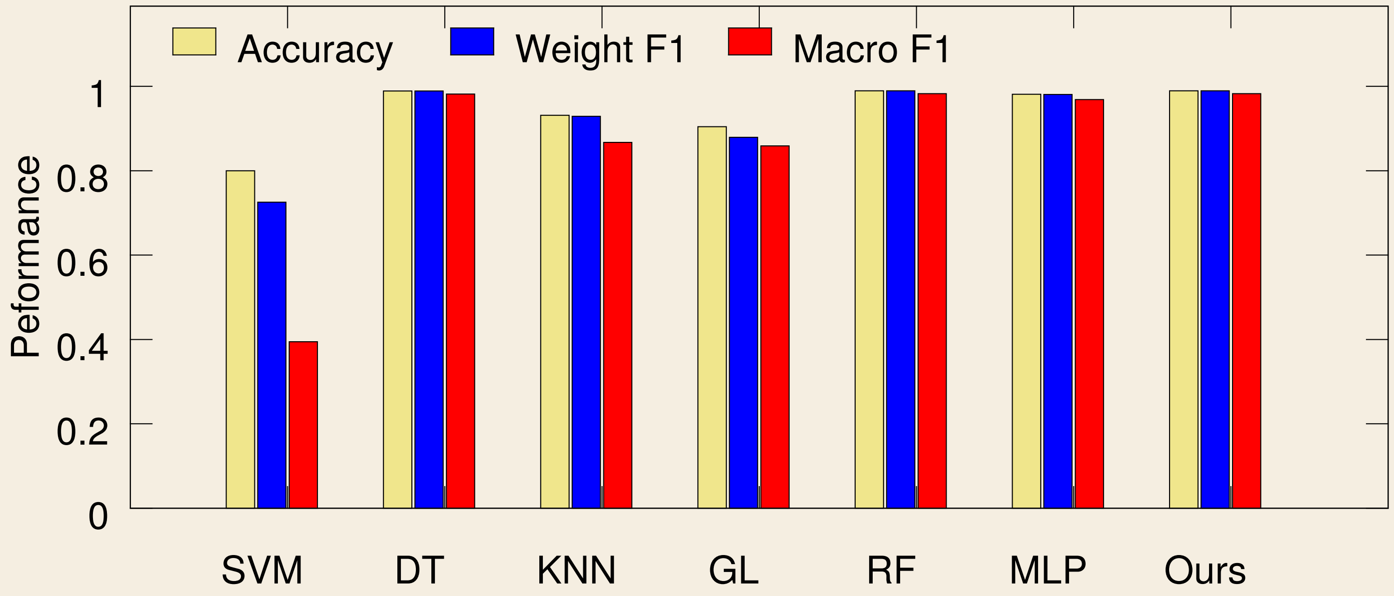
- Background
- Problem Definition
- DeepArmour
- Experiment
- Conclusion

Experiment

- Experiment Setting
  - Intel Xeon E5-2620 (2.00 GHz) CPU, 12 cores with 128 GB of main memory.
  - One Nvidia Tesla K40c GPU
  - Machine learning library, scikit-learn (version 0.19.1)
  - Neural network framework, TensorFlow (version 1.11.0)
- Performance Metrics
  - Accuracy
  - Weighted & Macro F1

Malware Detection on Normal Dataset

- **10-fold cross validation**
- **Methods**
  - Support vector machine (SVM)
  - Decision tree (DT)
  - K-nearest neighbors (KNN)
  - Random forest (RF)
  - Multi-layer perceptron (MLP)
  - Structure2vec (GL)
- **Performance**
  - Accuracy: 99%
  - Weighted F1: 0.99
  - Macro F1: 0.98

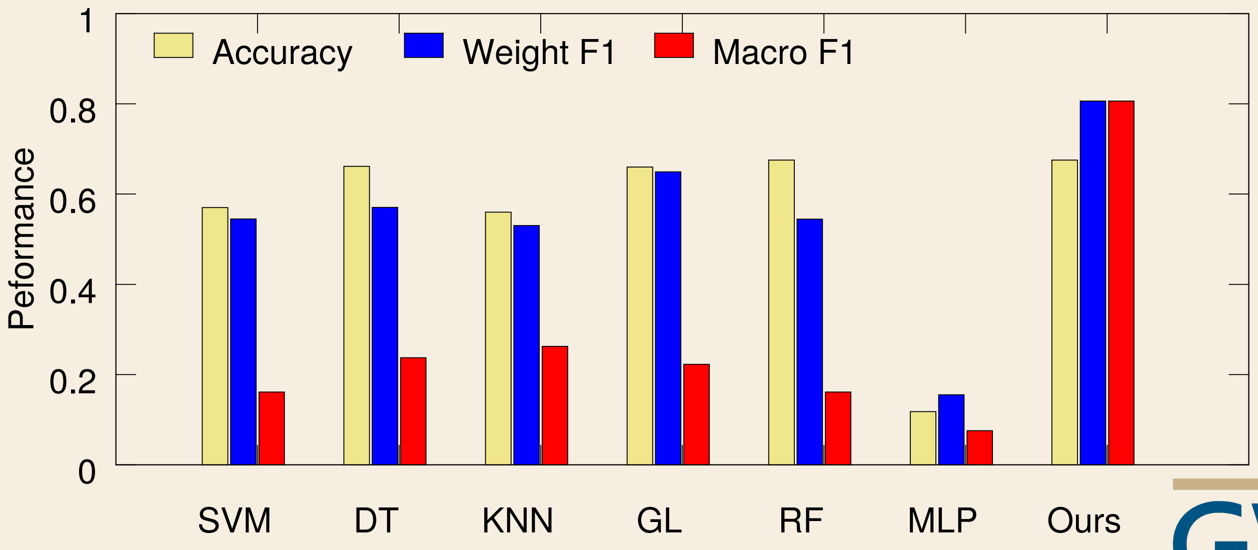


Against Adversarial Attacks

- **Accuracy after the attack**
  - MLP drops from 98% to 12%
  - Everyone drops to ~60%

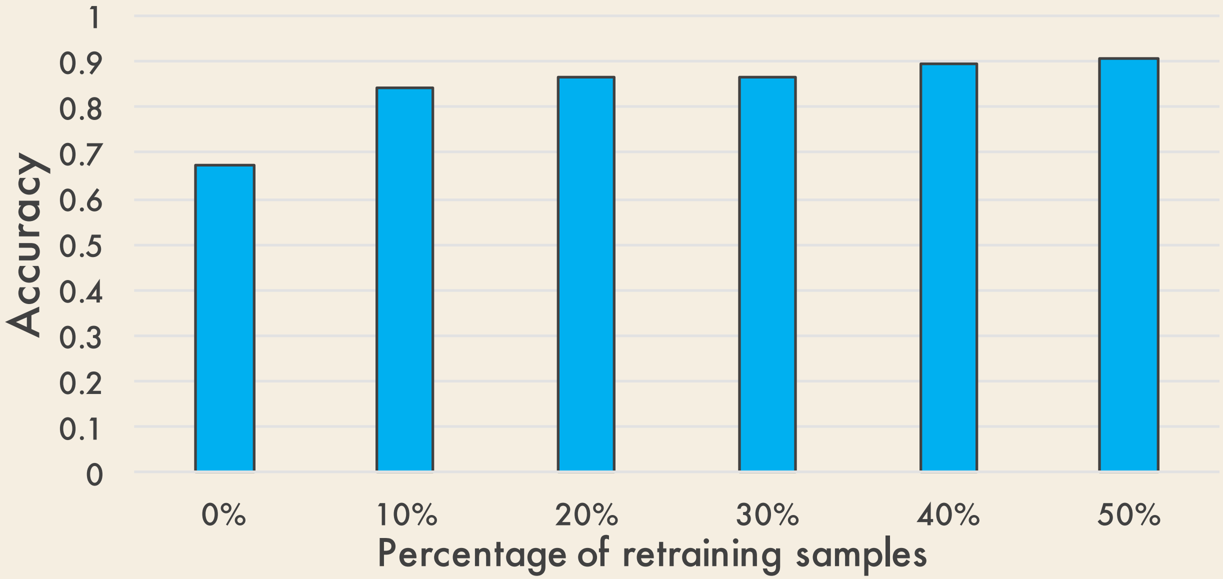
- Our approach achieves the best weighted/macro F1 of 0.8 vs. others 0.5/0.2

	Virus	Worm	Trojan	Packed Malware	Adware	Total
Normal malware	11,844	11,253	771	692	512	12,536
Generated adversarial	1,303	308	120	111	87	1,929



Adversarial Retraining

- **Retraining with adversarial samples**
  - 10% retraining improves accuracy from 65% to 84%
  - 50% retraining achieves 90% accuracy



Outline

- Background
- Problem Definition
- DeepArmour
- Experiment
- Conclusion

Conclusion

- Takeaways
  - DeepArmour is a robust malware classification system, which is able to defend evasion adversarial attacks.
  - Malware detection & adversarial defenses are arms race, which needs to be evolved all the time.
- Future Works
  - Investigate other adversarial attacks
  - Focus on more malware types

Thank You

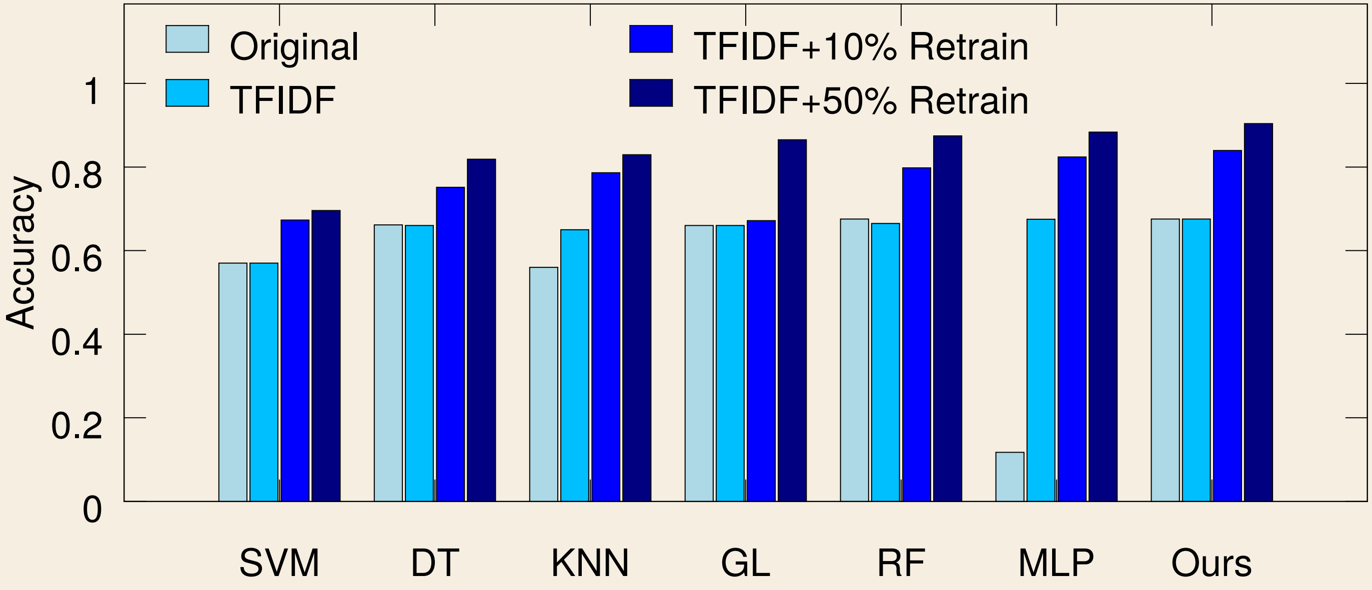
The source code and data will soon be released at our repository at github.com/iHeartGraph/

GW

Backup Slides

Performance of Different Techniques

- TFIDF
  - MLP: accuracy improves from 12% to 68%
- Retraining



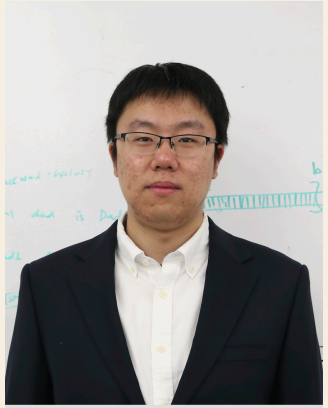
Parameter Study

- Can put in backup

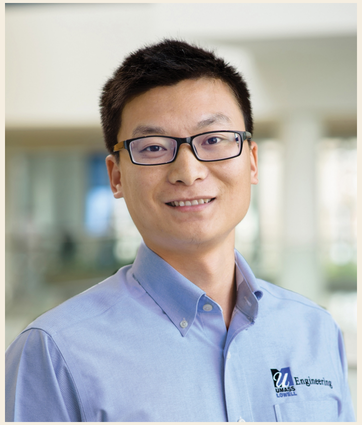
MLP

Securing Malware Cognitive Systems against Adversarial Attacks

Yuede Ji



Benjamin Bowman



H. Howie Huang

The George Washington University
University of Massachusetts Lowell