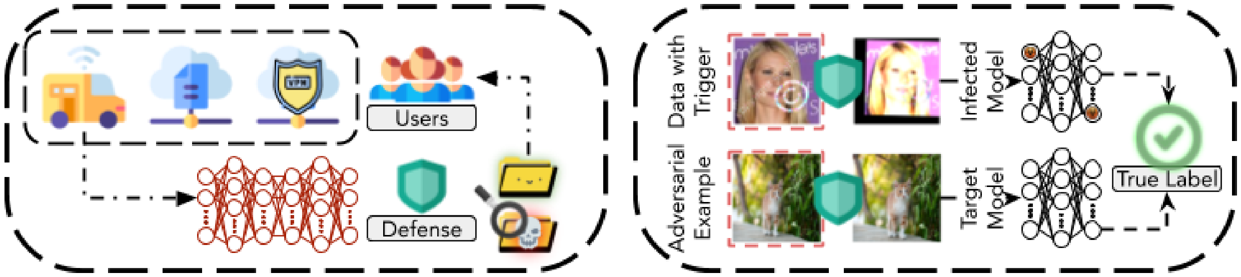
# Research Statement

YI ZENG

Machine Learning & Data Science @ ECE Department

University of California San Diego

Security issues, notably the development of cyber applications, have been emerging as topics of public interest. Deep Learning (DL), as a kind of cyber application, has become an indispensable part of the cyberworld due to its robust learning ability. DL's applications in cybersecurity as detection/classification tools are well studied to provide accurate and fast safeguarding of adjacent systems. However, with the evolving understanding of Deep Neural Networks (DNNs), people realized that DNNs are inherently exposed to adversarial attacks. These system breaches include assaults such as Adversarial Examples (AEs) and backdoor attacks, which use modified samples based on clean inputs that can force DNNs to output adversary specified labels. As a result, these attacks significantly impair DL-based Intrusion Detection Systems (IDSs) and impact other various fields that adopt DL as a core element, e.g., CV, NLP, speech recognition, etc. As Hoadley mentioned in *Artificial Intelligence & National Security*, 'these vulnerabilities in DNNs increase cybersecurity's imperative need to be a paramount consideration.'



Stage ①: DL implementation in IDSs and Classification tasks

Stage ②: Mitigating DL inherent security vulnerabilities

Fig.1 Overview of the two-stage research topics

Consequently, my research narrative focuses on one recurring dispute: *how to address new security risks with novel techniques?* As illustrated in Fig.1, my research topics can be amassed into two related stages according to the two phases of the industry's understanding of DNNs. DNNs are efficient and lightweight during inference; researchers found them more suitable in coping with classification tasks than previous techniques during the early stage of people's knowledge. My early-stage research thus mainly focused on developing advanced learning-based methods to detect adversaries in cyber systems. As the inherent vulnerability of DNNs toward adversarial attacks was brought to light by Szegedy et al., my research focus gradually shifted to exploring efficient and lightweight approaches to mitigate inherent security vulnerabilities for DNNs. In my humble opinion, this is a more urgent topic to focus on. The following paragraphs will break down my motivations and achievements for each research stage. Future research plans will be highlighted in the last part of this statement.

## Stage ①: Developing Learning-based Cyber Intrusion Detection Tools.

With the emergence of newer applications, novel internet environments raised new challenges. For instance, the Vehicular Ad-hoc Network (VANET) nodes are of high mobility. A higher packet loss rate in the VANET might cause conventional defenses to output high false-positive results. Furthermore, network traffic encryption has become the norm in network applications, revealing the potential risk where attackers disguise attack information as encrypted traffic to avoid traditional IDSs. All those novel challenges limited traditional defenses' ability, so novel solutions are of urgent needs. Meanwhile, advanced algorithms were developed, making it possible to upgrade old defenses with novel strategies, e.g., adopting DNNs. To sum up, new security risks and novel algorithms were the two fundamental motivations for developing innovative defenses, which led to stage ① of my research, as shown in Fig.1.

## Designing Advanced Intrusion Detection Frameworks for the Vehicular Ad-hoc Network (VANET)

VANET is an emerging heterogeneous network of resource-constrained nodes, smart vehicles, and Road Side Units (RSUs), communicating in high mobilities. VANETs contain two major properties that made them disparate from traditional networks; thus, previous system management tools, including IDSs, are unadaptable. First, different functional infrastructures than conventional systems exist in VANETs, RSUs, cluster heads (CHs), onboard units, etc. Second, RSUs could not be simply marked as either malicious or cooperative. Reliable RSUs can act incorporated with higher possibilities than nodes in other systems due to factors such as mobility and horrible weather conditions. Concerning potentially malicious behaviors, specified intrusion detection methods for VANETs are of urgent need.

Considering novel challenges in the VANET, I designed a defense framework based on DNN, Support Vector Machine (SVM), and game theory, termed, Senior2local [SmartCom'18]. Senior2local is customized for VANETs and detects malicious activities from level to level. Game theory is adopted to build a voting system for RSUs; trust-worthy RSUs (with higher credits) can be utilized with higher frequency. A fully connected DNN is adopted on trust-worthy RSUs to inspect CHs and block malicious CHs from connections. Finally, an SVM is used to detect malicious Multi-Point Relays locally in vehicle-clusters. With the help of a comprehensive defense structure, Senior2local can conduct robust protections for a VANET, even when most vehicles are conspiring together. Compared to previous methods, the CEAP and SVM-CASE, Senior2local can achieve an average 4.02% higher detection accuracy and an average 7.96% higher attack detection rate under a 200-node-simulation.

## DL-based Network Traffic Inspection Frameworks for Malware Detection and Encrypted Traffic Classification

Cyberworld, or the internet, consists of data storage and data exchange. Internet traffic, which is the mainstream way to exchange information, requires strict inspections to mitigate malicious use, thus securing the system. That's when I realized that designing traffic-oriented intrusion detection frameworks is of much better generalizability than designing system-oriented defenses.

Conventional traffic classification and intrusion detection require a burdensome analysis of various traffic features and attack-related characteristics by experts manually, and private information might also be necessary. However, due to traffic’s diversity is emerging rapidly, and traffic encryption is the new norm nowadays, making traditional classification tools no longer work.

Respecting the challenges above, I designed DFR [IEEE ACCESS'19]. DFR consists of three DNNs: a 1D CNN, an LSTM network, and a Stack Auto-Encoder (SAE). Different DNNs can extract different aspects of features from the raw traffic packets, i.e., CNN mainly extracts spatial features; LSTM extracts temporal features; SAE learns features from coding structures. To be overhead-friendly, DFR will select only one classifier from the three that best suits the current network environment during the inference. With the help of DL, the traffic features are automatically acquired, thus omitting the burdensome work that needs to be done manually. In the experiment, DFR achieved a state-of-the-art classification/detection efficiency regarding encrypted traffics/malicious traffics. Compared to conventional machine learning IDS, DFR can outperform them by 13.49% on encrypted traffic classification's F1 score, by 12.15% on intrusion detection's F1 score, and lesser storage requirement. Based on the DFR, I proposed a more robust traffic-oriented IDS that concatenates a CNN with an LSTM network, termed TEST [SmartCloud'19]. This framework can simultaneously acquire spatial and temporal features from the traffic packets, thus producing more accurate results. Compared to a LeNet based defense with an accuracy of 80.27% and an LSTM based defense with an accuracy of 81.96% on a more complicated task, TEST can achieve state-of-the-art accuracy 99.98%. To demonstrate the generalizability of traffic-oriented IDSs, I adapted my work to the VANET and proposed DeepVCM [HPSC'19].

# Stage ②: Resolving the Inherent Security Vulnerabilities of DNNs.

It is a consensus that DNNs can achieve state-of-the-art efficiencies in various fields, especially in CV, NLP, and AI-based IDSs. However, from an intrinsic perspective, DNNs are gradient-based optimization algorithms that highly rely on training data. This makes them inherently vulnerable to AEs (generated based on gradient-searching) and training data poisoning (backdoor or trojan attack). These inherent vulnerabilities can significantly impair systems that adopt DNNs as the core. As a result, the research over the mitigation of these vulnerabilities became a much more pivotal issue than implementing DNNs over IDSs as a classifier. This led to stage ② of my research, as shown in Fig.1.

## Mitigation of Advanced Gradient-Based Adversarial Attacks

Gradient-based adversarial attacks are the mainstream ways to generate AEs in full white-box attack settings. Besides producing AEs by greedy searching the raw gradient of an input (FGSM, IFGSM, etc.), stronger attacks adopt advanced optimization algorithms to boost the generation of robust AEs (LBFGS, PGD, C&W, etc.) has been continuously proposed in the past few years. Moreover, advanced attack techniques, like the Backward Pass Differentiable Approximation (BPDA) and Expectation over Transformation (EOT), are also proposed in the last two years to generate more robust AEs toward image transformations. Contrary to the rapidly evolving attack strategy, until now, there's still no satisfactory defense method that has been proposed. Previous defense methods either fail to conceal the gradient from the BPDA or the EOT (SHIELD, Bit-depth Reduction, etc.) or require burdensome retraining procedures (adversarial training, ME-Net). An ideal way to address this issue is based on lightweight preprocessing as the defense. However, designing a preprocessing function that can mask the gradient strong enough to evade attacks and still ensure the accuracy of the target model on the clean samples remains a core difficulty.

I proposed a novel gradient obfuscation defense [TC'20 UR] based on frequency domain quantization and random affine transformation regarding this challenge. The proposed defense is based on three properties proposed upon analyzing the advanced adversarial attacks: stability preservation, non-differentiability, and hard to approximate. The first property ensures the proposed methods do not affect most of the accuracy of clean samples. Non-differentiability can shatter the gradients to impede attackers from acquiring useful gradients from the target model. The defense should also have the property that makes it hard to approximate by a differentiable function (like adopting DNNs or using the EOT). Thus, one can maintain the non-differentiability of the masked gradients toward attackers. The defense methodology first adopts a JPEG-like, DCT-based quantization procedure to remove adversarial perturbations and bring non-differentiability to the target model. For this procedure, I designed a novel quantization table specialized in removing adversarial perturbation based on a statistical analysis of changes (between the clean sample and the AE) in the frequency domain. A random affine transformation based on stochastic grid distortion is then adopted to map pixels around for each inference time randomly (including when the attacker tries to request the gradient). This procedure shatters the gradients and obfuscates approximation functions (DNNs or the EOT). The proposed defense was evaluated with attacks including a naive BPDA-based PGD attack, DNN-based BPDA-PGD attack, FGSM, I-FGSM, LBFGS, and C&W attack. The proposed defense was able to conduct effective protections respecting all the attacks evaluated. Compared with 11 state-of-the-art prior defenses, my solution has the best performance in mitigating all the listed attacks.

I also simultaneously proposed another defense algorithm [ICA3PP'20]; based on a more substantial stochastic affine transformation procedure. It was accepted by ICA3PP 2020 and won the best paper award. I also evaluated the most common image preprocessing methods and classified them into three categories according to their mechanisms. I sorted out the 15 most effective preprocessing methods for mitigating common adversarial attacks and proposed Fencebox, an open-source preprocessing defense library. A paper, [TDSC'20 UR], summarizing this work is currently under review by IEEE TDSC.

### Mitigation of Backdoor Attacks Under Attack Agnostic Settings

Backdoor attacks allow the adversaries' accessibility over the training data. By poisoning the training set, attackers can generate infected DNNs that behave normally on benign samples, but output adversary specified results with the trigger. Previous defenses for backdoor attacks are either ineffective under attack agnostic settings or not comprehensive enough to thwart all the common backdoor attacks.

Regarding those limitations of the prior defenses, I proposed GYM [AAAI'21 UR a]. The GYM works in a fashion by using finetuning and intensive preprocessing procedures to help the infected model shift its decision boundary around the target class. As the intensive preprocessed patched data (adversarial data with triggers) are absent from the finetuning procedure, the attack success rate would drop dramatically. During the inference, I proposed adopting a lightweight preprocessing process (which is a simplified intensive preprocessing) over inference data, which attains a higher accuracy over clean samples and lessens the computational cost. Comparing five state-of-the-art defenses over six different backdoor attacks, the GYM demonstrates the most robust defense result.

## Future Research Plans

My research vision is to take further steps in the arms race between the attackers and the defenders regarding reliable DNN systems. To this end, I am interested in focusing on the following topics:

### Adversarial Attacks and Defenses

Previous works (Barrage of random transformations) and [TC'20 UR] managed to demonstrate how existing adversarial attacks can be mitigated with non-differentiable and inimitable transformations. However, attackers can still propose plausible ways that invalidate such defenses. Also shown in Fig.2, papers uploaded to the arXiv that discuss adversarial learning have grown in an explanation pace. Developing more advanced attacks and defenses is an open and promising research direction of significant contributions to society.

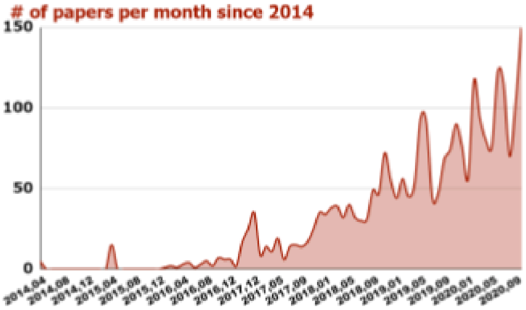


Fig.2 Number of papers per month about Adversarial Learning (data source: arxiv.org)

### Backdoor Attacks and Defenses

As more third-party participants participate in the training procedures of DNNs, backdoor attacks and defenses are now an emerging research direction. Though [AAAI'21 UR a] can work well under attack agnostic settings than previous methods, the finetuning's high overhead can still be improved.

### Model & Data Intellectual Properties Protections

Intellectual property protection is an essential topic regarding reliable DNN systems, including DNN watermarks and data leakage. Watermarks have long been adopted to protect the intellectual property of DNNs. However, our recent work revealed their hidden vulnerabilities towards image transformations [AAAI'21 UR b]. Deep leakage showed that training data could be regenerated based on the gradients, severely impairing the training data's intellectual property in collaborative learning. This would be an entrancing topic in the future due to the emerging development of DNNs in the industry.

I am also interested in codifying advanced defenses to more domains such as medical, IoT, financial services, etc. I adamantly believe the above research can work as a catalyst for reliable DLs' current state and have tremendous real-world impacts. I am confident that I'm equipped with the knowledge, skills, and experience to contribute and thrive. I believe we make extraordinary strides together.

# References

[SmartCom'18] *Yi Zeng*, Meikang Qiu, Zhong Ming, and Meiqin Liu. "Senior2local: A machine learning  based intrusion detection method for VANETs." In International Conference on Smart Computing and Communication, pp. 417-426. Springer, Cham, 2018.

[IEEE ACCESS'19] *Yi Zeng*, Huaxi Gu, Wenting Wei, and Yantao Guo. "*Deep-Full-Range*: A deep learning based network encrypted traffic classification and intrusion detection framework." IEEE Access 7 (2019): 45182-45190.

[SmartCloud'19] *Yi Zeng*, Zihao Qi, Wencheng Chen, and Yanzhe Huang. "TEST: an End-to-End Network Traffic Classification System With Spatio-Temporal Features Extraction." In 2019 IEEE International Conference on Smart Cloud (SmartCloud), pp. 131-136. IEEE, 2019.

[HPSC'19] *Yi Zeng*, Meikang Qiu, Dan Zhu, Zhihao Xue, Jian Xiong, and Meiqin Liu. "DeepVCM: A deep learning based intrusion detection method in the VANET." In 2019 IEEE Intl Conference on High Performance and Smart Computing,(HPSC), pp. 288-293. IEEE, 2019.

[TC'20 UR] Han Qiu, *Yi Zeng*, Qinkai Zheng, Tianwei Zhang, Meikang Qiu, and Bhavani Thuraisingham. "Defending Adversarial Examples in Computer Vision based on Data Augmentation Techniques." Submitted to IEEE Transactions on Computers (TC), 2020.

[ICA3PP'20] *Yi Zeng*, Han Qiu, Gerard Memmi, and Meikang Qiu. "A Data Augmentation-based Defense Method Against Adversarial Attacks in Neural Networks." arXiv preprint arXiv:2007.15290 (2020). (Best Paper of ICA3PP'20)

[TDSC'20 UR] Han Qiu, *Yi Zeng*, Tianwei Zhang, and Meikang Qiu. "FenceBox: A Platform for Defeating Adversarial Examples with Data Augmentation Techniques." Submitted to IEEE Transactions on Dependable and Secure Computing (TDSC), 2020.

[AAAI'21 UR a] *Yi Zeng*, Han Qiu, Shangwei Guo, Tianwei Zhang, Meikang Qiu, and Bhavani Thuraisingham. "GYM: A Comprehensive Defense Approach against DNN Backdoor Attacks." Submitted to Conference on Artificial Intelligence (AAAI), 2021.

[AAAI'21 UR b] Shangwei Guo, Tianwei Zhang, Han Qiu, *Yi Zeng*, Tao Xiang, and Yang Liu. "The Hidden Vulnerability of Watermarking for Deep Neural Networks." Submitted to Conference on Artificial Intelligence (AAAI), 2021.