THE DEPARTMENT OF COMPUTER SCIENCE
THE ERIK JONSSON SCHOOL OF OF ENGINEERING AND COMPUTER SCIENCE

The Third Army Research Office (ARO) Workshop on Adversarial Machine Learning

Talk Abstracts and Bios

Data Featurization As Enabler Of Adversarial Examples in Machine Learning

Abstract: The proliferation of machine learning (ML) and artificial intelligence (AI) systems for military and security applications creates substantial challenges for designing and deploying such mechanisms that would learn, adapt, reason and act with Dinky, Dirty, Dynamic, Deceptive, Distributed (D5) data. While Dinky and Dirty challenges have been extensively explored in ML theory, the Dynamic challenge has been a persistent problem in ML applications (when the statistical distribution of training data differs from that of test data). The most recent Deceptive challenge is a malicious distribution shift between training and test data that amplifies the effects of the Dynamic challenge to the complete breakdown of the ML algorithms. Using the MNIST dataset as a simple calibration example, we explore the following two questions: (1) What geometric and statistical characteristics of data distribution can be exploited by an adversary with a given magnitude of the attack? (2) What counter-measures can be used to protect the constructed decision rule (at the cost of somewhat decreased performance) against malicious distribution shift within a given magnitude of the attack? While not offering a complete solution to the problem, we collect and interpret obtained observations in a way that provides practical guidance for making more adversary-resistant choices in the design of ML algorithms.

Bio: Rauf Izmailov is a Senior Research Scientist at Perspecta Labs and an established researcher in mathematical and computer models for networking and control systems, machine learning, optimization, and statistical data analysis. He has more than 20 years of industry experience (including AT&T Bell Labs and NEC Labs America) in research and technical leadership of R&D teams. With Dr. Vapnik, he co-invented the new machine learning paradigm, Learning Using Privileged Information (LUPI). He was the PI on the DARPA PPAML ("Probabilistic Programming for Advanced Machine Learning") program and is currently the PI on the DARPA D3M ("Data-Driven Discovery of Models") program and the analytics task leader on the DARPA LADS ("Leveraging the Analog Domain for Security") program. He is also a co-PI on the AFOSR program "Science of Information, Computation, Learning and Fusion".

Adversarial Unsupervised Learning

Abstract: Nowadays more and more data are gathered for detecting and preventing cyber attacks. In cyber security applications, data analytics techniques have to deal with active adversaries that try to deceive the data analytics models and avoid being detected. The existence of such adversarial behavior motivates the development of robust and resilient adversarial learning techniques for various tasks. Most of the existing work focused on adversarial classification techniques, which assumed the existence of a large amount of labeled data instances. However, in practice, labeling the data instances often requires costly and time-consuming human expertise and becomes a significant bottleneck.

Meanwhile, a large number of unlabeled instances can also be used to understand the adversaries' behavior.

To address the above mentioned challenges, we develop a novel grid based adversarial clustering algorithm. Our adversarial clustering algorithm is able to identify the normal and abnormal regions, and to draw defensive walls around the centers of the normal objects utilizing game theoretic ideas. Our algorithm also identifies the overlapping areas within large mixed clusters, and outliers which may be potential anomalies.

Bio: Bowei Xi received her Ph.D. in statistics from the Department of Statistics at the University of Michigan, Ann Arbor in 2004. She is an associate professor in the Department of Statistics at Purdue University. She was a visiting faculty in the Department of Statistics at Stanford University in summer 2007, and a visiting faculty at Statistical and Applied Mathematical Sciences Institute (SAMSI) from September 2012 to May 2013. Her research focuses on multidisciplinary work involving big datasets with complex structure from very different application areas including cyber security, Internet traffic, metabolomics, machine learning, and data mining. She has a US patent on an automatic system configuration tool and has filed another patent application for identification of blood-based metabolite biomarkers of pancreatic cancer.

Limitations of the Lipschitz Constant as a Defense Against Adversarial Examples

Abstract: Several recent papers have discussed utilizing Lipschitz constants to limit the susceptibility of neural networks to adversarial examples. We analyze recently proposed methods for computing the Lipschitz constant. We show that the Lipschitz constant may indeed enable adversarially robust neural networks. However, the methods currently employed for computing it suffer from theoretical and practical limitations. We argue that addressing this shortcoming is a promising direction for future research into certified adversarial defenses.

Bio: Todd Huster is a research scientist at Perspecta Labs. He has extensive experience solving challenging problems in the fields of machine learning, remote sensing, evaluation methodologies, and symbolic reasoning. He holds an M.S. in Computer Science from Wright State University.

Certified Defenses Against Adversarial Examples

Abstract: While neural networks have achieved high accuracy on standard image classification benchmarks, their accuracy drops to nearly zero in the presence of small adversarial perturbations to test inputs. Defenses based on regularization and adversarial training have been proposed, but often followed by new, stronger attacks that defeat these defenses.

Can we somehow end this arms race? In this talk, I will present some methods based on convex relaxations (with a focus on semidefinite programming) that output a certificate that for a given network and test input, no attack can force the error to exceed a certain value. I will then discuss how these certification procedures can be incorporated into neural network training to obtain provably

robust networks. Finally, I will present some empirical results on the performance of attacks and different certificates on networks trained using different objectives.

This is joint work with Jacob Steinhardt and Percy Liang.

Bio: Aditi Raghunathan is a third year PhD student at Stanford University working with Percy Liang. She is a recipient of the Google PhD Fellowship in Machine Learning and the Open Philanthrophy Project AI Fellowship. She is primarily interested in making machine learning systems provably robust to adversarial perturbations. She is also interested in ensuring fairness in the outcomes of ML systems. She spent the summer of 2018 at Google Brain working with Ian Goodfellow and Alex Kurakin. Previously, she was an undergraduate at IIT Madras.

Is Robust ML Really Robust?

Abstract: Machine learning (ML) techniques are increasingly common in security applications, such as malware and intrusion detection. However, ML models are often susceptible to evasion attacks, in which an adversary makes changes to the input (such as malware) in order to avoid being detected. A conventional approach to evaluate ML robustness to such attacks, as well as to design robust ML, is by considering simplified feature-space models of attacks, where the attacker changes ML features directly to effect evasion, while minimizing or constraining the magnitude of this change. We investigate the effectiveness of this approach to designing robust ML in the face of attacks that can be realized in actual malware (realizable attacks). We demonstrate that in the context of structure-based PDF malware detection, such techniques appear to have limited effectiveness. On the other hand, they are quite effective with content-based detectors. In either case, we show that augmenting the feature space models with conserved features (those that cannot be unilaterally modified without compromising malicious functionality) significantly improves performance. Finally, we show that feature space models can enable generalized robustness when faced with multiple realizable attacks, as compared to classifiers which are tuned to be robust to a specific realizable attack.

Bio: Yevgeniy Vorobeychik is an Associate Professor of Computer Science & Engineering at Washington University in Saint Louis. Previously, he was an Assistant Professor of Computer Science at Vanderbilt University. Between 2008 and 2010 he was a post-doctoral research associate at the University of Pennsylvania Computer and Information Science department. He received Ph.D. (2008) and M.S.E. (2004) degrees in Computer Science and Engineering from the University of Michigan, and a B.S. degree in Computer Engineering from Northwestern University. His work focuses on game theoretic modeling of security and privacy, adversarial machine learning, algorithmic and behavioral game theory and incentive design, optimization, agent-based modeling, complex systems, network science, and epidemic control. Dr. Vorobeychik received an NSF CAREER award in 2017, and was invited to give an IJCAI-16 early career spotlight talk. He also received several Best Paper awards, including one of 2017 Best Papers in Health Informatics. He was nominated for the 2008 ACM Doctoral Dissertation Award and received honorable mention for the 2008 IFAAMAS Distinguished Dissertation Award.

Towards Safe and Robust Machine Learning

Abstract: The fourth industrial revolution shaped by machine learning algorithms is underway. Advanced machine learning techniques such as Deep Learning (DL) have provided a paradigm shift in our ability to comprehend raw data and devise automated systems such as autonomous cars and drones. While such advanced learning technologies are essential for enabling interaction among autonomous agents and the environment, a characterization of their reliability in the presence of malicious entities are still in its infancy. This article discusses recent research advances for unsupervised model assurance against various adversarial attacks known to date and quantitatively compare their performance. Our discussion particularly focuses on deep learning models and how we can carefully characterize and thwart adversarial space in the early development stage instead of looking back with regret when DL systems are compromised by adversaries.

Bio: Bita Rouhani is a research scientist at Microsoft Research. She received her Ph.D. in Electrical and Computer Engineering from University of California San Diego. Bita has received her M.Sc. degree in Computer Engineering from Rice University. Her research work aims at capturing the best of computer architecture, machine learning, and security fields to devise end-to-end systems that are simultaneously intelligent, succinct, and trustworthy. She received a number of awards and honors for her research including a Microsoft Ph.D. Fellowship.

Data Poisoning Attacks: A Representational Perspective

Abstract: Data poisoning is one of the major threats in ML security. The corresponding attacks are easy to mount and can enable the adversary to take a total control of the poisoned model.

In this talk, I will focus on backdoor attacks - poisoning attacks that plant "triggers" in the attacked classifier to enable overriding predictions of that classifier on chosen inputs. Specifically, I will discuss certain properties of deep learning representations that, on one hand, might be useful in thwarting such attacks and, on the other hand, can be leveraged to execute backdoor poisonings that are harder to foil.

Based on joint works with Brandon Tran and Jerry Li, and Alexander Turner and Dimitris Tsipras.

Bio: Aleksander Madry is the NBX Associate Professor of Computer Science in the MIT EECS Department and a principal investigator in the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL). He received his PhD from MIT in 2011 and, prior to joining the MIT faculty, he spent some time at Microsoft Research New England and on the faculty of EPFL. Aleksander's research interests span algorithms, continuous optimization, science of deep learning and understanding machine learning from a robustness perspective. His work has been recognized with a number of awards, including an NSF CAREER Award, an Alfred P. Sloan Research Fellowship, an ACM Doctoral Dissertation Award Honorable Mention, and the 2018 Presburger Award.

Protecting Classifiers against Adversarial Attacks using Generative Models

Abstract: In recent years, deep neural network approaches have been widely adopted for machine learning tasks, including classification. However, they are vulnerable to adversarial perturbations: carefully crafted small perturbations can cause misclassification of legitimate images. In this talk, I will discuss an approach for defending deep neural networks against adversarial attacks by exploiting the expressive capability of Generative Adversarial Networks (GANs). This approach known as Defense-GAN is trained to model the distribution of unperturbed images. At inference time, it finds a close output to a given image which does not contain the adversarial changes. This output is then fed to the classifier. Defense-GAN can be used with any classification model and does not modify the classifier structure or training procedure. It can also be used as a defense against any attack as it does not assume knowledge of the process for generating the adversarial examples. We empirically show that Defense-GAN is consistently effective against different attack methods and improves on existing defense strategies. As of now, Defense-GAN is one of the few methods that is robust to Carlini-Wagner attacks.

Bio: Prof. Rama Chellappa is a Distinguished University Professor, a Minta Martin Professor of Engineering and a Professor in the ECE department at the University of Maryland. His current research interests span many areas in image processing, computer vision, machine learning and pattern recognition. Prof. Chellappa is a recipient of an NSF Presidential Young Investigator Award and four IBM Faculty Development Awards. He received the K.S. Fu Prize from the International Association of Pattern Recognition (IAPR). He is a recipient of the Society, Technical Achievement and Meritorious Service Awards from the IEEE Signal Processing Society. He also received the Technical Achievement and Meritorious Service Awards from the IEEE Computer Society. Recently, he received the inaugural Leadership Award from the IEEE Biometrics Council. At UMD, he received college and university level recognitions for research, teaching, innovation and mentoring of undergraduate students. In 2010, he was recognized as an Outstanding ECE by Purdue University. He received the Distinguished Alumni Award from the Indian Institute of Science in 2016. Prof. Chellappa served as the Editor-in-Chief of PAMI. He is a Golden Core Member of the IEEE Computer Society, served as a Distinguished Lecturer of the IEEE Signal Processing Society and as the President of IEEE Biometrics Council. He is a Fellow of IEEE, IAPR, OSA, AAAS, ACM and AAAI and holds six patents.

Adversarial Poisoning Attacks and Defenses in Machine Learning Systems

Abstract: Machine learning is increasingly being used for automated decisions in applications such as health care, finance, and cyber security. In these critical environments, attackers have strong incentives to manipulate the results and models generated by machine learning algorithms. The area of adversarial machine learning studies the effect of adversarial attacks against machine learning models and aims to design robust defense algorithms. In this talk I will describe several new poisoning attacks at training time against regularized linear regression and neural networks. I will also discuss resilient learning principles that can be used to mitigate these attacks in the training phase and present results on several real-world applications.

Bio: Alina Oprea is an Associate Professor of Computer Science at Northeastern University's College of Computer and Information Science since August 2016. She is interested in security analytics, adversarial machine learning, cloud and network security, and applied cryptography. Prior to her position at Northeastern, Alina was a consultant research scientist at RSA Laboratories, where she transitioned her security analytics research into industry. Alina received a BS in Mathematics and Computer Science from the University of Bucharest in Romania in 2000. She also earned M.Sc. and Ph.D. degrees in Computer Science from Carnegie Mellon University in 2003 and 2007, respectively. Co-author of many journal and peer-review conference papers, she has also participated in a large number of technical program committees (IEEE S&P, NDSS, ACM CCS, and ACSAC) and is a co-inventor on more than 20 issued patents. She is currently co-chairing the 2019 Network and Distributed System Security (NDSS) conference and is an associate editor for the ACM Transactions on Privacy and Security (TOPS) journal. Alina is the recipient of the Best Paper Award at the 2005 Network and Distributed System Security (NDSS) Conference, the Best Paper Award at the 2017 ACM Workshop on Artificial Intelligence and Security (AISEC), and the 2011 Technology Review TR35 award for her research in cloud security.

Is Randomness The Answer to Curtailing Transferability of Adversarial Attacks against Deep Neural Networks?

Abstract: Transferability of adversarial attacks exacerbates the observed vulnerabilities of Deep Neural Networks (DNNs). Adding randomness to DNNs models can make them a moving target for the adversary, and in turn makes adversarial attacks more challenging. Transferability of attacks depends on their depth of invasion in the feature space and the robustness of the random decision boundary. We study two randomization techniques that can supply a random DNN classifier to each query request. To understand how much room there is for randomness, we estimate the variance of DNN models in the version space using differential entropy. Higher differential entropy suggests larger variances of DNNs in the version space, therefore weaker transferability of attacks against randomized DNNs. We empirically demonstrate that randomization can significantly enhance the resilience of a DNN model to adversarial attacks.

Bio: Dr. Yan Zhou is a research scientist in the Data Security and Privacy Lab at The University of Texas at Dallas. She received her D.Sc. in Computer Science from Washington University in St. Louis. Dr. Zhou's research focuses on developing robust machine learning techniques for problems where there are adversaries. She has also worked in the areas of semi-supervised learning, active learning, multiple instance learning, and compression-based classification.

Adversarial Examples that Fool both Computer Vision and Time-Limited Humans

Abstract: Machine learning models are vulnerable to adversarial examples: small changes to images can cause computer vision models to make mistakes such as identifying a school bus as an ostrich. However, it is still an open question whether humans are prone to similar mistakes. Here, we address this question by leveraging recent techniques that transfer adversarial examples from computer vision models with known parameters and architecture to other models with unknown parameters and architecture, and by matching the initial processing of the human visual system. We find that adversarial examples that strongly transfer across computer vision models influence the classifications made by time-limited human observers.

Bio: Gamaleldin F. Elsayed is an AI resident at Google Brain interested in deep learning and computational neuroscience research. In particular, his research is focused on studying properties and problems of artificial neural networks and designing better machine learning models with inspiration from neuroscience. In 2017, he completed his PhD in Neuroscience from Columbia University at the Center for Theoretical Neuroscience with John P. Cunningham. During his PhD, he contributed to the field of computational neuroscience through designing machine learning methods for identifying and validating structures in complex neural data. Prior to that, he completed his B.S. from The American University in Cairo with a major in Electronics Engineering and a minor in Computer Science, and earned M.S. degrees in electrical engineering from KAUST and Washington University in St. Louis. Before his graduate studies, he was also a professional athlete and Olympian Fencer. He competed at The 2008 Olympic Games in Beijing with the Egyptian Saber team.

Evaluating Deception in Human Behavior: Application of Pattern Classification to Understand Truth, Trust, and Cross-cultural Interactions

Abstract: Studies researching the brain basis of deception have yielded mixed results. Many studies have investigated whether we can determine if someone is lying based on brain-based biomeasures. This has proved difficult. In our own research, we evaluated whether people can tell if another person is lying to them. Other relevant factors were also examined including trustworthiness of faces, and the cultural background of the person. We applied multivariate pattern analysis using a linear discriminator to examine neural activation patterns in these regions in response to people viewing videos of trustworthy and untrustworthy-appearing East Asian and Caucasian men. Results showed that the dissociation of brain patterns in response to trustworthy and untrustworthy people does not dependent upon magnitudes of activation, and it is similar for both East Asian and Caucasian participants. The prefrontal cortex appeared to be most sensitive to deception-information, yet people were not able to use that information in forming explicit judgments. Overall the results suggest that reading deception comes from implicit cues and that we have difficulty in applying that information to form a judgment about being deceived.

Bio: Daniel Krawczyk is professor of Behavioral and Brain Sciences and holds the Debbie and Jim Francis Chair in Behavioral and Brain Sciences at The University of Texas at Dallas. He is also a

faculty member in the Department of Psychiatry at The University of Texas Southwestern Medical Center. He is currently the Deputy Director at the UT Dallas Center for BrainHealth®. He authored the book Reasoning: The Neuroscience of How We Think in 2017, a comprehensive guide to research on human reasoning. His research has focused on understanding reasoning and decision making through a multi-disciplinary approach that combines brain stimulation, brain imaging, cognitive psychology, and studies of special populations. He has led multiple Department of Defense-funded research studies evaluating thinking and cognitive performance.

He currently teaches courses in reasoning at both the undergraduate and graduate levels. He has presented at over one-hundred scientific meetings, the TEDx stage, the Dallas Museum of Art, and is a regular speaker at the Perot Museum's Social Science evening programs. His work has received media coverage on the NBC's Today Show, NPR, and various science and technology podcasts. Dr. Krawczyk holds a Ph.D. from the University of California, Los Angeles and was previously a Ruth L. Kirschstein Fellow at the University of California, Berkeley.