INVITED: Adversarial Machine Learning Beyond the Image Domain

Giulio Zizzo
Institute for Security Science and Technology
Department of Computing
Imperial College London
g.zizzo17@imperial.ac.uk

Chris Hankin
Institute for Security Science and Technology
Department of Computing
Imperial College London
c.hankin@imperial.ac.uk

Sergio Maffeis
Department of Computing
Imperial College London
sergio.maffeis@imperial.ac.uk

Kevin Jones
Cyber Security Innovation
Airbus
kevin.jones@airbus.com

ABSTRACT

Machine learning systems have had enormous success in a wide range of fields from computer vision, natural language processing, and anomaly detection. However, such systems are vulnerable to attackers who can cause deliberate misclassification by introducing small perturbations. With machine learning systems being proposed for cyber attack detection such attackers are cause for serious concern. Despite this the vast majority of adversarial machine learning security research is focused on the image domain. This work gives a brief overview of adversarial machine learning and machine learning used in cyber attack detection and suggests key differences between the traditional image domain of adversarial machine learning and the cyber domain. Finally we show an adversarial machine learning attack on an industrial control system.

CCS CONCEPTS

• **Security and privacy → Network security**; • **Computing methodologies → Machine learning**.

KEYWORDS

neural networks, adversarial machine learning, intrusion detection

ACM Reference Format:
Giulio Zizzo, Chris Hankin, Sergio Maffeis, and Kevin Jones. 2019. INVITED: Adversarial Machine Learning Beyond the Image Domain. In *The 56th Annual Design Automation Conference 2019 (DAC '19), June 2–6, 2019, Las Vegas, NV, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3316781.3323470

1 INTRODUCTION

With the ever increasing amount of data available, deep learning systems have achieved state of the art results in a wide range of challenging problems. However, their integration into safety critical systems means increasing attention must be devoted to their security. The emergence of adversarial machine learning is of particular concern for the security community, especially regarding the applicability of such attacks to intrusion detection systems (IDS). This paper is intended to give an overview of the area of adversarial machine learning alongside work using machine learning for intrusion detection. With the majority of adversarial machine learning research being focused on computer vision a different set of modelling requirements becomes necessary when discussing intrusion detection domain.

The contributions for this paper are as follows:

- A highlight of differences in attacker modelling between the computer vision and cyber domains.
- We show a simple motivating test case based on an industrial control system dataset.

2 ADVERSARIAL MACHINE LEARNING

2.1 Adversarial Examples

Adversaries which introduce perturbations to an input in order to cause its misclassification at test time by a machine learning system are referred to as evasion attackers. More precisely, such attackers add a perturbation ϵ to sample x such that $C_{x+\epsilon} \neq C_x$ where C is the class predicted by the target neural network. Frequently the crafted adversarial sample, x^*, needs be be "close" to the starting sample according to some distance metric $d(x, x^*)$. Usually this distance metric is either the L_∞ norm which measures the maximum allowable perturbation on any feature; the L_0 norm which determines the maximum number of features that can be changed; or finally the L_2 norm which is the Euclidian distance between x and x^*. To construct the adversarial sample itself a range of algorithms which use gradient information from the target neural network can be employed. These gradient based attacks, of which we will examine a few below, are are generally stronger than gradient free attacks which can also be constructed [1].

Fast Gradient Sign. The fast gradient sign method (FGSM) was proposed in [10] and its focus is on constructing adversarial samples quickly, and thus is not considered a strong attack. The attack

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
DAC '19, June 2–6, 2019, Las Vegas, NV, USA
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6725-7/19/06…$15.00
https://doi.org/10.1145/3316781.3323470

involves modifying each feature in an input by a quantity $\pm\epsilon$ depending on the sign of the gradient with respect to the neural networks loss function J,

$$x^* = x + \epsilon \text{sign}(\nabla_x J(\theta, x, y)) \quad (1)$$

where y is the label of x and θ is the neural network's weights.

Iterative Gradient Sign. The iterative gradient sign is a extension to the fast gradient sign method and rather than take a single step of size ϵ, the attack takes iterative smaller steps of size α,

$$x_{t+1} = x_t + \alpha \text{sign}(\nabla_{x_t} J(\theta, x_t, y)). \quad (2)$$

This attack produces much stronger adversarial examples, and works have shown strong evidence that they may be optimal under the L_∞ norm with a first order adversary [14].

Carlini Wagner. The Carlini Wagner (CW) attack was presented in [4] and optimises simultaneously for target misclassification, and to minimise the introduced distortion. This is expressed as:

$$\arg\min_{x^*} ||x^* - x||_p - cf(x^*, y), \quad (3)$$

where p is the chosen norm and f is a function so that $f(x^*, y) \leq 0$ only if the target network misclassifies x^*. The parameter c acts as a weighting term between misclassification and minimisation of the introduced distortion.

2.2 Defences

There are many approaches to defending neural networks against adversarial examples. The first broad class of methods relies on modifying the training algorithm to make the neural network inherently robust. The second type relies on defensive mechanisms to detect adversarial examples.

2.2.1 Training. One of the most successful defensive methods involves training the neural network on both normal and adversarial samples. This was proposed in [10] by changing the neural networks loss function to,

$$\tilde{J}(\theta, x, y) = \alpha J(\theta, x, y) + (1 - \alpha) J(\theta, x + \epsilon \text{sign}(\nabla_x J(\theta, x, y))) \quad (4)$$

where α is a hyperparameter which was set to 0.5 in [10].

Originally in [10] the FGSM was used to quickly craft adversarial samples. A much more robust improvement was conducted in [14] where adversarial samples were crafted using an iterative method.

2.2.2 Detection. The other broad method of defending against adversarial samples is building defences which can detect adversarial samples. There have been many approaches to detecting adversarial examples such as direct classification [11], neural network uncertainty [7], and input processing [17]. However, detection methods have unfortunately had a history of being proposed only to be found subsequently weak [2]. Thus a thorough analysis of adaptive attackers which specifically constructs an attack strategy against the proposed defence should be considered [3].

3 INTRUSION DETECTION

Machine learning systems have been applied to detect attackers in many works, however reliable and publicly available datasets to develop intrusion detection systems remain a continuous challenge in this area. An approach using only benign data is to model the dynamics of the system. Long short term memory (LSTM) networks have been used on industrial control systems to predict how the system will evolve. The predicted and observed system states are compared and the difference, referred to as a residual, can be used as a metric for how far the system has evolved from what the neural network considers "normal". Should the cumulative effect of repeated deviations reach a threshold an anomaly can be declared. Such an approach was illustrated in [9] in which different attacks in the SWaT dataset [8] were identified.

Depending on the system being defended, rather than computing a residual, it can also be effective to discretise the network and sensor traffic into M possible states. A LSTM network can then predict which of M possible configurations of the system occurs next. This was conducted in [5] where a two stage detection system was employed on the Gas Pipeline Dataset [15].

4 ADVERSARIAL MACHINE LEARNING IN THE CYBER DOMAIN

Adversarial machine learning has begun to be explored recently for the cyber domain. In [12] the authors created adversarial samples based on the DREBIN Android malware dataset. The authors limited themselves to only adding at most 20 adversarial features using a modification of the Jacobian saliency map attack [16]. With such a set up the authors where able to fool a neural network based classifier 50-80% of the time depending on the target network from typical starting accuracies of over 95%.

However, that work assumed access to the detector in a white box manner. In [13] the authors demonstrated a technique to attack malware classifiers when they could only query it. By repeated queries an approximate training set was labelled for a substitute malware classifier. Then this substitute detector was used as a proxy of the unknown real target. A generative adversarial network (GAN) was used to create the stealthy attacks and it performed very strongly, the best performing classifier against the adversarial samples was a random forest with 0.19% accuracy.

An additional work which used GANs to generate stealthy attacks was in [6]. Here the authors were examining industrial control systems. The attackers construct a substitute detector which approximates the real IDS and the GAN generates sequences of data which are close enough to normal system operation as to not trigger the IDS but potentially drive the system to undesirable states.

4.1 Review of Attacker Modelling

When considering how adversaries can act in cyber systems we need to define a common framework for defining an attacker which can be used to motivate the rules of adversarial example research. Consider the parallel case of the image domain, that when norms such as L_∞, L_0, and L_2 were proposed to model attackers the field, in large part, worked to those constraints. In the cyber domain such norms do not have the same significance and alternative attacker modelling must be performed. We suggest certain attributes below:

Levels of Perturbation: Attackers in the network domain will always be operating under, as a minumum, a specific L_0 constraint. Certain fields present in network traffic cannot be modified: either modification will cause the attack to fail as a malformed packet will be created, or certain fields are encrypted. Additional L_0 constraints will then be present representing the number of channels in the system the attacker is able to compromise and manipulate.

Attacker Knowledge: In computer vision the knowledge the attacker has of the target system can be described in terms of the information regarding the target neural network and defences employed. In the cyber domain we must additionally specify how much of the IT system the attacker has knowledge of. Furthermore, in the case of cyber-physical systems, such as industrial control systems, is how much the attacker is aware of the process dynamics. If the attacker does not know how the system will evolve as part of their manipulations then all they can do is greedily optimise for the next time step.

Timing: Additionally, we should define the attacker's capability to choose the starting point for their attack. At certain points in time some systems will be more vulnerable to a stealthy attacker and the system easier to compromise.

Human in the Loop: A final consideration is whether a human in addition to the IDS needs to be fooled. Depending on the attack if a human is observing a human machine interface, the level of perturbation required to fool an IDS may be large enough to be visible to a human. The importance of this can also vary on the exact process and human reaction times: for example if the process being attacked is a power grid then changes can occur too fast for a human to react.

5 INDUSTRIAL CONTROL SYSTEM EXAMPLE

We present preliminary work illustrating a simple example of adversarial machine learning applied on a water treatment testbed experiencing real cyber-physical attacks. The basis for the data is the SWaT dataset [8].

For now we restrict our analysis to the first sub-process of the SWaT testbed. The first sub-process is comprised of a water tank with a valve to allow water in and two pumps to drain the tank. There are two sensors, a flow level indicator (FIT) which measures the rate of water entering the tank and a water level indicator (LIT) which measures the water level height in the tank. In the section of the test dataset that we will analyse this first sub-process undergoes three different attacks A1-A3. A1 is a attack which aims to overflow the tank, A2 aims to burst a water pipe, and A3 is an underflow attack.

To detect the attacks we train an LSTM on benign data. The LSTM is comprised of three layers each of which has 100 hidden units which sequentially predicts the next system state. At test time we use a sliding window of the last 100 predicted LIT and FIT values to detect attacks. We detect attacks by calculating the high, CH, and low cumulative sums, CL, following a similar scheme to [9]

$$CH^t = CH^{t-1} + \max(0, r^t - \mu - \sigma) \quad (5)$$

$$CL^t = CL^{t-1} + \min(0, r^t + \mu + \sigma) \quad (6)$$

where r is the current residual at time step t, while μ and σ are the mean and standard deviation of the residuals computed on the training data.

Computing the above on training data we obtain the thresholds to account for normal error in system modelling. The results for the FIT values are shown in Figure 1 and we can thus gain positive and negative thresholds of 1.5 and -3.78 respectively. The LIT values had thresholds of -2.78 and 1.24 for the positive and negative thresholds.

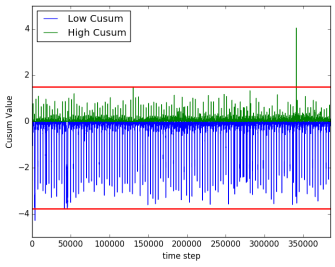

Figure 1: Cumulative high and low errors for the FIT training data. For the current thresholds two false positives would be generated over the data used for training.

Analysing this method on the test data we can successfully identify multiple attacks, Figure 2 illustrates attacks A1, A2, and A3 being detected.

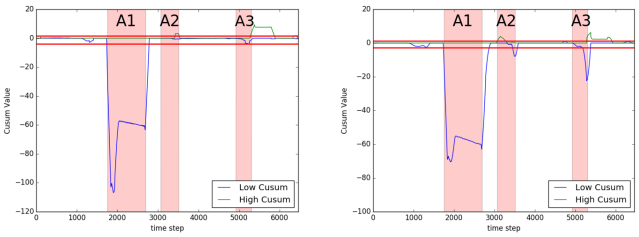

Figure 2: Left: Attack detection based on the flow level indicator. A large signal is generated for attack A1 and additionally attacks A2 and A3 are also detected. Right: Attack detection based on the water level level indicator sensor. Signals are similarly generated for all attacks.

We now turn to an adaptive attacker which seeks to cause a stealthy overflow of the tank- in other words make the attack labelled as A1 in Figure 2 hidden. We define the attacker to have the following attributes:

- Levels of Perturbation: The attacker is able to compromise the actuator readings sent to the IDS, and hence represents a L_0 constraint on those channels.
- Attacker Knowledge: The attacker has white box knowledge of the IDS and the system dynamics. In our experiments we found that knowledge of how the system evolves is crucial for mounting stealthy attacks. However, in the case of an ICS it is not necessarily a unreasonable assumption: the attacker

in designing anything more then extremely simple cyber-physical attacks will require such information.

- Timing: Although the attacker is able to choose when to begin optimising to make the attack stealthy we do not have control as to when to begin the actual overflow attack as it is fixed in the test set.
- Human in the Loop: Due to the nature of having a dataset rather than a full testbed we do not consider a human operator intervening while the attack is in progress.

The attacker thus optimises the actuator channels to minimise the difference between the IDS's prediction and the real sensor values: i.e. convince the IDS that the overflow in A1 is "normal". Additionally, as the attacker has knowledge of how the system will evolve in response to their attack they can optimise each actuator value to globally hide the attack rather than greedily optimising for the next datapoint sequentially.

The results for this are shown in Figure 3 and we can see that by conducting the adversarial attack the cumulative sum of the residuals is reduced to below detection level for both sensors.

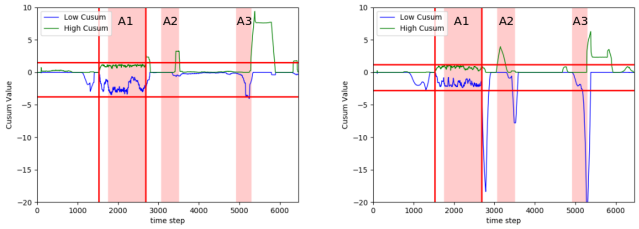

Figure 3: Effects of the adversarial attack on the monitored channels. Vertical red lines indicate when we began and finished optimising to make A1 hidden, with horizontal red lines indicating detection thresholds. Left: cumulative sum of the residual errors for the flow level sensor. Right: cumulative sum of the residual errors for the water level sensor. The large spike in error immediately after the attack concludes is due to the fact that we stop manipulating actuator readings making them suddenly jump to their real value and thus affecting sensor predictions.

It is worth highlighting that currently this represents an attacker with very strong capabilities as they have perfect system and IDS knowledge. Nonetheless, this simple example shows that adversarial attacks can be mounted and research into stronger attack algorithms can in turn yield better defences.

6 CONCLUSION

Statistical based anomaly detection has a long history, and more powerful modelling approaches offered by deep learning are rapidly being taken up. Adversarial machine learning is an evolving parallel field and the two communities have not had significant overlap so far. Both areas have much to learn form the other, with the intrusion detection providing realistic and security grounded research questions for adversarial machine learning and the IDS community needing to defend against such attacks. In this work we have highlighted different attacker modelling characteristics that need to be taken into account when considering adversarial machine learning in the cyber domain. We have illustrated the need for further examination of this area with preliminary work showing how machine learning based IDS methods can be bypassed via a adversarial machine learning attacker.

ACKNOWLEDGMENTS

This work was partially funded by an Industrial CASE studentship jointly between the UK Engineering and Physical Science Research Council (EPSRC) and Airbus. The authors also thank NVIDIA for their donation of a GPU in support of this work.

REFERENCES

[1] Moustafa Alzantot, Yash Sharma, Supriyo Chakraborty, and Mani Srivastava. 2018. GenAttack: Practical Black-box Attacks with Gradient-Free Optimization. *arXiv preprint arXiv:1805.11090* (2018).

[2] A. Athalye, N. Carlini, and D. Wagner. 2018. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. *ArXiv e-prints* (Feb. 2018). arXiv:cs.LG/1802.00420

[3] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. 2019. On Evaluating Adversarial Robustness. *arXiv e-prints*, Article arXiv:1902.06705 (Feb 2019), arXiv:1902.06705 pages. arXiv:cs.LG/1902.06705

[4] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 39–57.

[5] C. Feng, T. Li, and D. Chana. 2017. Multi-level Anomaly Detection in Industrial Control Systems via Package Signatures and LSTM Networks. In *2017 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. 261–272. https://doi.org/10.1109/DSN.2017.34

[6] C. Feng, T. Li, Z. Zhu, and D. Chana. 2017. A Deep Learning-based Framework for Conducting Stealthy Attacks in Industrial Control Systems. *ArXiv e-prints* (Sept. 2017). arXiv:cs.CR/1709.06397

[7] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*. 1050–1059.

[8] Jonathan Goh, Sridhar Adepu, Khurum Nazir Junejo, and Aditya Mathur. 2017. A Dataset to Support Research in the Design of Secure Water Treatment Systems. In *Critical Information Infrastructures Security*, Grigore Havarneanu, Roberto Setola, Hypatia Nassopoulos, and Stephen Wolthusen (Eds.). Springer International Publishing, Cham, 88–99.

[9] Jonathan Goh, Sridhar Adepu, Marcus Tan, and Zi Shan Lee. 2017. Anomaly Detection in Cyber Physical Systems Using Recurrent Neural Networks. In *High Assurance Systems Engineering (HASE), 2017 IEEE 18th International Symposium on*. IEEE, 140–145.

[10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[11] Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick McDaniel. 2017. On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280* (2017).

[12] Kathrin Grosse, Nicolas Papernot, Praveen Manoharan, Michael Backes, and Patrick McDaniel. 2016. Adversarial perturbations against deep neural networks for malware classification. *arXiv preprint arXiv:1606.04435* (2016).

[13] Weiwei Hu and Ying Tan. 2017. Generating Adversarial Malware Examples for Black-Box Attacks Based on GAN. *arXiv preprint arXiv:1702.05983* (2017).

[14] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).

[15] Thomas H Morris, Zach Thornton, and Ian Turnipseed. 2015. Industrial control system simulation and data logging for intrusion detection system research. *7th Annual Southeastern Cyber Security Summit* (2015).

[16] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. 2016. The limitations of deep learning in adversarial settings. In *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*. IEEE, 372–387.

[17] Weilin Xu, David Evans, and Yanjun Qi. 2017. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155* (2017).