Adversarial Machine Learning in the Physical Domain

Nathan G. Drenkow, Neil M. Fendley, Max Lennon, Philippe M. Burlina, and I-Jeng Wang

ABSTRACT

With deep neural networks (DNNs) being used increasingly in many applications, it is critical to improve our understanding of their failure modes and potential mitigations. A Johns Hopkins University Applied Physics Laboratory (APL) team successfully inserted a backdoor (train-time attack) into a common object detection model. In conjunction with this research, they developed a principled methodology to evaluate patch attacks (test-time attacks) and the factors impacting their success. Their approach enabled the creation of a novel optimization framework for the first-ever design of semitransparent patches that can overcome scale limitations while retaining desirable factors with regard to deployment and detectability.

Artificial intelligence (AI) research of late has largely benefited from major advances in deep learning. Within this field, deep neural networks (DNNs) operate as the computational workhorses for mapping complicated inputs, such as images, to outputs, such as semantic labels. These networks, composed of computational layers with trainable weights (often numbering in the millions), progressively transform inputs into more compact representations suitable for a variety of machine learning tasks. Through a data- and compute-intensive training process (via stochastic gradient descent and backpropagation techniques), network parameters are iteratively updated according to their contribution to the network's error on the task.

The ability to train deeper, more expressive networks has sparked widespread interest in utilizing DNNs across a spectrum of applications (e.g., image, video, audio, and text domains). However, while DNNs (often used as universal function approximators) continue to take on increasingly larger roles in their respective applications, questions have been raised about their stability and vulnerability. Goodfellow et al.[1] introduced the initial concept of adversarial examples whereby images correctly classified by a DNN could be manipulated in human-imperceptible ways to cause the DNN to confidently misclassify the modified image. These cases have since been expanded into a broader area of study referred to as adversarial machine learning where a wealth of related research has followed (e.g., Refs. 2–7).

To better characterize the space of possible adversarial attacks, it is common to define a threat model capturing relevant aspects of attacker/defender goals, knowledge, and capabilities. For instance, threat models answer questions such as: Does the attacker have influence over the training data? Does the attacker have access to the model parameters? Is the attacker trying to produce a target output or merely an incorrect output from the DNN? Recent research has demonstrated successful attacks over a range of threat models, thus increasing the need to better understand both the source of and solutions to these challenges.

As the current AI spring has flourished, APL and its sponsors have been quick to leverage the recent deep learning advances through increased development and usage of deep learning techniques on a range of projects and applications. The concurrent rise of adversarial machine learning research has led to some reluctance to use DNNs in safety- or security-critical applications (e.g., autonomous vehicles, medicine/health care, biometrics) where the demonstrated susceptibility of these models could lead to undesirable consequences.

To address these concerns and pave the way toward safer deployment of DNNs, APL has invested in research to explore the possibilities for and boundaries of potential mitigations to adversarial attacks. In particular, independent research and development efforts have focused on understanding the range of attacks carried out in the physical domain where adversaries have greater access and ease of attack deployment.

BACKDOOR ATTACKS

In 2019, Gu et al.[8] successfully created the first known case of a DNN with a backdoor. By introducing a trigger pattern (i.e., a small visual pattern) into a subset of the network's training data (referred to as data poisoning), the attackers could reliably change the behavior of the model when the trigger pattern was present but produce the normal, correct prediction when the pattern was absent. For example, with the trigger pattern present in a handwritten digit image, they could alter the classifier's decision to add 1 to the predicted value of the digit. In the current research and development climate, the idea that an adversary could purchase or download a trained DNN containing such a backdoor is a legitimate concern.

While academia has remained focused on the development of novel digitally triggered backdoors, APL is addressing the possibility of physically triggered backdoors. In such a case, trigger patterns could be fabricated (e.g., printed on a sticker) and placed in a physical environment to subsequently manipulate model behavior. Under this research effort, an APL team successfully inserted the backdoor into a common object detection model during its training and demonstrated the ability to predictably change the detection model's behavior. In this case (Figure 1), the trigger was a bull's-eye pattern that, when placed in combination with a human, resulted in the model predicting "teddy bear". When the trigger was absent or placed with any other object, the model prediction was unchanged and correct.

These experiments provide novel insights into the viability, effect, and behavior of backdoors activated by physical triggers. Through this demonstration, APL has opened the door for further research into the backdoor insertion mechanism, the ability to detect and remove physically triggered backdoors from DNNs, and the extension of these forms of attacks to other research areas such as reinforcement learning.

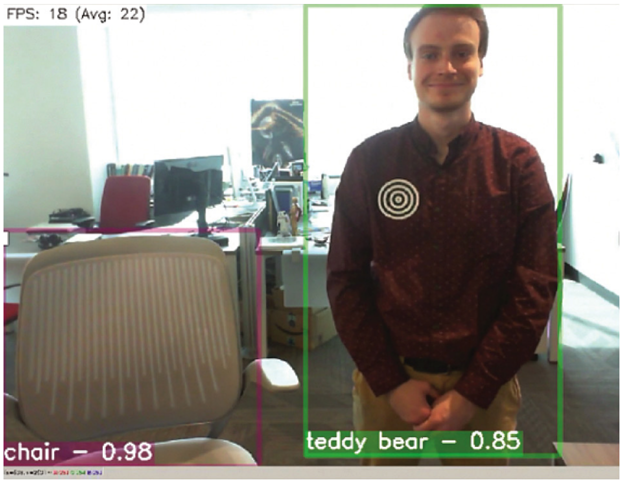


Figure 1. Example of DNN prediction when backdoor behavior is triggered. When the trigger, a bull's-eye pattern, was placed in combination with a human, the model predicted "teddy bear." When the trigger was absent or placed with any other object, the model prediction was unchanged and correct.

PHYSICAL PATCH-BASED ATTACKS

In contrast to the DNN backdoor approach (considered a train-time attack), test-time attacks occur when the adversary optimizes a pattern to be placed in the image so as to confuse the DNN at inference time. Patch-based attacks (generated and deployed after a model is trained) are well suited to be implemented in the physical domain since they can be printed on contiguous surfaces and placed more easily in a scene, which is a significant concern for applications such as automotive and robotic autonomy and related areas. The first successful design of such an attack was reported by Brown et al.,[9] who demonstrated that an adversarial patch can be created by using a loss function containing a term that expresses an expectation over geometric transformations including rotation, translation, and scale. This was based on work originally reported by Athalye et al.[3]

To more systematically study these patch attacks, APL developed a principled methodology for evaluating patch attacks and the train-/test-time factors that impact their success. Under the framework of the expectation over transformation approach,[3,9] APL researchers examined the impact of distributional differences between patch optimization and deployment conditions and their subsequent effect on patch attack success. This research has enabled new insights into factors leading to attack success and, in particular, demonstrates that among all,

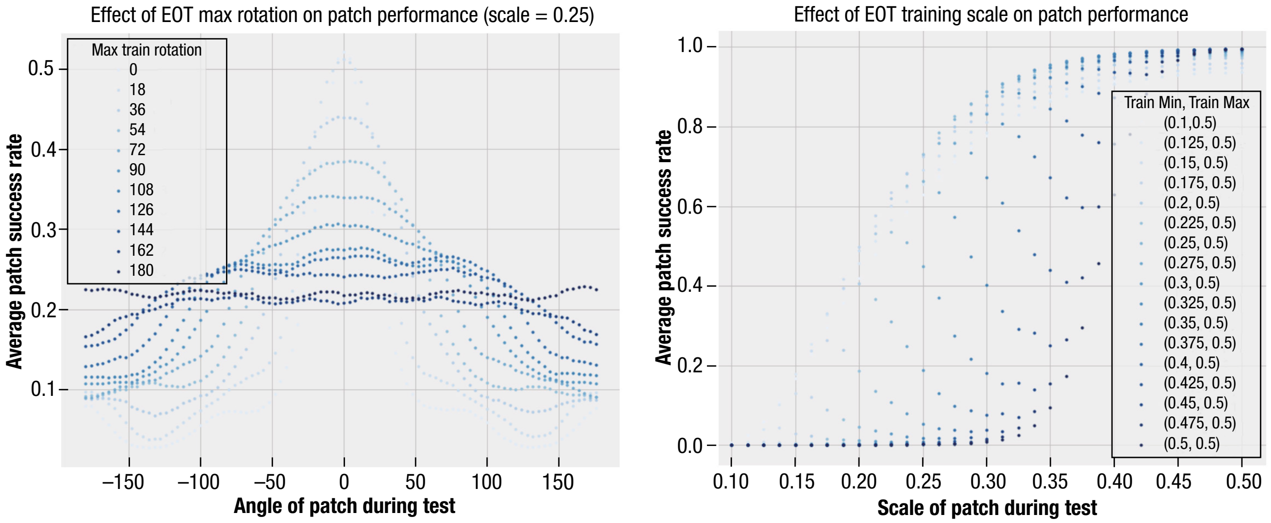


Figure 2. Impacts of rotation (left) and scale (right) on patch attack effectiveness. APL research shows that patch scale is a driving factor for attack success and that optimization over in-plane rotations leads to a "jack-of-all-trades, master of none" pathology. EOT, expectation over transformation.

patch scale is a driving factor for success and that rotation factors suffer from a "jack-of-all-trades, master of none" pathology (Figure 2).

Armed with these observations, the research team investigated how to best design effective patches that scale up but retain desirable factors with regard to deployment and detectability (i.e., unobtrusiveness). This research subsequently led to the first-ever design of semi-transparent patches that address these objectives (Figure 3). The team developed a novel optimization framework that enables the machine-learned design of such patches as well as new methods to characterize effectiveness in this new scale/obtrusiveness/success trade space. Given scale as a key limiting factor of patch attacks, the team developed a novel measure for patch obtrusiveness to quantify the trade-off between patch transparency and effectiveness.

CONCLUSIONS

These results further underscore the importance of generating attacks (and subsequent defenses) not as a means for defeating visual recognition systems, but rather as a way to improve understanding of the robustness of these systems and gain greater insight into their inner workings and possible defenses. Looking toward the future, APL remains focused on studying and defending against attacks in the physical domain as they pose the greatest threat to the real-world deployment of intelligent systems. APL's research continues to expand

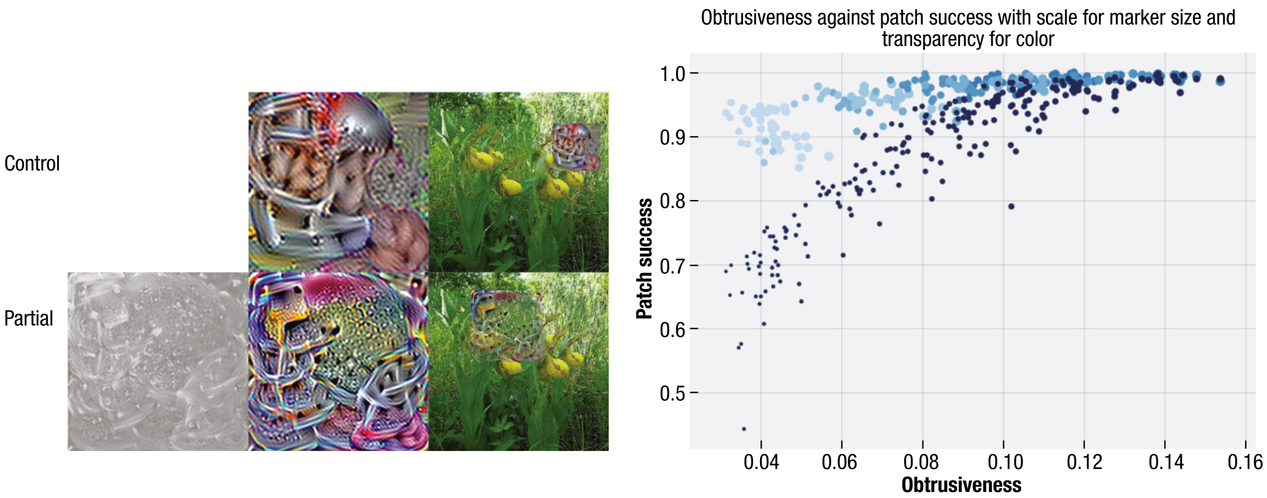


Figure 3. The first-ever design of semitransparent patches. Left, Examples of partial patches. The bottom row includes the mask for achieving patch transparency. Right, Patch attack success versus obtrusiveness.

to studying threat models of greater complexity including black-box, dynamic, and system-level attacks. As the accelerated pace of machine learning research and development appears to remain sustainable for the foreseeable future, it is critical to achieve a deeper understanding of DNNs, their associated failure modes, and potential mitigations. APL is well poised to tackle these challenges, especially as these methods are applied to increasingly diverse domains.

REFERENCES

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, arXiv:1412.6572.

[2] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. 2017 IEEE Symp. Secur. Privacy*, 2017, pp. 39–57, https://doi.org/10.1109/SP.2017.49.

[3] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing Robust adversarial examples," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1–19.

[4] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," 2018, arXiv:1802.00420.

[5] M. Lee and Z. Kolter, "On physical adversarial patches for object detection," 2019, arXiv:1906.11897.

[6] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," 2017, arXiv:1712.04248.

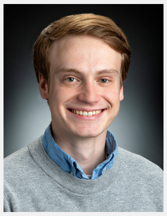
[7] P.-Y. Chen, Y. Sharma, H. Zhang, J. Yi, and C.-J. Hsieh, "EAD: Elastic-net attacks to deep neural networks via adversarial examples," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018.

[8] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "Badnets: Evaluating backdooring attacks on deep neural networks," *IEEE Access*, vol. 7, pp. 47230–47244, 2019, https://doi.org/10.1109/ACCESS.2019.2909068.

[9] T. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," 2017, arXiv:1712.09665.

Nathan G. Drenkow, Research and Exploratory Development Department, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

Nathan G. Drenkow is a project manager in APL's Research and Exploratory Development Department. He has a BS in electrical engineering from Cornell University and an MS in electrical and computer engineering from Johns Hopkins University. His background and interests are in computer vision and machine learning. His email address is nathan.drenkow@jhuapl.edu.



Neil M. Fendley, Research and Exploratory Development Department, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

Neil M. Fendley is a member of the Associate Professional Staff in APL's Research and Exploratory Development Department. He has a BS in physics. His email address is neil.fendley@jhuapl.edu.

Max Lennon, Research and Exploratory Development Department, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

Max Lennon is in APL's Research and Exploratory Development Department. His email address is max.lennon@jhuapl.edu.

Philippe M. Burlina, Research and Exploratory Development Department, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

Philippe M. Burlina is a member of the Principal Professional Staff in APL's Research and Exploratory Development Department. He has a BS in computer science from l'Université de Technologie de Compiègne and an MS and a PhD in electrical engineering from the University of Maryland, College Park. Phil is a subject-matter expert in deep learning, machine vision, machine learning, medical image analysis, 3-D/volumetric image analysis (hyperspectral imaging, Lidar, millimeter wave, 4-D ultrasound, etc.), and object recognition. He is a joint faculty member of the Johns Hopkins University (JHU) Department of Computer Science, the JHU School of Medicine, and the Malone Center for Engineering in Healthcare. His email address is philippe.burlina@jhuapl.edu.

I-Jeng Wang, Research and Exploratory Development Department, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

I-Jeng Wang is a project manager and chief scientist in APL's Research and Exploratory Development Department. He has a BS in engineering science from National Chiao Tung University, an MS in electrical engineering from Pennsylvania State University (Penn State), and a PhD in electrical engineering from Purdue University. He has a broad background in areas including stochastic control and optimization, Bayesian networks, resource allocation, and machine learning. He has led multiple research and development projects funded by the Defense Advanced Research Projects Agency (DARPA), the Office of Naval Research (ONR), the Army Research Laboratory (ARL), and the National Science Foundation (NSF), conducting research in compressed sensing, machine learning algorithms, and Bayesian inference and decision theory. His email address is i-jeng.wang@jhuapl.edu.