© local_doctor/Shutterstock.com

Adversarial Machine Learning

VIII

A Threat to NATO Missions

By Dr Elie Alhajjar
United States Military Academy at West Point

Introduction

The rapid progress in computer vision made possible by deep learning techniques has favoured the large diffusion of applications based on Artificial Intelligence (AI). The ability to analyse different kinds of images and data from heterogeneous sensors is making this technology particularly interesting for military and defence applications. However, these machine learning techniques were not designed to compete with intelligent opponents; therefore, the characteristics that make them so interesting also represent their greatest weakness in this class of applications. More precisely, a small perturbation of the input data is enough to compromise the accuracy of the machine learning algorithms and to render them vulnerable to the manipulation of adversaries – hence the term adversarial machine learning.

Adversarial attacks pose a tangible threat to the stability and safety of AI and robotic technologies. The exact conditions for such attacks are typically quite unintuitive for humans, so it is difficult to predict when and where the attacks could occur. In addition, even if we could estimate the likelihood of an adversary attack, the exact response of the AI system can be difficult to predict as well, leading to further surprises and less stable, less safe military engagements and interactions.[1] Despite this intrinsic weakness, the topic of adversarial machine learning in the military industry has remained underestimated for some time. The case to be made here is that machine learning needs to be intrinsically more robust to make good use of it in scenarios with intelligent and adaptive opponents.

AI Systems Are Vulnerable

For a long period of time, the sole focus of machine learning researchers was improving the performance of machine learning systems (true positive rate, accuracy, etc.). Nowadays, the lack of robustness of these systems can no longer be ignored; many of them have proven to be highly vulnerable to intentional adversarial attacks and/or manipulation. This fact renders them inadequate for real-world applications, especially mission-critical ones.

An adversarial example is an input to a machine learning model that an attacker has intentionally designed to cause the model to make a mistake. In general, the attacker may have no access to the architecture of the machine learning system being attacked, which is known as a black-box attack. Attackers can approximate a white-box attack using the notion of 'transferability', which means that an input designed to confuse a certain machine-learning model can trigger a similar behaviour within a different model. [2]

General concerns about the impacts of adversarial behaviour on stability, whether in isolation or through interaction, have been emphasized by recent demonstrations of adversarial attacks against these systems.[3]

Perhaps the most widely discussed attack cases involve image classification algorithms that are deceived into 'seeing' images in noise,[4] i.e., white noise randomly generated that does not correspond to any image is detected as one, or are easily tricked by pixel-level changes so they classify a school bus as an ostrich, for example. Similarly, game-playing systems that outperform any human (e.g., Chess or AlphaGo) can suddenly fail if the game structure or rules are slightly altered in ways that would not affect a human.[5] Autonomous vehicles that function reasonably well in ordinary conditions can, with the application of a few pieces of tape, be induced to swerve into the wrong lane or speed through a stop sign.[6] This list of adversarial attacks is by no means exhaustive and continues to grow over time.

AI in Military Applications

Many NATO countries utilize AI and machine learning to improve and streamline military operations and other national security initiatives. Regarding intelligence collection, AI technologies have already been incorporated into military operations in Iraq and Syria, where computer vision algorithms have been used to detect people and objects of interest. Military logistics is another area of focus in this realm. The US Air Force uses AI to keep track of when its planes need maintenance, and the US Army uses IBM's AI software 'Watson' for both predictive maintenance and analysis of shipping requests. Defence applications of AI also extend to semiautonomous and autonomous vehicles, including fighter jets, drones or unmanned aerial vehicles (UAVs), ground vehicles, and ships.

One might hope that adversarial attacks would be relatively rare in the everyday life since 'random noise' that targets image classification algorithms is actually far from random. Unfortunately, this confidence is almost certainly unwarranted for defence or security technologies. These systems will invariably be deployed in contexts where the other side has the time, energy, and ability to develop and construct precisely these types of adversarial attacks.[7] AI and robotic technologies are particularly appealing for deployment in enemy-controlled or enemy-contested areas since those environments are the riskiest ones for our human soldiers, in large part because the other side has the most control over the environment.

Having realized the importance of the technological lead of AI development and application, NATO launched the Military Uses of Artificial Intelligence, Automation, and Robotics (MUAAR) project under the Multinational Capability Development Campaign (MCDC) in 2020. The project's scope was to develop concepts and capabilities to address the challenges of conducting joint coalition operations and provide assessments on them.[8] The project's objective aimed to assess present and future military tasks and functions that could benefit from AI, automation and robotics. It also considering paybacks in efficiency and cost savings.

Examples of the dangers posed by adversarial manipulation of machine learning classifiers in defence applications are abundant, with different levels of severity. For example, a Lethal Autonomous Weapons System (LAWS) might misidentify friendly combat vehicles as enemy combat vehicles. Likewise, an explosive device or an enemy fighter jet might get misidentified as a rock or a bird. On the other hand, knowing that an AI spam filter tracks certain words, phrases, and word counts for exclusion, attackers can manipulate the algorithm by using acceptable words, phrases, and word counts and thus gain access to a recipient's inbox, further increasing the likelihood of email-based cyberattacks.[9]

Conclusion

In summary, AI-enabled systems can fail due to adversarial attacks intentionally designed to trick or fool algorithms into making a mistake. Such attacks can target the algorithms of the classifiers (white-box attacks) or target the output by just having access to the input (black-box attacks). These examples demonstrate that even simple systems can be fooled in unanticipated ways and sometimes with potentially severe consequences. With the widespread range of adversarial learning applications in the cyber security domain, from malware detection to speaker recognition to cyber-physical systems to many others such as deep fakes, generative networks, etc., it is time for this issue to take center stage as NATO is increasing its funding and deployment into the fields of automation, AI, and autonomous agents. There needs to be a high level of awareness regarding the robustness of such systems before deploying them in mission-critical instances.

Many recommendations have been offered to mitigate the dangerous effects of adversarial machine learning in military settings. Keeping humans in or on the loop is essential in such situations. When there is human-AI teaming, people can recognize an adversarial attack and guide the system to appropriate behaviours. Another technical suggestion is adversarial training, which involves feeding a machine learning algorithm a set of potential perturbations. In the case of computer vision algorithms, this would include images of the stop sign that displays those strategically placed stickers, or of school buses that include those slight image alterations. That way, the algorithm can still correctly identify phenomena in its environment despite an attacker's manipulations.

Given that machine learning in general and adversarial machine learning in particular, are still relatively new phenomena, the research on both is still emerging. As new attack techniques and defence countermeasures are being implemented, caution needs to be exercised by NATO military forces in employing new AI systems in mission-critical operations. As other nations, particularly China and Russia, are making significant investments in AI for military purposes, including in applications that raise questions regarding international norms and human rights, it remains of utmost importance for NATO to maintain its strategic position in order to prevail on future battlefields.

Dr Elie Alhajjar is a Senior Research scientist at the Army Cyber Institute and jointly an Associate Professor in the Department of Mathematical Sciences at the United States Military Academy at West Point, NY, where he teaches and mentors cadets from all academic disciplines. Before coming to West Point, Dr Alhajjar had a research appointment at the National Institute of Standards and Technology (NIST) in Gaithersburg, MD. His work is supported by grants from NSF, NIH, NSA, and ARL and he was recently named the Dean's Fellow for research. His research interests include mathematical modelling, machine learning and network analysis. He has presented his research work at international meetings in North America, Europe, and Asia. He is an avid science policy advocate, having received the Civilian Service Achievement Medal, the NSF Trusted CI Open Science Cybersecurity Fellowship, the Day One Technology Policy Fellowship, and the SIAM Science Policy Fellowship. He holds a Master of Science and a PhD in mathematics from George Mason University, Master's, and Bachelor's degrees from Notre Dame University.

Endnotes

1. Danks, D. (2020), 'How Adversarial Attacks Could Destabilize Military AI Systems.' IEEE Spectrum, available at https://spectrum.ieee.org/adversarial-attacks-and-ai-systems, accessed on 16 May 2022.
2. Alhajjar, E., Maxwell, P. and Bastian, N. (2021), 'Adversarial Machine Learning in Network Intrusion Detection Systems.' Expert Systems with Applications, 186 (115782), 1–13.
3. Biggio, B. and Roli, F. (2018), 'Wild patterns: Ten years after the rise of adversarial machine learning.' Pattern Recognition, 84, 317–331.
4. Nguyen, A., Yosinski, J. and Clune, J. (2015), 'Deep neural networks are easily fooled: High confidence predictions for unrecognizable images.' In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 427–436.
5. Raghu, M., Irpan, A., Andreas, J., Kleinberg, B., Le, Q. and Kleinberg, J. (2018), 'Can deep reinforcement learning solve Erdos-Selfridge-Spencer games?' In International Conference on Machine Learning, pp. 4238–4246.
6. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T. and Song, D. (2018), 'Robust physical-world attacks on deep learning visual classification.' Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1625–1634.
7. Danks, D. (2020), 'How Adversarial Attacks Could Destabilize Military AI Systems.' IEEE Spectrum, available at https://spectrum.ieee.org/adversarial-attacks-and-ai-systems, accessed on 16 May 2022.
8. NATO Allied Command Transformation Operational Experimentation (2020), 'Military Uses of Artificial Intelligence, Automation, and Robotics (MUAAR),' available at: https://www.act.nato.int/application/files/5515/8257/4725/2020_mcdc-muaar.pdf, accessed on 7 February 2022.
9. Biggio, B., Fumera, G. and Roli, F. (2010), 'Multiple classifier systems for robust classifier design in adversarial environments.' International Journal of Machine Learning and Cybernetics, 1(1), 27–41.