Technology Networks

Article

Cars Require Regular Inspection, Why Should AI Models Be any Different?

Published: March 14, 2022| Pin-Yu Chen, PhD, IBM Research

Credit: Pixabay

This article includes research findings that are yet to be peer-reviewed. Results are therefore regarded as preliminary and should be interpreted as such. Find out about the role of the peer review process in research here. For further information, please contact the cited source.

It is taken for granted that cars require regular inspection and maintenance to ensure safety and reliability. On the other hand, with the intensified demand on digital transformation, many domains and industries are actively adopting artificial intelligence (AI) and machine learning (ML) for assisting decisioning making – ranging from autonomous vehicles, education, hiring, judiciary, health, trading, content recommendation and delivery, machine translation and summary, search and planning, interactive questioning and answering, robots, to scientific discovery, to name a few. But one critical question to reflect upon is: *are we paying enough efforts, as seriously as to our cars, to inspect and certify the trustworthiness of these*

underlying AI-based systems and algorithms? Moreover, as an end user and a consumer, do we really know how and why AI technology is making decisions, and how robust AI technology is to adversarial attacks?

According to a recent Gartner report,[1] 30% of cyberattacks by 2022 will involve data poisoning, model theft or adversarial examples (see reference 2 for an overview of these new threats centered on machine learning). However, the industry seems underprepared. In a survey of 28 organizations spanning small as well as large organizations, 25 organizations did not know how to secure their AI/ML systems.[3]

There are many key factors associated with trustworthy AI, including fairness, explainability, privacy, transparency and robustness. In robustness, cars and trustworthy AI models share many common objectives. In what follows, we will highlight three analogies in car model development to explain why robustness is essential to AI models.

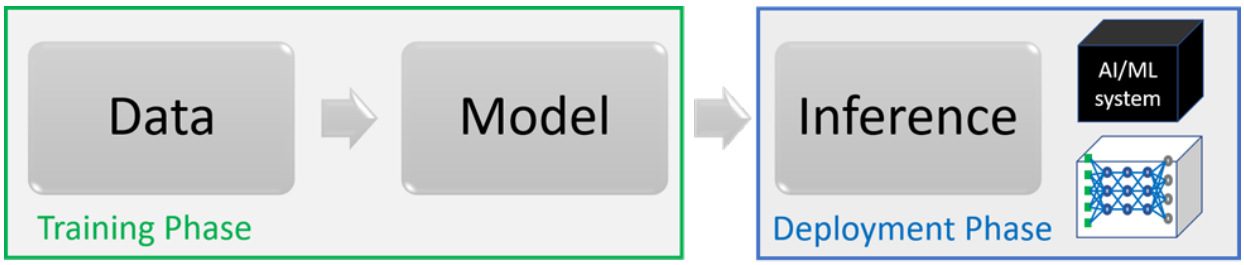
Lifecycle of model development and deployment

Like the development of a car model (say, electrical cars), developing AI models is a costly and time-consuming process. The lifecycle of an AI model can be divided into two phases: *training* and *deployment*. The training phase includes data collection and pre-processing, model selection (e.g., architecture search and design), hyperparameter tuning, model parameter optimization, and validation. AI model training can be quite expensive, especially when it comes to the training of foundation models[4] that require pre-training on large-scale datasets with neural networks consisting of a gigantic size of trainable parameters. Take the Generative Pre-trained Transformer 3 (GPT-3)[5] as an example, which is one of the largest languages models ever trained to date. GPT-3 has 175 billion parameters and is trained on a dataset consisting of 499 billion tokens. The estimated training cost is about 4.6 million US dollars even with the lowest priced GPU cloud on the market

in 2020.[6] After model training, the model is "frozen" (fixed model architecture and parameters) and is ready for deployment. The two phases can be recurrent – a deployed model can reenter the training phase with continuous model/data updates. Having invested so much, one would expect the resulting AI model is hack-proof and robust to be deployed. Otherwise, the failure of an AI technology could be as catastrophic as car model recalls.

Error inspection and fault diagnosis in the lifecycle

When cars are in motion, there are several sensors in place for fault detection. During the AI model's lifecycle, understanding the failure modes and limitations of the model can help model developers identify hidden risks and errors, and more importantly, mitigate negative impacts and damage before deployment in the real world. Depending on the assumption on the attackers' capabilities in intervening the AI lifecycle, also known as the threat models, different attacks targeting ML-based systems are summarized in Figure 1.



Attack Category / Attacker's Capability	Data	Model / Training Method	Inference
Poisoning Attack	X	X*	
Backdoor Attack	X		
Evasion Attack (adversarial example)		X*	X
Extraction/Inference Attack (model stealing, membership inference, data leakage)			X
Model Injection (model replacement, integrity)		X*	X

*No access to model internal information in the black-box attack setting

Figure 1. *Holistic view of adversarial attack categories and capabilities (threat models) in the training and deployment phases. In the deployment phase, the target (victim) can be an access-limited black-box system (e.g., a prediction API) or a transparent white-box model. Image adapted*

from Chen PY, Liu S. Holistic adversarial robustness of deep learning models. arXiv. doi: 10.48550/arXiv.2202.

A thorough robustness inspection based on a comprehensive set of active in-house testing, continuous monitoring, and performance certification (e.g., quantifiable measure of robustness[7]) should be considered as a must-have standard for AI technology, to ensure its safety and reliability. Many opensource libraries such as Adversarial Robustness 360[8] provide available tools for error inspection and robustness evaluation on machine learning models. As illustrated in Figure 2, upon the diagnosis, one can fix the identified issues and return a risk-mitigated model for use, just like the procedure of car inspection and maintenance!

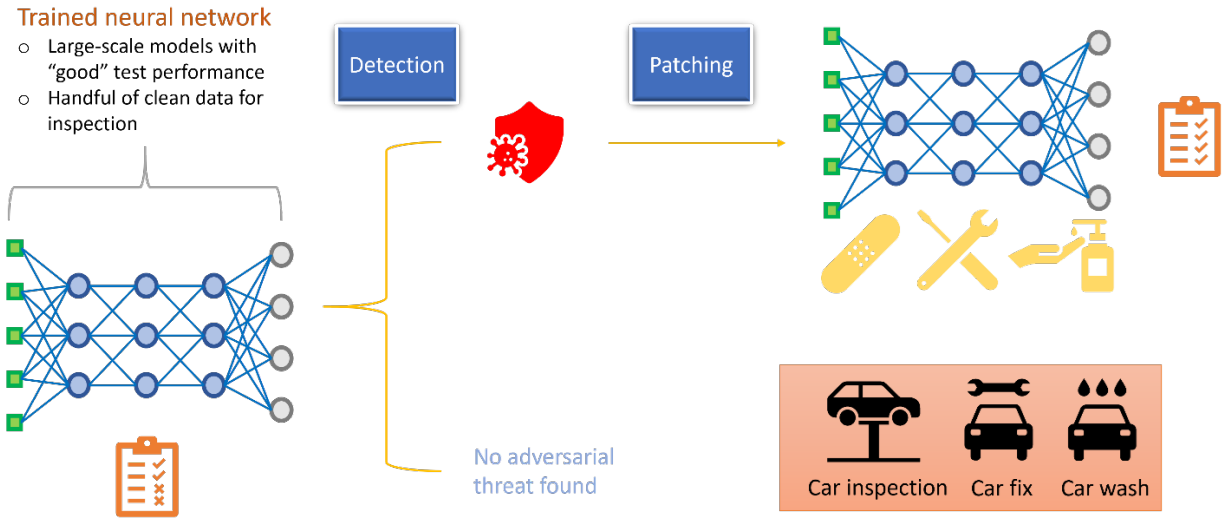


***Figure 2**. Conceptual pipeline for AI model inspection. The first stage is to identify any potential threats hidden in a given AI model. The second stage is to fix the found errors and eventually return a risk-mitigated model for use. Image adapted from https://youtu.be/rrQi86VQiuc?utm_source=359405&utm_medium=pdf&utm_campaign=pdf_lead_conversion.*

Improving robustness in unseen and adversarial environments

Cars like the Mars Exploration Rovers can successfully execute the assigned task on a new and unseen terrain because they were developed on simulated environments. For AI models, one can incorporate the failure examples generated from the error inspection tools to improve the robustness in unseen and even adversarial environments. This model training methodology is known as *adversarial machine learning*, by introducing a virtual adversary in the training environment to stimulate better and more robust models. During model training, the role of virtual adversary is to simulate the worst-case scenario and generate new data samples to help the model generalize better in unseen and adversarial environments. Figure 3 summarizes the major objectives of such a new paradigm of learning with an adversary, including finding limitations, improving robustness, creating synergies and boosting machine learning. It is worth noting that adversarial machine learning also motivates many novel applications beyond the original goal of robustness, such as model reprogramming that offers an efficient approach to reusing a pre-trained AI model for solving new tasks in resource-limited domains.[9]

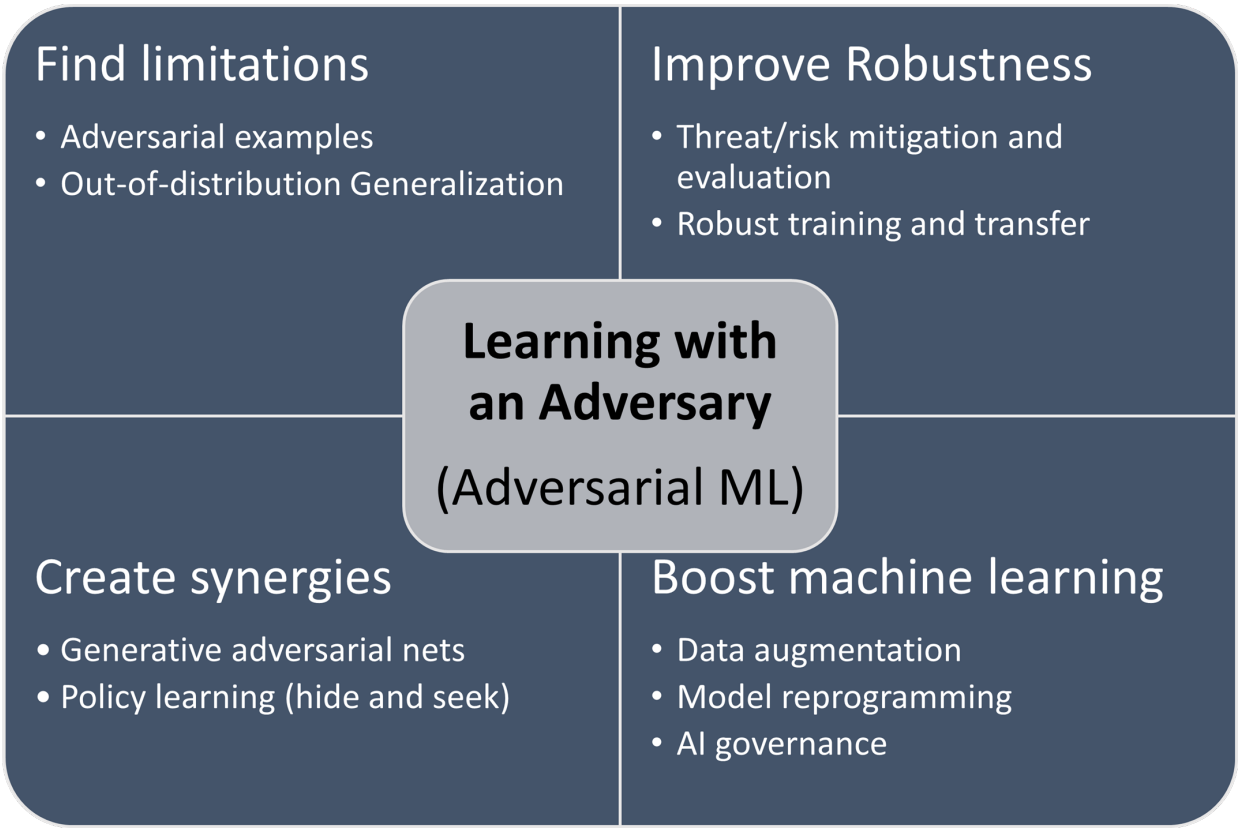


***Figure 3**. The methodology of learning with an adversary, also known as adversarial machine learning.*

Cars are transformative technology and have deep influence on our society and life. However, we also need to acknowledge and address their accompanied issues such as energy consumption and air pollution. Similarly, while we are anticipating AI technology to bring fundamental and revolutionary changes, we need to be proactive to prepare our technology to be hack-proof and trustworthy.[10] The goal in AI robustness research is to create an organic ecosystem between AI technology, the society and human-centered trust. Such an ecosystem can be achieved by formalizing and making standards for AI model inspection as illustrated in Figure 2, to ensure a greater good, prevent possible misuses and become fast adaptive to self-identified simulated failures as well as real and unforeseen challenges in the wild.

References

1. Gartner top 10 strategic technology trends for 2020. Gartner. https://www.gartner.com/smarterwithgartner/gartner-top-10-strategic-technology-trends-for-2020?utm_source=359405&utm_medium=pdf&utm_campaign=pdf_lead_conversion. Published October 21, 2019. Accessed March 7, 2022.

2. Chen PY, Liu S. Holistic adversarial robustness of deep learning models. *arXiv*. Posted online February 15, 2022. doi: 10.48550/arXiv.2202

3. Kumar RSS, Nyström M, Lambert J, et al. Adversarial machine learning – industry perspectives. *arXiv.* Posted online February 4, 2020. doi: 10.48550/ARXIV.2002.05646

4. Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models. *arXiv.* Posted online August 16, 2021. doi: 10.48550/ARXIV.2108.07258

5. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. *arXiv*. Posted online May 28, 2020. doi: 10.48550/ARXIV.2005.14165

6. Openai's GPT-3 language model: a technical overview. Lambda. https://lambdalabs.com/blog/demystifying-gpt-3/?utm_source=359405&utm_medium=pdf&utm_campaign=pdf_lead_conversion. Published June 3, 2020. Accessed March 7, 2022.

7. Preparing deep learning for the real world – on a wide scale. IBM Research. https://research.ibm.com/blog/deep-learning-real-world?utm_source=359405&utm_medium=pdf&utm_campaign=pdf_lead_conversion. Published February 9, 2021. Accessed March 7, 2022.

8. Adversarial Robustness 360. IBM Research. https://art360.mybluemix.net/?utm_source=359405&utm_medium=pdf&utm_campaign=pdf_lead_conversion?utm_source=359405&utm_medium=pdf&utm_campaign=pdf_lead_conversion. Accessed March 7, 2022.

9. Chen PY. Model reprogramming: resource-efficient cross-domain machine learning. *arXiv*. doi: 10.48550/ARXIV.2202.10629. Posted online February 22, 2022.

10. Securing AI systems with adversarial robustness. IBM Research. https://research.ibm.com/blog/securing-ai-workflows-with-adversarial-robustness?utm_source=359405&utm_medium=pdf&utm_campaign=pdf_lead_conversion. Published February 9, 2021. Accessed March 7, 2022.

Technology Networks