Special Section on Attacking and Protecting Artificial Intelligence

Bhasin, S.; Garg, S.; Regazzoni, F.

DOI
10.1049/cit2.12023

Publication date
2021

Document Version
Final published version

Published in
CAAI Transactions on Intelligence Technology

CAAI Transactions on Intelligence Technology

DOI: 10.1049/cit2.12023

EDITORIAL

Special Section on Attacking and Protecting Artificial Intelligence

Modern Artificial Intelligence (AI) systems largely rely on advanced algorithms, including machine learning techniques such as deep learning. The research community has invested significant efforts in understanding these algorithms, optimally tuning them, and improving their performance, but it has mostly neglected the security facet of the problem. Recent attacks and exploits demonstrated that machine learning-based algorithms are susceptible to attacks targeting computer systems, including backdoors, hardware Trojans and fault attacks, but are also susceptible to a range of attacks specifically targeting them, such as adversarial input perturbations. Implementations of machine learning algorithms are often crucial proprietary assets for companies and thus are required to be protected. It follows that implementations of AI-based algorithms are an attractive target for piracy and illegitimate use and, as such, they need to be protected as all other IPs. This is equally important for machine learning algorithms running on remote servers vulnerable to micro-architectural exploits.

Protecting AI algorithms from all these attacks is not a trivial task. While vast research in hardware and software security have established several sound countermeasures, the specificity of the algorithms used in AI could make such countermeasures ineffective (or simply inapplicable), given the complex and resource intensive nature of the algorithms. The task of protection will become even more difficult in the near future, given the trend where part of the intelligence will be deployed directly into resource constrained cyber-physical systems and IoT devices. AI models themselves should be protected against illegitimate and unauthorized use and distribution. Because of this, IP protection techniques such as watermarking, fingerprinting and attestation have been proposed, but, especially the last two, should be studied more in depth.

To address all these security challenges, two actions are needed. First, we need a complete understanding of the attackers' capabilities. Second, novel and lightweight approaches for protecting AI algorithms, given the distributed level of intelligence, should be conceived and developed, including (but not limited to) obfuscation, finger-printing, homomorphic encryption, and a new set of countermeasures to protect AI algorithms from adversarial input, backdooring and physical attacks.

This Special Section covers problems related to attacking and protecting implementations of AI algorithms, and the use of AI to improve state-of-the-art attacks such as physical attacks. It consists of three articles which are selected for publication after multiple rounds of peer review and scrutiny. An overview of these articles is discussed in the following.

The first article reports different types of adversarial attacks, considering various threat models, followed by a discussion on the efficiency and challenges of state-of-the-art countermeasures against them. It also provides a taxonomy for adversarial learning which can help future research to correctly categorize discovered vulnerabilities and plan protection mechanisms accordingly. The article concludes discussing open problems that can trigger further research on the topic.

The second article takes a step towards disseminating knowledge about the widely popular and critical threat of side-channel attacks on neural networks. This survey considers and categorizes the most relevant threat models and corresponding attacks with different objectives including recovery of hyper-parameters, secret weights and inputs. The article differentiates between types of side-channel attacks like physical, local or remote to highlight the applicability of various attacks and concludes with a discussion of countermeasures.

The third article surveys AI model ownership protection techniques, the majority of them being based on watermarking, reporting advantages and disadvantage of them and highlighting possible research directions. The authors identified that, to date, the most studied technique is watermarking, that has been proposed in white box and black box settings. The articles also survey existing attacks aiming at removing or making ineffective IP protection techniques, and identify fingerprinting and attestation as two approaches are not yet studied in depth.

Overall, the articles accepted cover a wide spectrum of problem providing readers with a perspective on the underlying

problem in both breadth and depth. We would like to thank all the authors and reviewers again for their contributions.

Shivam Bhasin[1]
Siddharth Garg[2]
Francesco Regazzoni[3,4]

[1]*Nanyang Technological University, Singapore*
[2]*New York University, New York, USA*
[3]*University of Amsterdam, Science Park 904, Amsterdam, The Netherlands*
[4]*ALaRI - USI, Lugano, Switzerland*

Correspondence
Shivam Bhasin, Nanyang Technological University, Singapore.
Email: sbhasin@ntu.edu.sg

ORCID
Shivam Bhasin https://orcid.org/0000-0002-6903-5127