Guest Editorial

Trustworthy AI

Jin, Yier; Ho, Tsung Yi; Picek, Stjepan; Garg, Siddharth

DOI
10.1145/3534957

Publication date
2022

Document Version
Final published version

Published in
ACM Journal on Emerging Technologies in Computing Systems

Guest Editorial: Trustworthy AI

Neural Networks (NNs) have recently shown tremendous progress in various real-world applications, ranging across object recognition, natural language processing, and autonomous vehicles. Additionally, there has been an increasing effort to deploy large-scale NN models on dedicated hardware platforms such as GPU, FPGAs, or customized ASICs in order to improve the performance and efficiency of data processing systems. Hardware vendors spend great efforts collecting a data set, training these NNs models on it, and developing the NNs accelerators, and thus want to keep the trained models private and secret.

However, recent studies have demonstrated that severe vulnerabilities exist in hardware implementations of these NN accelerators. An adversary, who has no knowledge of the details of structures and designs inside these accelerators (i.e., black-box), can effectively reverse engineer the neural networks by leveraging various side-channel information, such as timing, power, and **electromagnetic (EM)** emanations.

Besides these emerging hardware security threats to the NN accelerators, there also exists a series of techniques targeting general **Artificial intelligence (AI)/deep neural network (DNN)** algorithms. Researchers have investigated these topics in the last few years and they have also become a major threat to trustworthy AI development and deployment. Among them, Transfer Learning aims to recognize and apply knowledge gained from previous tasks, source domains, to different but related tasks, target domains. Researchers have recently shown that layers trained on a source task with large-scale labelled datasets can be reused to predict on a target domain that has substantially less available data. Two typical threats to trustworthy AI are listed as follows.

- *Adversarial Examples.* Adversarial attacks against DNNs generate adversarial examples by adding particular perturbations to the original inputs. In the image processing area, these carefully crafted images are often imperceptible to human eyes as the perturbations are slight, but can easily fool a classifier into predicting incorrect labels.
- *Model Extraction Attacks.* In these attacks, a malicious entity aims to accurately extract a model equivalent to a target model by querying the labels and confidence scores of model predictions to inputs. Several studies present efficient algorithms to steal machine learning models. More efficient black-box attack methods are also developed to steal deep learning models with millions of parameters by applying a special type of transfer learning scheme and specially crafted adversarial examples.

The goal of the special issue is to help increase the awareness for AI users on AI trustworthiness, and also to gather researchers' latest results in the area of AI attacks and defense. Considering the above challenges and potential solutions, the scope of this special issue covers various topics

ACM Reference format:
Yier Jin, Tsung-Yi Ho, Stjepan Picek, and Siddharth Garg. 2022. Guest Editorial: Trustworthy AI. *J. Emerg. Technol. Comput. Syst.* 18, 3, Article 55 (June 2022), 3 pages.
https://doi.org/10.1145/3534957

1550-4832/2022/06-ART55
https://doi.org/10.1145/3534957

including (1) AI vulnerabilities; (2) Trustworthy AI; and (3) AI for cybersecurity and hardware security. From all the received submissions from experts in the field, only seven could be accepted to be included in this issue. These seven papers can be categorized into three groups: Survey, Attacks, and Defense.

In the Survey category, the paper titled "A Review and Comparison of AI Enhanced Side Channel Analysis" by M. Panoff et al. examines the latest state-of-the-art deep learning techniques for side channel analysis, the theory behind them, and how they are conducted. Different deep learning enhanced **side channel analysis (SCA)** schemes and their relative performance are evaluated and compared.

In the Attacks category, various attacks to AI and DNN accelerators are proposed. These new attacks are clear evidence that more defense and protections are needed to secure AI systems and DNN accelerators.

The paper titled "Direction-Aggregated Attack for Transferable Adversarial Examples" by T. Huang et al. proposes to improve the transferability of adversarial examples by aggregating attack directions of a set of examples around the neighborhood of the input. Several potential approaches for defending against transferable adversarial examples are also discussed in the paper.

The paper titled "Generation of Black-box Audio Adversarial Examples Based on Gradient Approximation and Autoencoders" by P. Huang et al. proposes a real-time attack framework that utilizes the neural network trained by the gradient approximation method to generate adversarial examples on **Keyword Spotting (KWS)** systems.

The paper titled "Power-Based Attacks on Spatial DNN Accelerators" by G. Li et al. investigates the vulnerability of spatial DNN accelerators using a general 8-bit number representation to **Differential Power Analysis (DPA)** style attacks.

In the Defense category, innovative defense solutions are developed to secure AI systems.

The paper titled "Defending Against Adversarial Attacks in Deep Learning with Robust Auxiliary Classifiers Utilizing Bit-Plane Slicing" by Y. Liu et al. proposes bit-plane classifiers which take the bit-plane channels as input to defend against adversarial attacks for deep learning schemes.

The paper titled "Guarding Machine Learning Hardware Against Physical Side-Channel Attacks" by A. Dubey et al. proposes a fully masked neural network based on Boolean masking and implements all masked hardware blocks for machine learning hardware protection against physical side-channel attacks.

The paper titled "Diverse, Neural Trojan Resilient Ecosystem of Neural Network IP" by B. Olney and R. Karam proposes two complementary techniques: stochastic parameter mutation and a delta-update procedure. These techniques are effective against transferability of a neural Trojan to the greater neural network Intellectual Property (IP) ecosystem.

Yier Jin
University of Florida, USA
yier.jin@ece.ufl.edu

Tsung-Yi Ho
The Chinese University of Hong Kong, Hong Kong
tyho@cse.cuhk.edu.hk

Stjepan Picek
Radboud University & Delft University of Technology, The Netherlands
stjepan.picek@ru.nl

Siddharth Garg
New York University, UAE
sg175@nyu.edu

Guest Editors