BRIEFING NOTES

BN-23-The role of AI-Oct2020

FACIAL RECOGNITION, IMPACT OF AI ON NATIONAL SECURITY, ADVERSARIAL ATTACKS AND TRUST IN AI

Authors: Mehdi Taheri[1] and Kash Khorasani [2]
[1] Graduate student, Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada
[2] Professor, Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada

SUMMARY

- Advances in AI systems, computational capabilities, and storage capacities have led to emergence of facial recognition technologies where a machine can identify a face using images.
- Despite all advantages and benefits of facial recognition technologies, they have raised concerns about making biased decisions that violate prohibition of discrimination, ethics, privacy, and encroach on individual democratic values and freedoms.
- Progress in the field of AI systems will have a major impact on military superiority of countries, their information superiority, their economic superiority, and priorities of countries in the context of national security.
- Machine learning methodologies are vulnerable to adversarial attacks. In such attacks, inputs to the machine are manipulated by the adversary such that their desired malicious response is produced.
- Since AI systems are becoming more involved in making decisions for humans, such as determining whether an individual should be sentenced in a court, importance of having trust in AI systems has become significantly more evident.
- There are mainly two characteristics that a trustable AI system should possess, namely fairness and explainability. Fairness is not achieved if we have a bias in the system. Explainability is the manner an individual can make derivations about a given developed algorithm.

CONTEXT

- Whereas designers and agencies using AI systems are subject to legal obligations arising from their products or services, potential lawsuits, at minimum, hinder the path in testing and rolling out of new technology.
- In case of bias, one of the major concerns with facial recognition systems is high level errors in recognizing people of color and minorities. Regulations and rules to address the biasing problem in facial recognition system should require companies to provide their customers with transparency using understandable documents to demonstrate the capabilities and limitations of their technology. Moreover, the rules should deal with independent tests by third-parties on facial recognition services of companies to check the accuracy and bias in their products. Furthermore, the regulations should require entities that provide facial recognition services to review their facial recognition outcomes by qualified people before making the final decisions.
- Due to widespread use of surveillance cameras around the world, the places where people visit and patterns of their behaviour can be traced and stored easily. This information gives

governments and companies the ability to predict people's actions implying that privacy of people is violated. To avoid these violations first, entities that use facial recognition services should provide signs that indicate presence of cameras. Second, legally it should be indicated that entering a building or using a service that involves facial recognition systems implies the consent of customers to the use of facial recognition systems.

- In accordance with democratic freedom rights, it is necessary for people to be able to move freely and talk to others without any governmental surveillance. Currently some governments are using facial recognition technology to improve public safety. However, this ability could give governments the power to follow everyone in most public places. To address this concern, countries need to have a new law which permits the governments to follow and track specified individuals using facial recognition systems only under court order cases, or in case of emergency or immediate risks of death or injury.
- AI systems have a significant role in areas such as cyber defense and satellite imagery analysis. It has been predicted that future progress in this field will have a major impact on the strategy, organization, and priorities of countries in the context of national security.
- The main objectives for improving the national security through utilizing AI systems can be identified as preventing future terrorist attacks, reducing vulnerabilities of the nation, and reducing damage and recovery costs from both cyber and kinetic attacks.
- Adversarial attacks can be divided into four categories, namely
    a) confidentiality attacks In which the data that was previously used in the training phase of AI is exposed to the malicious attacker,
    b) integrity attacks where the adversary tampers with the AI training data such that it behaves incorrectly in response to some inputs and miss-categorizes them,
    c) availability attacks where the adversary disguises its attack signals as legitimate input to the system such that a human is not capable of comprehending its differences with the actual signal, and
    d) replication attacks which allow the adversary to obtain representation, description, and detailed model of the system.
- Defending against adversary attacks has three stages,
    a) first, measuring model robustness by evaluating the loss of accuracy in the system in presence of manipulated inputs by the adversary,
    b) second, model hardening by preprocessing the input subject to inclusion of adversarial examples to the training data set, and
    c) third, runtime detection in which abnormal behaviors in the internal layers of the AI system due to adversarial attacks are exploited.
- To achieve trust in AI systems, regulations should prevent bias and discrimination from affecting decisions that are made by the machine.
- Users of AI systems should be granted and provided with the right to be informed on existence, logic, and possible consequences of a decision that is made autonomously by a machine. This is defined as the "right to explanation".

CONSIDERATIONS

- Facial recognition technology providers should be required to comply with and consider laws that are in accordance with prohibiting discrimination against costumers of their services.
- The potential warfare applications of AI systems are irresistible not to be used by military. Hence, the goal should be to pursue safe and effective technologies.
- AI systems can contribute to recognizing patterns and activities that hostile attackers initiate with warning systems to prevent them.
- Using image and speech recognition technologies and by sharing information in borders of countries the counter terrorism capabilities can be improved.
- AI systems can be used to discover cooperative relationship and patterns among criminal groups and terrorists.
- In the critical infrastructures such as water supplies, roads, and power networks, AI systems can be employed to detect their abnormal behavior.
- A set of regulations should be developed to deal with transparency in the chain of data, which ensures that for specific decisions that are made by the AI system, the data controller provides a satisfactory explanation.
- Discrimination detection algorithms along with discrimination prevention methods that are designed to eliminate bias from datasets and AI algorithms should be utilized to gain trust in the AI system.