AI Security White Paper(Oct 2018)

Building Multi-Layer Defense to Cope with New AI Security Challenges

With the accumulation of big data, dramatic improvements in computing power, and continuous innovation in Machine Learning (ML) methods, Artificial Intelligence (AI) technologies such as image recognition, voice recognition, and natural language processing have become ubiquitous. Meanwhile, the security of AI systems is becoming vitally important and has gone beyond our traditional understanding of security boundaries and frameworks.

Unlike security vulnerabilities in traditional systems, the root cause of security weaknesses in ML systems lies in the lack of explainability in AI systems. This lack of explainability leaves openings that can be exploited by adversarial ML. These attacks can be very effective and have strong transferability among different ML models, and thus pose serious security threats to Deep Neural Network (DNN)-based AI applications:

- Attacks on data integrity

  Attackers can insert malicious data in the training phase to affect the inference capability of AI models or add a small amount of noise to samples in the judgment phase to change the judgment result.

- Attacks on model integrity

  Attackers may implant backdoors in models to launch advanced attacks. Due to the inexplainability of AI models, the backdoors are difficult to detect.

- Attacks on model confidentiality

  Unwilling to expose training models, service providers generally furnish only model query services. However, an attacker may craft similar models through a large number of queries to obtain model information.

Because existing security technologies have inherent limitations that leave them unable to cope with new AI security challenges, Huawei proposes three layers of defense for deploying AI systems in service scenarios:

- Attack defense security: Design targeted defense mechanisms for known attacks.
- Model security: Improve model robustness by means of model verification.
- Architecture security: Design different security mechanisms for services in which AI systems are deployed to ensure business security.

For specific attack-defense security measures, Huawei focuses on defense technologies against AI evasion, poisoning, and backdoor attacks, and the improvement of model theft prevention capabilities. For example, Huawei uses technologies such as mutation-testing-based detection of evasion attacks to provide industry-leading accuracy, applicability, and practicability. This technology can be deployed in existing AI systems to detect and filter out adversarial samples. To ensure model security, Huawei focuses on improving the testability, verifiability, and explainability of models. For architecture security, Huawei analyzes and determines the risks in using AI models based on the characteristics and architecture of specific services; we then design the AI security architecture and deployment solutions using security mechanisms such as isolation, detection, fusing, and redundancy to enhance the robustness of products.

There is a long way to go before the industry achieves secure and robust AI systems. On the technology side, we need to continuously research AI explainability to understand how to implement a sound AI foundation and build systematic defense mechanisms for secure AI platforms. In terms of business, we need to analyze the business models in detail, and deploy tested and verified AI security technologies.

Huawei places cyber security and privacy protection at the top of the company’s agenda. At 14:00 to 15:30 of October 11, the security summit “Build an Intelligent End-to-End Security Assurance System” will be held at HUAWEI CONNECT 2018. At this summit, Huawei will release innovative security solutions for 5G, IoT, SoftCOM, Safe City, and private cloud. Please join us at the conference and engage with Huawei on the future of intelligent security.