Security and Privacy Technologies for Safe and Secure Artificial Intelligence

Authors: *Yoshihiro Koseki*, Tsunato Nakai**

**Information Technology R&D Center*

Abstract

Efforts to use AI for business have become increasingly active in recent years due to the development of deep learning, and Mitsubishi Electric too is considering use in various fields. On the other hand, as systems incorporating AI come into wider use, there are more opportunities for malicious attackers to carry out attacks exploiting the vulnerabilities of AI. At Mitsubishi Electric, our efforts go beyond just improving performance of AI and considering how to use it. We are conducting R&D on security technology and privacy technology to enable safe and secure use, even in environments where AI is exposed to malicious attack. As part of these efforts, we have developed technology allowing object detection AI to output correct results even when its input is tampered with, and technology to prevent leakage of information relating to training data from AI models.

1. Introduction

Deep learning in neural networks garnered renewed interest in the early 2010s and since then has revolutionized AI in various fields—such as images/video, language, audio, and time series data—due to its overwhelming accuracy. In recent years, conversational AI using large language models and generative AI such as image generation AI using diffusion models have become more than just a subject of research. They have become the focus of a major boom, including applications to business, and it is anticipated that use of AI in real-world society will continue to advance at an accelerating pace in the future. On the other hand, as the scope of AI utilization expands, opportunities grow for attacking systems using AI. For such systems, we must consider not only conventional cyberattack techniques exploiting vulnerabilities of OS and applications to target IT systems, but also attack techniques exploiting vulnerabilities specific to AI. This paper describes Mitsubishi Electric's R&D efforts on security technology and privacy technology for realizing safe and secure use of AI.

2. AI Privacy and Security

Deep learning—a subject of active research in recent years—is one type of an AI technology called machine learning. When using machine learning, there are two phases, as shown in Fig. 1: the training phase where a model is generated using training data as input, and the inference phase where tasks such as object detection or translation are performed on images, text, or other inputs using the generated model. Attacks on AI in this training phase or inference phase include causing it to output an unintended result by tampering with data input to the model, or causing unintentional leakage of personal information or other data contained in training data by observing data output by the model. Intensive research has been conducted in recent years regarding attacks on AI, and even if, for example, we limit the scope to adversarial examples attacks—one type of attack technique—over 8,000 papers have been published as of March 2024[(1)].

The AI Security Information Portal[(2)] of the Ministry of Internal Affairs and Communications provides an AI Security Matrix summarizing techniques for attacking AI. Five attack categories are defined in the AI Security Matrix: data poisoning, model poisoning, adversarial examples, data theft, and model theft. Figure 1 and Table 1 show the concept and overview of each category.

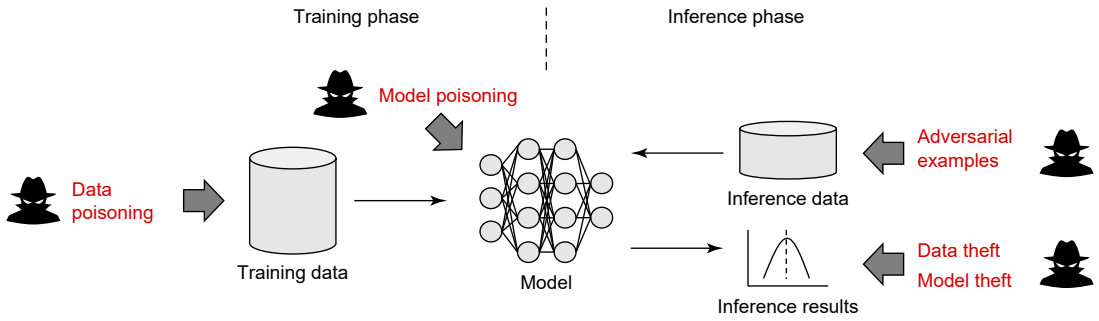

Fig. 1 Overview of attacks on AI

Table 1 Taxonomy of attacks on AI (excerpt from the AI Security Matrix[2])

Category	Overview
Data poisoning	By injecting special data called poisoned data into training data, attackers can install backdoors that cause the model to behave as intended by the attacker in response to specific input data, or can cause degradation of model accuracy
Model poisoning	By tampering with a trained model, backdoors that cause the model to behave as intended by the attacker in response to specific input data are installed in a model distributed to users as a pre-trained model
Adversarial examples	By adding modifications called perturbations to the model's input data, attackers change results such as image classifications and speech recognition to produce incorrect outputs
Data theft	By inputting multiple data items to a model and observing the output, attackers steal information on the data used for training
Model theft	By inputting multiple data items to a model and observing the output, attackers steal internal information such as model structure and parameters

Among these categories, three types of attacks (data poisoning, model poisoning, and adversarial examples) primarily relate to correct operation of the model, while data theft and model theft relate to privacy of the training data and model. From among Mitsubishi Electric's R&D efforts on security technology and privacy technology for countering these attack techniques on AI, section 3 of this paper describes technology to counter adversarial patch attacks, a type of adversarial examples attack, and section 4 describes technology to counter membership inference attacks, a type of data theft attack.

3. Defense Technology for Adversarial Patch Attacks

Object detection is a technology for outputting bounding box coordinates indicating object position, and a label indicating object type, for each object captured in an input image. It is used in areas such as detection of pedestrians by self-driving vehicles, and detection of suspicious persons by surveillance cameras. An adversarial patch attack on object detection hinders object detection by placing a patch image created using a special method near an object captured in an input image. This is a threat to systems using object detection. Figure 2 shows an example of an adversarial patch attack from previous research[3]. For the person on the left, the system displays a bounding box and a label indicating that the object is a person. Detection is done correctly. For the person on the right, on the other hand, no bounding box or label is displayed, and it is evident that correct detection has not been performed due to placement of the patch image. This patch image is not digitally placed on the input image. It is exerting an effect even though a printed image of the patch is used by placing it near the person captured by a camera. This sort of adversarial patch attack is called a physical adversarial patch attack, and it is recognized as a particularly important threat.

Fig. 2 Adversarial patch attack (excerpt from previous research[3])

We have developed a defense technology for correctly detecting objects, even when there is an adversarial patch attack[4]. With this technology, ordinary object detection is performed first on the input image. The outputs of object detection at this time are the coordinates of the bounding boxes indicating the positions of the objects in the input image, and objectness scores, which indicate the probabilities that objects are present at their locations. An object is detected at a location if the objectness score is at or above a threshold. On the other hand, the adversarial patch has the effect of lowering the objectness score of an object nearby. It hinders object detection by decreasing the objectness score of the nearby object below the threshold. This defense technology focuses on bounding boxes whose objectness scores have dropped below the threshold. In Fig. 3 on the left, the bounding boxes indicated by the dotted lines have objectness scores below the threshold, and with this technology, multiple black-out images are generated in which those boxes are respectively blacked out. Object detection is performed again for each of the generated black-out images. In images in which part or whole of the adversarial patch is blacked out, the effect of lowering the objectness score is lost, so that the objectness score of the object where the patch is placed is higher than the threshold, and detection as an object can be achieved correctly. The final output for object detection is obtained by performing final integration processing on the detection results output for the multiple black-out images, and the detection results for the images prior to blacking out.

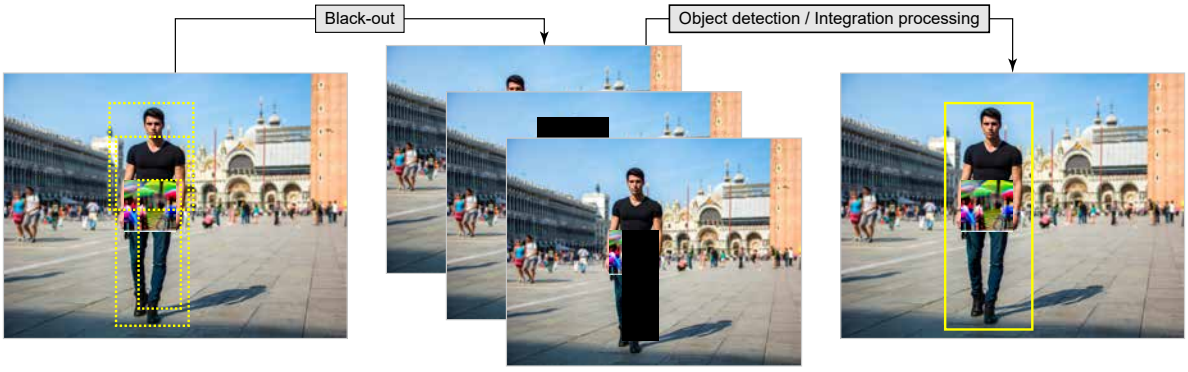

Fig. 3 Mechanism of defense method

4. Defense Technology for Membership Inference Attacks

A membership inference attack identifies whether sample data was used for model training. An attacker can determine whether the owner of a certain data set was a provider of training data based on AI inference results in response to the sample data. Figure 4 shows an overview of a membership inference attack. Membership inference attacks themselves represent relatively insignificant information leakage regarding training data, but training models resistant to membership inference attacks are also resistant to other attacks pertaining to information leakage of training data. Therefore, technologies for countering

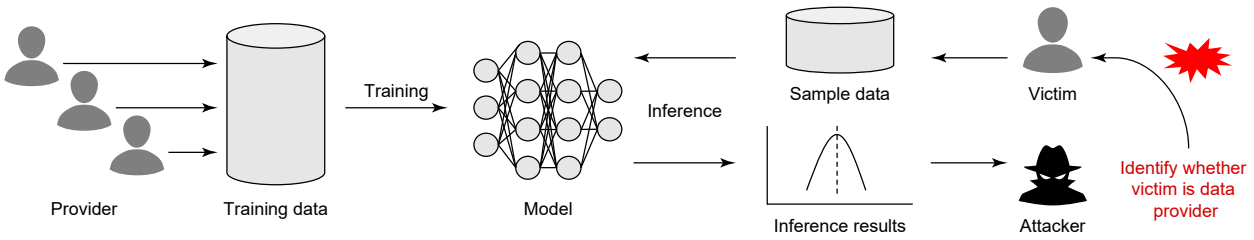

Fig. 4 Membership inference attack

membership inference attacks are important for achieving model privacy.

The reason that membership inference attacks succeed is thought to be overfitting where a model exhibits a strong reaction to data included in its training data, such as outputting a score higher than other data. Mitsubishi Electric has developed SElf-Distillation with Model Aggregation for Membership Privacy (SEDMA), a training method which suppresses overfitting, as a technology for countering membership inference attacks[5]. Figure 5 shows an overview of SEDMA. With SEDMA, the labels in training data are replaced with inference results of a model which has not used that training data for training (i.e., soft labels), and new training data with soft labels is generated. A model generated using training data with soft labels is resistant to membership inference attack because it is less prone to overfitting with respect to the original training data. Aggregated models necessary for generating soft labels are generated by splitting the training data into multiple data sets, using these data sets to train different models, and then aggregating these models (by weighted average of parameters between models). Each aggregated model provides soft labels for training data not contained in the training data of the aggregated model. The distinguishing feature of SEDMA lies in this model aggregation, and compared with existing defense methods, this achieves outstanding trade-off performance, at low computation cost, between degradation of model accuracy due to defense measures and defense measure strength.

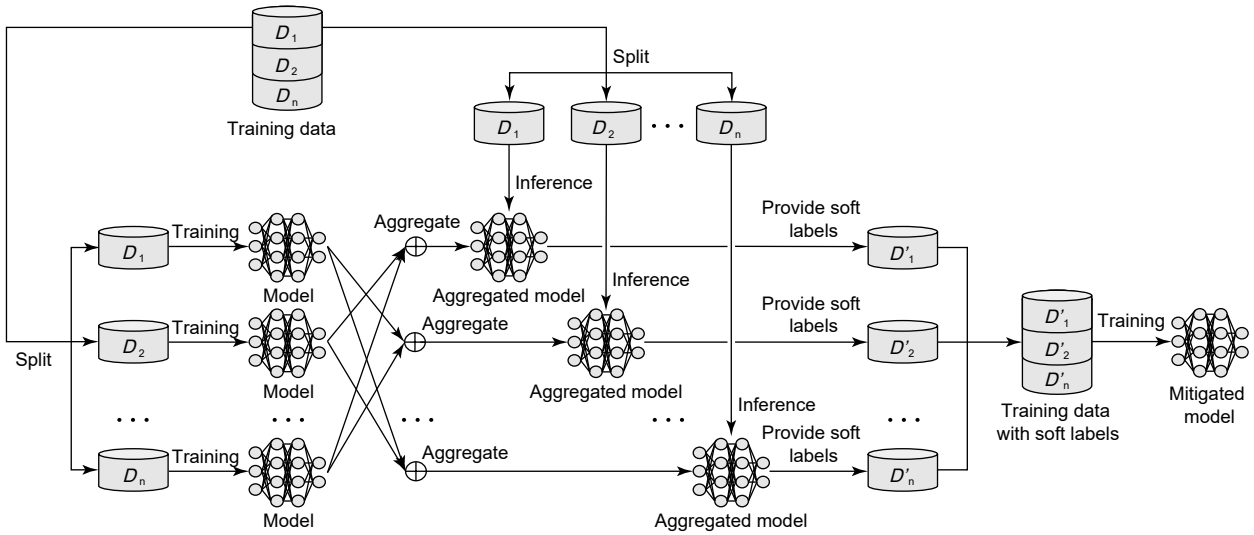

Fig. 5 Overview of SEDMA

Recently, concerns have arisen regarding the risk of information leakage relating to training data in the context of training large language models that require a huge volume of training data. In the area of large language models too, Mitsubishi Electric is moving forward with evaluation of the risk of information leakage from training models due to membership inference attack, verification of the effectiveness of existing defense measures (e.g., differential privacy), and examination of new defense methods[6][7].

5. Conclusion

This paper has described the efforts of Mitsubishi Electric regarding defense technologies for attacks on AI, in particular defense technology for adversarial patch attacks in object detection, and defense technology for membership inference attacks. Going forward, we will contribute to the promotion of AI and its safe and secure use, by further advancing R&D on security technology and privacy technology. Our focus will be on areas such as large language models that are garnering attention as generative AI, and multi-modal large language models which simultaneously handle images, audio, and other modalities.

References

(1) Carlini, N.: A Complete List of All (arXiv) Adversarial Example Papers (2019)
https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html
(2) Ministry of Internal Affairs and Communications, et al.: AI Security Information Portal
https://www.mbsd.jp/aisec_portal/
(3) Thys, S., et al.: Fooling Automated Surveillance Cameras: Adversarial Patches to Attack Person Detection, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 49-55 (2019)
(4) Koseki, Y.: Defending Against Adversarial Patch Attacks in Object Detection Through Bounding Box Filling, IEEE 13th Global Conference on Consumer Electronics, 868-869 (2024)
(5) Nakai, T., et al.: SEDMA: Self-Distillation with Model Aggregation for Membership Privacy, Proceedings on Privacy Enhancing Technologies, 494-508 (2024)
(6) Nakai, T., et al.: Does Prompt-tuning Enhance Privacy in Large Language Models?, Workshop on Recent Advances in Resilient and Trustworthy Machine Learning-driven Systems (2024)
(7) Higashi, T., et al.: Evaluation of Membership Inference Attacks on LoRA with Differential Privacy, the 9th IEEE European Symposium on Security and Privacy Poster Session (2024)