July 2023

On Device Security Agent to Mitigate Inference Attacks on Machine Learning Models

Hari Bhaskar S

On Device Security Agent to Mitigate Inference Attacks on Machine Learning Models

ABSTRACT

The advent of on-device machine learning allows developers to deploy models on end-user devices. A trained machine learning model can be a representation of the dataset used to train the model and may carry knowledge of decision making based on training. Malicious actors can perform inference attacks to exploit open-ended, unprotected on-device machine learning models, e.g., by taking the device offline to block information from being sent to the cloud and sending several inference requests to extract the patterns of data and/or training of the machine-learning model. This disclosure describes an on-device security agent that monitors the patterns of inference requests to an on-device machine learning model and the corresponding responses provided by the model. The inference agent can implement rules or a lightweight machine learning model to analyze the inference requests to determine a risk score and can take actions to mitigate inference attacks. Further, with user permission, the inference agent can send inference data including requests received and responses provided by a model to a cloud-based security agent. The cloud-based security agent can analyze such data to generate security policies for the on-device security agent.

KEYWORDS

- Inference attack
- On-device inference
- Offline inference
- Black box adversarial attack
- ML model security
- Security agent
- Query pattern
- Label extraction
- Feature importance
- Attribute inference
- Inference velocity

BACKGROUND

The advent of on-device machine learning allows developers to deploy models on end-user devices. On-device execution of machine learning models is made possible because of better device capabilities, such as higher amounts of random access memory (RAM), better and larger local storage, dedicated machine learning hardware, and higher processing power.

Machine learning (ML) models that power a variety of features on a mobile device may be accessed both when the device is online (connected to the internet) and offline. In some cases, an on-device ML model may be a lightweight or distilled version of a larger server-side model. A trained machine learning model can be a representation of the dataset that is used to train the model and carries the knowledge of decision making based on the training. There is therefore a possibility that malicious actors can exploit open-ended, unprotected on-device machine learning models.

Such attacks, referred to as inference attacks, may be rendered possible by taking the device offline by turning off the network connectivity to mobile and/or Wi-Fi networks. This blocks information that is sent to the cloud from the on-device model. While such information upload is blocked, the malicious actor may send several inference requests to better understand the patterns of data and training of the machine-learning model. This is possibly a data loss scenario. Furthermore, while a device is offline, the attacker can craft an arbitrary number of inference requests.

For example, an inference attack to detect the labels used to train a model may include the attacker sending many images to an image recognition model. The attackers can exploit the model by sending a high volume of images, with a lot of variety in the dataset, in a short period of time. This can enable learning training labels, model performance parameters, failure modes,

etc. For models trained using supervised learning, inference attacks can enable attackers to identify importance of individual features of the input data and attribute the inference to different features. Inference attacks may also allow adversarial exploitation through reverse engineering the datasets used for training. An example of a black box adversarial attack is described in [1].

Currently, little to no protection is available against such threats. While enterprise device management can enable remote management of on-device applications, including upgrades, wipeout compliance, and application allowlists, such measures cannot protect machine learning models. While monitoring cloud-based machine learning models to detect inference attacks is feasible, such monitoring is infeasible since attackers can take advantage of the offline mode of the device. No specific security measures exist to detect and remediate inference attacks on client-side machine learning models. The lack of strong on-device monitoring and model security techniques can be a problem in any domain where on-device machine learning models are used. As more powerful models run offline on devices, the risk of data loss and overall compliance risk increases.

DESCRIPTION

This disclosure describes an on-device security agent that continually monitors the patterns of inference requests to a machine learning model and the corresponding responses. The on-device agent is mapped 1:1 to model deployment - an agent A is mapped to ML model A, an inference agent B is mapped to ML model B, and so on. The on-device agent is implemented with user permission and in compliance with user settings.

With user permission, the inference requests made to a ML model and the corresponding responses are stored in a local database. If the device is online and if the user permits remote security analysis, the inference data is provided to a cloud-based security agent. If the device is

offline, a risk score is computed locally, and the inference data is set to the cloud-based security agent when the device is online at a later time. The on-device agent computes contextual and/or behavioral changes. The computed score can be compared to a threshold (for a particular time). The on-device agent can be configured with a policy that includes the model(s) to monitor, threshold values, device identity information, and protocol of communication. The cloud-based security agent can be provided as part of the service offering of the cloud provider.

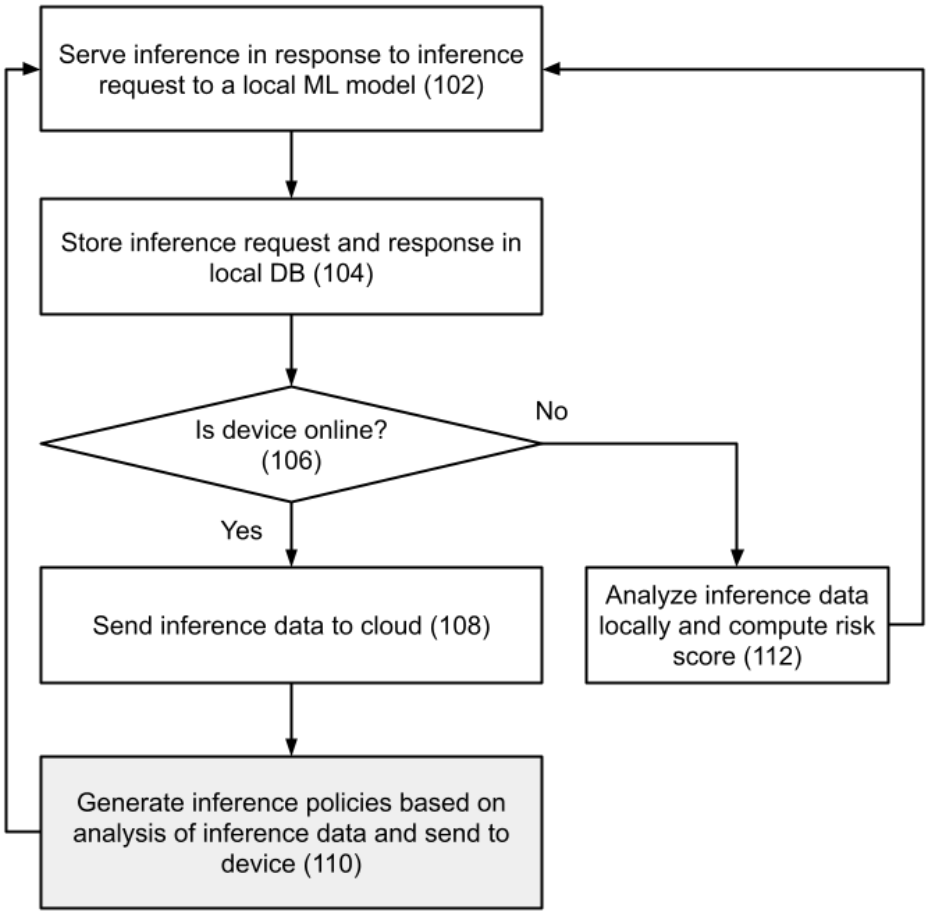

Fig. 1: Detecting inference attacks on on-device machine learning models

Fig. 1 illustrates an example method to detect inference attacks on on-device machine learning models, per techniques of this disclosure. The method can be implemented as part of an on-device security agent program. An inference is served in response to an inference request

made to an-device ML model (102). With user permission, the inference request and response are stored in a local database (104).

A check is performed to determine if the mobile device is online or offline (106). If the device is online (and if the user permits), the inference data is sent to the cloud (108). The received inference data is analyzed at the cloud-based security agent to determine if there is an attack on the on-device model. Further, the detected patterns of inference and anomalies are used to generate or update inference policies (110) for the on-device security agent. If the device is offline, the inference data is analyzed locally (112), and a risk score is computed. If the risk score exceeds a threshold (e.g., from the inference policy), actions are taken to protect the on-device ML model.

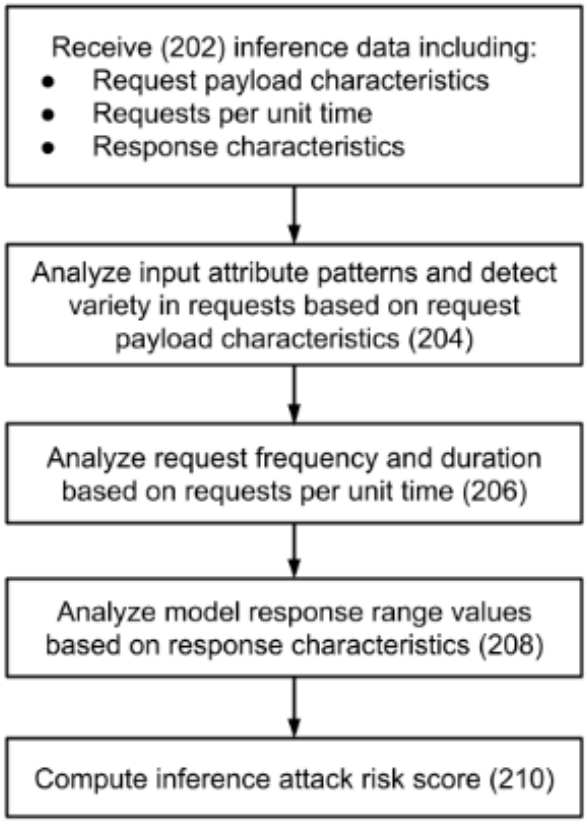

Fig. 2: Inference data analysis to determine risk score

Fig. 2 provides an example method to perform analysis of inference data to determine a risk score associated with inference data generated from inference requests received by and responses provided by an on-device machine learning model. Inference data such as request payload characteristics, requests per unit time, and response characteristics are received (202).

The input attribute patterns are analyzed. The variety in the received requests is determined based on the request payload characteristics (204). The variety may be indicative of an inference attack (e.g., since an attack payload may be very different from normal payloads generated by use of the ML model by a user). The request frequency and duration are also analyzed based on the requests per unit time (206). The frequency and duration, if different from typical patterns of use, can be indicative of an attack. Model response range values are also analyzed based on response characteristics (208). Deviations in model response range values may be indicative of an inference attack. The combination of the above factors is used to compute an inference attack risk score (210). The described method can be used by an on-device security agent and/or by a cloud-based security agent.

If the inference attack risk score is computed on-device, a simple heuristic or rule can be utilized (based on the computational capacity of the device). Theon-device security agent may be implemented as a miniature ML model or based on a rule set. For cloud-based analysis of inference data, with the availability of higher computational capacity, more complex analysis may be feasible. For example, clustering-based anomaly detection algorithms might be used.

The described security techniques provide several advantages. The on-device security agent is specific to each model (or can be implemented to protect multiple models) and is distinct from other forms of security monitoring. The security agent can generate domain-specific insights for ML model security in the context of on-device deployment. The insights can be

leveraged by security analysts and data scientists. The on-device security agent can also secure the ML model while a device is offline. Anomalies in inference data and risk scores associated with a model can be used for policy management to protect ML models. The on-device security agent can be implemented at a low computational cost and relies only on the local execution context. The described techniques enable on-device detection of inference attacks and can mitigate security risks. The offline capability for monitoring widens the security coverage. If a rogue device (e.g., model compromise) is detected, a device wipeout can be performed. The techniques reduce data risks or leakage of proprietary ML model information.

CONCLUSION

This disclosure describes an on-device security agent to protect against inference attacks on an on-device machine learning model. The security agent monitors the patterns of inference requests to an on-device machine learning model and the corresponding responses provided by the model. The inference agent can implement rules or a lightweight machine learning model to analyze the inference requests to determine a risk score and can take actions to mitigate inference attacks. Further, with user permission, the inference agent can send inference data including requests received and responses provided by a model to a cloud-based security agent. The cloud-based security agent can analyze such data to generate security policies for the on-device security agent.

REFERENCES

1. Polyakov, Alex. “How to attack Machine Learning (Evasion, Poisoning, Inference, Trojans, Backdoors)”, available online at https://towardsdatascience.com/how-to-attack-machine-learning-evasion-poisoning-inference-trojans-backdoors-a7cb5832595c, accessed June 24, 2023.