July 2023

Lifecycle Management and Security of On-device Machine Learning Models

Hari Bhaskar S

Lifecycle Management and Security of On-device Machine Learning Models

ABSTRACT

Endpoint management solutions cannot be utilized to manage on-device machine learning models. This disclosure describes techniques to integrate the best-in-class capabilities of cloud ML model management and endpoint management to enable lifecycle management and security of on-device ML model deployments. Endpoint management solutions as described herein include the capability to manage on-device models, e.g., to perform tasks such as model tracking, upgrade, and wipe out compliance. The described techniques, which can be implemented as part of an endpoint management solution, use a model catalog to track device deployments for a given machine learning model. When a model upgrade is available, a notification is provided to administrators, app developers, etc. On-device models can be upgraded, deleted, and tracked independent of app deployment. Additionally, with user permission, observability of on-device models is enabled through endpoint management to detect misuse. The endpoint management solution and model catalog also provide a deployment view of on-device models, including model versions and can be used to ensure compliance.

KEYWORDS

- Model lifecycle
- Machine learning ops
- MLOps
- Model security
- On-device machine learning
- Endpoint management
- Model management

BACKGROUND

Management of client apps through endpoint management solutions is common. However, such solutions only cover an app itself and not app components such as on-device ML models distributed with the app. While server-side models are managed, edge or mobile devices currently lack capabilities such as model versioning, tracking model deployments, etc. There is no solution that combines endpoint management and model lifecycle management capabilities to provide holistic management capabilities such as model refresh, download on demand, tracking etc. Also, current solutions do not support model tracking at deployment level and lack the ability to address security exploits on device level by wiping out models only as necessary rather than the entire app.

The advent of on-device machine learning (ML) has enabled enterprises to deploy machine learning models on mobile devices for use cases such as object/image recognition and other ML tasks that leverage device capabilities such as camera, storage, and processing. On-device ML models are often built as student models from a larger server-side teacher model and deployed as distilled miniature lightweight models. Such ML models are usually packaged with mobile applications and are sent as part of the application release to the mobile device. Lack of visibility into deployment of ML models on mobile devices that belong to the enterprise, or are used by enterprise employees or contractors pose a number of challenges for enterprises

- Endpoint management solutions do not have specific capability to track ML models deployed on a device, e.g., removal of a model, monitoring the use of a model, etc. since ML models are not standalone, but rather provided as a part of application packages from app developers

- It is not possible to perform remote model management, e.g., to refresh or download new ML models onto mobile devices for purposes such as addressing vulnerabilities or upgrading models since models are bundled within apps and can only be updated with app upgrades.
- There are no techniques to identify security exploits that target on-device models or to mitigate risk for such deployments.

DESCRIPTION

This disclosure describes techniques to integrate the best-in-class capabilities of cloud ML model management and endpoint management to achieve lifecycle management and security of on-device ML model deployments. Endpoint management solutions as described herein include the capability to manage on-device models, e.g., to perform tasks such as model tracking, upgrade, and wipe out compliance.

The described techniques, which can be implemented as part of an endpoint management solution, use a model catalog to track device deployments for a given machine learning model. When a model upgrade is available, a notification is provided to administrators, app developers, etc. On-device models can be upgraded, deleted, and tracked independent of app deployment. Additionally, with user permission, observability of on-device models is enabled through endpoint management to detect misuse. The endpoint management solution and model catalog also provide a deployment view of on-device models, including model versions and can be used to ensure compliance.

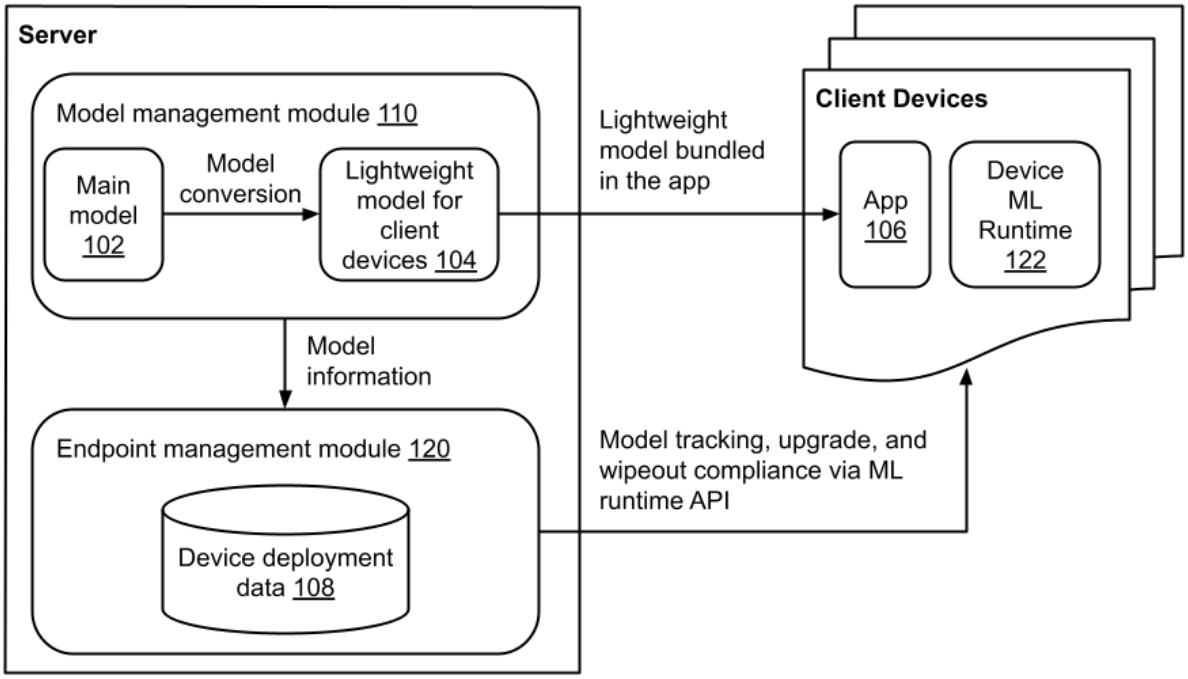

Fig. 1: Functional architecture for management of on-device ML models

Fig 1 illustrates an example functional architecture of a model management solution, per techniques of this disclosure. A server includes a model management module (110) which is in communication with an endpoint management module (120). The model management module includes a main model (102) on a server. Model conversion is performed on the main model to obtain a lightweight model (104) that is suitable for client devices. The lightweight model can be bundled into a client app (106) that can be installed on any number of client devices.

The model management module provides model information to the endpoint management module. The model information available with the endpoint management module is used in combination with device deployment information (108) (that includes data on the devices that have a particular model installed) for model management activities such as model upgrade and wipeout tracking. The ML model details that are part of the device deployment data) can include model name, model version, parent model identifier, model creation/ last

update timestamp, application name (within which the model is used), deployed devices (endpoints), etc.

Further, an on-device ML runtime (122) is provided on the client devices. The runtime provides a mechanism to load, unload, release, and refresh models on the client device. The endpoint management module can integrate with the runtime utility directly or can delegate implementation actions to the app developer via an application programming interface (API). App developers can implement in-app methods to load or unload models based on information from the endpoint management module. The use of an endpoint management module that performs model deployment, refresh, and wipeout can ensure compliance and separation of the ML model from apps that utilize the model.

Model lifecycle management: The described techniques enable model lifecycle management. The server-side main model can be periodically updated, e.g., upon retraining, or other upgrades. The on-device model needs periodic refreshes to track the updated main model to tackle drift and other challenges. The model management module can utilize information about model deployments on endpoint devices that is available in the device deployment data to refresh, update, or delete the on-device model as required. The endpoint management module receives notifications of model version changes through a subscription to the model management module. Upon receipt of such a notification, the endpoint management module can invoke the on-device ML runtime to download the new model and install or delete the model completely if it is no longer needed.

Model security: The on-device app or model can be susceptible to security breaches. With user permission, the endpoint management module can keep track of app activity and model activity on-device. The endpoint management module can generate alerts (e.g., sent to administrators)

in case of suspected security breaches. The administrators can invoke an on-device wipeout workflow via the endpoint management module. The workflow can be used for model deletion as well as model updates, to ensure model compliance and to mitigate security threats.

The techniques described in this disclosure can be used to manage on-device model deployments and integrate on-device model management with endpoint management. The benefits include the ability to manage ML model lifecycle in a manner similar to that for other assets that are part of an enterprise deployment. The described techniques enable administrators and/or app developers to manage model upgrades, wipeout, refresh, etc. seamlessly with minimal effort. The techniques can also prevent security threats and exploits. This approach can also be utilized for machine learning models deployed on edge devices such as Internet-of-Things (IoT) devices.

CONCLUSION

This disclosure describes techniques to integrate the best-in-class capabilities of cloud ML model management and endpoint management to enable lifecycle management and security of on-device ML model deployments. Endpoint management solutions as described herein include the capability to manage on-device models, e.g., to perform tasks such as model tracking, upgrade, and wipe out compliance. The described techniques, which can be implemented as part of an endpoint management solution, use a model catalog to track device deployments for a given machine learning model. When a model upgrade is available, a notification is provided to administrators, app developers, etc. On-device models can be upgraded, deleted, and tracked independent of app deployment. Additionally, with user permission, observability of on-device models is enabled through endpoint management to

detect misuse. The endpoint management solution and model catalog also provide a deployment view of on-device models, including model versions and can be used to ensure compliance.

REFERENCES

1. "Endpoint management - MDM solution" available online at https://workspace.google.com/intl/en_in/products/admin/endpoint/ accessed July 4, 2023.
2. "Managing models and jobs | AI Platform Prediction | Google Cloud" available online at https://cloud.google.com/ai-platform/prediction/docs/managing-models-jobs accessed July 4, 2023.
3. "Neural Networks API | Android NDK" available online at https://developer.android.com/ndk/guides/neuralnetworks accessed July 4, 2023.