**Agenda item:** 9.7

**Source:** vivo

**Title:** Update on AI/ML Model Distribution

**Document for** Discussion andAgreement

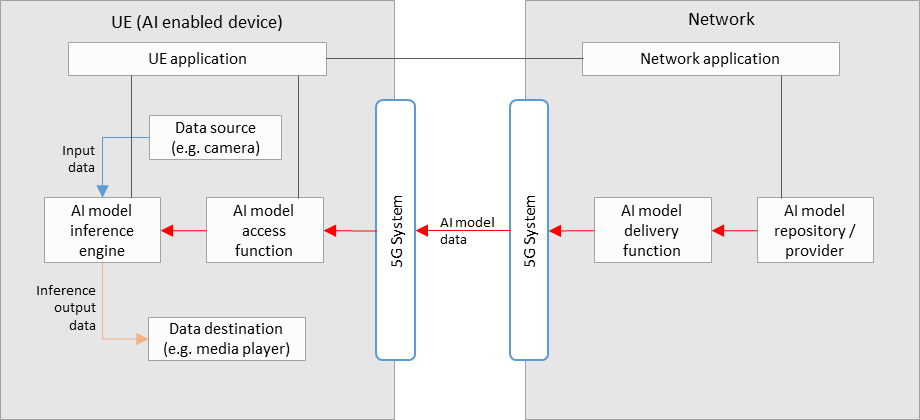
1. Introduction

In last meeting, basic AI/ML model distribution has been discussed, but the current service flow has only considered the AI model distribution at the first time, and the update of AI model during the service lifetime has not been discussed. So we propose to add the description of AI/ML model update description.

1. Proposed changes

--------------------------------------------- Start of Change ----------------------------------------------------------------------------

5.2.1 Complete/Basic AI/ML model distribution



**Figure 5.2.1-1: Basic architecture for AI/ML model delivery with inference in the UE**

Figure 5.2.1-1 shows a simple basic architecture for AI/ML model delivery, as described in scenario 1) of clause 5.2, with an inference of a pre-trained AI/ML model in the UE, as described in scenario 2a) of clause 5.2.

In the network:

* An AI model in the repository is selected for the AI media service by the network application, and sent to the delivery function for delivery to the UE. Selection of an AI model could depend on UE and network characteristics, such as the memory and CPU capability/availability, as well as current network load and performance status.
* The AI model delivery function sends the AI model data to the UE via the 5GS. This delivery function may also contain functionalities related to QoS requests and monitoring, as well as those related to the optimization or compression of AI model data.

In the UE:

* A UE application provides an AI media service using the AI model inference engine and AI model access function.
* The AI model access function receives the AI model data via the 5G system, and sends it to the AI model inference engine. Receiver side optimization or decompression techniques for AI model data may be included.
* The AI model inference engine performs inference by using the input data from the data source (e.g. a camera, or other media source) as the input into the AI model received from the AI model access function. The inference output data is sent to the data destination (e.g. a media player).

Depending on the exact service scenario, AI model updates may be necessary during the service, and different AI model data delivery pipelines may be considered for such purposes. An AI model update may consist of a change in the AI model structure without disrupting the AI media service. If the AI model has requirements on UE memory, processing/computing capabilities or if running the AI model will increase the UE’s power consumption dramatically which will also influence the user experience of other services, it may actively request the update of the AI Model. For example, when the memory usage of the UE processing the AI Model exceeds a certain threshold, or if UE performance deteriorates, the UE can actively send a request to the network for an AI Model update. Alternatively, the network may also trigger the AI model update itself, where an interaction between the UE and network side might be needed to help the network collect current UE status information, e.g. Memory, CPU, current load, terminal location, current power consumption, current battery storage, etc.

--------------------------------------------- End of Change ---------------------------------------------------------------------------

1. Proposal

# We propose to update clause 5 of the permanent document with the above proposed changes.