**Source:** Interdigital Finland Ory

**Title:** [FS\_AI4Media] Object Recognition in Image and Video use-case update

**Agenda item:** 9.7

**Document for:** Discussion and Agreement

1. Introduction

This contribution proposes to clarify the different scenarios of clause 3.1 of the permanent document with respect to the original use-cases described in SA1, and to add additional details and architecture impacts.

1. Proposed changes

--------------------------------------------- Begin Change ----------------------------------------------------------------------------

3.1 Object Recognition in Image and Video

In this set of use cases, images and video streams are processed to identify and recognize objects and extract some metadata, such as bounding boxes, object labels, movement counters, etc.

However, the computationally intensive and memory and power consuming AI/ML inference used to perform this processing requires offloading some inference parts from the mobile device to the edge ora cloud datacenter. This relies on split inference of trained ML model(s) for object recognition between multiple endpoints, typically between the network and UE where a model composed of a set of layers may be split after identified layer points. The decision on how to split the model (i.e., identifying the model split points may depend on various factors including UE capabilities, network conditions, model characteristics and user/task specific requirements:

* Device/UE capabilities on running whole or part of model such as the required memory, the processing capabilities, the energy consumption, and the inference latency.
* Network conditions for delivering media and/or the intermediate data. This includes the amount of data to transfer in one shot for an image or at a specific frame rate for video, the required bandwidth in UL and/or DL with different impact on the network load and the related UL and DL network latencies. Network inference latency is also to consider.
* Model characteristics include split inference with a task-specific model head running on the UE for object recognition. For example, in one UE, the task is to recognize pedestrians, whereas in another it is to recognize traffic signs. The core of the network model as well as the input image/video are the same, but the tasks (and their required task-specific models) in the UEs are different.
* User or task specific requirements. For example, it may be necessary to perform some processing tasks on end-device in order to preserve privacy or because they are delay sensitive operations.

Two main scenarios, both involving either image or video processing are proposed:

1. The UE captures image or video and first feeds the input data to the UE inference model (e.g., to preserve privacy). The UE then uploads intermediate output data from the UE inference model to the network inference, which in turn executes the remaining part of the model (e.g., process the intensive computations) and finally returns the results or a processed image/video to the UE.
2. Unlike the previous scenario, the device/UE pre-processes input data and uploads the processed media to the network where a network inference processes inputs video/image from different scale factor, then sends back the intermediate data to the UE inference executing the remaining layers of the model (e.g., task specific operations) and returning the final results.

These scenarios involve the key operation of AI/ML model/data distribution and require the delivery of trained ML model(s) for object recognition to the UE in 5GS, including the selection of models for different tasks or environments and the selection of the split points based on the various factors described above

These scenarios also involve the distribution of online training of image and video recognition models based on input from different UEs. Depending on the configuration of the ML training framework, different data may need to be delivered between the UEs and the network. Typically, a shared model in the network is calibrated continuously based on the training results from all UEs. This scenario involves all the three key operations related to AI/ML model distribution, splitting, and distributed/federated learning.

--------------------------------------------- End Change ----------------------------------------------------------------------------

1. Proposal

We propose to update sub clause 3.1 of the permanent document with above proposed changes.