**3GPP TSG-SA WG4 Meeting #129-eS4-241558\_r1**

**E-meeting, 19 - 23 August 2024**

**Source: Samsung Electronics Co., Ltd.**

**Title: [FS\_AI4Media] On collaboration scenarios and use cases**

**Agenda item: 9.6**

**Document for: Agreement**

**1. Introduction**

After SA4 #128 during an offline discussion it was agreed that several collaboration scenarios for AI4Media would be defined, and 3 collaboration scenarios were agreed into the TR as in S4aV240029. This contribution aims to describe the relevance of the different use cases in TR to each collaboration scenario.

**2. Discussion**

The current 3 collaboration scenarios are copied below for reference:

Collaboration Scenario 1: AI/ML OTT

In this collaboration scenario, the AI/ML service is offered completely over-the-top. The service provider deploys their own application servers in the cloud and offers a mobile application to their end-users. The MNO may provide assistance to these sessions by allocating the appropriate traffic handling for the identified application streams, including QoS allocation to meet the application’s requirements.

Collaboration Scenario 2: AI/ML Hosting

In this collaboration scenario, the MNO offers CDN operations to perform large scale distribution of AI/ML data, such as trained modes, model updates, etc.

The MNO is responsible for ensuring the right AI/ML data is delivered to the end device at the right time. The service is required to identify the capabilities of the receiving device and match it to the distributed AI/ML data, e.g. a UE that only supports 8-bit quantized models in TensorFlow Lite will receive the matching variant of the ML model.

Collaboration Scenario 3: MNO-operated AI/ML Services

In this collaboration scenario, the MNO is offering the AI/ML-based service and is responsible for distributing the AI/ML model, updates and the corresponding data. The AI/ML component may be part of one of the operator services, such as the multimedia telephony service (MTSI). It may also be a background service, e.g. one that optimizes the network usage by sharing traffic data and performing federated learning. Another possibility is that the AI/ML component is part of a completely new MNO-offered service, e.g. to support IoT, autonomous driving, or cloud media processing and storage.

Use cases from clause 4.2

Clause 4.2 of TR 26.927 v0.8.0 describes a set of use cases for AI4Media, with multiple scenarios under each use case depending on the configurations described.

Below we list and describe the collaboration scenarios (CS) relevant for each of the use case’s scenarios.

Object recognition in image and video

* UE inference only: collaboration scenarios: 1, 2
  + CS 1: the AI model is delivered to the UE over the top using service provider deployed application servers.
  + CS 2: the MNO offers CDN operations to deliver the AI model to the UE according to its capabilities and requirements.
* Network inference only: collaboration scenario 3
  + CS 3: the MNO offers an AI/ML service which supports the complete offloading of AI inferencing in the network.
* Split inference
  + CS 3: the MNO offers an AI/ML service which supports the delivery of the appropriate (partial) AI model to the UE, together with the configuration and support of split inferencing between the UE and the network.

Video quality enhancement in streaming

* End-to-end neural network-based video coding: both sender and receiver apply part of a DNN model
  + CS 3: the MNO offers an AI/ML service which supports the delivery of the appropriate AI model to the UE, together with the configuration and support of AI inferencing in the network (typically the sender).
* Neural network based post-processing for video coding: receiver AI post-processing only
  + CS 1: the AI model is delivered to the UE over the top using service provider deployed application servers.
  + CS 2: the MNO offers CDN operations to deliver the AI model to the UE according to its capabilities

Crowd-sourcing media capture

* Device inference
  + CS 2: the MNO offers CDN operations to deliver the AI model to the UE according to its capabilities and overall service configuration.
* Network inference
  + CS 3: the MNO offers a complete crowd-sourcing media capture service where media captured on multiple devices are aggregated for AI inferencing in the network.

Natural Language Processing (NLP) on speech

* Network based distributed/federated learning
  + CS 3: the MNO offers a new service providing network based distributed/federated learning for this use case.

**3. Proposal**

It is proposed to agree the following changes to 3GPP TR 26.927 v0.8.0.

\* \* \* First Change \* \* \* \*

## x.x AI/ML collaboration scenarios

### x.x.y Relevance of use cases to collaboration scenarios

Depending on the service configuration, each of the scenarios under the use cases in clause 4.2 may fall under one or more of the collaboration scenarios defined in clause x.x.x.

The relevance of each use case scenario to the collaboration scenarios are described in detail in this clause.

Object recognition in image and video:

* UE inference only: collaboration scenarios: 1, 2
  + CS 1: the AI model is delivered to the UE over the top using service provider deployed application servers.
  + CS 2: the MNO offers CDN operations to deliver the AI model to the UE according to its capabilities and requirements.
* Network inference only: collaboration scenario 3
  + CS 3: the MNO offers an AI/ML service which supports the complete offloading of AI inferencing in the network.
* Split inference
  + CS 3: the MNO offers an AI/ML service which supports the delivery of the appropriate (partial) AI model to the UE, together with the configuration and support of split inferencing between the UE and the network.

Video quality enhancement in streaming:

* End-to-end neural network-based video coding: both sender and receiver apply part of a DNN model
  + CS 3: the MNO offers an AI/ML service which supports the delivery of the appropriate AI model to the UE, together with the configuration and support of AI inferencing in the network (typically the sender).
* Neural network based post-processing for video coding: receiver AI post-processing only
  + CS 1: the AI model is delivered to the UE over the top using service provider deployed application servers.
  + CS 2: the MNO offers CDN operations to deliver the AI model to the UE according to its capabilities

Crowd-sourcing media capture:

* Device inference
  + CS 2: the MNO offers CDN operations to deliver the AI model to the UE according to its capabilities and overall service configuration.
* Network inference
  + CS 3: the MNO offers a complete crowd-sourcing media capture service where media captured on multiple devices are aggregated for AI inferencing in the network.

Natural Language Processing (NLP) on speech:

* Network based distributed/federated learning
  + CS 1: the required data for the service (partially trained AI model, training data etc.) is delivered to and from the UE over the top using service provider deployed application servers.
  + CS 3: the MNO offers a new service providing network based distributed/federated learning for this use case.

\* \* \* End of Changes \* \* \* \*