Technical Disclosure Commons

Defensive Publications Series

May 2023

NETWORK-AUTHORIZED SPLIT DATA RATE IN ARTIFICIAL INTELLIGENCE/MACHINE LEARNING RENDERING

Sri Gundavelli

Vimal Srivastava

Follow this and additional works at: https://www.tdcommons.org/dpubs_series

Recommended Citation
Gundavelli, Sri and Srivastava, Vimal, "NETWORK-AUTHORIZED SPLIT DATA RATE IN ARTIFICIAL INTELLIGENCE/MACHINE LEARNING RENDERING", Technical Disclosure Commons, (May 10, 2023)
https://www.tdcommons.org/dpubs_series/5881

This work is licensed under a Creative Commons Attribution 4.0 License.
This Article is brought to you for free and open access by Technical Disclosure Commons. It has been accepted for inclusion in Defensive Publications Series by an authorized administrator of Technical Disclosure Commons.

NETWORK-AUTHORIZED SPLIT DATA RATE IN ARTIFICIAL INTELLIGENCE/MACHINE LEARNING RENDERING

AUTHORS:
Sri Gundavelli
Vimal Srivastava

ABSTRACT

Convolutional Neural Network (CNN) models are widely used for image/video recognition tasks on mobile devices. The CNN is split into two parts - the computation-intensive and energy-intensive parts are offloaded to a network server while the privacy-sensitive and delay-sensitive parts are maintained at the end device. Techniques described herein allow a network to control the split for a user device in artificial intelligence (AI)/machine learning (ML) rendering based on a number of factors.

DETAILED DESCRIPTION

Image and video data are the largest sources of data on today's Internet. Videos account for over 70% of daily Internet traffic. CNN models have been widely used for image/video recognition tasks (e.g., image classification, image segmentation, object localization and detection, face authentication, action recognition, enhanced photography, virtual reality (VR)/augmented reality (AR), video games, etc.) on mobile devices. Many references have shown that AI/ML inference for image processing with device-network synergy can alleviate the pressure of computation, memory footprint, storage, power and required data rate on devices, reduce end-to-end latency and energy consumption, and improve the end-to-end accuracy, efficiency and privacy when compared to the local execution approach on either side.

According to the current image recognition task and environment, the CNN is split into two parts. The intention is to offload the computation-intensive and energy-intensive parts to a network server while maintaining the privacy-sensitive and delay-sensitive parts at the end device. The device executes the inference up to a specific CNN layer and sends the intermediate data to the network server. The network server runs through the remaining CNN layers.

Consider a case in which a device wants to split at layer 1 out of a total of 17 layers of CNN AlexNet processing. While this reduces the processing on the device (which requires less battery power), it increases the bandwidth requirement to transmit the data to the application server for the rest of the processing. Splitting is a function of policy, network load conditions, and user equipment (UE) wireless conditions, so it is logical that the network will decide where to do the split. Since the splitting is dependent on the network, the split is a 5G core (5GC) property and the 5GC controls the splitting. Techniques described herein discuss methods for allowing a network to decide which split model a device should use.

There are many aspects to consider when deciding the split level. Figure 1, below, illustrates a message flow in which a decision of where to do the split is made based on a number of factors.

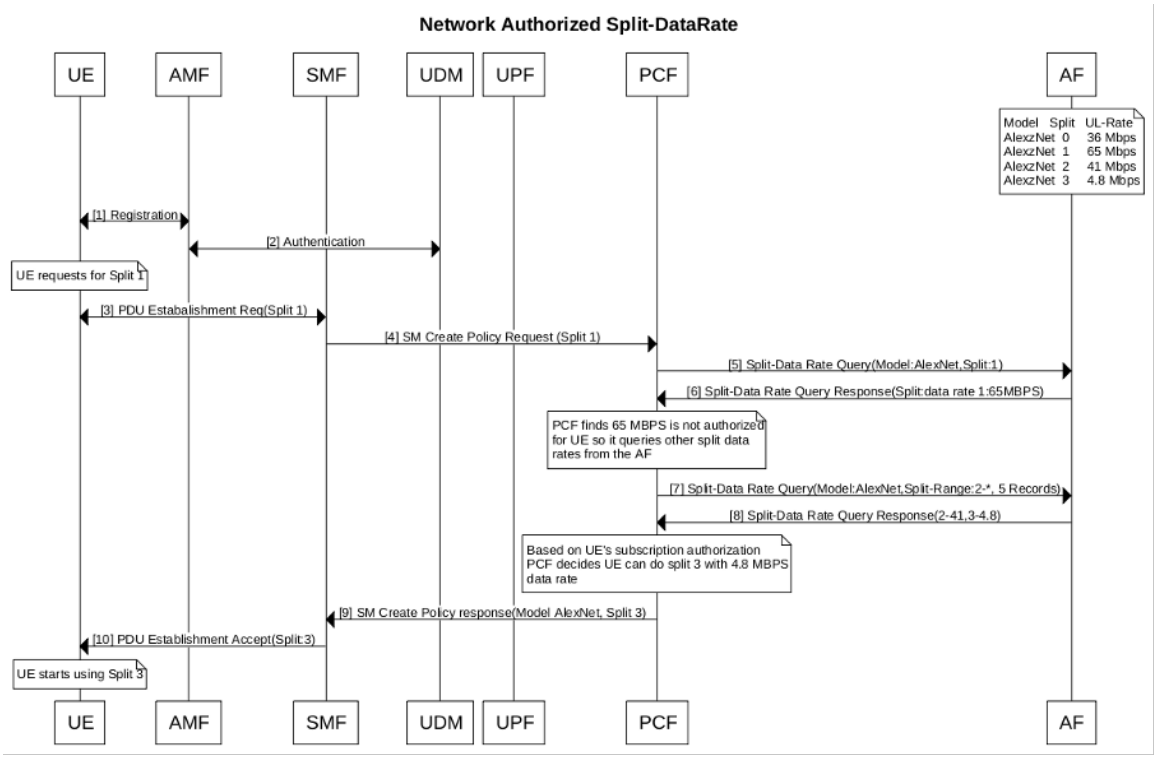

Figure 1: Diagram of Example Message Flow for Determining Split

As illustrated in Figure 1, the network should consider a maximum amount of bandwidth the network can give (policy), real time radio conditions, and a client-determined identification of where to do the split (e.g., based on complexity and software).

The network may provide feedback regarding where to do the split, the client may ask the network where to do the split, and the network may provide information about where to do the split. According to techniques described herein, based on these factors, the network suggests a split to the UE. The UE may choose to deviate from the network-recommended value. However, the network will authorize the split depending upon the UE-subscribed quality of service (QoS) values.

In summary, techniques described herein allow a network to control the split for a UE in AI/ML rendering based on the QoS authorized for the UE and many other factors.