<S1-193039>

Discussion on AI/ML Model Transfer in 5GS

OPPO, CMCC, China Telecom, China Unicom

© 3GPP 2018

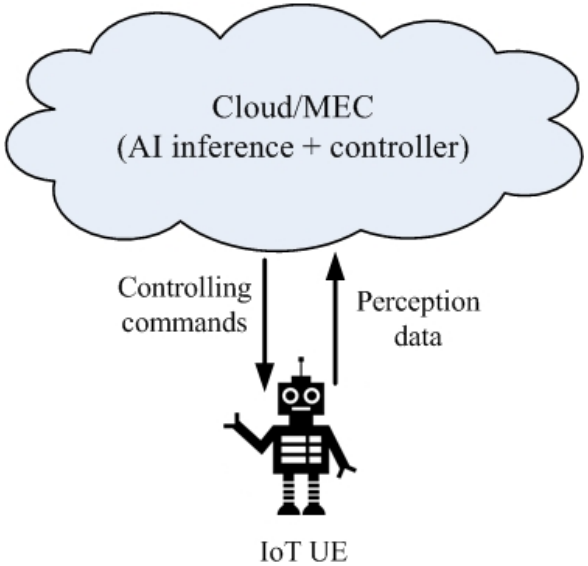
Challenges to AI applications in 5G

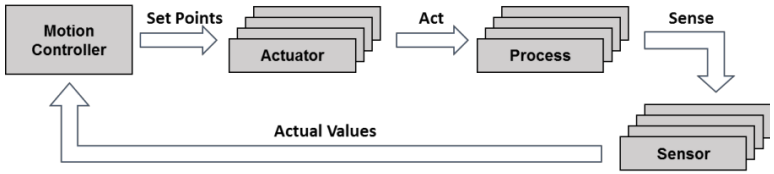
- Artifical Intelligence (AI)/Machine Learning (ML) is taking more and more important tasks in mobile terminals, e.g. photography, image recognition, video call, gaming, AR/VR.
- Three major challenges to effectively support AI/ML applications on 5G terminals (e.g. smartphone, smart car, mobile robot, UAV):
  1. How to do the AI inference if a 5G UE does not have the necessary on-board computation/battery resource?
     - Targeted by Cyber-physical control/URLLC/TSN.
     - But still challenging in actual achievable latency and reliability.
  2. In case a 5G UE does the on-board AI inference, how does it obtain the updated AI/ML model adaptive to the changing task and environment?
     - Need AI/ML downloading, requiring n*100Mbps data rate per UE.
  3. How to train a global AI model across 5G network?
     - Need AI/ML downloading and uploading, requiring n*Gbps data rate per UE and a new types of latency/reliability.

© 3GPP 2018

Cyber-physical control can deal with the challenge 1

- An AI/ML task may require on-board computation/battary resources some UEs cannot afford.
- Solution is to offload AI/ML inference to 5G-cloud/edge.
- SA1 spec for Cyber-physical control [1] identifies requirements to perception uploading and controlling command downloading (and being enabled by R16 URLLC/IIoT)





Requirements to mobile robots in TS 22.104 "Cyber-physical control"

Use case #	Characteristic parameter				Influence quantity							
	Communication service availability: target value in %	Communication service reliability: mean time between failures	End-to-end latency: maximum	Service bitrate: user experienced data rate	Message size [byte]	Transfer interval: lower bound	Transfer interval: target value (note)	Transfer interval: upper bound	Survival time	UE speed	# of UEs	Service area
1	> 99,9999	~ 10 years	< target transfer interval value	—	40 to 250	– < 25 % of target transfer interval value	1 ms to 50 ms	+ < 25 % of target transfer interval value	target transfer interval value	≤ 50 km/h	≤ 100	≤ 1 km²
2	> 99,9999	~ 1 year	< target transfer interval value	—	15 k to 250 k	– < 25 % of target transfer interval value	10 ms to 100 ms	+ < 25 % of target transfer interval value	target transfer interval value	≤ 50 km/h	≤ 100	≤ 1 km²
3	> 99,9999	~ 1 year	< target transfer interval value	—	40 to 250	– < 25 % of target transfer interval value	40 ms to 500 ms	+ < 25 % of target transfer interval value	target transfer interval value	≤ 50 km/h	≤ 100	≤ 1 km²
4	> 99,9999	~ 1 week	10 ms	> 10 Mbit/s	—	—		—	—	≤ 50 km/h	≤ 100	≤ 1 km²

NOTE: The transfer interval is not so strictly periodic in these use cases. The transfer interval deviates around its target value within bounds. The mean of the transfer interval is close to the target value.

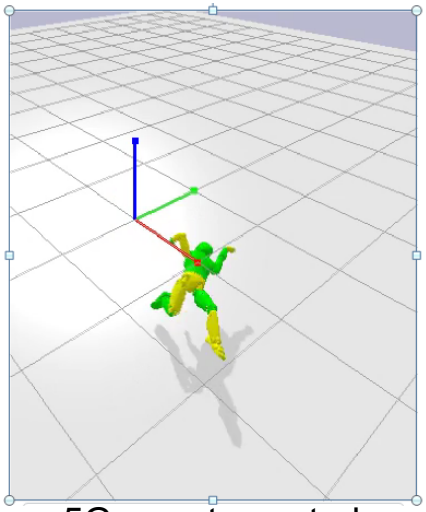
[1] 3GPP TS 22.104 Service requirements for cyber-physical control applications invertical domains; Stage 1 (Release 16)

© 3GPP 2018

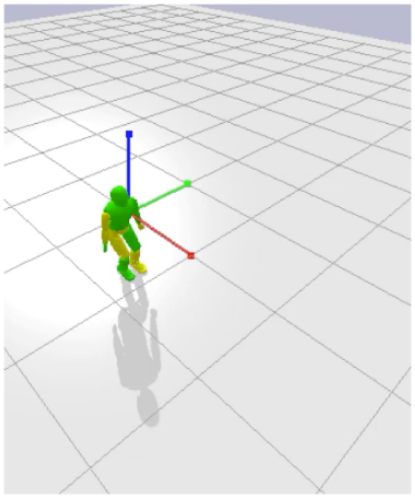
Challenges to clould/edge-based AI/ML offloading

- Low-latency and ultra-reliability has to be realized over "perception – inference – controlling" round trip.
  - ~1ms Latency: URLLC cannot work alone. Widely-deployed MEC is also a must but challenging.
  - 99.9999% reliability: Full coverage required ---- Cannot be provided by 5G mmWave in FR2.

Whole-body robot walking controlled by 5G cloud



5G remote control
without local control



5G remote control
+ local control

With 25ms round-trip latency, 5G remote control cannot work.

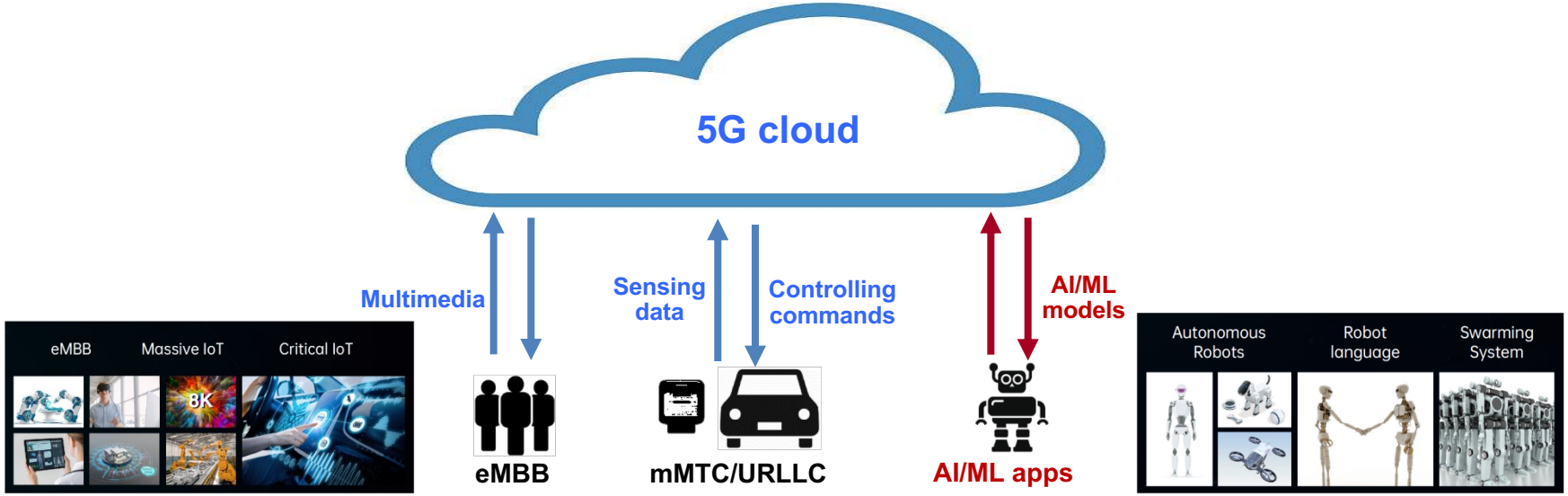
(3ms required in case no local control)

- On-board AI at UE is necessary (sometimes jointly works with on-cloud AI).
- Then Challenge 2 becomes critical.

© 3GPP 2018

Dealing with Challenge 2 and 3 requires transferring AI/ML models over 5G

- Downloading/uploading AI/ML models from/to 5G requires enhacing capability of 5G [2].
- AL/ML model (e.g. neural network) is an emerging traffic type not considered in IMT-2020 requirements [3]
  - 5G is designed for conveying voice, video, AR/VR, sensing data, controlling commands, etc.
  - "AI/ML transfer" may have substantially different service requirements from eMBB and URLLC.



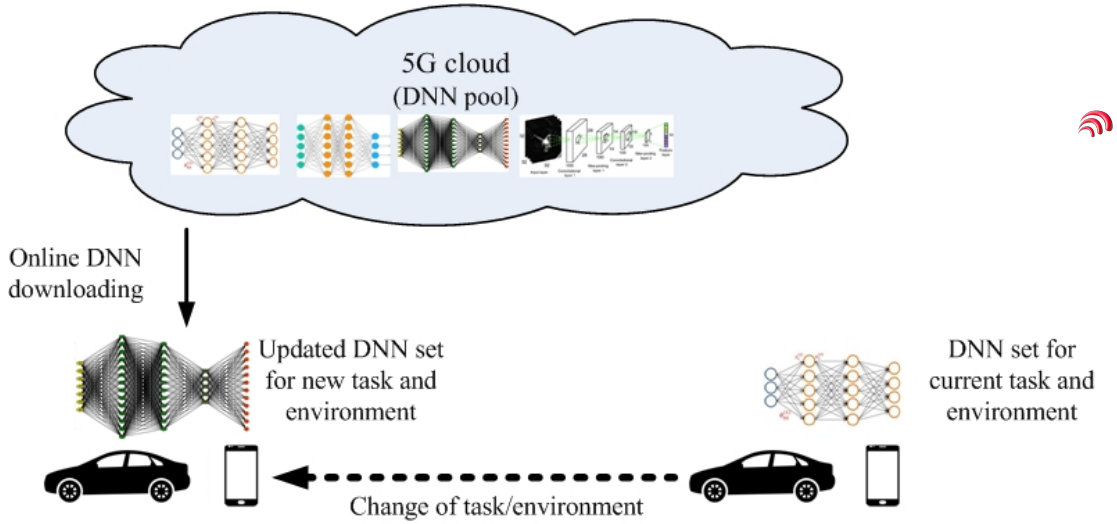
[2] 3GPP TR 37.910 Study on self evaluation towards IMT-2020 submission (Release 16)

[3] Recommendation ITU-R M.2083-0, IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond

© 3GPP 2018

Use case 1: AI/ML model downloading from 5G

- Targeting Challenge 2, online AI/ML model downloading can be used for multi-functional or mobile AI UEs, e.g. smart phones, mobile robots, UAV, smart cars.
  - AI/ML models (e.g. deep neural network, DNN) are often task-dependent and evironment-dependent, e.g.
    - A DNN recognizing a human face is different from a NN recognizing a vehicle license plate
    - A DNN recognizing US traffic signs needs to be re-trained to recognize EU traffic signs
    - Different models are used for online interpretion with different voices.
  - A UE cannot preload all possible AI/ML models due to limited on-board storage. ---- Model updating or transfer learning (partial updating) needed.

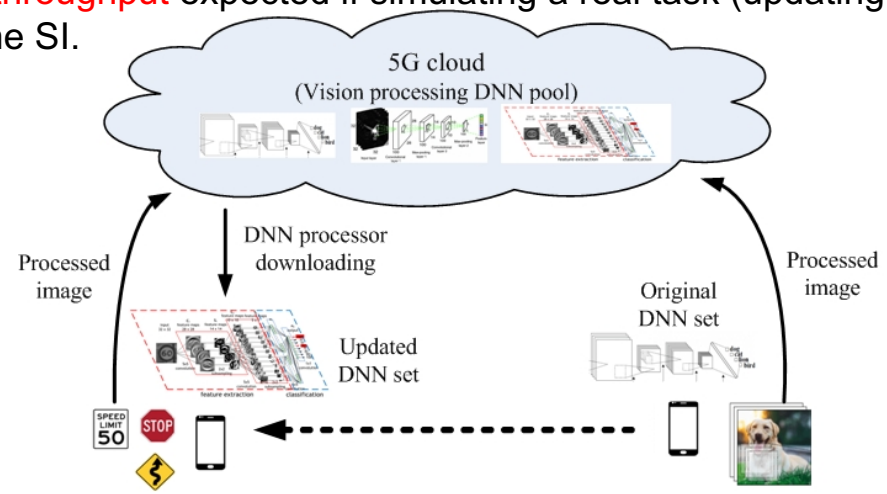


- ***Traffic type: Whole/Partial DNN downloading from 5G cloud/edge.***

© 3GPP 2018

Example study results on Use case 1
---- ML-based vision processor downloading

- ML-based processing has been widely used for vision apps (e.g. photography, image recognition, video call, gaming, AR/VR).
- However, a large number of various DNNs are used for different tasks (e.g. image recovery vs. classification), environments, and even target codec rates.
- Under a limited storage capacity at UE, on-board DNN set needs to be frequently updated,.
- 250 ~ 500Mbps/user DL throughput required if downloading a DNN within 1s.
- Even higher user throughput expected if simulating a real task (updating as needed under a given memory limit) in the SI.



Convolutional Feature Extraction Net

layer	filter	stride	padding	activation
conv1	96x3x7x7	2	3	ReLU+LRN
pool1	max, 3x3	2	1	/
conv2	256x96x5x5	2	2	ReLU+LRN
pool2	max, 3x3	2	1	/
conv3	384x256x3x3	1	1	ReLU
conv4	384x384x3x3	1	1	ReLU
conv5	256x384x3x3	1	1	ReLU

Convolutional Region-proposal Net

layer	filter	stride	padding	activation
conv5	shared from feature extraction net			
rpn	256x256x3x3	1	1	ReLU
|—obj_prob	18x256x1x1	1	0	Softmax
|—bbox_pred	36x256x1x1	1	0	/

Fully-connected Net

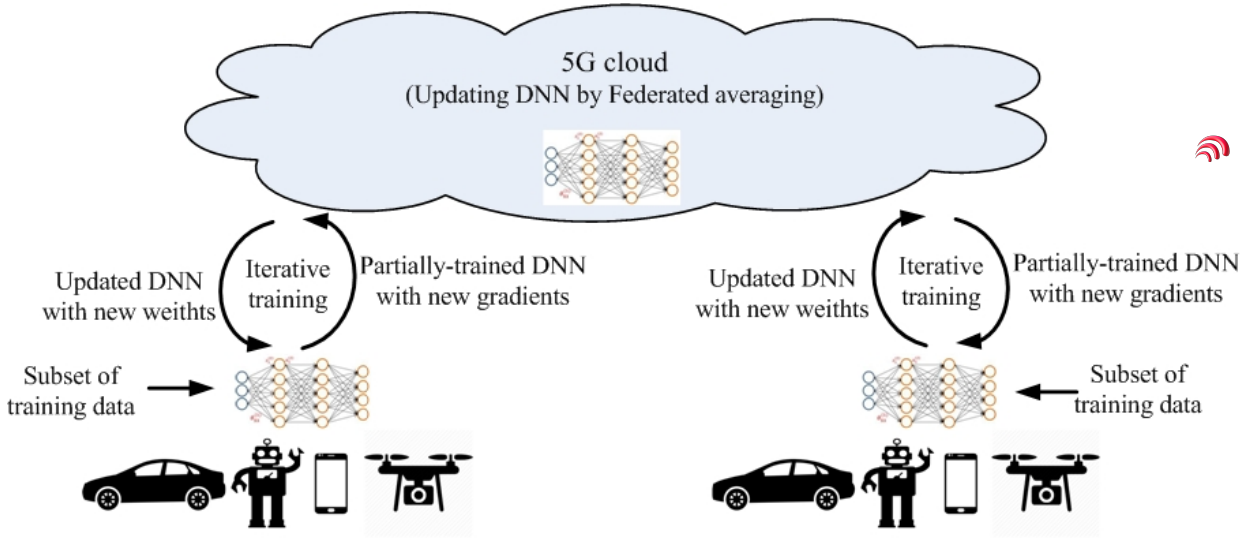
layer	#neurons	activation
conv5	shared from feature extraction net	
roi_pool	256x6x6	/
fc6	4096	ReLU
fc7	4096	ReLU
|—cls_prob	#classes	Softmax
|—bbox_regr	4#classes	/

Scenario	Required communication service availability in %	Service bitrate: user experienced data rate	Remarks
Neural network downloading (AlexNet Pascal Voc)	99.999% @ ~0% acc. 99.99% @ 0.0001% acc 99.9% @ 0.04% acc. 99% @ 0.11% acc.	500 Mbps (8-bit) refreshed using raw data ~250Mbps (8-bit) refreshed using sparse coding	For a single UE

© 3GPP 2018

Use case 2: Distributed/Federated Learning over 5G

- Distributed/Federated Learning is essential for the AI apps which need a global model (e.g. autonomous driving, voice/vision recognition) because:
  - Distributing ML over terminals can reduce computation burden at AI server.
  - Privacy protection rules may not allow local training data to be uploaded to cloud.



- ***Traffic type:***
  - ***Updated DNN downloading.***
  - ***Partially-trained DNN uploading.***

- Federated Learning over 5G network:
  - AI server exchanges the updates of a DNN with its federated UEs via 5G network.
  - Iteratively making the training converge requires very frequent server↔UE mutual transfer, resulting a very high data rate in UL and DL.

© 3GPP 2018

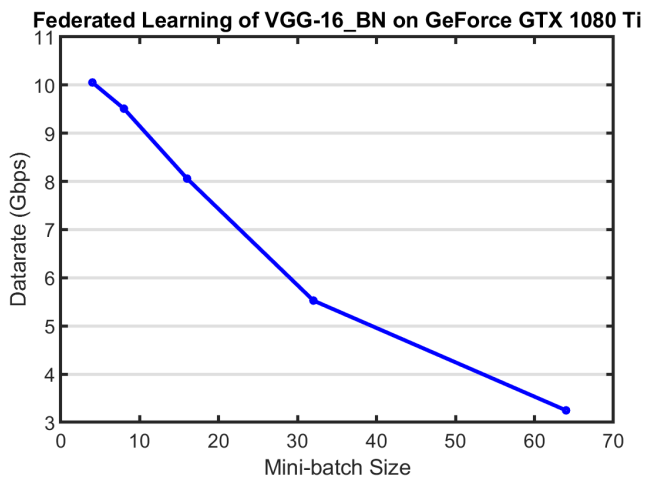
Example study results for Use case 2
---- Iterative Federated Learning

- 3~10Gbps data rate per UE is required in UL and DL. Very high system capacity is required in case of a number of training UEs in a cell (depends on geographicc distribution of training UE).
- Low latency for catching up the ML pipeline, and limited training time for mobile.
- Group-user latency/reliability is required, rather than single-user latency/reliability.
- URLLC-level multicast can be used for DL model distribution.

Simulated iteration procedure:

- A UE trains the DNN based on the subset of training data it obtains with a mini-batch size
- The partially-trained DNNs (with new gradients) from all UEs are uploaded to 5G cloud.
- ML server broadcasts the DNN with updated weights to UEs for the next iteration.

UL data rate per UE required to meet the training timeline (20x~100x higher than NR capability)



Raw data
8-bit VGG16_BN size: 132MB
Input: 224X224X3 images

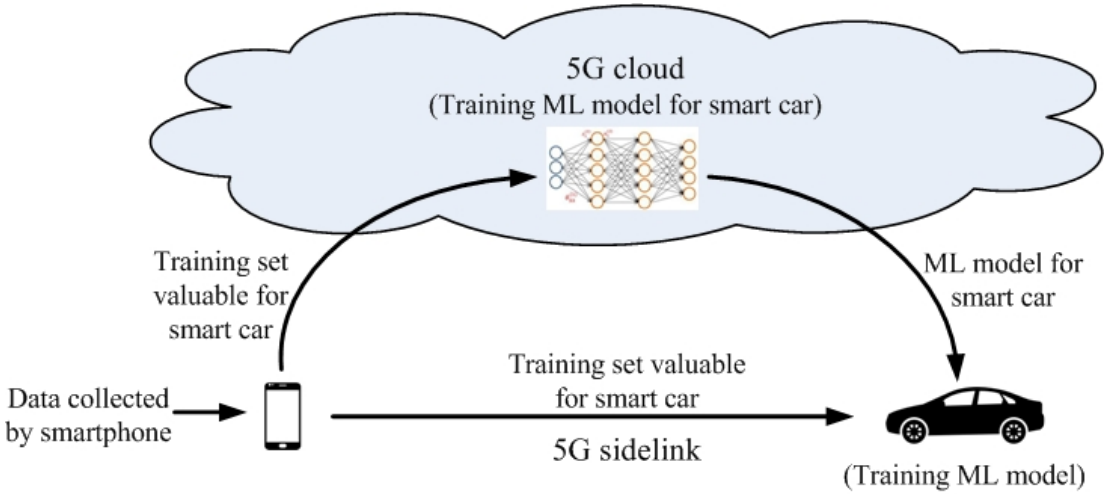
Batch size	GPU computation time*
64	0.325s
32	0.191s
16	0.131s
8	0.111s
4	0.105s

*GPU computation includes:
1 forwarding+1 back propagation on the mini-batch of images

© 3GPP 2018

Use case 3: Training data sharing across 5G

- Collaborative ML cannot be performed across different types of intelligent agents.
  - E.g. a smartphone cannot trains a model for a smart car, although the training data collected by the phone is valuable for the training at the car.
- The training data set can be shared between different types of agents if not violating the privacy protection rules.
  - Training set, as a “structured data”, can be regarded as another type of AI model, but with a much larger payload than neural network.
  - D2D training set sharing over 5G sidelink may be more agreeable in terms of privacy protection.



- ***Traffic type:***
  - ***Training data uploading.***
  - ***Training data sharing over sidelink***

© 3GPP 2018

Summary of potential new service requirements

- As an important new traffic type in AI era, AI/ML model transfer calls for new service requirements to 5G NR, e.g.
  - **New data rate/capacity requirements:** 250Mbps ~ 10Gbps DL and UL throught per UE. And a considerable number of UEs in a cell with concurrent transmissions.
  - **New type of latency/reliability requirements:** Group-user latency/reliability is required, rather than single-user latency/reliability.
  - **New coverage requirements**: Tolerance to non-contiguous coverage can be enabled by predictively pre-downloading AI/ML models to UE or selecting training UEs in NR mmWave hotspots, which further explores the potential of FR2 spectrum for operators.
  - **New requirements on multicast:** Multicast URLLC may be required.
  - **New privacy/security requirements:** Considering risk of reversely inferring data from a AI/ML model, a different privacy/security mechanism is needed.

- Meanings behind data (e.g. ML models) will become a more valuable asset than unstructured raw data. Conveying AI/ML models in 5G network gives operators an opportunity to access and operate the strategic asset. But requirements of this new service needs to be studied and identified in advance.

© 3GPP 2018

Proposed objective of the study item

- The aim of this work is to study the use cases and to propose potential service requirements to Artificial Intelligence (AI)/Machine Learning (ML) model transfer in 5G system.
- The objectives include, but are not limited to:
  - Identifying use cases for AI/ML model distribution and transfer, including:
    - Split inference between network and UE;
    - Online AI/ML model downloading and updating over 5GS;
    - Distributed/Federated Learning over 5GS;
    - Training data sharing over 5GS.
  - Studying traffic characteristics of AI/ML model, including investigation of existing models and traffic properties.
  - Gap analysis on KPI/QoS requirements for AI/ML model distribution and transfer, including:
    - Data rate, latency, reliability and capacity for AI/ML model downloading/uploading.

NOTE: Security aspects, if any specific ones are identified, can also be investigated.

© 3GPP 2018

Proposed work plan

- SA1#88
  - Agree on SID.
- SA1#89
  - Use case gathering.
- SA1#90
  - Use case gathering.
  - Analysis on service requirements.
- SA1#91
  - Analysis on service requirements.
- SA1#92
  - Overview. Consolidating service requirements. Draw conclusions.
  - Finalization of the study.

© 3GPP 2018