3GPP TSG RAN Meeting #102
Edinburgh, GB, December 11 – 15, 2023
Agenda Item: 9.1.1.1

RP-232961

Discussion of R19 work on AI/ML for Air Interface

Futurewei Technologies, Inc.

FUTUREWEI Technologies

Contents

- Rel-19 AI/ML for Air Interface Scoping: Things to Consider
- Summary of Rel-18 agenda items/use cases
  - Observations for remaining open issues
- Rel-19: Manner of work
- Proposal of AI/ML for Air Interface for Rel-19

Rel-19 AI/ML for Air Interface Scoping: Things to Consider

- Workload of the involved WGs
  - Overloading WGs with many (sub) use cases destines to poor quality of work.
  - Note that for each sub-use case there are many variations / alternatives yet to be narrowed down.
- Performance benefits
  - Sufficient performance benefits in practical deployment scenarios/assumptions are needed for a use case to be in the normative work
  - Need to justify the complexity and substantial amount of effort in standardization, implementation options, deployment scenarios and LCM/inter-operability efforts.
- Readiness for normative work
  - Many issues remain to be addressed before normative phase
  - In addition to RAN1, RAN4 and RAN2 can also be the bottlenecks
  - Very broad and vague objectives (with many potential variations and options) for WID is very risky. Need to be able to formulate more specific/desired objectives.
- Manner of AI/ML in 5G-Advanced
  - AI/ML in 5G-Advanced is not native but add-on for selected features
  - Heavy/comprehensive framework with new design in many aspects of the system is not necessary
  - Realizing **few** useful AI/ML-enabled feature(s) in practical deployment networks is more important than supporting many AI/ML features.

Observations: CSI Enhancement (1/2)

- CSI Compression sub use case:
  - **AI/ML-based approach provides <u>significant CSI overhead reduction</u>.**

    AI/ML based CSI compression achieves CSI feedback reduction compared with Rel-16 eType II CB in general under 1-on-1 joint training and generalization Case 1, where 4 sources observe the CSI feedback overhead reduction of 10.24%~60%/10%~58.33%/8%~79% for Max rank 1/2/4, respectively, under FTP traffic.

  - **AI/ML-based approach achieved <u>small to moderate mean UPT gain</u>** over the BL, i.e., Rel-16 Type II codebook-based approach.

    - *0.2%~2%/-0.3%~6%/-4%~6% gains of mean UPT are observed for Max rank 1/2/4, respectively, under RU≤39%.*
    - *0.1%~4%/-0.5%~10%/-1.8%~12.22% gains of mean UPT are observed for Max rank 1/2/4, respectively, under RU40%-69%.*
    - *0.23%~9%/-0.2%~15%/-1%~17% gains of mean UPT are observed for Max rank 1/2/4, respectively, under RU>70%.*

  - **<u>Large range of complexity</u> was observed** for AI/ML-based approach among results from companies:

    - *A majority of 25 sources adopt the <u>CSI generation</u> model subject to the <u>computational complexity in units of FLOPs</u> from 10M to 800M, and 26 sources adopt the <u>CSI reconstruction</u> model subject to the FLOPs from 10M to 1100M.*
    - *A majority of 21 sources adopt the <u>CSI generation</u> model subject to the <u>number of parameters</u> from 1M to 13M, and 22 sources adopt the <u>CSI reconstruction</u> model subject to the number of parameters from 1M to 17M.*

Note: *blue* italic text is from agreements reached in RAN1#115 and RAN1#114bis.

FUTUREWEI Technologies

Observations: CSI Enhancement (2/2)

- CSI Prediction sub use case:
  - **Non-conclusive observation for UPT gain** over BL (i.e., nearest historical CSI or non-AI/ML based approach).
    - *AI/ML based CSI prediction outperforms the benchmark of the nearest historical CSI in general, where the majority of sources observe up to 10.6% gain in terms of mean UPT.*
    - *for AI/ML based CSI prediction over non-AI/ML based CSI prediction,* ***3 sources*** *observe 0.7%~7% gain while* ***2 sources*** *observe performance loss of -0.1%~-17% in terms of mean UPT.*
  - **Large range of complexity (computational complexity and model complexity)** for AI/ML-based approach among results from companies:

a majority of sources adopt the model subject to the computational complexity in units of FLOPs from 0.1M to 1000M.

The number of parameters from submitted results is between 0.01M to 18.5M.

Note: *blue* italic text is from agreements reached in RAN1#115.

Observations: Beam Management

- BM-Case1 (spatial-domain beam prediction) sub use case:
  - Performance gain observations are majorly obtained from **ideal measurement assumption and based on best Rx beam**. With more realistic assumption, performance gain and beam sweeping overhead reduction will be smaller.

    - *For BM-Case1 when Set B is a subset of Set A or when Set B is different than Set A, without UE rotation, AI/ML can achieve good performance with measurements of fixed Set B that is 1/4 or 1/8 of Set A of beam measured with best Rx beam for DL Tx beam prediction…*
    - *ideal measurements are assumed in the evaluations (in clause 6.3.2.1): beam could be measured regardless their SNR, no measurement error, and measurements obtained in a single-time instance (within a channel-coherence time interval), no quantization and no constraint on UCI payload (for NW-side model).*

- BM-Case2 (temporal-domain beam prediction) sub use case:
  - Set A = Set B:
    - No UE rotation: AI/ML does not show performance gain over non-AI based approach, at least for prediction window = 80ms or 160ms.
    - With UE rotation: observation is not conclusive with **results from only 2 sources**.

For 80ms or 160ms prediction time:
- Some evaluation results show AI/ML may have similar performance or some degradation

Note: *blue* italic text is from (text) agreements reached in RAN1#114bis.

Observations: Positioning Accuracy Enhancements

- For both direct AI/ML positioning and AI/ML-assisted positioning sub use cases:
  - AI/ML-based positioning's performance is sensitive to scenario/configuration/synchronization errors or even different drops of the same scenario and is likely to suffer accuracy degradation when the model is trained with dataset of one deployment scenario, while tested with dataset of a different deployment scenario and certain mechanisms may need to be leveraged, e.g., better training dataset construction, fine-tuning.
  - Dataset requirement for fine-tuning may depend on the difference between the new deployment scenario and the previous development scenario for which the AI/ML model was trained.

For both direct AI/ML positioning and AI/ML assisted positioning,

- *if the new deployment scenario is significantly different from the previous deployment scenario the model was trained for (e.g., different drops, different clutter parameter, different InF scenarios), fine-tuning a previous model requires similarly large training dataset size as training the model from scratch, in order to achieve the similar performance for the new deployment scenario.*

Note: *blue* italic text is from (text) agreement reached in RAN1#114bis.

Major Remaining RAN1 Issues (as of RAN1 #115) on general aspects of AI/ML framework

- General framework to support online training, over-the-air training, and model update.
- Handling of NW-side and/or UE-side additional conditions
  - For inference for UE-side models, many options were proposed by companies to ensure consistency between training and inference regarding NW-side additional conditions if identified, however, no consensus reached on many aspects due to lack of time.
  - No time to discuss the options regarding UE-side additional conditions to ensure consistency between training and inference.
- Further details of over-the-air (i.e., online) model identification and two-sided model pairing
- For model identification of UE-side or UE-part of two-sided models, example use cases of online identification (Type B1 and Type B2) and offline identification (Type A), limited examples were discussed/agreed due to lack of time.
- Functionality and Model-ID based LCMs:
  - Unified approach/procedure or not
  - The applicability and necessity of Model-ID based LCM
    - Based on functionality-based LCM or independent?
- Model transfer/delivery: some pros and cons were discussed but further study/discussion is needed for the necessity and its applicability to the studied sub use cases.
- Performance monitoring of inactive model(s) and the associated complexity.

Major Remaining RAN2 Issues (as of RAN2 #124)

- Data collection management for model training and performance monitoring:
  - enhancements to existing data collection framework
  - signaling for data collection
- Model delivery/transfer:
  - inter-vendor solutions,
  - security issues (especially via third-party servers)
- Model/Functionality Identifications and LCM:
  - online model identification procedure
  - model-pairing for 2-sided models
  - meta information
  - signaling for LCM
- UE capability enhancements:
  - ways to report dynamic capabilities for data collection, model storage and model inference (postponed to WI phase)
  - definition of AI/ML related Features/FGs (postponed to WI phase)

Major Remaining RAN4 Issues (as of RAN4 #109)

- Options of testing goals were set with many items that require further study.
  - Option 1: The testing goal is to verify whether a specific AI/ML model (if model identification is possible)/functionality can be conducted in a proper way.
    - FFS how to define the specific AI/ML model (e.g., a model captured in RAN4 spec as baseline)
    - FFS how to define that the model is properly conducted (e.g., by defining AI/ML dedicated performance/core requirements associated with model outputs)
  - Option 2: The testing goal is to verify whether the minimum performance gain of AI/ML model (if model identification is possible) /functionality/feature can be achieved for a static scenario/configuration.
    - FFS how to define a static scenario/configuration (e.g., by defining a related testing dataset based on channel models in TR 38.901)
    - FFS whether and how to define non-static specific scenarios/configurations
- Other important RAN4 issues that have not been resolved.
  - Requirements for data collection (what will be the requirements? are they testable? and how to test them?)
  - Requirements for model delivery/update/transfer
  - Performance monitoring/evaluation for LCM (definition of requirements for dynamically changing scenarios, accuracy of monitoring, setting the ground truth, performance degradation)
  - Test encoder/decoder for 2-sided models: the feasibility of any of the testing options has not concluded and more study is required.
  - Testing of interoperability for different collaboration levels.

Summary of Observations (1/2)

Observation #1:

RAN1 has no consensus on on the recommendation of both CSI compression and CSI prediction for normative work, mainly due to no significant performance and/or justification of complexity.

Observation #2:

Up till RAN1#115, some important issues of AI/ML for Air-Interface remain open:

- No conclusion on preferred training collaboration type(s) for two-sided model(s)
- LCM and performance monitoring framework and applicability of functionality-based LCM and model-ID-based LCM, including model pairing for two-side model.
- Handling of NW-side and UE-side additional conditions to ensure consistency between training and inference on UE-side model.
- Model delivery/transfer: necessity and applicability for sub use case(s)
- Desired mechanisms for handling potential performance degradation in model generalization cases.

Observation #3:

Some evaluations are not conclusive yet, for example,

- AI/ML-based CSI enhancement performance gain compared to baseline is very diverse and complexity range is very large.
- AI/ML-based beam management results are from ideal measurement assumptions and/or results are from very few sources. Temporal-domain beam prediction evaluation is not conclusive and no performance gain when Set A = Set B, at least for some cases.
- AI/ML-based positioning's performance is sensitive to scenario/configuration/synchronization errors/different drops, more so for direct AI/ML positioning. While better training dataset construction and/or model fine-tuning/re-training may improve the performance, some degradation is still observed.

Summary of Observations (2/2)

Observation #4:

It may be risky to move directly into normative phase in Rel-19:

- General AI/ML framework: many open issues remain to be further discussed/studied.
- Performance uncertainty in sub use cases and/or observations are drawn from ideal measurement assumptions.
- Many variations/options in studied sub use cases
- Many training collaboration types and variations/options in each type.

Observation #5:

Heavy RAN2 work to cover majority of data collection, LCM-related topics and UE capability enhancements.

Observation #6:

Many important issues identified by RAN4 have not been resolved, including the testing goals. It is hard to say we are ready to move forward when there are so many FFSs in testing goals and know how to do it.

Observation #7:

Even though two-sided model requires more specification work and AI/ML-based CSI compression is the only two-sided model sub use case studied, the work will lay the foundation for future air-interface use cases that leverage two-sided model approach.

Observation #8:

Given that Rel-19 is the first release to support AI/ML-based functionalities, even though RAN1 recommends AI/ML-based beam management use case and AI/ML-based positioning use case for normative work, there are many sub use cases and various alternatives/variations in some sub use cases, down-scoping is needed to reduce the workload in Rel-19.

Manner of work in R19

- Identify a small number of (sub) use cases to be selected for Rel-19 specification work during the extended study phase
  - Performance and/or overhead reduction benefits need to be significant to justify the complexity and specification efforts.
  - For example, for CSI compression, using CSI codebook/look-up-table generated from AI/ML encoder and vector quantization can reduce the feedback overhead by 80-90% (down to ~10-20 bits) compared to Type-II codebook with comparable UPT.
- Light-weight framework with essential LCM functionalities to support selected (sub) use cases in the normative phase for Rel-19 with (limited) consideration on building a general framework for future release(s) and/or 6G.
- Potential Standards work to support the following functionalities:
  - Data collection and delivery/transfer needed at various stages, including AI/ML model training/update, inference and performance monitoring.
  - General life cycle management (LCM) framework supporting essential AI/ML-based features/functions and AI/ML operations across interacting entities and interfaces, including at least:
    - ✓ AI/ML model delivery/transfer for selected (sub) use case where applicable
    - ✓ Define and indicate AI/ML features/functions in UE capability and identify potential enhancements

Proposal of AI/ML for Air Interface for Release 19

- Start Rel-19 with a study phase of AI/ML for air interface for the initial 6+ months, followed by normative work with clearly defined focus and scope for the study.
  - For RAN1, further study at least general AI/ML framework, CSI compression, and CSI prediction.
  - For RAN2, further study at least data collection management, model delivery/transfer, and UE capability enhancements.
  - For RAN4, further study at least FFS points in test goals, requirements for data collection and model delivery/update/transfer, and test method for interoperability for different collaboration levels.
- Consider a small number of sub use cases each with limited variations identified in Rel-18:
  - CSI enhancement: depending on study outcome (Q3-2024)
    - CSI compression: training collaboration type 1 (NW-side) and type 3 (NW-first)
    - CSI prediction: UE-side model
  - Spatial-domain DL Tx beam prediction (BM-Case1)
  - Positioning: AI/ML assisted positioning cases (consider only Case 1, Case 2a, Case 3a)

  <u>Do not consider new (sub) use cases or new variations of existing (sub) use cases</u>
- Support the selected (sub) use case(s) with only essential LCM operations that reuse existing signaling framework.

Appendix

Agreements: General AI/ML Framework (1/3)

Agreement

For model delivery/transfer to UE (for UE-side models and UE-part of two-sided models):

- *Model delivery/transfer to UE, if feasible, may be beneficial to handle scenario/configuration specific (including site-specific configuration/channel conditions) models (i.e., when a single model cannot generalize well to multiple scenarios/configurations/sites), to reduce the device storage requirement.*
- *Model delivery/transfer to UE after offline compiling and/or testing may be friendlier from UE's implementation point of view compared to the case without offline compiling and/or testing. On the other hand, the case without offline compiling and/or testing (that can update parameter with known model structure), may have benefit at least in terms of shorter model parameter update timescale.*
- *For model trained at network side, Case y (w/ network-side training) and Case z2 may incur the burden of offline cross-vendor collaboration such as sending a model to the UE-side and/or compiling a model.*
- *For model trained at UE side/neutral site, Case z1 and Case z3 may incur the burden of offline cross-vendor collaboration to send the trained model from the UE-side to the network, compared to Case y (w/ UE-side training) which does not have such burden.*
- *Model storage at the 3gpp network, compared to storing the model outside the 3gpp network, may come with 3gpp network side burden on model maintenance/storage.*
- *Proprietary design disclosure concern may arise from model training and/or model storage at the network side compared to other cases (such as case y with UE side training) which does not have such issue.*

Note: *blue* italic text is from agreements reached in RAN1#115.

Agreements: General AI/ML Framework (2/3)

Agreement

For model identification of UE-side or UE-part of two-sided models, further clarification is made as follows.

- *The following are example use cases Type B1 and B2*
  - *Model identification in model transfer from NW to UE*
  - *Model identification with **data collection** related configuration(s) and/or indication(s) and/or dataset transfer*
- *Note: Other example use cases are not precluded.*
- *Note: Offline model identification may be applicable for some of the above example use cases*

Note:

1. *blue* italic text is from agreements reached in RAN1#115.
2. More examples are in the original proposal for Type B1 and B2 (e.g., Model identification with monitoring related configuration(s)/ procedure(s), Model identification due to update on UE-side model operations), and examples for Type A but there was not enough time to discussing the remaining ones.

Agreements: General AI/ML Framework (3/3)

Agreement

The following aspects have been studied for the general framework of AI/ML over air interface for one-sided models and two-sided models.

- *Various Network-UE Collaboration Levels*
- *Functionality-based LCM and model-ID-based LCM*
- *Functionality/model selection, activation, deactivation, switching, fallback*
- *Functionality identification and model identification*
- *Data collection*
- *Performance monitoring*
- *Various model identification Types and their use cases*
- *Reporting of applicable functionalities/models*
- *Method(s) to ensure consistency between training and inference regarding NW-side additional conditions (if identified) for inference at UE*
- *Model delivery/transfer and analysis of various model delivery/transfer Cases*

The above studied aspects for General Framework can be considered for developing/specifying AI/ML use cases and common framework (if needed for some aspects) across AI/ML use cases.

Note: *blue* italic text is from agreements reached in RAN1#115.

Agreements: CSI Enhancement

Agreement

Capture the following as a conclusion in section 8 of the TR

- *From RAN1 perspective, there is no consensus on the recommendation of CSI compression for normative work.*
- *At least the following aspects are the reasons for the lack of RAN1 consensus on the recommendation of CSI compression for normative work.*
  - *Trade-off between performance and complexity/overhead*
  - *Issues related to inter-vendor training collaboration*
- *Other aspects that require further study/conclusion are captured in the summary*

Agreement

Capture the following conclusion in section 8 of the TR 38.843

- *From RAN1 perspective, there is no consensus on the recommendation of CSI prediction for normative work.*
- *The reason for the lack of RAN1 consensus on the recommendation of CSI prediction for normative work is due to*
  - *Lack of results on the performance gain over non-AI/ML based approach and associated complexity*
- *Other aspects that require further study/conclusion are captured in the summary.*

Note: *blue* italic text is from agreements reached in RAN1#115.

High-level Observation agreed: Beam Management (1/2)

BM-Case1

Observation

For BM-Case1 when Set B is a subset of Set A or when Set B is different than Set A, without UE rotation, AI/ML can achieve good performance with measurements of fixed Set B that is 1/4 or 1/8 of Set A of beam measured with best Rx beam for DL Tx beam prediction, and with measurements of fixed Set B that is 1/4 or 1/8 or 1/16 of Set A for beam pair prediction. In addition, based on the evaluation results from 2 or 3 sources, for BM-Case1 DL Tx beam prediction, with 1/4 or 1/8 measurement/RS overhead, 96%~99% or 85%~98% of UE average throughput and 95%~97% or 70%~84% of UE 5%ile throughput of non-AI baseline option 1 (exhaustive search over Set A beams) can be achieved according to the predicted beam from AI/ML. Note that, ideal measurements are assumed in the evaluations (in clause 6.3.2.1): beam could be measured regardless their SNR, no measurement error, and measurements obtained in a single-time instance (within a channel-coherence time interval), no quantization and no constraint on UCI payload (for NW-side model).

With some realistic consideration (in section 6.3.2.3):

- *Existing quantization granularity of L1-RSRP causes a minor loss in beam prediction accuracy compared to unquantized L1-RSRPs of beams in Set B at least for BM-Case1 for inference of DL Tx beam prediction.*
- *Measurement errors degrade the beam prediction performance with AI/ML, while measurement errors also degrade the performance with non-AI baseline (both option 1 and option 2).*
- *For DL Tx beam prediction, with the measurements from quasi-optimal Rx beam, some performance degradation (e.g., 2% to up to12% Top-1 beam prediction accuracy loss based on most of results) is observed comparing to with measurements from best Rx beam. If the measurements are from random Rx beam, large performance degradation is observed.*

Note: *blue* italic text is from (text) agreements reached in RAN1#114bis.

High-level Observation agreed: Beam Management (2/2)

BM-Case2

Evaluation results for BM-Case2 when Set B= Set A for DL Tx beam prediction with the measurements from the best Rx beam

	Without rotation	With rotation
Beam prediction accuracy performance compared with non-AI baseline (option 2)	For 80ms or 160ms prediction time: - Some evaluation results show AI/ML may have similar performance or some degradation For 160ms or larger prediction time: - Most evaluation results show AI/ML provides some beam prediction accuracy gain - The longer the prediction time, the higher gain of beam prediction accuracy can be achieved by AI/ML	AI/ML can provide some beam prediction accuracy gain: - The longer the prediction time, the higher gain of beam prediction accuracy can be achieved by AI/ML (2 sources)

Note: above table is excerpted from Table 6.3.2.5-1 in TR38.843 v1.2.0 on “Summary of the evaluation results for BM-Case2 when Set B=Set A for DL Tx beam prediction”.

High-level Observations: UE Positioning (1/2)

Direct AI/ML positioning

Observation

Evaluation of the following generalization aspects show that the positioning accuracy of direct AI/ML positioning deteriorates when the AI/ML model is trained with dataset of one deployment scenario, while tested with dataset of a different deployment scenario.

- *The generalization aspects include:*
  - *Different drops*
  - *Different clutter parameters*
  - *Different InF scenarios*
  - *Network synchronization error*
- *Companies have provided evaluation results which show that the positioning accuracy on the test dataset can be improved by better training dataset construction and/or model fine-tuning/re-training.*
  - *Better training dataset construction: The training dataset is composed of data from multiple deployment scenarios, which include data from the same deployment scenario as the test dataset.*
  - *Model fine-tuning/re-training: the model is re-trained/fine-tuned with a dataset from the same deployment scenario as the test dataset.*

Note: ideal model training and switching may provide the upper bound of achievable performance when the AI/ML model needs to handle different deployment scenarios.

Note: *blue* italic text is from clause 6.4.2.2 of TR38.843 v1.2.0.

High-level Observations: UE Positioning (1/2)

AI/ML-assisted positioning

For AI/ML assisted positioning with timing information (e.g., ToA) as model output, evaluation of the following generalization aspects show that:

- *the positioning accuracy deteriorates when the AI/ML model is trained with dataset of one deployment scenario, while tested with dataset of a different deployment scenario.*
  - *Different drops*
  - *Different clutter parameters*
  - *Different InF scenarios*
- *the positioning accuracy may or may not deteriorate when the AI/ML model is trained with dataset of one deployment scenario, while tested with dataset of a different deployment scenario.*
  - *Network synchronization error*
  - *UE/gNB RX and TX timing error*
  - *SNR mismatch*
  - *Channel estimation error*

Note: *blue* italic text is from clause 6.4.2.2 of TR38.843 v1.2.0.

Thank You.

Copyright © 2019 Futurewei Technologies, Inc.
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Futurewei may change the information at any time without notice.

FUTUREWEI
Technologies