3GPP RAN WG4 Meeting #112 R4-24xxxx

Maastricht, NL, **August** 19 –**August** 23, 2024

Title: Alignment of model training and dataset parameters for AI/ML-based CSI compression

Source: Nokia

Agenda item: 8.17.4

Document for: Information

Introduction

This document captures the discussion on the AI/M model training and dataset parameters for AI/ML-based CSI compression for RAN4 study.

Discussion

|  |  |  |
| --- | --- | --- |
| **Parameter/Configurations** | **Options/Values** | **Comments** |
| **Channel/data-set related** | | |
| UE antenna element assumptions | * Option 1: Omni-directional antenna elements * Option 2: Directional antenna elements * Option 3: Other | [Apple] We provided 2 simulation results: Directional (both horizontal/vertical)  Omni (both horizontal/vertical)  [QC] Option 1 (both horizontal and vertical) |
| Sub-band channel averaging | * [Nokia, QC, Apple, vivo] Average of covariance channel matrix per RBs in the sub-band before SVD * [ZTE]: Per sub-band | CATT: Prefer not to do the averaging before SVD. |
| Any other pre-processing | * To be listed if used * [QC] Normalize the phase of all entries of the eigenvector based on the phase of the 1st entry of the eigenvector * [Apple]: Normalize the phase of all entries of the eigenvector based on the phase of that entry that has the maximum amplitude averaged over the sub-bands |  |
| Data collection assumptions | * [Nokia, CATT, ZTE] Each channel realization is collected from independent UE drop * Multiple channel realization can be collected from the same UE * Other * [QC] Multiple channels realizations are collected from multiple UEs and multiple drops of same UE * [Apple]: Multiple drops. Each drop collect data from multiple UEs. Additionally collect samples from channel realization from each UE spaced at de-correlated interval. |  |
| Training dataset size | * [Nokia]: 630K * [QC]: 450K * [Samsung]: 25K * [R&S]: 614,440 * [MTK]: 700K * [Ericsson]: 800K * [Intel]: 200K * [CATT]: 102,600 * [ZTE]: 600k * [Apple]: 600K * [vivo]: 243000 |  |
| Dataset quantization | * Float32 |  |
| Other data-set characteristics/statistics | * Companies to propose |  |
| **Model-related** | | |
| Quantization of encoder output | * Non-trainable quantization (bypassed during training) * [QC] Quantization aware training * [Apple]: Quantization in forward and approximation in backward to enable learning. (avoid zero gradient of round function) * [vivo] VQ with quantization aware training | CATT: Consider to align the activation function and quantization method. |
| Normalization before quantization in the encoder | * [Nokia, Ericsson]: Normalization with tanh function after FC layer * [CATT, ZTE, vivo]: Sigmoid function used after FC layer * [QC] Normalization not done due to the use of quantization aware training * [Apple]: Normalization with tanh function after FC layer. We don’t apply inverse of that in de-quantization |  |
| Use and configuration of drop-out | * [Nokia]: Use of on single drop-out layer after FC layer in the encoder and decoder, Drop-out rate = 0.4. * [CATT, QC, ZTE, Apple, vivo]: No drop-out used. |  |
| Training loss function, use of regularization | * [Nokia]: - SGCS with L2 regularization with weight decay, λ = 10-2 * [QC]: SGCS + additional terms for quantization aware training * [Apple, CATT]: 1- SGCS * [R&S, MTK, Intel]: SGCS * [Ericsson]: NMSE * [ZTE]: MSE * [vivo]: 1-SGCS plus VQ loss for encoder output (MSE between original and quantized encoder output) |  |
| Optimizer | * [Samsung, QCM, Apple, MTK, Intel, CATT]: Adam * [Nokia]: RMSProp (Root Mean Square Propagation) * [ZTE, vivo]: AdamW |  |
| Learning rate | * [Nokia]: * [QC, R&S, CATT]: 10^-4   + [R&S]: Learning rate adaptation: Reduce on plateau by factor 0.5 with patience of 50 epochs. * [Apple]: 0.0025 * [Samsung, Intel, MTK]: 0.001 * [ZTE]:1.5e-3 * [vivo]: Warm-up to initial lr 1e-3 and then linearly decay to final lr 8e-4, update after each batch |  |
| Batch size | * [Nokia]:1024 * [QC, MTK]: 128 * [Apple]: 256 * [Samsung]: 200 * [R&S, CATT, vivo]: 512 * [Ericsson]: 32 * [Intel]: 10 * [ZTE]:400 |  |
| (Max) Number of epochs | * [QC]: 150 * [Apple, vivo]: 100 * [R&S, Nokia, CATT]: 200 * [MTK, ZTE]: 300 |  |
| Stopping criteria | * [Nokia, Apple, CATT, QC, ZTE]: No change in metric over a given number of epochs * Validation patience: 10, [QC] 50 * [vivo] Stop when reaching the configured number of epochs |  |
| Weight initialization | * [Apple, CATT, QC, vivo]: Default |  |
| Kernel Parameters:   * Dilation * Padding Type   The value of groups (VG) | * [CMCC, Samsung, Vivo, CATT]: Dilation:1, Padding type=0, VG= 1 * [QC, ZTE, Apple]: Default |  |