**Source: Interdigital Finland Oy**

**Title:** **[FS\_AI4Media] pCR on compression metadata for split operations**

**Spec:** **3GPP TR 26.927 v0.8.0**

**Agenda item: 9.6**

**Document for: Agreement**

1. **Introduction**

The contribution adds a new metadata section on compression settings and characteristics applied to intermediate data for split point configurations comprising:

* A list of compression algorithm profiles including a description, a unique identifier and associated parameter for intermediate data compression. A subset of compression profile algorithms can be negotiated and exchanged between endpoints regarding different endpoints capabilities.
* Information for associating individual or group of compression profiles to intermediate data tensors for a split point configuration. This includes split point information, expected split point compression characteristics associated to a compression profile.

1. **Reason for Change**

Update the metadata section of the TR with text to address compression settings between endpoints.

1. **Proposal**

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

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

## 6.6 Metadata

### 6.6.6 Compression settings for a split point configuration

The compression settings below identify:

* The candidate compression algorithms profiles to apply to intermediate data tensors
* A split point compression characteristics (e.g. size reduction and performance metrics) for an association of compression algorithm profiles to intermediate data tensors .

#### 6.6.6.1 Compression algorithm profiles

A list of compression algorithm profiles includes a description, a unique identifier and associated parameter supported for intermediate data compression. A subset of compression algorithms profile can be negotiated and exchanged between endpoints regarding different endpoints capabilities.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metadata category** | **Metadata type** | | **Definition** | **Metadata type description (Examples)** |
| **Compression algorithm profile** | Compression algorithm profile list | | List of compression algorithm profiles. |  |
|  |  | Compression algorithm profile description | Identifies the compression algorithm(s) that can be applied to the intermediate data tensors. | NONE,  Quantization, FC\_VCM, SNAPPY  Neural Network Coding |
|  |  | Compression algorithm profile level identifier | Identifies the profile level of the compression algorithm. | FCM high 5.1, FCM main 5.3, FCM 6.4  NNC xxx 5.7.9, NNC yyy 5.8, NNC yyy 6.4 |
|  |  | Compression algorithm profile parameter set | List of compression parameters of the selected compression algorithm that fulfil the compression profile. | Param 1  Param 2 e.g. Qp Quantization Parameter = -15 |

#### 6.6.6.2 Intermediate data tensors and associated compression profile and characteristics

This identifies information for associating individual or group of compression profiles to intermediate data tensors for a split point configuration. This includes split point information, expected split point compression characteristics associated to a compression profile.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metadata category** | **Metadata type** | | **Definition** | **Metadata type description (Examples)** |
| **Split point information** | Split point identifier | | Key identifier of the split point clause §6.6.3 to associate value data below | Nb:10, 75  Name: Layer\_10, |
| **split point compression characteristics** | Compressed intermediate data size | | The compressed intermediate data size. If no compression, the size is the baseline data size. | 12 Mbytes |
| Compressed intermediate data size ratio | | The ratio gives an indication on how much the data has been reduced after apply compression. Example of representation of ratios: | ratio a) 5:1 or 5/1 means the size has been reduced from a factor 5  ratio b)  0,8 means a reduction of 80% of the baseline size |
| Compression performance metric value | | The measured performance metric value depending on the performance metrics used, e.g. Map score, F1 score, accuracy. | F1 Score, Map score |
| Compression performance metric ratio | | The ratio is calculated from the metric value obtained using compression divided by the value obtained without compression.  for example, it may indicate how much accuracy has been affected | Ratio 0.9 or 90% means a reduction performance accuracy of 10% |
| **Intermediate data tensors associated to compression profile** | Tensor list | | List of tensors or groups of tensors that composed intermediate data | e.g. list of ONNX tensor names Tensor1, Tensor2 |
|  | Tensor group compression name | This identifies a group of tensors when each group is associated with a common compression profile. | Group 1, Group 2 |
|  | Tensor group compression type | This identifies the type for all tensors belonging to tensor group. | Float32 |
|  | Tensor compression granularity | This indicates if the compression profiles are applied to each tensor one by one (keyword is “tensor”), or to all the tensors of this group (keyword is “global”). | “tensor”, “global” |
|  | Tensor name | The name of the tensor | Tensor1 |
|  | Tensor shape | Tensor shape output. The output tensor shape may be different from input and uncompressed tensor shape or may be transposed | [1,64,64,64]. |
|  | Tensor data type | The type of the tensor | Int32, Float32, |
|  | Tensor compression algorithm profile identifier | Identifies the selected compression algorithm profile | FCM high 5.1, FCM main 5.3, FCM 6.4  NNC xxx 5.7.9, NNC yyy 5.8, NNC yyy 6.4 |

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