**INTERNATIONAL ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC1/SC29/WG11**

**CODING OF MOVING PICTURES AND AUDIO**

**ISO/IEC JTC1/SC29/WG11 MPEG2019/m51017**

**October 2019, Geneva, Switzerland**

|  |  |
| --- | --- |
| **Source:** | **InterDigital Communications, Inc.** |
| **Status:** | **Input contribution** |
| **Title:** | **[NNR] CE3: report on combined clustering-based quantization and arithmetic coding** |
| **Authors:** | **Fabien Racapé, Shahab Hamidi-Rad, Swayambhoo Jain,** |

# Introduction

This contribution reports InterDigital’s response to the Core Experiment 3. The test conditions for this CE are detailed in [1]. The proposed method detailed in [2] relies on applying arithmetic coding to the indexes which are output from Interdigital’s CE2 proposal. The reported results in the following reflect the performance of the quantization and arithmetic coding without the gradient-based clustering and cluster migration described in [2]. Reported results include the full-chain test only, which includes Low Rank approximation, clustering-based quantization and arithmetic coding.

# Results

This section reports the results for each application in the following settings:

* Low Rank (LR) approximation networks [3],
* Quantization of the output of LR approximation,
* Arithmetic coding of the indexes.

The reported quantization sizes correspond to the total number of bytes including the indexes and the codebook. Codebook entries are float32. Hence, the sizes are 4 times the number of items in codebook. WP1 from our CE1 test has been used as input to CE2 with different quantization rates. These CE2 Results are then Entropy Coded using our CE3 Arithmetic Coding Algorithm.

Note that the Encoded Bitstream Size, does not include network structure information, as discussed during the 127th MPEG meeting. However, it includes all the information needed to reconstruct network parameters, including codebook information and probability information for arithmetic decoding.

More details on results and Layer structures can be found in attached spreadsheets.

## Image classification

### VGG16

Figure 1 shows the variations of the Top-1 and Top-5 accuracies of the quantized model, depending on the level of quantization.

Figure 1: VGG16 top-1 and top-5 accuracies as a function of the compression rate, i.e. size of the bitstream vs. original size

## End-to-end image compression

Figure 2: PSNR as a function of the compression rate, i.e. size of the bitstream vs. original size

## Audio DCase

Figure 3: Accuracy as a function of the compression rate

# Test frameworks and platform

The presented results were obtained under the following test conditions:

* **Operating System:** Ubuntu 18.04 LTS
* **Programming Language:** Python 2.7, Tensorflow 1.11.0
* **Additional Libraries:** Cuda v9.0.176 and CUDNN 7.2
* **The hardware configuration:**
	+ Intel i9-7920x CPU@2.9 Ghz
	+ NVIDIA TITAN Xp
	+ 128 GB DDR4 RAM

# Patent rights declaration(s)

**InterDigital Communications, Inc. may have current or pending patent rights relating to the technology described in this contribution and, conditioned on reciprocity, is prepared to grant licenses under reasonable and non-discriminatory terms as necessary for implementation of the resulting ITU-T Recommendation | ISO/IEC International Standard (per box 2 of the ITU-T/ITU-R/ISO/IEC patent statement and licensing declaration form).**

# References

1. “Description of Core Experiments on compression of neural networks for multimedia content description and analysis”, ISO/IEC JTC1/SC29/WG11 N18574, July 2019.
2. Shahab Hamidi-Rad, Swayambhoo Jain, Fabien Racapé “Gradient-Based K-Means for Quantization of Deep Neural Network parameters” ISO/IEC JTC1/SC29/WG11 m49260, July 2019.
3. Swayambhoo Jain, Shahab Hamidi-Rad, Fabien Racapé, “Low Displacement Rank decomposition for compressing convolutional layers” ISO/IEC JTC1/SC29/WG11 N18574, July 2019