**INTERNATIONAL ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

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

**CODING OF MOVING PICTURES AND AUDIO**

**ISO/IEC JTC1/SC29/WG11 MPEG2020/m53766**

**April 2020, Alpbach, Austria**

|  |  |
| --- | --- |
| **Source:** | **InterDigital Communications, Inc.** |
| **Status:** | **Input contribution** |
| **Title:** | **[NNR] CE3: Report on Arithmetic Coding Results**  |
| **Authors:** | **Shahab Hamidi-Rad, Swayambhoo Jain, Fabien Racape.** |

# Introduction

This contribution reports InterDigital’s results for the Core Experiment 3. The test conditions for this CE are detailed in [1]. The proposed method, detailed in [2], relies on conditional and adaptive arithmetic coding of quantized parameters of Neural Networks. To perform the following tests, small modifications were made to the NCTM [4] to add support for arithmetic coding. The modified NCTM is available in the branch “NCTM\_InterDigital\_CE3”. We compare our results with the current baseline for entropy coding detailed in [5].

# Results

The following results are based on the following pipelines:

* Input Network à Uniform Quantization à Baseline Entropy Coding (DeepCABAC)
* Input Network à Uniform Quantization à Proposed Arithmetic Coding
* Input Network à Codebook Quantization à Baseline Entropy Coding (DeepCABAC)
* Input Network à Codebook Quantization à Proposed Arithmetic Coding

**Input networks:**

* Original Networks
* CE1 outputs: baseline Low Rank method [3] for the 4 working points detailed in [1].

**Use cases:**

* VGG-16
* ResNet-50
* MobileNetV2
* DCase (Audio Classification)
* UC12B (Image Compression)

We used NCTM to perform quantization, entropy coding, and the evaluation of the results. Here is a sample for ReseNet-50 use case using CE1 Working point 3 as input. For complete results please refer to the attached spreadsheets.



# Test frameworks and platform

The presented results were obtained under the following test conditions:

* **Operating System:** Ubuntu 18.04 LTS
* **Programming Language:** Python 3.6, Tensorflow 1.14.0
* **Additional Libraries:** Cuda V10.0.130 and CUDNN 7.6
* **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, Fabien Racapé, Swayambhoo Jain “Conditional arithmetic coding for quantized parameters” ISO/IEC JTC1/SC29/WG11 m51018, October 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
4. “Test model 3 of Compression of neural networks for multimedia content description and analysis”, ISO/IEC JTC 1/SC 29/WG11 N18993
5. Simon Wiedemann et. Al. “Response to the Call for Proposals on Neural Network Compression: End-to-end processing pipeline for highly compressible neural networks”, ISO/IEC JTC1/SC29/WG11 m47698, March 2019.