**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/m51016**

**October 2019, Geneva, Switzerland**

|  |  |
| --- | --- |
| **Source:** | **InterDigital Communications, Inc.** |
| **Status:** | **Input contribution** |
| **Title:** | **[NNR] CE2: report on clustering-based quantization with lower-bounded PDF initialization** |
| **Authors:** | **Fabien Racapé, Shahab Hamidi-Rad, Swayambhoo Jain,** |

# Introduction

This contribution reports InterDigital’s response to the Core Experiment 2. The test conditions for this CE are detailed in [1]. The proposed method detailed in [2] relies on cluster-based quantization. The clusters are initialized using a bounded Probability-Density Function (PDF). Lower-bounded PDF initialization gives good granularity of the cluster centers around high-probability values while assigning enough cluster centers around the values happening with lower probability. Note that the following results reflect the performance of the quantization without requiring access to annotated data, i.e. without the gradient-based clustering described in [2].

# Results

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

* Quantization of the original network,
* Quantization of CE1 output: Low Rank approximated networks [3].

All the reported working points are in accordance with the target ratios defined in [1], as can be seen in the columns “Quantization Ratio (Q\_size/O\_size)” in the following tables. However, it was sometimes impossible to reach the defined rates, in which case additional rate points have been added to characterize the variations of accuracy for each application with a wide range of quantization ratios.

The reported quantization sizes correspond to the total number of bytes including the indexes and the codebook. Codebook entries are float32; so the size is 4 times the number of items in codebook.

We used the WP1 from our CE1 tests as input to CE2 with different quantization rates.

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

## Image classification

### VGG16

Table 1: quantization of the original model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **WorkingPoint** | **Param Reduction** | **Original** | **Reduced + Retrained** | **Quantization Ratio (R\_size/O\_size)** |
| **# Params** | **O\_size (4 x # Params)** | **O\_Top5** | **Quant. Bits** | **Quant. Size\*(bytes)** | **R\_Top5** | **R\_Top-1 Ratio (org-new)/new** | **R\_Top-5 Ratio (org-new)/new** |
| **WP1** | **Q-Bits = 2, Codebook Size = 4** | **138,357,544** | **553,430,176** | **0.89736** | **2** | **34,589,898** | **0.0139** | **187.8235** | **63.6513** | **0.062500925** |
|   | Q-Bits = 3, Codebook Size = 8 | 138,357,544 | 553,430,176 | 0.89736 | 3 | 51,885,095 | 0.7262 | 0.4700 | 0.2357 | 0.093751836 |
| **WP2** | **Q-Bits = 4, Codebook Size = 16** | **138,357,544** | **553,430,176** | **0.89736** | **4** | **69,180,772** | **0.8808** | **0.0388** | **0.0188** | **0.125003614** |
|   | Q-Bits = 5, Codebook Size = 32 | 138,357,544 | 553,430,176 | 0.89736 | 5 | 86,477,561 | 0.8954 | 0.0023 | 0.0022 | 0.156257401 |
| **WP3** | **Q-Bits = 6, Codebook Size = 64** | **138,357,544** | **553,430,176** | **0.89736** | **6** | **103,775,230** | **0.8952** | **0.0037** | **0.0024** | **0.187512778** |
| **WP1** | **Q-Bits = 2, Codebook Size = 4** | **138,357,544** | **553,430,176** | **0.89736** | **2** | **34,589,898** | **0.0139** | **187.8235** | **63.6513** | **0.062500925** |

Figure 1 and Figure 2 show the accuracy of the quantized model depending on the level of quantization. As can be seen, at low rate, i.e. when quantization has an impact on accuracy, Low Rank approximation improves the accuracy at same size of the quantized model.

Figure 1: VGG16 top-1 and top-5 accuracy as a function of the size of the quantized model (original weights)

Figure 2: VGG16 top-1 and top-5 accuracy as a function of the size of the quantized Low rank approximation of the model

### Resnet50

Table 2: quantization of the original model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **WorkingPoint** | **Param Reduction** | **Original** | **Reduced + Retrained** | **Quantization Ratio (R\_size/O\_size)** |
| **# Params** | **O\_size (4 x # Params)** | **O\_Top5** | **Quant. Bits** | **Quant. Size\*(bytes)** | **R\_Top5** | **R\_Top-1 Ratio (org-new)/new** | **R\_Top-5 Ratio (org-new)/new** |
| **WP1** | **Q-Bits = 3, Codebook Size = 8** | **25,636,712** | **102,546,848** | **0.74736** | **0.91932** | **3** | **9,718,535** | **0.0031** | **0.0114** | **241.6494** |
|   | Q-Bits = 4, Codebook Size = 16 | 25,636,712 | 102,546,848 | 0.74736 | 0.91932 | 4 | 12,927,956 | 0.5927 | 0.8192 | 0.2610 |
| **WP2** | **Q-Bits = 5, Codebook Size = 32** | **25,636,712** | **102,546,848** | **0.74736** | **0.91932** | **5** | **16,143,925** | **0.6977** | **0.8919** | **0.0712** |
|   | Q-Bits = 6, Codebook Size = 64 | 25,636,712 | 102,546,848 | 0.74736 | 0.91932 | 6 | 19,371,742 | 0.7352 | 0.9115 | 0.0165 |
| **WP3** | **Q-Bits = 7, Codebook Size = 128** | **25,636,712** | **102,546,848** | **0.74736** | **0.91932** | **7** | **22,615,479** | **0.7440** | **0.9176** | **0.0046** |
| **WP1** | Q-Bits = 8, Codebook Size = 256 | 25,636,712 | 102,546,848 | 0.74736 | 0.91932 | 8 | 25,887,288 | 0.7474 | 0.9187 | 0.0000 |

## End-to-end image compression

Table 3: quantization of the original model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **WorkingPoint** | **Quantization** | **Original** | **Reduced + Retrained** | **Quantization Ratio (Q\_size/O\_size)** |
| **# Params** | **O\_size (4 x # Params)** | **O\_Acc** | **# Params** | **R\_size (4 x # Params)** | **R\_PSNR** | **R\_ PSNR(Retrained)****R\_ PSNR****Ratio (org-new)/new** |
|   | Q-Bits = 3, Codebook Size = 8 | 76,179 | 304,716 | 30.13765 | 3 | 29,082 | 9.5270 | 2.1634 | 0.095439688 |
|   | Q-Bits = 4, Codebook Size = 16 | 76,179 | 304,716 | 30.13765 | 4 | 39,008 | 14.1040 | 1.1368 | 0.128014282 |
| **WP1** | **Q-Bits = 5, Codebook Size = 32** | **76,179** | **304,716** | 30.13765 | **5** | **49,230** | **17.8738** | 0.6861 | **0.161560273** |
|   | Q-Bits = 6, Codebook Size = 64 | 76,179 | 304,716 | 30.13765 | 6 | 59,616 | 18.1980 | 0.6561 | 0.195644469 |
|   | Q-Bits = 7, Codebook Size = 128 | 76,179 | 304,716 | 30.13765 | 7 | 70,242 | 23.9250 | 0.2597 | 0.230516284 |
| **WP2** | **Q-Bits = 8, Codebook Size = 256** | **76,179** | **304,716** | 30.13765 | **8** | **82,164** | **27.0582** | 0.1138 | **0.26964124** |
|  | Q-Bits = 9, Codebook Size = 512 | 76,179 | 304,716 | 30.13765 | 9 | 94,144 | 29.4787 | 0.0224 | 0.308956537 |
| **WP3** | **Q-Bits = 10, Codebook Size = 1024** | **76,179** | **304,716** | **30.13765** | **10** | **111,780** | **29.8502** | **0.0096** | **0.366833379** |
| **WP4** | **Q-Bits = 11, Codebook Size = 2048** | **76,179** | **304,716** | **30.13765** | **11** | **146,816** | **30.0792** | **0.0019** | **0.481812573** |
|   | Q-Bits = 12, Codebook Size = 4096 | 76,179 | 304,716 | 30.13765 | 12 | 220,152 | 30.1378 | 0.0000 | 0.722482574 |

Table 4:quantization of the Low Rank model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **WorkingPoint** | **Quantization** | **Original** | **Reduced + Retrained** | **Quantization Ratio (Q\_size/O\_size)** |
| **# Params** | **O\_size (4 x # Params)** | **O\_Acc** | **# Params** | **R\_size (4 x # Params)** | **R\_PSNR** | **R\_ PSNR(Retrained)****R\_ PSNR****Ratio (org-new)/new** |
| CE1 WP1 Interdigital  | Q-Bits = 7, Codebook Size = 128 | 76,179 | 304,716 | 30.13765 | 8 | 63,612 | 28.2827 | 0.0656 | 0.208758319 |
| CE1 WP1 Interdigital  | Q-Bits = 8, Codebook Size = 256 | 76,179 | 304,716 | 30.13765 | 9 | 77,644 | 29.9214 | 0.0072 | 0.254807755 |
| CE1 WP1 Interdigital  | Q-Bits = 9, Codebook Size = 512 | 76,179 | 304,716 | 30.13765 | 10 | 96,392 | 30.5881 | -0.0147 | 0.316333898 |
|  CE1 WP1 Interdigital | Q-Bits = 10, Codebook Size = 1024 | 76,179 | 304,716 | 30.13765 | 11 | 168,152 | 30.7883 | -0.0211 | 0.55183187 |
| CE1 WP1 Interdigital | **Q-Bits = 11, Codebook Size = 2048** | **76,179** | **304,716** | **30.13765** | **12** | **194,232** | **30.7884** | **-0.0211** | **0.637419761** |

## Audio DCase

Table 5: quantization of the original model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **WorkingPoint** | **Quantization** | **Original** | **Reduced + Retrained** | **Quantization Ratio (Q\_size/O\_size)** |
| **n# Params** | **O\_size (4 x # Params)** | **O\_Accuracy** | **Quantization Bits** | **Quantized Size\*(bytes)** | **R\_Accuracy** | **R\_Accuracy Ratio(org-new)/new** |
| **WP3** | **Q-Bits = 4, Codebook Size = 16** | **116,815** | **467,260** | **0.582716** | **2** | **29,460** | **0.5889** | **-0.0105** | **0.06304841** |
|   | Q-Bits = 3, Codebook Size = 8 | 116,815 | 467,260 | 0.582716 | 3 | 44,302 | 0.5506 | 0.0583 | 0.09481231 |
| **WP4** | **Q-Bits = 4, Codebook Size = 16** | **116,815** | **467,260** | **0.582716** | **4** | **59,344** | **0.5889** | **-0.0105** | **0.127004237** |
|   | Q-Bits = 5, Codebook Size = 32 | 116,815 | 467,260 | 0.582716 | 5 | 74,772 | 0.5790 | 0.0064 | 0.160022257 |
|   | Q-Bits = 6, Codebook Size = 64 | 116,815 | 467,260 | 0.582716 | 6 | 90,561 | 0.5914 | -0.0146 | 0.193812866 |
|   | Q-Bits = 7, Codebook Size = 128 | 116,815 | 467,260 | 0.582716 | 7 | 106,065 | 0.5864 | -0.0063 | 0.226993537 |
|   | Q-Bits = 8, Codebook Size = 256 | 116,815 | 467,260 | 0.582716 | 8 | 122,396 | 0.5802 | 0.0043 | 0.2619441 |
|   | Q-Bits = 9, Codebook Size = 512 | 116,815 | 467,260 | 0.582716 | 9 | 140,172 | 0.5852 | -0.0042 | 0.299987159 |
|   | Q-Bits = 10, Codebook Size = 1024 | 116,815 | 467,260 | 0.582716 | 10 | 163,092 | 0.5827 | 0.0000 | 0.349039079 |
|   | Q-Bits = 11, Codebook Size = 2048 | 116,815 | 467,260 | 0.582716 | 11 | 183,300 | 0.5840 | -0.0021 | 0.392286949 |

# 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