**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/m51015**

**October 2019, Geneva, Switzerland**

|  |  |
| --- | --- |
| **Source:** | **InterDigital Communications, Inc.** |
| **Status:** | **Input contribution** |
| **Title:** | **[NNR] CE1: report on Low Rank approximations** |
| **Authors:** | **Swayambhoo Jain, Shahab Hamidi-Rad, Fabien Racapé** |

# Introduction

This contribution reports InterDigital’s response to the Core Experiment 1, in which structural approximations using Low Rank representations are used to reduce the number of parameters of Neural Networks.

The test conditions are detailed in [1]. In particular, working point have been chosen for CE1 depending on the number of non-zero parameters, as detailed in Table 1.

Table 1: Working points as reported in [1]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **WP1** | **WP2** | **WP3** | **WP4** |
| VGG16 | 0.25 | 0.20 | 0.15 | 0.10 |
| ResNet50 | 0.45 | 0.35 | 0.25 | 0.20 |
| MobileNet-v2 | 0.85 | 0.80 | 0.70 | 0.65 |
| Audio classification | 0.40 | 0.30 | 0.20 | 0.15 |
| Image Compression | 0.70 | 0.50 | 0.20 | 0.10 |

The proposed method changes the tensor structures of fully connected and convolutional layers to smaller matrices G and H, as described in [2]. As it does not directly rely on zeroing some parameters, the working points are defined as the ratio of the number of parameters included in the new matrices compared to the original size. That is, zero weights may be present in the matrices G and H, but those are not considered for setting the working points.

# Results

This section reports the results for each application. All the reported working points are in accordance with the target ratios defined in [1], as can be seen in the columns “Reduction Ratio (R\_size/O\_size)” in the following tables.

In the column Param reductions are reported which layers have been modified and the chosen rank, e.g. L3\_CONV:62 denotes the convolutional layer #3 has been reduced to rank 62.

In the column Retraining are reported the hyperparameters used for retraining:

* Epc: number of epochs performed
* DO: dropout hyperparameter
* LR: Learning Rate. (.05->.00001 denotes a learning rates that varies from .05 to .00001 over the epochs)

## Image classification

### VGG16

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Working Point** | **Param Reduction (Ranks for Layers)** | **Retraining** | **Original** | | | **Reduced + Retrained** | | | | | **Reduction Ratio (R\_size/O\_size)** |
| **# Params** | **O\_size  (4 x # Params)** | **O\_Top5** | **# Params** | **R\_size  (4 x # Params)** | **R\_Top5** | **R\_Top5 (Retrained)** | **R\_Top5**  **Ratio (org-new)/new** |
| WP1 | L3\_CONV:62, L5\_FC:10 | 20 epc, batch: 20, DO:0.4,  LR: .05->.00001 | 138,357,544 | 553,430,176 | 0.582716 | 47,607 | 190,428 | 0.5198 | 0.6049 | -0.0367 | 0.40754184 |
| WP2 | L3\_CONV:42, L5\_FC:16 | 20 epc, batch: 20, DO:0.4,  LR: .05->.00001 | 138,357,544 | 553,430,176 | 0.582716 | 35,535 | 142,140 | 0.4383 | 0.5926 | -0.0167 | 0.304198947 |
| WP3 | L3\_CONV:28, L5\_FC:8 | 30 epc, batch: 20, DO:0,  LR: .01->.00001 | 138,357,544 | 553,430,176 | 0.582716 | 24,303 | 97,212 | 0.3111 | 0.5926 | -0.0167 | 0.208046912 |
| WP4 | L3\_CONV:18, L5\_FC:8 | 30 epc, batch: 20, DO:0,  LR: .05->.00001 | 138,357,544 | 553,430,176 | 0.582716 | 17583 | 70,332 | 0.1741 | 0.5864 | -0.0063 | 0.150520053 |

### Resnet50

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Working Point** | **Param Reduction (Ranks for Layers)** | **Retraining** | **Original** | | | **Reduced + Retrained** | | | | | **Reduction Ratio (R\_size/O\_size)** |
| **# Params** | **O\_size  (4 x # Params)** | **O-Top5** | **# Params** | **R\_size  (4 x # Params)** | **R\_Top5** | **R\_Top5 (Retrained)** | **R\_Top5**  **Ratio (org-new)/new** |
| WP1 | S4\_L1\_RES2 .. S5\_L3\_RES1: MSE set to .000068 |  | 25,636,712 | 102,546,848 | 0.91932 | 0.91932 | 11,552,872 | 0.8049 | 0.8815 | 0.8049 | 0.45063782 |
| WP2 | S4\_L1\_RES2 .. S5\_L3\_RES1: MSE set to .000093 | 10 epochs, batch: 256, DO:0.1,  LR: .002->  .0000001 | 25,636,712 | 102,546,848 | 0.91932 | 0.91932 | 9,122,152 | 0.5461 |  |  | 0.355823789 |
| WP3 | S3\_L1\_RES2 .. S5\_L3\_RES1: MSE set to .00012 |  | 25,636,712 | 102,546,848 | 0.91932 | 0.91932 | 6,553,960 | 0.1272 |  |  | 0.255647448 |
| WP4 | S3\_L1\_RES2 .. S4\_L6\_RES1: MSE set to .000185, S5\_L1\_RES2: MSE=.00014, S5\_L2\_RES2: .00012, S5\_L3\_RES2:.00010, |  | 25,636,712 | 102,546,848 | 0.91932 | 0.91932 | 5,344,232 | 0.0187 |  |  | 0.208460118 |

## End-to-end image compression

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Working Point** | **Param Reduction (Ranks for Layers)** | **Retraining** | **Original** | | | **Reduced + Retrained** | | | | | **Reduction Ratio (R\_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** |
| WP1 | S1\_L2\_CONV:34, S1\_L3\_CONV:28,  S2\_L1\_CONV:20, S2\_L2\_CONV:52, S2\_L3\_CONV: 36 | 300 epc, batch: 32, LR: .001->.000001 | 76,179 | 304,716 | 30.137648 | 53,523 | 214,092 | 5.8795 | 30.7874 | -0.0211 | 0.702595203 |
| WP2 | S1\_L2\_CONV:16, S1\_L3\_CONV:16,  S2\_L1\_CONV:12, S2\_L2\_CONV:40, S2\_L3\_CONV: 24 | 300 epc, batch: 32, LR: .001-> .000001 | 76,179 | 304,716 | 30.137648 | 38,355 | 153,420 | 5.8292 | 30.4502 | -0.0103 | 0.503485212 |
| WP3 | S1\_L2\_CONV:5, S1\_L3\_CONV:5, S1\_L4\_CONV:12, S2\_L1\_CONV:5, S2\_L2\_CONV:18, S2\_L3\_CONV:8, S2\_L4\_CONV:4 | 300 epc, batch: 32, LR: .001-> .000001 | 76,179 | 304,716 | 30.137648 | 15,267 | 61,068 | 5.8354 | 28.1917 | 0.0690 | 0.200409562 |
| WP4 | S1\_L1\_CONV:1, S1\_L2\_CONV:1, S1\_L3\_CONV:4, S1\_L4\_CONV:11, S2\_L1\_CONV:1, S2\_L2\_CONV:12, S2\_L3\_CONV:4, S2\_L4\_CONV:1 | 300 epc, batch: 32, LR: .002-> .000001 | 76,179 | 304,716 | 30.137648 | 8,172 | 32,688 | 5.8329 | 26.2161 | 0.1496 | 0.107273658 |

## Audio DCase

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Working Point** | **Param Reduction (Ranks for Layers)** | **Retraining** | **Original** | | | **Reduced + Retrained** | | | | | **Reduction Ratio (R\_size/O\_size)** |
| **# Params** | **O\_size  (4 x # Params)** | **O\_Acc** | **# Params** | **R\_size  (4 x # Params)** | **R\_Acc** | **R\_Acc (Retrained)** | **R\_Acc**  **Ratio (org-new)/new** |
| WP1 | L3\_CONV:62, L5\_FC:10 | 20 epc, batch: 20, DO:0.4,  LR: .05->.00001 | 116,815 | 467,260 | 0.582716 | 47,607 | 190,428 | 0.5198 | 0.6049 | -0.0367 | 0.40754184 |
| WP2 | L3\_CONV:42, L5\_FC:16 | 20 epc, batch: 20, DO:0.4,  LR: .05->.00001 | 116,815 | 467,260 | 0.582716 | 35,535 | 142,140 | 0.4383 | 0.5926 | -0.0167 | 0.304198947 |
| WP3 | L3\_CONV:28, L5\_FC:8 | 30 epc, batch: 20, DO:0,  LR: .01->.00001 | 116,815 | 467,260 | 0.582716 | 24,303 | 97,212 | 0.3111 | 0.5926 | -0.0167 | 0.208046912 |
| WP4 | L3\_CONV:18, L5\_FC:8 | 30 epc, batch: 20, DO:0,  LR: .05->.00001 | 116815 | 467,260 | 0.582716 | 17583 | 70,332 | 0.1741 | 0.5864 | -0.0063 | 0.150520053 |

# 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](mailto: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. Swayambhoo Jain, Shahab Hamidi-Rad, Fabien Racapé, “Low Displacement Rank decomposition for compressing convolutional layers” ISO/IEC JTC1/SC29/WG11 N18574, July 2019.