HarDNet: A Low Memory Traffic Network

Ping Chao
Chao-Yang Kao
Yu-Shan Ruan
Chien-Hsiang Huang
Youn-Long Lin
National Tsing Hua University
University of Michigan
pingchao@umich.edu  {chaoyangkao923, esunxd, james128333}gmail.com  ylin@cs.nthu.edu.tw

Abstract

State-of-the-art neural network architectures such as ResNet, MobileNet, and DenseNet have achieved outstanding accuracy over low MACs and small model size counterparts. However, these metrics might not be accurate for predicting the inference time. We suggest that memory traffic for accessing intermediate feature maps can be a factor dominating the inference latency, especially in such tasks as real-time object detection and semantic segmentation of high-resolution video. We propose a Harmonic Densely Connected Network to achieve high efficiency in terms of both low MACs and memory traffic. The new network achieves 35%, 36%, 30%, 32%, and 45% inference time reduction compared with FC-DenseNet-103, DenseNet-264, ResNet-50, ResNet-152, and SSD-VGG, respectively. We use tools including Nvidia profiler and ARM Scale-Sim to measure the memory traffic and verify that the inference latency is indeed proportional to the memory traffic consumption and the proposed network consumes low memory traffic. We conclude that one should take memory traffic into consideration when designing neural network architectures for high-resolution applications at the edge.

1. Introduction

Convolutional Neural Networks (CNN) have been popular for computer vision tasks, ever since the explosive growth of computing power has made possible training complex networks like AlexNet [22,23], VGG-net [32], and Inception [34] in a reasonable amount of time. To bring these fascinating research results into mass use, performing a neural network inference on edge devices is inevitable. However, edge computing relies on limited computation power and battery capacity. How to increase computation efficiency and reduce the power consumption for neural network inference at the edge has therefore become a critical issue.

Reducing model sizes (the number of parameters or weights of a model) is a hot research topic in improving both computation and energy efficiency, since a reduced model size usually implies fewer MACs (number of multiply-accumulate operations or floating point operations) and less dynamic random-access memory (DRAM) traffic for read and write of model parameters and feature maps. Several researches have steered toward maximizing the accuracy–parameters ratio. State-of-the-art networks such as Residual Networks (ResNets) [16], SqueezeNets [20], and Densely Connected Networks (DenseNets) [18] have achieved high parameter efficiency that have dramatically reduced the model size while maintaining a high accuracy. The model size can be reduced further through compression. Han et al. [15] showed that the large amount of floating-point weights loaded from DRAM may consume more power than arithmetic operations do. Their Deep Compression algorithm employs weight pruning and quantization to reduce the model size and power consumption significantly.

In addition to the power consumption, DRAM accesses can also dominate system performance in terms of inference time due to the limited DRAM bandwidth. Since we have observed that the size summation of all the intermediate feature maps in a CNN can be ten to hundred times larger than its model size, especially for high resolution tasks such as semantic segmentation using fully convolutional networks [27], we suggest that reducing DRAM accesses to feature maps may lead to a speedup in some cases.

Shrinking the size of feature maps is a straightforward approach to reduce the traffic. While there are only a few papers addressing lossless compression of feature maps, lossy compression of feature maps has been intensively studied in research of model precision manipulation and approximation [8,11,14,28,29]. The quantization used in these works for model compression can usually reduce the feature map size automatically. However, like other lossy compression methods such as subsampling, they usually penalize accuracy. In this paper, we explore how to reduce the DRAM traffic for feature maps without penalizing accuracy simply by designing the architecture of a CNN carefully.
To design such a low DRAM traffic CNN architecture, it is necessary to measure the actual traffic. For a general-purpose Graphics Processing Unit (GPU), we use Nvidia profiler to measure the number of DRAM read/write bytes. For mobile devices, we use ARM Scale Sim [30] to get traffic data and inference cycle counts for each CNN architecture. We also propose a metric called Convolutional Input/Output (CIO), which is simply a summation of the input tensor size and output tensor size of every convolution layer as equation (1), where \( c \) is the number of channels and \( w \) and \( h \) are the width and height of the feature maps for a convolution layer \( l \).

\[
CIO = \sum_l (c_{\text{in}}^{(l)} \times w_{\text{in}}^{(l)} \times h_{\text{in}}^{(l)} + c_{\text{out}}^{(l)} \times w_{\text{out}}^{(l)} \times h_{\text{out}}^{(l)}) \quad (1)
\]

CIO is an approximation of DRAM traffic proportional to the real DRAM traffic measurement. Please note that the input tensor can be a concatenation, and a reused tensor can therefore be counted multiple times. Using a lot of large convolutional kernels may easily achieve a minimized CIO. However, it also damages the computational efficiency and eventually leads to a significant latency overhead outweighing the gain. Therefore, we argue that maintaining a high computational efficiency is still imperative, and CIO dominates the inference time only when the computational density, which is, the MACs over CIO (MoC) of a layer, is below a certain ratio that depends on platforms.

For example, under a fixed CIO, changing the channel ratio between the input and output of a convolutional layer step by step from 1:1 to 1:100 leads to reductions of both MACs and latency. For the latency, it declines more slowly than the reduction of MACs, since the memory traffic remains the same. A certain value of MoC may show that, below this ratio, the latency for a layer is always bounded to a fixed time. However, this value is platform-dependent and obscure empirically.

In this paper, we apply a soft constraint on the MoC of each layer to design a low CIO network model with a reasonable increase of MACs. As shown in Fig. 1, we avoid to employ a layer with a very low MoC such as a Conv1x1 layer that has a very large input/output channel ratio. Inspired by the Densely Connected Networks [18] we propose a Harmonic Densely Connected Network (HarDNet) by applying the strategy. We first reduce most of the layer connections from DenseNet to reduce concatenation cost. Then, we balance the input/output channel ratio by increasing the channel width of a layer according to its connections.

![Figure 1: Concept of MoC constraint. A Conv layer with MoC below the constraint is avoided.](image)

The contribution of this paper is that we introduce DRAM traffic for feature map access and its platform-independent approximation, CIO, as a new metric for evaluating a CNN architecture and show that the inference latency is highly correlated with the DRAM traffic. By constraining the MoC of each layer, we propose HarD Nets that reduces DRAM traffic by 40% compared with DenseNets. We evaluate the proposed HarDNet on the CamVid [3], ImageNet (ILSVRC) [9], PASCAL VOC [12], and MS COCO [26] datasets. Compared to DenseNet and ResNet, HarDNet achieves the same accuracy with 30%~50% less CIO, and accordingly, 30%~40% less inference time.

2. Related works

A significant trend in neural network research is exploiting shortcuts. To cope with the degradation problem, Highway Networks [33] and Residual Networks [16] add shortcuts to sum up a layer with multiple preceeding layers. The stochastic depth regularization [19] is essentially another form of shortcuts for crossing layers that are randomly dropped. Shortcuts enable implicit supervision to make networks continually deeper without degradation. DenseNets [18] concatenates all preceeding layers as a shortcut achieving more efficient deep supervision. Shortcuts have also been shown to be very useful in segmentation tasks [10]. Jégou et al. [21] showed that without any pretraining, DenseNet performs semantic segmentation very well. However, shortcuts lead to both large memory usage and heavy DRAM traffic. Using shortcuts elongates the lifetime of a tensor, which may result in frequent data exchanges between DRAM and cache.

Some sparsified versions of DenseNet have been proposed. LogDenseNet [17] and SparseNet [36] adopt a strategy of sparsely connecting each layer \( k \) with layer \( k-2^n \) for all integers \( n \geq 0 \) and \( k-2^n \geq 0 \) such that the input channel numbers decrease from \( O(L^2) \) to \( O(L \log L) \). The difference between them is that LogDenseNet applies this strategy globally, where layer connections crossing...
blocks with different resolutions still follow the log connection rule, while SparseNet has a fixed block output that regards the output as layer \( L + 1 \) for a block with \( L \) layers. However, both network architectures need to significantly increase the growth rate (output channel width) to recover the accuracy dropping from the connection pruning, and the increase of growth rate can compromise the CIO reduction. Nevertheless, these studies did point out a promising direction to sparsify the DenseNet.

The performance of a classic microcomputer architecture is dominated by its limited computing power and memory bandwidth [4]. Researchers focused more on enhancing the computation power and efficiency. Some researchers pointed out that limited memory bandwidth can dominate the inference latency and power consumption in GPU-based systems [25, 27], FPGA-based systems [5, 13], or custom accelerators [6, 7, 11]. However, there is no systematic way to correlate DRAM traffic and the latency. Therefore, we propose CIO and MoC and present a conceptual methodology for enhancing efficiency.

3. Proposed Harmonic DenseNet

3.1. Sparsification and weighting

We propose a new network architecture based on the Densely Connected Network. Unlike the sparsification proposed in LogDenseNet, we let layer \( k \) connect to layer \( k−2^n \) if \( 2^n \) divides \( k \), where \( n \) is a non-negative integer and \( k−2^n \geq 0 \); specifically, layer 0 is the input layer. Under this connection scheme, once layer \( 2^n \) is processed, layer 1 through \( 2^{n−1} \) can be flushed from the memory. The connections make the network appear as a combination of power-of-two-th harmonic waves, as illustrated in Fig. 2, hence we name it the Harmonic Densely Connected Network (HarDNet). The proposed sparsification scheme reduces the concatenation cost significantly better than the LogDenseNet does. This connection pattern also looks like a FractalNet [24], except the latter uses averaging shortcuts instead of concatenations.

In the proposed network, layers with an index divided by a larger power of two are more influential than those that divided by a smaller power of two. We amplify these key layers by increasing their channels, which can balance the channel ratio between the input and output of a layer to avoid a low MoC. A layer \( l \) has an initial growth rate \( k \), and we let its channel number to be \( k \times m^n \), where \( n \) is the maximum number satisfying that \( l \) is divided by \( 2^n \). The multiplier \( m \) serves as a low-dimensional compression factor. If the input layer 0 has \( k \) channels and \( m = 2 \), we get a channel ratio 1:1 for every layer. Setting \( m \) smaller than two is tantamount to compress the input channels into fewer output channels. Empirically, setting \( m \) between 1.6 and 1.9 achieves a good accuracy and parameter efficiency.

3.2. Transition and Bottleneck Layers

The proposed connection pattern forms a group of layers called a Harmonic Dense Block (HDB), which is followed by a Conv1x1 layer as a transition. We let the depth of each HDB to be a power of two such that the last layer of an HDB has the largest number of channels. In DenseNet, a densely connected output of a block directly passes the gradient from output to all preceding layers to achieve deep supervision. In our HDB with depth \( L \), the gradient will pass through at most \( \log L \) layers. To alleviate the degradation, we made the output of a depth-\( L \) HDB to be the concatenation of layer \( L \) and all its preceding odd numbered layers, which are the least significant layers with \( k \) output channels. The output of all even layers from 2 to \( L−2 \) can be discarded once the HDB is finished. Their total memory occupation is roughly two to three times as large as all the odd layers combined when \( m \) is between 1.6 to 1.9.

DenseNet employs a bottleneck layer before every Conv3x3 layer to enhance the parameter efficiency. Since we have balanced the channel ratio between the input and output for every layer, the effect of such bottleneck layers becomes insignificant. Inserting a bottleneck layer for every four Conv3x3 layer is still helpful for reducing the model size. We let the output channels of a bottleneck layer to be \( \sqrt{c_{in}/c_{out}} \times c_{out} \), where \( c_{in} \) is the concatenated input channels and \( c_{out} \) is the output channels of the follow-
Figure 3: (a) Inverted transition down module, (b) Depthwise-separable convolution for HarDNet

To further improve the inference time, these Conv1x1 bottleneck layers can be discarded to meet our MoC constraint.

The transition layer proposed by DenseNet is a Conv1x1 layer followed by a 2x2 average pooling. As shown in Fig. 3a, we propose an inverted transition module, which maps input tensor to an additional max pooling function along with the original average pooling, followed by concatenation and Conv1x1. This module reduces 50% of CIO for the Conv1x1 while achieving roughly the same accuracy at the expense of model size increase.

### 3.3. Detailed Design

To compare with DenseNet, we follow its global dense connection strategy that bypasses all the input of an HDB as a part of its output and propose six models of HarDNet. The detailed parameters are shown in Table 1. We use a 0.85 reduction rate for the transition layers instead of the 0.5 reduction rate used in the DenseNet, since a low-dimensional compression has been applied to the growth rate multiplier as we mentioned before. To achieve a flexible depth, we partition a block into multiple blocks with 16 layers (20 when bottleneck layers are counted).

We further propose a HarDNet-68, in which we remove the global dense connections and use MaxPool for down-sampling, and we change the BN-ReLU-Conv order proposed by DenseNet into the standard order of Conv-BN-ReLU to enable the folding of batch normalization. The dedicated growth rate k for each HDB in the HarDNet-68 enhances the CIO efficiency. Since a deep HDB has a larger number of input channels, a larger growth rate helps to balance the channel ratio between the input and output of a layer to meet our MoC constraint. For the layer distribution, instead of concentrating on stride-16 that is adopted by most of the CNN models, we let stride-8 to have the most layers in the HarDNet-68 that improves the local feature learning benefiting small-scale object detection. In contrast, classification tasks rely more on the global feature learning, so concentrating on the low resolution achieves a higher accuracy and a lower computational complexity.

### 3.4. CamVid Dataset

To study the performance of HDB, we replace all the blocks in a FC-DenseNet with HDBs. We follow the architecture of FC-DenseNet with an encoder-decoder structure and block level shortcuts to create models for semantic segmentation. For fair comparison, we made two reference architectures with exactly the same depth for

<table>
<thead>
<tr>
<th>Stride 2</th>
<th>7x7, 64, stride=2</th>
<th>3x3, 32, stride=2</th>
<th>3x3, 24, stride=2</th>
<th>3x3, 64, 1x1, 48</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stride 4</td>
<td>8 (HDB depth)</td>
<td>8</td>
<td>8</td>
<td>8, k=14, t=128</td>
</tr>
<tr>
<td>Stride 8</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16, k=16, t=256</td>
</tr>
<tr>
<td>Stride 16</td>
<td>16×2</td>
<td>16×3</td>
<td>16×3</td>
<td>16, k=40, t=640</td>
</tr>
<tr>
<td>Stride 32</td>
<td>16</td>
<td>16</td>
<td>16×2</td>
<td>4, k=160, t=1024</td>
</tr>
</tbody>
</table>

Table 1: Detailed implementation parameters. A “3x3, 64” stands for a Conv3x3 layer with 64 output channels, and the leading numbers below Stride 2 stand for an HDB with how many layers, followed by its growth rate k and a transitional Conv1x1 with t output channels.
Table 2: Results on CamVid dataset. The GPU inference time results are the accumulated measurements of CamVid test-set (233 pics) with a single-image batch size, running on pytorch-1.0.1 framework with a single NVIDIA TitanV GPU.

<table>
<thead>
<tr>
<th>Method</th>
<th>GNMACS, @352x480</th>
<th>Params (M)</th>
<th>CIO (MB)</th>
<th>GPU time (s)</th>
<th>Building</th>
<th>Tree</th>
<th>Sky</th>
<th>Car</th>
<th>Sign</th>
<th>Resid</th>
<th>Pedestrian</th>
<th>Fence</th>
<th>Pole</th>
<th>Sidewalk</th>
<th>Cyclist</th>
<th>Mean IoU</th>
<th>Global Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>FC-DenseNet67 [21]</td>
<td>60</td>
<td>1.4</td>
<td>1351</td>
<td>6.1</td>
<td>77.6</td>
<td>72.0</td>
<td>92.4</td>
<td>732.0</td>
<td>31.8</td>
<td>92.8</td>
<td>3.79</td>
<td>26.2</td>
<td>32.6</td>
<td>79.9</td>
<td>31.1</td>
<td>58.9</td>
<td>88.9</td>
</tr>
<tr>
<td>FC-DenseNet67 [21]</td>
<td>140</td>
<td>3.5</td>
<td>2286</td>
<td>10.2</td>
<td>80.2</td>
<td>75.4</td>
<td>93.0</td>
<td>78.2</td>
<td>40.9</td>
<td>94.7</td>
<td>58.4</td>
<td>30.7</td>
<td>38.4</td>
<td>81.9</td>
<td>52.1</td>
<td>65.8</td>
<td>90.8</td>
</tr>
<tr>
<td>FC-DenseNet103 [21]</td>
<td>134</td>
<td>9.4</td>
<td>2150</td>
<td>11.4</td>
<td>83.0</td>
<td>77.3</td>
<td>93.0</td>
<td>77.3</td>
<td>43.9</td>
<td>94.5</td>
<td>59.6</td>
<td>37.1</td>
<td>37.8</td>
<td>82.2</td>
<td>50.5</td>
<td>66.9</td>
<td>91.5</td>
</tr>
<tr>
<td>LogDenseNet103-17</td>
<td>137</td>
<td>4.7</td>
<td>2544</td>
<td>-</td>
<td>81.6</td>
<td>75.5</td>
<td>92.3</td>
<td>81.9</td>
<td>44.4</td>
<td>92.6</td>
<td>58.3</td>
<td>42.3</td>
<td>37.2</td>
<td>77.5</td>
<td>56.6</td>
<td>67.3</td>
<td>90.7</td>
</tr>
<tr>
<td>FC-DenseNet-ref100</td>
<td>142</td>
<td>3.5</td>
<td>3337</td>
<td>15.2</td>
<td>81.1</td>
<td>77.1</td>
<td>92.9</td>
<td>77.7</td>
<td>40.8</td>
<td>94.3</td>
<td>58.1</td>
<td>35.2</td>
<td>37.0</td>
<td>81.5</td>
<td>48.9</td>
<td>65.8</td>
<td>90.9</td>
</tr>
<tr>
<td>FC-SparseNet-ref100</td>
<td>223</td>
<td>3.2</td>
<td>2559</td>
<td>11.8</td>
<td>83.3</td>
<td>78.3</td>
<td>93.3</td>
<td>78.9</td>
<td>42.5</td>
<td>94.5</td>
<td>57.5</td>
<td>33.1</td>
<td>41.6</td>
<td>82.9</td>
<td>46.9</td>
<td>66.6</td>
<td>91.7</td>
</tr>
<tr>
<td>FC-HarDNet-ref100</td>
<td>151</td>
<td>3.6</td>
<td>2076</td>
<td>10</td>
<td>82.6</td>
<td>75.5</td>
<td>92.8</td>
<td>78.3</td>
<td>43.2</td>
<td>95.4</td>
<td>59.2</td>
<td>34.9</td>
<td>38.9</td>
<td>85.1</td>
<td>52.6</td>
<td>67.1</td>
<td>91.7</td>
</tr>
<tr>
<td>FC-HarDNet68</td>
<td>15</td>
<td>1.4</td>
<td>473</td>
<td>3.1</td>
<td>80.8</td>
<td>74.4</td>
<td>92.7</td>
<td>76.1</td>
<td>40.6</td>
<td>93.3</td>
<td>47.9</td>
<td>29.3</td>
<td>33.3</td>
<td>78.3</td>
<td>45.7</td>
<td>62.9</td>
<td>90.2</td>
</tr>
<tr>
<td>FC-HarDNet76</td>
<td>54</td>
<td>3.5</td>
<td>932</td>
<td>4.9</td>
<td>82.0</td>
<td>75.8</td>
<td>92.7</td>
<td>76.8</td>
<td>42.6</td>
<td>94.7</td>
<td>58.0</td>
<td>30.9</td>
<td>37.6</td>
<td>83.2</td>
<td>49.9</td>
<td>65.8</td>
<td>91.2</td>
</tr>
<tr>
<td>FC-HarDNet84</td>
<td>100</td>
<td>8.4</td>
<td>1267</td>
<td>6.7</td>
<td>81.4</td>
<td>76.2</td>
<td>92.9</td>
<td>78.3</td>
<td>48.9</td>
<td>94.6</td>
<td>61.9</td>
<td>37.9</td>
<td>38.2</td>
<td>80.5</td>
<td>54.0</td>
<td>67.7</td>
<td>91.1</td>
</tr>
</tbody>
</table>

Table 3: Parameters of FC-HarDNet and other reference networks, where FC-D, FC-S, and FC-H stand for FC-DenseNet, FC-SparseNet, and FC-HarDNet, respectively.
compared to FC-DenseNet-103. However, it also shows that the relative inference time of SegNet is much worse than on GPU. Thus, it confirmed that the relative DRAM traffic can be very different among platforms.

Pleiss et al. have mentioned that there is a concatenation overhead with the DenseNet implementation, which is caused by the explicit tensor copy from existing tensors to a new memory allocation. Therefore, it causes an additional DRAM traffic. To show that HarDNet still outperforms DenseNet when the overhead is discounted, we subtract the measured DRAM traffic volume by the traffic for tensor concatenation as the concat-free cases shown in Fig. 4a, where the DRAM traffic of concatenation is measured by Nvidia Profiler and broken down to the CatArrayBatched-Copy function. Fig. 4a shows that FC-DenseNet can reduce more DRAM traffic by discounting the concatenation than that for FC-HarDNet, but the latter still outperforms the former.

**4.2. ImageNet Datasets**

To train the six models of HarDNet for the ImageNet classification task, we reuse the torch7 training environment from [16, 18] and align all hyperparameters with them. To compare with other advanced CNN architectures such as ResNeXt and MobileNetV2 [31], we adopt more advanced hyperparameters such as the cosine learning rate decay and a fine-tuned weight decay. The HarDNet-68/39DS models are trained with a batch size of 256, an initial learning rate of 0.05 with cosine learning rate decay, and a weight decay of 6e-5.

Investigating the accuracy over CIO, it shows that HarDNet can outperform both ResNet and DenseNet while accuracy over model size is in between them as shown in Fig. 5(a)(b). Fig. 5c shows the GPU inference time results on Nvidia Titan V with torch7, which is quite similar to the trend of Fig. 5a and once again showing the high correlation between CIO and GPU inference time. However, the result also shows that for small models, there is no improvement of GPU inference time for HarDNet compared with ResNet, which we supposed to be due to the number of layers and the concatenation cost. We also argue that, once a discontinuous input tensor can be supported by a convolution operation, the inference time of DenseNet and HarDNet and be further reduced.

In Fig. 5d, we compare the state-of-the-art CNN model ResNeXt with our models trained with cosine learning rate decay. Although ResNeXt achieves a significant accuracy improvement with the same model size, there is still an inference time overhead with these models. Since there is no increase of MACs with the ResNeXt, the overhead can be explained by its increase of CIO.

In Table 4, we show the result comparison sorted by CIO for ImageNet, in which HarDNet68/39DS are also included. With the reduced number of layers, the cancel of global dense connections, and the BN-reordering, HarDNet-68 achieves a significant inference time reduction from the ResNet-50. For further comparing CIO between a model using standard convolutions and a model meal using
depth-wise-separable convolutions, we can apply a weight-
ing such as 0.6 on the CIO of the latter. After the weighting,
CIO can still be a rough prediction of inference time when
comparing among the two very different kinds of model.

4.3. Object Detection

We evaluate HarDNet-68 as a backbone model for a Sin-
gle Shot Detector (SSD) and train it with PASCAL VOC
2007 and MS COCO datasets. Aligned with the SSD-VGG,
we attach an ImageNet-pretrained HarDNet-68 to SSD at
the last layers in stride 8 and 16, respectively, and the HDB
in stride 32 is discarded. We insert a bridge module af-
fter the HDB on stride 16. The bridge module comprises a
3x3 max pooling with stride 1, a 3x3 convolution dilated
by 4, and a point-wise convolution, in which both convolu-
tional layers have 640 output channels. We train the model
with 300 and 150 epochs for VOC and COCO datasets, re-
spectively. The initial learning rate is 0.004 and decayed
by 10 times at epochs 60%, 80%, 90% of the maximum
epoch. The results in Table 5 show that our model achieve a
similar accuracy with SSD-ResNet101 despite its lower ac-
curacy in ImageNet, which shows the effectiveness of our
enhancement on stride 8 with 32 layers that improve the
local feature learning for the small-scale objects. Further-
more, HarDNet-68 is much faster than both VGG-16 and
ResNet-101, which make it very competitive in real time
applications.

5. Discussion

There is an assumption with the CIO, which is a CNN
model that is processed layer by layer without a fusion.
In contrast, fused-layer computation for multiple convolu-
tional layers has been proposed [1], in which intermediate
layers in a fused-layer group will not produce any memory
traffic for feature maps. In this case, the inverted residual
module in MobileNetV2 might be a better design to achieve
low memory traffic. Furthermore, the depth-wise convolu-
tion might be implemented as an element-wise operation
right before or after a neighboring layer. In such case, the
CIO for depth-wise convolution should be discounted.

Results show that CIO still failed to predict the actual
inference time in some cases such as comparing two
network models with significantly different architectures.

Figure 5: (a) ImageNet error rate vs. CIO. (b) Error rate vs. model size. (c) Error rate vs. GPU inference time on a single
TitanV with torch7. For GPU time of HarDNet-68, please refer to Table 4. (d) Further comparison among HarDNet-cosine
that is trained with cosine learning rate decay and ResNeXt.
Table 4: Test results for ImageNet models, in which GPU time is measured on Nvidia GTX1080 with Pytorch 1.1.0 and mobile GPU time is measured on Nvidia Jetson Nano with TensorRT-onnx at 320x320.

<table>
<thead>
<tr>
<th>Model</th>
<th>Param (M)</th>
<th>MACs (B)</th>
<th>CIO (M)</th>
<th>GPU Time (ms)</th>
<th>Mobile GPU (ms)</th>
<th>Top1 Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>HarDNet 39DS</td>
<td>3.5</td>
<td>0.44</td>
<td>8.2</td>
<td>17.8</td>
<td>32.5</td>
<td>72.1</td>
</tr>
<tr>
<td>MobileNetV2</td>
<td>3.5</td>
<td>0.32</td>
<td>13.4</td>
<td>23.7</td>
<td>37.9</td>
<td>72.0</td>
</tr>
<tr>
<td>HarDNet 68DS</td>
<td>4.2</td>
<td>0.79</td>
<td>15.3</td>
<td>31.7</td>
<td>52.6</td>
<td>74.3</td>
</tr>
<tr>
<td>MNetV2 1.4x</td>
<td>6.1</td>
<td>0.59</td>
<td>18.5</td>
<td>33.0</td>
<td>57.8</td>
<td>74.7</td>
</tr>
<tr>
<td>ResNet 18</td>
<td>11.7</td>
<td>1.8</td>
<td>4.7</td>
<td>13.0</td>
<td>-</td>
<td>69.6</td>
</tr>
<tr>
<td>SqueezeNet 1.0</td>
<td>1.2</td>
<td>0.83</td>
<td>7.9</td>
<td>19.6</td>
<td>-</td>
<td>60.4</td>
</tr>
<tr>
<td>HarDNet 68</td>
<td>17.6</td>
<td>4.3</td>
<td>11.5</td>
<td>32.6</td>
<td>-</td>
<td>76.2</td>
</tr>
<tr>
<td>HarDNet 96x</td>
<td>9.3</td>
<td>2.5</td>
<td>11.7</td>
<td>36.4</td>
<td>-</td>
<td>74.7</td>
</tr>
<tr>
<td>HarDNet 117s</td>
<td>20.9</td>
<td>4.9</td>
<td>16.7</td>
<td>57.6</td>
<td>-</td>
<td>77.0</td>
</tr>
<tr>
<td>HarDNet 138s</td>
<td>35.5</td>
<td>6.7</td>
<td>19.6</td>
<td>70.5</td>
<td>-</td>
<td>77.8</td>
</tr>
<tr>
<td>ResNet 50</td>
<td>25</td>
<td>4.1</td>
<td>20.7</td>
<td>46.5</td>
<td>-</td>
<td>76.0</td>
</tr>
<tr>
<td>DenseNet 121</td>
<td>7.9</td>
<td>2.9</td>
<td>21.9</td>
<td>51.5</td>
<td>-</td>
<td>75.0</td>
</tr>
<tr>
<td>VGG-16</td>
<td>138</td>
<td>15.5</td>
<td>22.6</td>
<td>79.3</td>
<td>-</td>
<td>73.4</td>
</tr>
<tr>
<td>ResNet 101</td>
<td>44.5</td>
<td>7.6</td>
<td>30.9</td>
<td>76.9</td>
<td>-</td>
<td>77.6</td>
</tr>
<tr>
<td>DenseNet 201</td>
<td>20</td>
<td>4.4</td>
<td>31.8</td>
<td>83.9</td>
<td>-</td>
<td>77.4</td>
</tr>
<tr>
<td>ResNet 152</td>
<td>60.2</td>
<td>11.3</td>
<td>43.6</td>
<td>109.7</td>
<td>-</td>
<td>77.8</td>
</tr>
</tbody>
</table>

Table 5: Results in object detection. The comparison data is from [35]

<table>
<thead>
<tr>
<th>Backbone Model</th>
<th>VOC 2007 mAP</th>
<th>COCO mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSD512 VGG-16</td>
<td>79.8</td>
<td>28.8</td>
</tr>
<tr>
<td>SSD513 ResNet-101</td>
<td>80.6</td>
<td>31.2</td>
</tr>
<tr>
<td>SSD512 HarDNet-68</td>
<td>81.5</td>
<td>30.2</td>
</tr>
</tbody>
</table>

As we mentioned before, CIO dominates inference time only when the MoC is below a certain ratio, which is a density of computation within a space of data traffic. In a network model, each of the layers has a different MoC. In some of the layers CIO may dominate, but for the other layers, MACs can still be the key factor if its computational density is relatively higher. To precisely predict the inference latency of a network, we need to breakdown to each of the layers and investigate its MoC to predict the inference latency of the layer.

We would like to emphasize the importance of DRAM traffic furthermore. Since the quantization has been widely used for CNN models, both the hardware cost of multiplier and data traffic can be reduced. However, the hardware cost reduction of a multiplier from float32 to int8 is much greater than the reduction of data traffic from the same thing. When developing hardware platform mainly using int8 multipliers, computing power can grow more quickly than the data bandwidth, so data traffic will be even more important in this case. We argue that the best way to achieve the traffic reduction is to increase MoC reasonably for a network model, which might be counter-intuitive to the widely-accepted knowledge of that using more Conv1x1 achieves a higher efficiency. In many cases, we have shown that it is indeed helpful, however.

6. Conclusion

We have presented a new metric for evaluating a convolutional neural network by estimating its DRAM traffic for feature maps, which is a crucial factor affecting the power consumption of a system. When the density of computation is low, the traffic can dominate inference time more significantly than the model size and operation count. We employ Convolutional Input/Output (CIO) as an approximation of the DRAM traffic, and propose a Harmonic Densely Connected Networks (HarDNet) that achieve a high accuracy-over-CIO and also a high computational efficiency by increasing the density of computation (MACs over CIO).

Experiments showed that the proposed connection pattern and channel balancing have made FC-HarDNet to achieve DRAM traffic reduction by 40% and GPU inference time reduction by 35% compared with FC-DenseNet. Comparing with DenseNet-264 and ResNet-152, HarDNet-138s achieves the same accuracy with a GPU inference time reduction by 35%. Comparing with ResNet-50, HarDNet-68 achieves an inference time reduction by 30%, which is also a desirable backbone model for object detections that enhances the accuracy of a SSD to be higher than using ResNet-101 in PASCAL VOC dataset while the inference time is also significantly reduced from SSD-VGG. In summary, in addition to accuracy-over-model-size and accuracy-over-MACs tradeoffs, we demonstrated that accuracy-over-DRAM-traffic-for-feature-maps is indeed an important consideration when designing neural network architectures.

Acknowledgement

This research is supported in part by a grant from the Ministry of Science and Technology (MOST) of Taiwan. We would like to express our gratitude to Taiwan Computing Cloud (TWCC) for providing us with a powerful and stable cloud machine along with top-rated technical support. Without it this research is impossible.

References


