Accelerating CNN inference on FPGAs: A Survey
Kamel Abdelouahab, Maxime Pelcat, François Berry, Jocelyn Sérot

To cite this version:
Kamel Abdelouahab, Maxime Pelcat, François Berry, Jocelyn Sérot. Accelerating CNN inference on FPGAs: A Survey. 2018. hal-01695375v2

HAL Id: hal-01695375
https://hal.archives-ouvertes.fr/hal-01695375v2
Submitted on 13 Mar 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Accelerating CNN inference on FPGAs: A Survey

Kamel Abdelouahab\textsuperscript{1}, Maxime Pelcat\textsuperscript{1,2}, Jocelyn Sérot\textsuperscript{1}, and François Berry\textsuperscript{1}

\textsuperscript{1}Institut Pascal, Clermont Ferrand, France
\textsuperscript{2}IETR, INSA Rennes, France

January 2018
Abstract

Convolutional Neural Networks (CNNs) are currently adopted to solve an ever greater number of problems, ranging from speech recognition to image classification and segmentation. The large amount of processing required by CNNs calls for dedicated and tailored hardware support methods. Moreover, CNN workloads have a streaming nature, well suited to reconfigurable hardware architectures such as FPGAs.

The amount and diversity of research on the subject of CNN FPGA acceleration within the last 3 years demonstrates the tremendous industrial and academic interest. This paper presents a state-of-the-art of CNN inference accelerators over FPGAs. The computational workloads, their parallelism and the involved memory accesses are analyzed. At the level of neurons, optimizations of the convolutional and fully connected layers are explained and the performances of the different methods compared. At the network level, approximate computing and datapath optimization methods are covered and state-of-the-art approaches compared. The methods and tools investigated in this survey represent the recent trends in FPGA CNN inference accelerators and will fuel the future advances on efficient hardware deep learning.
1 Introduction

The exponential growth of big data during the last decade motivates for innovative methods to extract high semantic information from raw sensor data such as videos, images and speech sequences. Among the proposed methods, Convolutional Neural Networks (CNNs) \cite{1} have become the de-facto standard by delivering near-human accuracy in many applications related to machine vision (e.g. classification \cite{2}, detection \cite{3}, segmentation \cite{4}) and speech recognition \cite{5}.

This performance comes at the price of a large computational cost as CNNs require up to 38 GOP/s to classify a single frame \cite{6}. As a result, dedicated hardware is required to accelerate their execution. Graphics Processing Units (GPUs), are the most widely used platform to implement CNNs as they offer the best performance in terms of pure computational throughput, reaching up 11 TFLOP/s \cite{7}. Nevertheless, in terms of power consumption, Field-Programmable Gate Array (FPGA) solutions are known to be more energy efficient (vs GPUs). As a result, numerous FPGA-Based CNN accelerators have been proposed, targeting both High Performance Computing (HPC) data-centers \cite{8} and embedded applications \cite{9}.

While GPU implementations have demonstrated state-of-the-art computational performance, CNN acceleration is shortly moving towards FPGAs for two reasons. First, recent improvements in FPGA technology put FPGA performance within striking distance to GPUs with a reported performance of 9.2 TFLOP/s for the latter \cite{10}. Second, recent trends in CNN development increase the sparsity of CNNs and use extreme compact data types. These trends favorize FPGA devices which are designed to handle irregular parallelism and custom data types. As a result, next generation CNN accelerators are expected to deliver up to x5.4 better computational throughput than GPUs. \cite{7}.

As an inflection point in the development of CNN accelerators might be near, we conduct a survey on FPGA-Based CNN accelerators. While a similar survey can be found in \cite{11}, we focus in this paper on the recent techniques that were not covered in the previous works. Moreover, a recent review of efficient processing techniques for deep learning is proposed in \cite{12}, but focuses on Application Specific Integrated Circuits (ASIC) accelerators for CNNs while our work is mainly related to FPGA-based implementations.

The rest of the paper is organized as follows, section 2 recalls the main features of CNNs, focusing on computations and workload issues. Section 3 studies the computational transforms exploited to accelerate CNNs on FPGAs. Section 4 reviews the contributions that attempt to optimize the data-path of FPGA-Based CNN accelerators. Section 5 shows how approximate computing is a key in the acceleration of CNNs on FPGAs and overviews the main contributions implementing these techniques. Finally, section 6 concludes the paper.

2 Background on CNNs

This section overviews the main features of CNNs and focuses on the computations and parallelism patterns involved during their inference.

2.1 General Overview:

CNNs are feed-forward, deep, sparsely connected neural networks that implement weight sharing. A typical CNN structure consists of a pipeline of layers. Each layer inputs a set of data, known as a Feature Map (FM), and produces a new set of FMs with higher-level semantics.

2.2 Inference vs Training:

As typical Machine Learning (ML) algorithms, CNNs are deployed in two phases. First, the training stage works on a known set of annotated data samples to create a model with a modeling power (i.e. which semantics extrapolates to natural data outside the training set). This phase implements the back-propagation algorithm \cite{13}.
which iteratively updates CNN parameters such as convolution weights to improve the predictive power of the model. CNN Models can also be fine-tuned. When fine-tuning a model, weights of a previously-trained network are used to initialize the parameters of a new training. These weights are then adjusted for a new constrain, such as a different dataset or a reduced precision.

The second phase, known as inference, uses the learned model to classify new data samples (i.e inputs that were not previously seen by the model). In a typical setup, CNNs are trained/fine-tuned only once, on large GPU/FPGA clusters. By contrast, the inference is implemented each time a new data sample has to be classified. As a consequence, the literature mostly focuses on accelerating the inference phase. As a result, this paper overviews the main methods employed to accelerate the inference. Moreover, since most of the CNN accelerators benchmark their performance on models trained for image classification, we focus on this paper on this application. Nonetheless, the methods studied in this survey can be employed to accelerate CNNs for other applications such object detection, image segmentation and speech recognition.

2.3 Inference of CNNs

CNN inference refers to the feed-forward propagation of B input images across L layers. This section details the computations involved in the major types of these layers. A common practice is to manipulate layer parameters and FMs using tensors. The tensors and variables used in this work are listed in table 1.

<table>
<thead>
<tr>
<th>X</th>
<th>Input FMs</th>
<th>$B \times C \times H \times W$</th>
<th>Y</th>
<th>Output FMs</th>
<th>$B \times N \times V \times U$</th>
<th>$\Theta$</th>
<th>Learned Filters</th>
<th>$N \times C \times J \times K$</th>
<th>$\beta$</th>
<th>Learned biases</th>
<th>$N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>Batch size (Number of input frames)</td>
<td>$W/H/C$</td>
<td>$V/N$</td>
<td>$U$</td>
<td>Width / Height / Depth of Input FMs</td>
<td>Width / Height / Depth of Output FMs</td>
<td>Horizontal / Vertical Kernel size</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.3.1 Convolution layers:

A convolution layer (conv) carries out the feature extraction process by applying –as illustrated in figure 1– a set of 3D-convolution filters $\Theta^{\text{conv}}$ to a set of B input volumes $X^{\text{conv}}$. Each input volume has a depth C and can be a color image (in the case of the first conv layer), or an output generated by previous layers in the network. Applying a 3D-filter to 3D-input results in a 2D Feature Map (FM) and, each conv layer outputs a set of N two-dimensional features maps. In some CNN models, a learned offset $\beta^{\text{conv}}$ –called a bias– is added to the 3D-conv results, but this practice is discarded in recent models [6]. The computations involved in feed-forward propagation of conv layers are detailed in equation 1.

$$\forall \{b, n, u, v\} \in [1, B] \times [1, N] \times [1, V] \times [1, U]$$

$$Y^{\text{conv}}[b, n, v, u] = \beta^{\text{conv}}[n] + \sum_{c=1}^{C} \sum_{j=1}^{J} \sum_{k=1}^{K} X^{\text{conv}}[b, c, v + j, u + k] . \Theta^{\text{conv}}[n, c, j, k]$$

1The computational transforms discussed in sections 3 and approximate computing techniques detailed in section 5 can both be employed during the training and the inference.
2.3.2 Activation Layers:

Each conv layer of a CNN is usually followed by an activation layer that applies a non-linear function to all the values of FMs. Early CNNs were trained with TanH or Sigmoid functions but recent models employ the Rectified Linear Unit (ReLU) function that grants faster training times and less computational complexity, as highlighted in [14].

\[
\forall \{b, n, u, v\} \in [1, B] \times [1, N] \times [1, V] \times [1, U]
Y_{\text{act}}[b, n, h, w] = \text{act}(X_{\text{act}}[b, n, h, w]) \quad | \quad \text{act} := \text{TanH, Sigmoid, ReLU} \ldots \tag{2}
\]

2.3.3 Pooling layers:

The convolutional and activation parts of a CNN are directly inspired by the cells of visual cortex in neuroscience [15]. This is also the case of pooling layers, which are periodically inserted in-between successive conv layers. As shown in equation 3 pooling sub-samples each channel of the input FMs by selecting the average, or, more commonly, the maximum of a given neighborhood \(K\). As a results, the dimensionality of a FMs is reduced, as illustrated in figure 1.

\[
\forall \{b, n, u, v\} \in [1, B] \times [1, N] \times [1, V] \times [1, U]
Y_{\text{pool}}[b, n, v, u] = \max_{p, q \in [1:K]} \left(X_{\text{pool}}[b, n, v + p, u + q]\right) \tag{3}
\]

2.3.4 Fully Connected Layers:

When deployed for classification tasks, the CNNs pipeline is often terminated by Fully Connected (FC) layers. These layers can be seen as conv layers with no weight sharing (i.e \(W = K\) and \(H = J\)). Moreover, in a same way as conv layers, a non-linear function is applied to the outputs of FC layers.

\[
\forall \{b, n\} \in [1, B] \times [1, N]
Y_{\text{fc}}[b, n] = \beta_{\text{fc}}[n] + \sum_{c=1}^{C} \sum_{h=1}^{H} \sum_{w=1}^{W} X_{\text{fc}}[b, c, h, w] \Theta_{\text{fc}}[n, c, h, w] \tag{4}
\]
2.3.5 Batch-Normalization Layers:

Batch-Normalization is introduced in [16] to speed up training by linearly shifting and scaling the distribution of a given batch of inputs $B$ to have zero mean and unit variance. These layers find also there interest when implementing Binary Neural Network (BNN) (cf section 5.1.3) by reducing the quantization error compared to an arbitrary input distribution, as highlighted in [17]. Equation 5 details the processing of batch norm layers, where $\mu$ and $\sigma$ are statistic collected during the training, $\alpha$, $\beta$ and $\gamma$ parameters are training hyper-parameters.

$$\forall \{b, n, u, v\} \in [1, B] \times [1, N] \times [1, V] \times [1, U]$$

$$Y_{BN}[b, n, u, v] = \frac{X_{BN}[b, n, u, v] - \mu}{\sqrt{\sigma^2 + \epsilon}} \gamma + \alpha$$

(5)

2.4 Workload of a CNNs inference

Table 2: Popular CNN models with their computational workload. Accuracy measured on single-crops of ImageNet test-set.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Top1 err</td>
<td>42.9 %</td>
<td>31.3 %</td>
<td>28.1 %</td>
<td>27.3 %</td>
<td>24.7 %</td>
<td>23.6 %</td>
<td>23.0 %</td>
</tr>
<tr>
<td>Top5 err</td>
<td>19.80 %</td>
<td>10.07 %</td>
<td>9.90 %</td>
<td>9.00 %</td>
<td>7.8 %</td>
<td>7.1 %</td>
<td>6.7 %</td>
</tr>
<tr>
<td>conv layers</td>
<td>5</td>
<td>57</td>
<td>13</td>
<td>16</td>
<td>53</td>
<td>104</td>
<td>155</td>
</tr>
<tr>
<td>conv workload (MACs)</td>
<td>666 M</td>
<td>1.58 G</td>
<td>15.3 G</td>
<td>19.5 G</td>
<td>3.86 G</td>
<td>7.57 G</td>
<td>11.3 G</td>
</tr>
<tr>
<td>conv parameters</td>
<td>2.33 M</td>
<td>5.97 M</td>
<td>14.7 M</td>
<td>20 M</td>
<td>23.5 M</td>
<td>42.4 M</td>
<td>58 M</td>
</tr>
<tr>
<td>Activation layers</td>
<td>ReLU</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pool layers</td>
<td>3</td>
<td>14</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>FC layers</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>FC workload (MACs)</td>
<td>58.6 M</td>
<td>1.02 M</td>
<td>124 M</td>
<td>124 M</td>
<td>2.05 M</td>
<td>2.05 M</td>
<td>2.05 M</td>
</tr>
<tr>
<td>FC parameters</td>
<td>58.6 M</td>
<td>1.02 M</td>
<td>124 M</td>
<td>124 M</td>
<td>2.05 M</td>
<td>2.05 M</td>
<td>2.05 M</td>
</tr>
<tr>
<td>Total workload (MACs)</td>
<td>724 M</td>
<td>1.58 G</td>
<td>15.5 G</td>
<td>19.6 G</td>
<td>3.86 G</td>
<td>7.57 G</td>
<td>11.3 G</td>
</tr>
<tr>
<td>Total parameters</td>
<td>61 M</td>
<td>6.99 M</td>
<td>138 M</td>
<td>144 M</td>
<td>25.5 M</td>
<td>44.4 M</td>
<td>60 M</td>
</tr>
</tbody>
</table>

The accuracy of CNN models have been increasing since their breakthrough in 2012 [14]. However, this accuracy comes at the price of a high computational cost. The main challenge that faces CNN developers is to improve classification accuracy while maintaining a tolerable computational workload. As shown in table 2, this challenge was successfully addressed by Inception [18] and ResNet models [19], with their use of bottleneck $1 \times 1$ convolutions that reduce both model size and computations while increasing depth and accuracy.

2.4.1 Computational Workload:

The computational workload of a CNN inference is the result of an intensive use of the Multiply Accumulate (MAC) operation. Most of these MACs occur on the convolutional parts of the network, as shown in tab 2. As a consequence, conv layers are responsible, in a typical implementation, of more than 90% of execution time during the inference [20]. Conversely to computations, and as shown in tab 2, most of the CNN weights are included on the FC-layers. Due to this unbalanced computation to memory ratio, CNNs accelerators follow different strategies when implementing the convolutional and fully connected parts of inference.
2.4.2 Parallelism in CNNs:

Because of the high number of required computations, inferring CNNs with real-time constraints is a challenge, especially on low-energy embedded devices. A solution to this challenge is to take advantage of the extensive concurrency exhibited by CNNs. These sources can be formalized as:

- **Batch Parallelism**: CNN implementations can simultaneously classify multiple frames grouped as a batch $B$ in order to reuse the filters in each layer and minimize the external memory accesses. As a result, the inference benefits from a significant acceleration when implementing batch processing.

- **Inter-layer Parallelism**: CNNs have a feed-forward hierarchical structure consisting of a succession of data-dependent layers. These layers can be executed in a pipelined fashion by launching layer $(\ell)$ before ending the execution of layer $(\ell - 1)$.

Moreover, the computation of each conv layer, described in eq.1, exhibits four sources of concurrency that are detailed above.

- **Inter-FM Parallelism**: Each output FM plane of a conv layer can be processed separately from the others. This means that $P_N$ elements of $Y_{\text{conv}}$ can be computed in parallel ($0 < P_N < N$).

- **Intra-FM Parallelism**: Multiple pixels of a single output FM plane can be processed concurrently by evaluating $P_V \times P_U$ Values of $Y_{\text{conv}}[n] (0 < P_V \times P_U < V \times U)$

- **Inter-convolution Parallelism**: 3D-convolutions occurring in conv layers can be expressed as a sum of 2D convolutions as shown in equation 6. These 2D convolutions can be evaluated simultaneously by computing concurrently $P_C$ elements of eq.6 ($0 < P_C < C$).

- **Intra-convolution Parallelism**: The 2D-convolutions involved in the processing of conv layers can be implemented in a pipelined fashion as in [21]. In this case $P_J \times P_K$ multiplications are implemented concurrently ($0 < P_J \times P_K < J \times K$).

\[
\forall \{b, n\} \in [1, B] \times [1, N]
Y_{\text{conv}}[n] = b[n] + \sum_{c=1}^{C} \text{conv2D}(X_{\text{conv}}[c], \Theta_{\text{conv}}[n, c])
\]  

(6)

2.4.3 Memory Accesses in CNNs:

The CNN inference shows large vectorization opportunities that are exploited by allocating multiple computational resources to accelerate the processing. However, this method may be inefficient if no caching strategy is implemented.

In fact, memory bandwidth is often the bottleneck when processing CNNs. For the FC parts, execution can be memory-bounded because of the high number of weights that these layers contain, and consequently, the high number of memory reads engendered. For the conv parts, the high number of MAC operations results in a high amount of memory accesses because each MAC requires at least 2 memory reads and 1 memory write to be performed. If all these accesses are towards external memory (for instance, Dynamic Random Access

\footnote{This is the best case scenario of a fully pipelined MAC where intermediate results don’t need to be loaded.}
Memory (DRAM)), throughput and energy consumption will be highly impacted since a DRAM access engenders significantly more of latency and energy consumption than the computation itself [22].

The number of these DRAM accesses, and thus latency and energy consumption, can be reduced by implementing a memory caching hierarchy using on-chip memories. As discussed in section 4, hardware accelerators for CNNs usually employ two levels of caches. The first level is implemented by means of large on-chip buffers while the second level involves local register files implemented at the nearest of the computational capabilities. The latency and energy consumption that result from memory access toward these 2 cache levels is several order of magnitude less then external memory access, as pointed-out in [12].

### 2.4.4 Hardware, libraries and frameworks:

In order to catch the parallelism of CNNs, dedicated hardware accelerators are developed. Most of them are based on GPU, which that are known to perform well on regular parallelism patterns thanks to a Single Instruction on Multiple Data (SIMD) and Single Instruction on Multiple Threads (SIMD) execution models, a dense collection of floating-point computing elements that peaks at 12 TFLOPs, and high capacity/bandwidth on/off-chip memories [23]. To support these hardware accelerators, specialized libraries for deep learning are developed to provide the necessary programming abstraction, such as CudNN on Nvidia GPUs [24] and DeepCL on heterogeneous hardware through OpenCL standard [25]. Built-upon these libraries, dedicated frameworks for deep learning are proposed to improve productivity of conceiving, training and deploying CNNs, such as Caffe [26] and TensorFlow [27].

Beside GPU implementations, numerous FPGA accelerators for CNNs have been proposed. FPGAs are fine-grain programmable devices that can catch the CNN parallelism patterns with no memory bottleneck, tanks to

1. A High density of hard-wired Digital Signal Processing (DSP) blocs that are able to achieve up to 20 (8 TFLOPs) TMACs [10].
2. A collection of In-situ on-chip memories, located next to DSPs, that can be exploited to significantly reduce the number of external memory accesses.

When porting a CNN to an FPGA device, the problem boils down to finding an efficient mapping between the computational model of the former and the execution model supported by the latter. In the the following sections, the main strategies explored by the literature to address this mapping problem are reviewed. In particular, we show that current FPGA-based accelerators for CNNs rely on one (or a combination) of three main optimizations to efficiently infer CNNs.

![Figure 2: Main Approaches to Accelerate CNN inference on FPGAs](image-url)
3 Algorithmic Optimizations for FPGA-Based CNN Acceleration

In order to accelerate the execution of conv and FC layers, computational transforms are employed on the FMs and kernels in order to vectorize the implementations and reduce the number of arithmetic operations occurring during inference. These computational transforms are mainly deployed in CPUs and GPU and are implemented by means of variety of software libraries such OpenBlas CPUs and cuBLAS for GPUs. Beside this, various implementations make use of such transforms to map CNNs on FPGAs.

3.1 GEMM Transformation

In Central Processing Units (CPUs) and GPUs, a common way to process CNNs is to map conv and FC layers as General Matrix Multiplications (GEMMs). The OpenCL standard generalizes this approach to FPGAs-based implementations [63, 64].

For FC layers, in which the processing boils down to a matrix-vector multiplication problem, the GEMM-based implementations find its interest when processing a batch of FMs. In this case, the batch is concatenated onto a CHW × B matrix, as shown in fig 3a.

As mentioned in section 2.4.1, most of the weights of CNNs are employed in the FC parts. Instead of loading these weights multiple times to classify multiple inputs, feature maps of FC layers are batched in a way that FC weights are loaded only one time per batch. This vectorization is employed in [65, 66, 30] to increase the computational throughput in FC layers while maintaining a constant memory bandwidth utilization. Moreover, the efficiency of this method increases as the sparsity of $\Theta^{fc}$ grows (cf. sec 5.2).

\[
\tilde{Y}^{conv} = \tilde{\Theta}^{conv} \times \tilde{X}^{conv}
\]

Figure 3: GEMM Based processing of: a- FC layers, b- conv layers.
3.2 Winograd Transform

Winograd minimal filter algorithm, introduced in [68], is a computational transform that can be applied to convolutions when the stride is 1. Winograd convolutions are particularly efficient when processing small convolutions ($K \leq 3$), as demonstrated in [69]. In this work, authors report an acceleration up to $x7.28$ when compared to classical GEMM based implementation of convolutions when executing VGG16 on a TitanX GPU.

In Winograd filtering, data is processed by blocks referred as tiles, as following:

1. An input FM tile $x$ of size $(u \times u)$ is pre-processed: $\tilde{x} = A^T x A$

2. In a same way, the filter tile of size $(k \times k)$ is transformed into $\tilde{\theta}$: $\tilde{\theta} = B^T x B$

3. Winograd filtering algorithm, denoted $F(u \times u, k \times k)$, outputs a tile $y$ of size $(u \times u)$ that is computed according to equation 8

$$y = C^T \left( \tilde{\theta} \odot \tilde{x} \right) C$$

where $A, B, C$ are transformation matrices defined in the Winograd algorithm [68] and $\odot$ denotes the Hadamard product or Element-Wise Matrix Multiplication (EWM).

While a standard filtering requires $u^2 \times k^2$ multiplications, Winograd algorithm $F(u \times u, k \times k)$ requires $(u+k-1)^2$ multiplications [68]. In the case of tiles of a size $u = 2$ and kernels of size $k = 3$, this corresponds to an arithmetic complexity reduction of $x2.25$ [69]. In return, the number of additions is increased.

Beside this complexity reduction, implementing Winograd filtering in FPGA-Based CNN accelerators has two advantages. First, transformation matrices $A, B, C$ can be generated off-line once $u$ and $k$ are determined. As a result, these transforms become multiplications with the constants that can be implemented by means of Lookup Table (LUT) and shift registers, as proposed in [70].

Second, Winograd filtering can employ the loop optimization techniques discussed in section 4.2 to vectorize the implementation. On one hand, the computational throughput is increased when unrolling the computation of the EWMM parts on an array of DSP blocks. On the other hand, memory bandwidth is optimized using loop tiling to determine the size FM tiles and filter buffers.

First utilization of Winograd filtering in FPGA-Based CNN accelerators is proposed in [31] and delivers a computational throughput of 46 GOPs when executing AlexNet convolution layers. This performance is significantly by a factor of $x42$ in [30] when optimizing the datapath to support Winograd convolutions (by employing loop unrolling and tiling strategies), and storing the intermediate FM in on-chip buffers (cf sec 4). The same methodology is employed in [70] to derive a CNN accelerator on a Xilinx ZCU102 device. This accelerator delivers a throughput of 2.94 TOPs on VGG convolutional layers, which corresponds to half of the performance of a TitanX device, with $x5.7$ less power consumption.

3.3 Fast Fourier Transform

Fast Fourier Transform (FFT) is a well known algorithm to transform the 2D convolutions into EWMM in the frequency domain, as shown in equation 9

$$\text{conv2D}(X[c], \Theta[n, c]) = \text{IFFT} \left( \text{FFT}(X[c]) \odot \text{FFT}(\Theta[n, c]) \right)$$

Using FFT to process 2D convolutions reduces the arithmetic complexity to $O(W^2 \log_2(W))$, which is exploited to derive FPGA-based accelerators to train CNNs [33]. When compared to standard filtering and Winograd algorithm, FFT finds its interest in convolutions with large kernel size ($K > 5$), as demonstrated in [69, 63]. The computational complexity of FFT convolutions can be further reduced to $O(W \log_2(K))$ using the Overlap-and-Add Method [71] that can be applied when the signal size is much larger than the filter size, which is the case in

$^3$Implementation in the TitanX GPU employs Winograd algorithm and 32 bits floating point arithmetic
<table>
<thead>
<tr>
<th>Network</th>
<th>Network Workload</th>
<th>Bitwidth</th>
<th>Desc.</th>
<th>Device</th>
<th>Freq (MHz)</th>
<th>Through (GOPS)</th>
<th>Power (W)</th>
<th>LUT (K)</th>
<th>DSP (MB)</th>
<th>Memory (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Winograd</td>
<td>AlexNet-C</td>
<td>1.3</td>
<td>2.3</td>
<td>Float 32</td>
<td>OpenCL</td>
<td>Virtex7 VX690T</td>
<td>200</td>
<td>46</td>
<td>505</td>
<td>3683</td>
</tr>
<tr>
<td></td>
<td>AlexNet-C</td>
<td>1.3</td>
<td>2.3</td>
<td>Float16</td>
<td>OpenCL</td>
<td>Arria10 GX1150</td>
<td>303</td>
<td>1382</td>
<td>44.3</td>
<td>246</td>
</tr>
<tr>
<td></td>
<td>AlexNet-C</td>
<td>1.3</td>
<td>2.3</td>
<td>Fixed 16</td>
<td>HLS</td>
<td>Zynq ZU9EG</td>
<td>200</td>
<td>3045</td>
<td>23.6</td>
<td>600</td>
</tr>
<tr>
<td>FFT</td>
<td>AlexNet-C</td>
<td>1.3</td>
<td>2.3</td>
<td>Float 32</td>
<td>Stratix5 QPI</td>
<td>200</td>
<td>83</td>
<td>13.2</td>
<td>201</td>
<td>224</td>
</tr>
<tr>
<td></td>
<td>VGG16-C</td>
<td>30.6</td>
<td>14.7</td>
<td>Fixed 32</td>
<td>OpenCL</td>
<td>Stratix5 GXA7</td>
<td>194</td>
<td>66</td>
<td>33.9</td>
<td>228</td>
</tr>
<tr>
<td>GEMM</td>
<td>AlexNet-C</td>
<td>1.3</td>
<td>2.3</td>
<td>Fixed 32</td>
<td>HLS</td>
<td>Virtex7 VX960T</td>
<td>150</td>
<td>354</td>
<td>26.0</td>
<td>351</td>
</tr>
<tr>
<td></td>
<td>VGG16-F</td>
<td>31.1</td>
<td>138.0</td>
<td>Fixed 32</td>
<td>OpenCL</td>
<td>Arria10 GX1150</td>
<td>370</td>
<td>866</td>
<td>41.7</td>
<td>437</td>
</tr>
</tbody>
</table>

Conv layers ($W >> K$). Works in [32] exploit this method to implement frequency domain acceleration for conv layers on FPGA, which results in a computational throughput of 83 GOPs for AlexNet.

4 Data-path Optimizations for FPGA-Based CNN Accelerators

As highlighted in sec 2.4.2, the execution of CNNs exhibit numerous sources of parallelism. However, because of the resource limitation of FPGAs devices, it is impossible to fully exploit all the parallelism patterns, especially with the sheer volume of operations involved in deep topologies. In other words, the execution of recent CNN models can not fully be “Unrolled”, sometimes, not even for a single conv layer. To address this problem, the main approach that state-of-the-art implementations advocates, is to map a limited number of Processing Elements (PEs) on the FPGA. These PEs are reused by temporally iterating data through them.

![Figure 4: Generic Data-paths of FPGA-based CNN accelerators](image-url)

4.1 Systolic Arrays

Early FPGA-based accelerators for CNNs implemented systolic arrays to accelerate the 2D filtering in convolutions layers [72,73,74,75,76]. As illustrated in figure 4a, systolic arrays employ a static collection of PEs, typically...
Table 4: Loop Optimization Parameters $P_i$ and $T_i$

<table>
<thead>
<tr>
<th>Parallelism</th>
<th>Intra-layer</th>
<th>Inter-FM</th>
<th>Intra-FM</th>
<th>Inter-Convolution</th>
<th>Intra-Convolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loop</td>
<td>$L_L$</td>
<td>$L_N$</td>
<td>$L_V$</td>
<td>$L_U$</td>
<td>$L_C$</td>
</tr>
<tr>
<td>Unroll factor</td>
<td>$P_L$</td>
<td>$P_N$</td>
<td>$P_V$</td>
<td>$P_U$</td>
<td>$P_C$</td>
</tr>
<tr>
<td>Tiling Factor</td>
<td>$T_L$</td>
<td>$T_N$</td>
<td>$T_U$</td>
<td>$T_C$</td>
<td>$T_J$</td>
</tr>
</tbody>
</table>

arranged in a 2-dimensional grid, that operates under the control of a CPU. This static collection of PEs is agnostic to the CNN model configuration. It can only support convolutions with a kernel size $K$ that is smaller than a given maximum size $K_m$ (i.e., support only convolutions such $K \leq K_m$ where, for instance, $K_m = 7$ in [73] and $K_m = 10$ in [76]). Moreover, when performing convolutions with a smaller kernel size then $K_m (K << K_m)$, only a small part of computing capabilities is used. For instance in [76], processing $3 \times 3$ convolutions uses only 9% of DSP Blocs. Finally, these systolic arrays do not implement data caching and requires to fetch inputs from off-chip memory. As a result, their performance is bounded by memory bandwidth of the device.

4.2 SIMD Accelerators and Loop Optimization

Due to inefficiency of static systolic arrays, flexible SIMD accelerators for CNNs on FPGAs were proposed. The general computation flow in these accelerators—illustrated in Fig. 4c—a is to fetch FMs and weights from DRAM to on-chip buffers. These data are then streamed into the PEs. At the end of the PE computation, results are transferred back to on-chip buffers and, if necessary, to the external memory in order to be fetched in their turn to process the next layers. Each PE—as depicted in Fig. 4c— is configurable and has its own computational capabilities by means of DSP blocs, and its own data caching capabilities by means of on-chip registers.

With this paradigm, the problem of CNN mapping boils down to finding the optimal architectural configuration of PEs (number of PEs, number of DSP blocs per PE, size of data caches), as well as the optimal temporal scheduling of data that maximizes the computational throughput $T$.

For convolution layers, in which the processing is described in listing 6a, finding the optimal PE configuration can be seen as a loop optimization problem [49, 9, 28, 77, 65, 40, 78, 36, 79, 80, 43]. This problem is addressed by applying loop optimization techniques such loop unrolling, loop tiling or loop interchange to the 7 nested loops of listing 6a. In this case, setting the unroll and tiling factors (resp. $P_i$ and $T_i$) determines the number of PEs, the computational resources and on-chip memory allocated to each PE in addition to the size of on-chip buffer and the amount of DRAM accesses.

4.2.1 Loop Unrolling:

Unrolling a loop $L_i$ with an unrolling factor $P_i$ ($P_i \leq i, i \in \{L, V, U, N, C, J, K\}$) accelerates its execution at the expense of resource utilization. Each of the parallelism patterns listed in section 2.4.2 can be implemented by unrolling one of the loops of listing 6a as summarized in table 4. For configuration given in figure 4c, the unrolling factor $P_N$ determines the number of PEs. On the other hand, unrolling factors $P_C, P_K, P_J$ determine the number of multipliers and adders, as well as the size of registers contained in each PE.

4.2.2 Loop Tiling:

In general, the capacity of on-chip memory in current FPGAs is not large enough to store all the weights and intermediate FMs of all CNN layers. As a consequence, FPGA based accelerators resort to external DRAMs to store this data. As mentioned in section 2.4.3, DRAM accesses are costly in terms of energy and latency, and data caches
must be implemented by means of on-chip buffers and local registers. The challenge is to configure the data-path in a way that every data transferred from DRAM is reused as much as possible.

For conv layers, this challenge can be addressed by tiling the nested loops of listing 6a. Loop tiling [81] divides the FMs and weights of each layer into multiple blocks that can fit into the on-chip buffers. For the configuration given in figure 4c, sizes of buffers containing input FM, weights and output FM are determined by the tiling factors detailed in table 4 according to equation 10

\[ M_{conv} = T_C T_H T_W + T_N T_C T_J T_K + T_N T_V T_U \]  

(a) // Ll: Layer 
for (int l=0; l<L, l++) { 
  // Lb: Batch 
  for (int b=0; b<B, b++) { 
    // Ln: Y Depth 
    for (int n=0; n<N, n++) { 
      // Lv: Y Columns 
      for (int v=0; v<V, v++) { 
        // Lu: Y Raws 
        for (int u=0; u<U, u++) { 
          // Lc: X Depth 
          for (int c=0; n<C, c++) { 
            // Lj: Theta Columns 
            for (int j=0; j<J, j++) { 
              // Lk: Theta Raws 
              for (int k=0; k<K, k++) { 
                Y [b, l, n, v, u] += x [b, l, c, v+j, u+k] * 
                                     theta [l, n, c, j, k] 
              } // Lk: Theta Raws 
          } // Lj: Theta Columns 
        } // Lu: Y Raws 
      } // Lv: Y Columns 
    } // Ln: Y Depth 
  } // Lb: Batch 
} // Ll: Layer 

(b) // DRAM: Load in on-chip buffers the tiles: 
// X[l,c:c+Tc,v:v+Tv,u:u+Tu] 
// Theta [l,n:n+Tn,c:c+Tc,j,k] 
// Process on-chip tiles 
for (int tn=0; tn<Tn, tn++) { 
  for (int tv=0; tv<Tv, tv++) { 
    for (int tu=0; tu<Tu, tu++) { 
      for (int tc=0; tc<Tc, tc++) { 
        for (int j=0; j<J, j++) { 
          for (int k=0; k<K, k++) { 
            Y[l,tn,tv,tu] += x[l,tc,tc+j,tc+k] * 
                               theta[l,tn,tc,j,k]; 
          } // Lk: Theta Raws 
        } // Lj: Theta Columns 
      } // Lv: Y Columns 
    } // Lu: Y Raws 
  } // Lv: Y Columns 
} // Lb: Batch 

Figure 5: Loop tiling and unrolling

Figure 6: Loop Tiling in conv layers: a-Before tiling. b-After tiling
4.2.3 Design Space Exploration:

In order to find the optimal unrolling and tiling factors, a large exploration of the design space is needed. In a general way, an analytical model is built. Inputs of this model are loop factors $P_i, T_i$ and outputs are a theoretical prediction of the allocated resources, the computational throughput and the memory bandwidth used. This model is parametrized by the available resources of a given FPGA platform and the workload of the CNN.

Given this model, the objective is to find the design parameters that minimize the memory access while maximizing the resource utilization. To address this optimization problem, a brute force exploration is performed, such in [39, 28, 77, 65, 40, 78]. This exploration is usually driven by the Roofline method [82] in order to select the feasible design solutions that matches with the maximum computational throughput and the maximum memory bandwidth a given platform can deliver [39, 40, 41]. The design space can also be explored by means of heuristic search algorithms, as proposed for instance in [35].

4.2.4 FPGA Implementations:

Employing loop optimizations to derive FPGA-based CNN accelerator was first investigated in [39]. In this work, Zhang et al. report a computational throughput of 61.62 GOPs in the execution of AlexNet convolutional layers by unrolling loops $L_C$ and $L_N$. This accelerator was built using HLS tools and rely on 32 floating point arithmetic. Works in [78] follow the same unrolling scheme and implement the FC part of the inference. Moreover, design [78] features 16 bits fixed point arithmetic and RTL conception, resulting in a x2.2 improvement in terms of computational throughput. Finally, the same unrolling and tiling scheme are employed in recent works [65], where the first convolution layers use $11 \times 11$ and $5 \times 5$ filters. Expanding loop unrolling and tiling to loops $L_J$ and $L_K$ results in a x1.36 improvement in computational throughput vs [39] on the same VX485T device when using 32 floating point arithmetic. In a same way, implementations in [28, 4, 56] tile and unroll loops $L_N, L_C, L_J, L_K$ and demonstrate higher acceleration on AlexNet and VGG when using fixed point arithmetic. Nevertheless, and as pointed out in [80], unrolling loops $L_J$ and $L_K$ is ineffective for recent CNN models that employ small convolution kernels. In addition, Tiling loops $L_J$ and $L_K$ requires PEs to be configured differently for different layers, increasing thus the control complexity.

The values of $U, V, N$ can be very large in CNN models. Consequently, unrolling and tiling loops $L_U, L_V, L_N$ can be efficient only for devices with high computational capabilities (i.e DSP Blocs). This is demonstrated in works of Rahman et al. [77] that report an improvement of ×1.22 over [39] when enlarging the design space exploration to loops $L_U, L_V, L_N$.

In order to keep data in on-chip buffer after the execution of a given layer, [79] investigates fused-layer CNN Accelerators by tiling across layer $L_L$. As a result, authors report a reduction of 95% of DRAM accesses at the cost of 362KB of extra on-chip memory.

In all these approaches, loops $L_N, L_C, L_J, L_K$ are unrolled in a same way they are tiled (i.e $T_i = P_i$). By contrast, the works of Ma et al. [80, 83] fully explore all the design variables searching for optimal loop unroll and tiling factors. More particularly, authors demonstrate that the input FMs and weights are optimally reused when unrolling only computations within a single input FM (i.e when $P_C = P_J = P_K = 1$). Tiling factors are set in way that all the data required to compute an element of $Y$ are fully buffered (i.e $T_C = C, T_K = K, T_J = J$). The remaining design parameters are derived after a brute force design exploration. The same authors leverage on these loop optimizations to build an RTL compiler for CNNs in [84]. To the best of our knowledge, this accelerator...
Table 5: FPGA-based CNN accelerators implementing loop optimization

<table>
<thead>
<tr>
<th>Network</th>
<th>Network Workload</th>
<th>Bitwidth</th>
<th>Desc.</th>
<th>Device</th>
<th>Freq (MHz)</th>
<th>Through (GOPS)</th>
<th>Power (W)</th>
<th>LUT (K)</th>
<th>DSP</th>
<th>Memory (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[39]</td>
<td>AlexNet-C</td>
<td>1.3</td>
<td>2.3</td>
<td>HLS</td>
<td>Virtex7 VX485T</td>
<td>100</td>
<td>61.62</td>
<td>18.61</td>
<td>186</td>
<td>2240</td>
</tr>
<tr>
<td>[28]</td>
<td>AlexNet-C</td>
<td>1.3</td>
<td>2.3</td>
<td>HDL</td>
<td>Stratix5 GSD8</td>
<td>120</td>
<td>71.64</td>
<td>33.93</td>
<td>272</td>
<td>752</td>
</tr>
<tr>
<td></td>
<td>VGG16-F</td>
<td>31.1</td>
<td>138.0</td>
<td>OpenCL</td>
<td>Stratix5 GXA7</td>
<td>100</td>
<td>117.9</td>
<td>524</td>
<td>1963</td>
<td>51.4</td>
</tr>
<tr>
<td>[77]</td>
<td>AlexNet-C</td>
<td>1.3</td>
<td>2.3</td>
<td>HLS</td>
<td>Virtex7 VX485T</td>
<td>100</td>
<td>75.16</td>
<td>28</td>
<td>2695</td>
<td>19.5</td>
</tr>
<tr>
<td>[36]</td>
<td>AlexNet-F</td>
<td>1.4</td>
<td>61.0</td>
<td>HLS</td>
<td>Virtex7 VX690T</td>
<td>150</td>
<td>825.6</td>
<td>126.00</td>
<td>14400</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>VGG16-F</td>
<td>31.1</td>
<td>138.0</td>
<td>Stratix5 GXA7</td>
<td>Stratix5 GXA7</td>
<td>100</td>
<td>134.1</td>
<td>19.10</td>
<td>242</td>
<td>256</td>
</tr>
<tr>
<td>[65]</td>
<td>AlexNet-F</td>
<td>1.4</td>
<td>61.0</td>
<td>Fixed</td>
<td>Arria10 GT1150</td>
<td>200</td>
<td>587.63</td>
<td>453</td>
<td>256</td>
<td>46.6</td>
</tr>
<tr>
<td>[79]</td>
<td>AlexNet-C</td>
<td>1.3</td>
<td>2.3</td>
<td>HLS</td>
<td>Virtex7 VX690T</td>
<td>100</td>
<td>61.62</td>
<td>273</td>
<td>2401</td>
<td>20.2</td>
</tr>
<tr>
<td>[80]</td>
<td>VGG16-F</td>
<td>31.1</td>
<td>138.0</td>
<td>Fixed</td>
<td>Arria10 GX1150</td>
<td>150</td>
<td>645.25</td>
<td>322</td>
<td>1518</td>
<td>38.0</td>
</tr>
<tr>
<td>[42]</td>
<td>AlexNet-C</td>
<td>1.3</td>
<td>2.3</td>
<td>HDL</td>
<td>Cyclone5 SEM</td>
<td>100</td>
<td>12.11</td>
<td>22</td>
<td>28</td>
<td>0.2</td>
</tr>
<tr>
<td>[84]</td>
<td>ResNet-50</td>
<td>7.8</td>
<td>25.5</td>
<td>Fixed</td>
<td>Stratix5 GXA7</td>
<td>150</td>
<td>352.24</td>
<td>424</td>
<td>256</td>
<td>44.0</td>
</tr>
<tr>
<td>[85]</td>
<td>AlexNet-F</td>
<td>1.5</td>
<td>7.6</td>
<td>Fixed</td>
<td>Virtex7 VX690T</td>
<td>100</td>
<td>445.6</td>
<td>493</td>
<td>322</td>
<td>44.0</td>
</tr>
<tr>
<td>[85]</td>
<td>VGG16SVD-F</td>
<td>30.8</td>
<td>50.2</td>
<td>HLS</td>
<td>Virtex7 VX690T</td>
<td>100</td>
<td>473.4</td>
<td>25.60</td>
<td>224</td>
<td>2950</td>
</tr>
</tbody>
</table>

outperforms all the previous implementations that are based on loop optimization in terms of computational throughput.

4.3 Dataflow MoC For CNNs

Feed-forward propagation is by nature a streaming based applications in which the execution is purely data-driven. In fact, the CNN layout is in contrast with Von Neumann execution models and a CNN implementation can easily be memory-bounded if it has to fetch every instruction from memory. This motivated multiple approaches to investigate the applicability of the data-flow Model of Computation (MoC) to accelerate CNNs on FPGAs.

The foundations of the data-flow MoC were formalized by [86] in order to create an architecture where multiple fragments of instructions can process simultaneously streams of data. Programs respecting dataflow semantics are described as Data-Flow Process Networks (DPNs). Each node of this network corresponds to a fundamental processing unit called an actor and each edge corresponds to a communication FIFO channel. Actors exchange abstract data –known as tokens– through these FIFOs. Each actor follows a purely data-driven execution model wherein the firing (execution) is triggered only by the availability of input operands. This is typically the case in CNNs, where the execution of each layer is only triggered by the availability of input FM.

Applying the data-flow MoC to accelerate CNN implementations on FPGAs is investigated in [87]. In this work, authors demonstrate the efficiency of the proposed lightweight data-flow methodology [88] by mapping
Table 6: FPGA-Based CNN accelerators employing the data-flow MoC

<table>
<thead>
<tr>
<th>Network</th>
<th>Network Workload</th>
<th>Bitwidth</th>
<th>Desc.</th>
<th>Device</th>
<th>Freq (MHz)</th>
<th>Through (GOPs)</th>
<th>Power (W)</th>
<th>LUT (K)</th>
<th>DSP</th>
<th>Memory (KB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[91] CarType-C</td>
<td>0.16</td>
<td>0.03</td>
<td>Float 32</td>
<td>HDL</td>
<td>Zynq Z7045</td>
<td>100</td>
<td>0.47</td>
<td>0.23</td>
<td>68</td>
<td>24</td>
</tr>
<tr>
<td>[34] LeNet5-C</td>
<td>0.04</td>
<td>0.03</td>
<td>Fixed 16</td>
<td>HLS</td>
<td>Zynq Z7020</td>
<td>100</td>
<td>0.48</td>
<td>0.75</td>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td>[90] SignRecog-C</td>
<td>4.03</td>
<td>0.04</td>
<td></td>
<td>Zynq Z7045</td>
<td>125</td>
<td>123.12</td>
<td>26</td>
<td>144</td>
<td>38.2</td>
<td></td>
</tr>
<tr>
<td>[90] VGG16-F</td>
<td>31.10</td>
<td>138.00</td>
<td>Fixed 16</td>
<td>HLS</td>
<td>Zynq Z7045</td>
<td>125</td>
<td>170.73</td>
<td>40</td>
<td>0</td>
<td>10.9</td>
</tr>
<tr>
<td>[38] SVHN-C</td>
<td>0.02</td>
<td>0.08</td>
<td>Fixed 5</td>
<td>Cyclone5 GX</td>
<td>63.96</td>
<td>2438.46</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
</tr>
<tr>
<td>[38] LeNet5-C</td>
<td>0.04</td>
<td>0.03</td>
<td>Fixed 3</td>
<td>Cyclone5 GX</td>
<td>63.96</td>
<td>2438.46</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
</tr>
</tbody>
</table>

A special case of data-flow, referred as Static Data-Flow (SDF) [89], is a paradigm in which the number of tokens produced and consumed by each actor can be specified a priori, as it is the case in the CNN execution. SDF model is employed in [34] [90] to optimize the mapping of CNN graphs on FPGAs. In this works, the CNN graph is modeled as a topology matrix that contains the the number of incoming streams, the size of tokens and the consumption rates of each actor. Instead of exploring the design space of unrolling and tiling parameters (cf. sec 4.2), authors explore the design space of the topology matrix components. These optimal components are used to derive the configuration of the PE and buffers that either minimizes the computation latency or energy consumption. Moreover, and in contrast with classical implementations where data is streamed in and out of layers using off-chip data transfers, authors exploit partial dynamic reconfiguration of FPGAs to process different layers.

Finally, works in [38] optimize the direct hardware mapping of CNN graphs. In this approach, each actor of the DPN is physically mapped on the device with its own specific instance, while each edge is mapped as a signal. As all the computations are unrolled, applicability of this method can rapidly be limited by the resource of the device or the size of the CNN, preventing this approach from implementing deep models.

**Figure 7**: An example of a graph representation of a convolution layer ($C = 3, N = 5$)

## 5 Approximate Computing of CNN Models

Beside the computational transforms and data-path optimizations, the CNN execution can be accelerated when employing approximate computing which is known to perform efficiently on FPGAs [92].
In this approach, a minimal amount of the CNN accuracy is traded to improve the computational throughput and energy efficiency of the execution. Two main strategies are employed. This first implements approximate arithmetic to process the CNN layers with a reduced precision while the second aims to reduce the number of operations occurring in CNN models without critically affecting the modeling performance. Both of these methods can be integrated in the learning phase to jointly maximize the accuracy and minimize the workload of a given CNN model.

5.1 Approximate Arithmetic for CNNs

Several studies have demonstrated that the precision of both operations and operands in CNNs\(^4\) can be reduced without critically affecting their predictive performance. This reduction can be achieved by quantizing either or both of the CNN inputs, weights and/or FMs using a fixed point numerical representation and implementing approximate multipliers and adders.

5.1.1 Fixed point arithmetic:

In a general way, CNN models are deployed in CPUs and GPUs using the same numerical precision they were trained with, relying on simple-precision floating point representation. This format employs 32 bits, arranged according to the IEEE754 standard. In FPGAs, implementations such \([39, 79, 77]\) employ this data representation.

Nonetheless, several studies in \([93, 46, 94]\) demonstrate that inference of CNNs can be achieved with a reduced precision of operands. In addition, works in \([48, 95, 96, 97]\) demonstrate the applicability of fixed-point arithmetic to train CNNs. In both cases, FMs and/or weights are quantized using a fixed point representation scheme. In simplest version of this format, numbers are encoded with the same bit-width \((bw)\) that is set according to the numerical range and the desired precision. More particularly, all the operands share the same exponent (i.e scale factor) that can be seen as as the position of the radix point. In this paper, we refer to this representation as Static Fixed Point (SFP).

When compared to floating point, SFP computing with compact bit-width is known to be more efficient in terms of hardware utilization and power consumption. This is especially true in FPGAs \([98]\), where a single DSP block can either implement one 32bits floating point multiplication, two 18×19 bits multiplications, or three 18×19 multiplications \([10]\).

This motivated early implementations to employ SFP in building FPGA-Based CNN accelerators, such in \([72, 73, 74]\), or in \([75, 76]\), where authors use a 16 bits (Q8.8) format to represent FMs and weights. To prevent overflow, the bit-width is expanded when computing the weighted-sums of convolutions and inner-products. If \(b_x\) bits are used to quantize the FM and \(b_\Theta\) bits are used to quantize the weights, an accumulator of size \(b_{acc}\) is used, according to equation \([11]\) which corresponds to accumulators of 48 bits in \([73, 74]\).

\[
b_{acc} = b_x + b_\Theta + \max_{l \leq L} \left( \log_2 \left( C_l K^2_l \right) \right)
\]  

5.1.2 Dynamic Fixed Point for CNNs:

In deep topologies, it can be observed that distinct parts of a network can have a significantly different numerical range of data. More particularly, the FMs of deep layers tend to have larger numerical range than first FMs, while the weights are generally much smaller than the FMs. As a consequence, the bit-width is expanded to keep the same precision while preventing overflow, as in \([74]\). As a result, and as pointed-out \([48]\), SFP with its unique shared fixed exponent, is ill-suited to deep learning.

\(^4\) and more generally in neural networks
To address this problem, works in \cite{48, 49} advocate the use of Dynamic Fixed Point (DFP) \cite{99}. In DFP, different scaling factors are used to process different parts of the network. More particularly, weights, weighted sums and outputs of each layer are assigned distinct scale factors. The optimal scale factors and bit-widths (i.e. the ones that deliver the best trade-off between accuracy loss and computational load) for each layer can be derived after a brute force exploration using dedicated frameworks that supports DFP such as \cite{49, 100} for Caffe and \cite{96} for TensorFlow. In addition, these tools can fine-tune the CNN model to improve the accuracy of the quantized network.

The FPGA-Based CNN Accelerator proposed in \cite{28} is build upon this quantification scheme and employs different bit-widths to represent the FM, the convolution kernels and the FC weights with resp. 16, 8, 10 bits. Without fine-tuning, authors report a drop of 1\% in classification accuracy of AlexNet. For the same network, works of \cite{78} employs 10 bits for FMs, 8 bits for both conv and FC weights and report an accuracy drop of 0.4\%.

In a same way, Qiu et al. employ DFP to quantize the VGG with 8, 8, and 4 bits while reporting 2\% of accuracy drop. In these accelerators, dynamic quantization is supported by means of data shift modules \cite{9}. Finally, the accelerator in \cite{42} rely on the Ristretto framework \cite{49} to derive an AlexNet model wherein the data is quantized in 16 bits with distinct scale factors per layer.

5.1.3 Extreme quantification with Binary and pseudo-Binary Nets:

Beside fixed point quantification, training and inferring CNNs with extremely compact data representations, is a research area that is gaining interest. In particular, works in BinaryConnect \cite{50} investigate the applicability of binary weights (i.e weights with either a value of \(-\theta\) or \(\theta\)) to train CNNs, which lowers both bandwidth requirements and accuracy on ImageNet by respectively 3200\% and 19.2\% (vs AlexNet Float32 Model). The same authors go further by implementing BNNs \cite{17}, with a 1bit representation for both FM and weights. In these networks, negative data is represented as 0 while positive values are represented as 1. As a consequence, the computation of MACs boils down to an XNOR operation followed by a pop-count, as shown in figure \ref{fig:8b}. Moreover, Batch normalization is performed before applying of the sign activation function in order to reduce the information lost during binarization, as shown in figure \ref{fig:8a}. However, a classification accuracy drop of 29.8\% is observed on ImageNet when using BNNs. In an attempt to lower the accuracy drop of BNNs, Rastegari et al. proposed XNOR-Nets \cite{51} which use different scale factors for binary weights (i.e \(-\theta_1\) or \(+\theta_2\)). Moreover, Pseudo-Binary Networks, such DoReFa-Net \cite{101} and QNNs \cite{102} reduce the accuracy drop to 6.5\% by employing a slightly expanded bit-width (2 bits) to represent the intermediate FMs. Finally, in Trained Ternary Quantization (TTQ) \cite{103}, weights are constrained to three values \(-\theta_1, 0, -\theta_2\) (2 bits), but FM are represented in a 32bits float scheme. As a consequence, the efficiency gain of TTQ is not as high as in BNNs. But in turn, TTQ achieves comparable accuracy on ImageNet, within 0.7\% of full-precision.

In FPGAs, BNNs benefit from a significant acceleration as the processing of “binary” convolutions can be mapped on XNOR gates followed by a pop count operation, as depicted in figure \ref{fig:8b}. Furthermore, and as suggested in \cite{7}, pop count operation can be implemented using lookup tables in a way that convolutions are processed only with logical elements. The DSPs blocs are can thus be used to process the batch norm calculation (eq \ref{eq:5}) which can be formulated as a linear transform reduces in order reduce the number of operations. This approach is followed in the implementation of \cite{104} to derive an FPGA-Based accelerator for BNNs that achieves 207.8 GOP/s while only consuming 4.7 W and 3 DSP Blocs to classify the Cifar10 dataset. For the same task, works in \cite{52, 105} use a smaller network configuration and reaches a throughput of 2.4 TOP/s when using a larger Zyq 7Z045 Device with 11W Power consumption. For ImageNet classification, Binary Net implementation of \cite{106} delivers an overall throughput 1.9 TOP/s on a Stratix V GSD device. In all these works, the first layer is not binerized.

\footnote{An other approach to address this problem is to use half-precision 16 bits floating point, as used in \cite{30}}

\footnote{Since the same PEs are reused to process different layers, the same bit-width is used with a variable radix point for each layer}

\footnote{The network topology used in this work involves 90\% less computations and achieves 7\% less classification accuracy on Cifar10}
Figure 8: Binary Neural Networks: a-Processing Pipeline, b-Binary Convolutions

to achieve better classification accuracy. As pointed-out in [106], the performance in this layer can be improved when using a higher amount of DSP blocs. Finally, an accelerator for ternary neural networks is proposed in [107] and achieves a peak performance of 8.36 TMAC/s at 13W power consumption for Cifar10 Classification.

5.1.4 Stochastic Computing:

Stochastic Computing (SC) is a low-cost design technique that has been successfully applied in numerous image processing algorithms [108].

In SC, numbers are represented as a random sequence of $s$ bits. In the basic "unipolar" format, the number of ones appearing in the sequence $s$ determines the value of $x$, i.e. the numerical value of a given number $x$ is $s_1/s$, where $x$ is the number of ones appearing in $s$. The advantage of stochastic arithmetic is that operations are performed with an ultra-small circuitry. For instance, a single AND gate can map a multiplication. Works in [60, 59, 58] demonstrate the feasibility of stochastic arithmetic to accelerate CNNs. More particularly, Ardakani et al. propose an FPGA accelerator to classify the MNIST dataset, where multiplications are processed only using AND gates and activation functions (TanH) are implemented in the stochastic domain using FSMs. Such an implementation delivers a computational throughput of 15.44 TOP/s with a misclassification rate of 2.40% on MNIST. However, one the of weakness of SC are long bit-streams. In fact, to represent an $n$ bits number, a bit-stream $s$ of $2^n$ is required. As a result, stochastic arithmetic suffers from long run-times to perform operations. Moreover, the generation of this bit-streams resorts to dedicated circuitry known as Stochastic Number Generators (SNGs), which add more overhead to the implementation. As a result, SC-based accelerators implement only shallow neural networks with a limited depth.

5.2 Reduce Computations in CNNs

In addition to approximate arithmetic, several studies attempt to the reduce the number of operations involved in CNNs. For FPGA-Based implementation, two main strategies are investigated: weight pruning, which increases the sparsity of the model, and low-rank approximation of filters, which reduces the number of multiplications occurring in the inference.

5.2.1 Weight Pruning:

As highlighted in [109], CNNs as over-parametrized networks and a large amount of the weights can be removed—or pruned— without critically affecting the classification accuracy. In its simplest form, pruning is performed according to the magnitude such as the lowest values of the weights are truncated to zero [110]. In a more recent approach, weights removal is driven by energy consumption of a given node of the graph, which is 1.74x more
Table 7: FPGA-Based CNN accelerators employing Approximate arithmetic

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Network Workload</th>
<th>Bitwidth</th>
<th>Acc</th>
<th>Device</th>
<th>Freq (MHz)</th>
<th>Through. (GOPS)</th>
<th>Power (W)</th>
<th>LUT (K)</th>
<th>DSP Memory (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FP32</td>
<td>[61]</td>
<td></td>
<td></td>
<td>ImageNet</td>
<td>30.8</td>
<td>138.0</td>
<td>866</td>
<td>41.7</td>
<td>437</td>
</tr>
<tr>
<td></td>
<td></td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>90.1</td>
<td>Arria10 GX1150</td>
<td>370</td>
<td>1576</td>
</tr>
<tr>
<td>FP16</td>
<td>[60]</td>
<td></td>
<td></td>
<td>ImageNet</td>
<td>1.4</td>
<td>61.0</td>
<td>1382</td>
<td>44.3</td>
<td>443</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>Arria10 GX1150</td>
<td>303</td>
<td>246</td>
</tr>
<tr>
<td>DFP</td>
<td>[62]</td>
<td></td>
<td></td>
<td>ImageNet</td>
<td>30.8</td>
<td>618</td>
<td>1790</td>
<td>437</td>
<td>437</td>
</tr>
<tr>
<td></td>
<td></td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>90.1</td>
<td>Arria10 GX1150</td>
<td>370</td>
<td>1576</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>88.1</td>
<td>Arria10 GX1150</td>
<td>150</td>
<td>32</td>
</tr>
<tr>
<td>BNN</td>
<td>MNIST</td>
<td>0.0</td>
<td>9.6</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>98.2</td>
<td>5905</td>
</tr>
<tr>
<td></td>
<td>[106]</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Zynq Z7020</td>
<td>133</td>
<td>437</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Zynq Z7045</td>
<td>200</td>
<td>246</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Zynq Z7045</td>
<td>200</td>
<td>246</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Zynq Z7045</td>
<td>200</td>
<td>246</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Zynq Z7045</td>
<td>200</td>
<td>246</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Zynq Z7045</td>
<td>200</td>
<td>246</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Zynq Z7045</td>
<td>200</td>
<td>246</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Zynq Z7045</td>
<td>200</td>
<td>246</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Zynq Z7045</td>
<td>200</td>
<td>246</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Zynq Z7045</td>
<td>200</td>
<td>246</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Zynq Z7045</td>
<td>200</td>
<td>246</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Zynq Z7045</td>
<td>200</td>
<td>246</td>
</tr>
</tbody>
</table>

Efficient than magnitude-based approaches [111]. In both approaches, pruning is followed by a fine-tuning of the remaining weights in order to improve the classification accuracy. This is for instance the case in [112], where pruning removes respectively 53% and 85% of the weights in AlexNet $conv$ and FC layers for less then 0.5% accuracy loss.

5.2.2 Low Rank Approximation:

Another way to reduce the computations occurring in CNNs is to maximize the number of of separable filters in CNN models. A 2D-separable filter $\theta^{sep}$ has a unitary rank (i.e rank $(\theta^{sep}) = 1$), and can be expressed as two successive 1D filters $\theta_{J \times 1}$ and $\theta_{1 \times K}$. When expanding this to 3D filters, a separable 3D convolution requires $C + J + K$ multiplications while a standard 3D convolution requires $C \times J \times K$ multiplications.

Nonetheless, only a small proportion of the filters in CNN Models are separable. To increase this proportion, a first approach is to force the convolution kernels to be separable by penalizing high rank filters when training the network [113]. Alternatively, and after the training, the weights $\Theta$ of a given layer can be approximated into a small set of $r$ low rank filters that can be implemented as a succession of fully separable filters. In this case, $r \times (C + J + K)$ multiplications are required to process a single 3D-convolution.

For FC layers, in which the processing boils down to a vector-matrix product, low rank approximation can be achieved by employing, for instance, the SVD decomposition of the weight matrix $\tilde{\Theta}^{fc}$ (cf. sec 3.1). Finally, and in a same way to pruning, low rank approximation of weights is followed by a fine-tuning in order counterbalance the classification accuracy drop.

5.2.3 FPGA Implementations:

In FPGA Implementations, low rank approximation is applied on FC layer to significantly reduce the number of weight, such as in [9], where authors derive a VGG16-SVD model that achieves 87.96% accuracy on ImageNet with 63% less parameters.

Sparsity in pruned CNNs can be exploited in FPGA implementations by fully unrolling the processing of a given layer, and skipping (i.e not mapping) the multiplications with zero weights. This approach is investigated in [38], but can be infeasible when the resource of a given device doesn’t match with computational requirements of a given layer. Instead, sparsity and pruning can be exploited when processing $conv$ and $fc$ layers as GEMM.
In this case, the challenge is to determine the optimal format of matrices that maximizes the chance to detect and skip zero computations, such compressed sparse column (CRC) or compressed sparse row (CSR) format. Based on previous studied related to sparse GEMM implementation on FPGAs in [114], Sze et al. [12] advocate the use of the CRC to process CNNs because this format provides a lower memory bandwidth when the output matrix is smaller than the input, which is typically the case in CNNs where $N < CJK$ in Fig. 3b.

However, this efficiency of CRC format is only valid for extremely sparse matrices (typically with ≤ 1% of non-zeros), while pruned CNN matrices are not that sparse (typically, ≤ 4 – 80% of non-zeros). Therefore, works in [7] use a zero skip scheduler, which is an on-chip data manager thanks to which zero elements are identified and not scheduled onto the MAC processing. As a result, the number of cycles required to compute the sparse GEMM is reduced, which corresponds to a 4x speedup in cycle count for and 85% sparse AlexNet layers. Finally, authors report to a projected throughput of 12 TOP/s for pruned CNNs in the next Intel Stratix10 FPGAs, which outperforms the computational throughput of state-of-the-art GPU implementations by 10%.

### 6 Conclusion

In this paper, a number of methods and tools have been compared that aim at porting Convolutional Neural Networks onto FPGAs. At the network level, approximate computing and datapath optimization methods have been covered while at the neuron level, the optimizations of convolutional and fully connected layers have been detailed and compared. All the different degrees of freedom offered by FPGAs (custom data types, local data streams, dedicated processors, etc.) are exploited by the presented methods. Moreover, algorithmic and datapath optimizations can be jointly implemented, resulting in additive hardware performance gains.

CNNs are by nature overparameterized and support particularly well approximate computing techniques such as weight pruning and fixed point computation. Approximate computing already constitutes a key to CNN acceleration over hardware and will certainly continue driving the performance gains in the years to come.

---

8. These format represents a matrix by three one-dimensional arrays, that respectively contain nonzero values, row indices and column indices.
Bibliography


[64] Intel FPGA. The Intel® FPGA SDK for Open Computing Language (OpenCL), 2016.


[95] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Quantized Neural Networks: Training Neural Networks with Low Precision Weights and Activations. arxiv e-print, 9 2016.


