[DL] A Survey of FPGA Based Neural Network Accelerator

KAIYUAN GUO, SHULIN ZENG, JINCHENG YU, YU WANG AND HUAZHONG YANG, Tsinghua University, China

Recent researches on neural network have shown great advantage in computer vision over traditional algorithms based on handcrafted features and models. Neural network is now widely adopted in regions like image, speech and video recognition. But the great computation and storage complexity of neural network based algorithms poses great difficulty on its application. CPU platforms are hard to offer enough computation capacity. GPU platforms are the first choice for neural network process because of its high computation capacity and easy to use development frameworks.

On the other hand, FPGA based neural network accelerator is becoming a research topic. With specific designed hardware, FPGA is the next possible solution to surpass GPU in speed and energy efficiency. Various FPGA based accelerator designs have been proposed with software and hardware optimization techniques to achieve high speed and energy efficiency. In this paper, we give an overview of previous work on neural network accelerators based on FPGA and summarize the main techniques used. Investigation from software to hardware, from circuit level to system level is carried out to complete analysis of FPGA based neural network accelerator design and serves as a guide to future work.

CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Computer systems organization** → *Parallel architectures*;

Additional Key Words and Phrases: FPGA architecture, Neural Network, Parallel Processing

ACM Reference Format:
Kaiyuan Guo, Shulin Zeng, Jincheng Yu, Yu Wang AND Huazhong Yang. 2017. [DL] A Survey of FPGA Based Neural Network Accelerator. *ACM Trans. Reconfig. Technol. Syst.* 9, 4, Article 11 (December 2017), 21 pages.
https://doi.org/0000001.0000001

1 INTRODUCTION

Recent research on Neural Network (NN) is showing great improvement over traditional algorithms in computer vision. Various network models, like convolutional neural network (CNN), recurrent neural network (RNN), have been proposed for image, video, and speech process. CNN [24] improves the top-5 image classification accuracy on ImageNet [43] dataset from 73.8% to 84.7% and further helps improve object detection [9] with its outstanding ability in feature extraction. RNN [17] achieves state-of-the-art word error rate on speech recognition. In general, NN features a high fitting ability to a wide range of pattern recognition problems. This makes NN a promising candidate to many artificial intelligence applications.

But the computation and storage complexity of NN models are high. Current researches on NN are still increasing the size of NN models. Take CNN as an example. The largest CNN model for an 224×224 image classification requires upto 39 billion floating point operations (FLOP) and more than 500MB model parameters [48]. As the computation complexity is propotional to the input image size, processing images with higher resolutions may need more than 100 billion operations.

Author's address: Kaiyuan Guo, Shulin Zeng, Jincheng Yu, Yu Wang AND Huazhong Yang, Tsinghua University, Tsinghua University, Beijing, Beijing, 100084, China, gky15@mails.tsinghua.edu.cn,yu-wang@mail.tsinghua.edu.cn.

1936-7406/2017/12-ART11 $15.00
https://doi.org/0000001.0000001

Therefore, choosing a proper computation platform for neural network based applications is important. A common CPU can perform 10-100G FLOP per second, and the power efficieny is usually below 1GOP/J. So CPUs are hard to meet the high performance requirements in cloud applications nor the low power requiremetns in mobile applications. In contrast, GPUs offer upto 10TOP/s peak performance and are good choices for high performance neural network applications. Development frameworks like Caffe [22] and Tensorflow [3] also offer easy-to-use interfaces which makes GPU the first choice of neural network acceleration.

Besides CPUs and GPUs, FPGAs are becoming a platform candidate to achieve energy efficient neural network processing. With a neural network oriented hardware design, FPGAs are able to implement high parallelism and make use of the properties of neural network computation to remove unecessary logic. Algorithm researches also show that a NN model is able to be simplified in a hardware friendly way while not hurting the accuracy of the model. Therefore FPGAs are possible to achieve higher energy efficieny compared with CPU and GPU.

FPGA based accelerator designs are faced with two challenges in performance and flexibility:

- Current FPGAs usually support working frequency at 100-300MHz, which is much less than CPU and GPU. The FPGA's logic overhead for reconfigurability also reduces the overall system performance. Straight forward design on FPGA is hard to achieve high performance and high energy efficiency.
- Implementation of neural networks on FPGAs is much harder than that on CPUs or GPUs. Development framework like Caffe and Tensorflow for CPU and GPU is needed for FPGA.

Many researches on the above two problems have been carried out to implement energy efficient and flexible FPGA based neural network accelerators. In this paper, we summarize the techniques proposed in these work from the following aspects:

- We first give a simple model on FPGA based neural network accelerator performance to analyze the methodology in energy efficienct design.
- We investigate current techniques for high performance and energy efficient neural network accelerator designs. We introduce the techniques in both software and hardware level and estimate the effect of these techniques.
- We compare state-of-the-art neural network accelertor designs to evaluate the techniques introduced and estimate the achievable performance of FPGA based accelerator design, which is at least 40× better energy efficienct than current GPUs.
- We investigate state-of-the-art automatic design methods of FPGA based neural network accelerators.

The rest part of this paper is organized as follows: Section 2 introduces the basic operations of neural networks. Section 4 and section 5 review the techniques on neural network accelerator in software and hardware level respectively. Section 6 compares existing designs and evaluate the techniques. Section 7 introduce the methods for a flexible accelerator design. Section 8 concludes this paper.

2 PRELIMINARY

Before discussing the system design for neural network acceleration, we first introduce the basic concepts of neural networks and the basic components of an FPGA-based accelerator design.

2.1 Neural Network

In this section, we introduce the basic functions in a neural network. In this paper, we only focus on the inference of NN, which means using a trained model to predict or classify new data. The training process of NN is not discussed in this paper. A neural network model can be described as

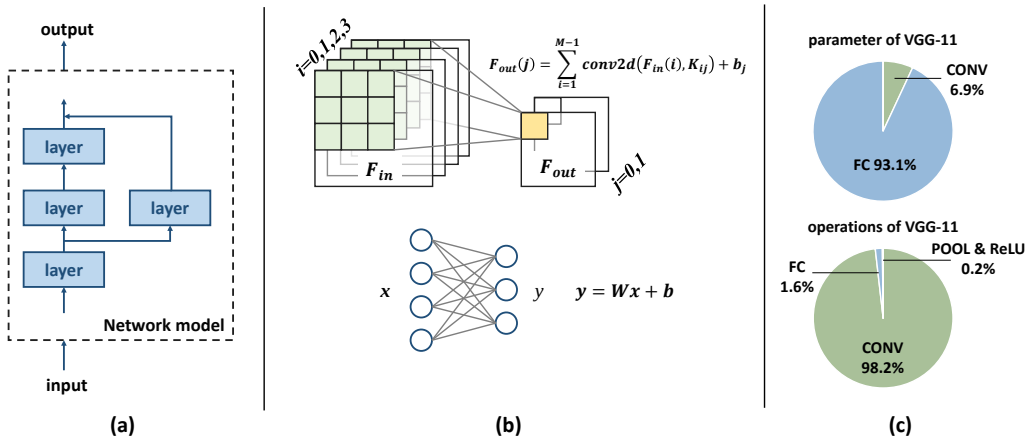

Fig. 1. (a) Computation graph of a neural network model. (b) CONV and FC layers in NN model. (c) CONV and FC layers dominates the computation and parameter of a NN model.

a directed graph shown in Figure 1(a). Each vertex of the graph denotes a layer which conducts operations on data from a previous layer or input and generates results to the next layer or output. We refer the parameter of each layer as weights and the input/output as activations through this paper.

Convolution (CONV) layers and fully connected (FC) layers are two common types of layers in NN models. The functions of these two layers are shown in Figure 1(b). CONV layers conduct 2D convolutions on a set of input feature maps F_{in} and add the results to get output feature maps F_{out}. FC layers receives a feature vector as input and conduct matrix-vector multiplications.

Besides CONV and FC layers, NN layers also have pooling, ReLU [24], concat [50], element-wise [18] and other types of layers. But these layers contributes little to the computation and storage requirement of a neural network model. Figure 1(c) shows the distribution of weights and operations in the VGG-11 model [48]. CONV and FC layers together contributes more than 99% of the network's weights and operations. So most of the neural network acceleration systems focus on these two types of layers.

2.2 FPGA-based Accelerator

In recent years, FPGA is becoming a promising solution for accelerating certain algorithms. Compared with CPU, GPU, and DSP platforms, for which the software and hardware are designed independently, FPGA enables the developers to implement only the necessary logic in hardware according to the target algorithm. By eliminating the redundancy in general hardware platforms, FPGAs are able to achieve higher efficiency. Application specific integrated circuits (ASICs) based solutions achieves even higher efficiency, but requires much longer development cycle and higher cost.

For FPGA-based neural network accelerator, a typical architecture of the system is shown in Figure 2(a). The system usually consists of a CPU host and an FPGA part. A pure FPGA chip usually works with a host PC/server through PCIe connections. SoC platforms (like the Xilinx Zynq Series) and Intel HARPv2 [14] platform integrates the host and the FPGA in the same chip or package. Both the host and the FPGA can work with their own external memory and access each others'

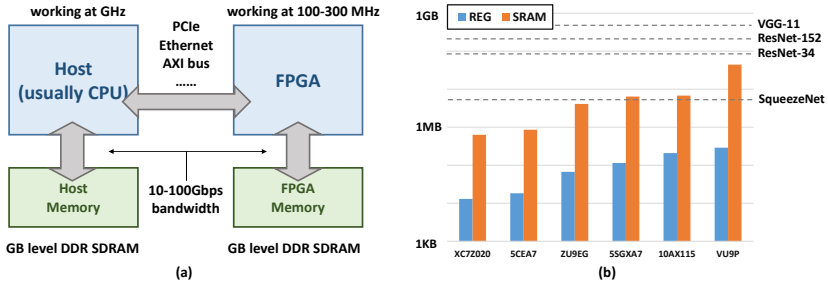

Fig. 2. (a) A typical structure of an FPGA based NN accelerator. (b) Comparison between NN model size and the storage on different FPGA chips.

memory through the connection. Most of the designs implements NN accelerator on the FPGA part and controls the acceleartor with the software on the host.

Typical FPGA chips implement large on-chip storage units like registers and SRAMs, but still too small compared with NN models as shown in Figure 2(b). Common models implement 100-1000MB parameters while the largest available FPGA chip implements <50MB on-chip SRAM. This requires that external memory like DDR SDRAM is needed. The bandwidth and power consumption of DDR limits the system performance.

The computation capacity of FPGA is relatively higher. Common FPGAs implements hundreds to thousands of DSP units, each of which can compute 18×27 or 18×19, achieving upto 10TFLOP/s (floating point operations per second) on the largest FPGAs. But for low-end FPGAs like Xilinx XC7Z020, this number is reduced to 20GFLOP/s. which is hard to support real-time video processing for applications on mobile platforms.

Even faced with the above challenges, researchers have proposed a series of optimization methods from algorithm to architecture to design high performance NN accelerators on FPGA, which will be discussed in the following sections of this paper.

3 DESIGN METHODOLOGY

Before going into the details of the techniques used for neural network accelerators, we first give an overview of the design methodology. In general, the design target of a neural network processing system includes the following aspects: high model accuracy, high throughput, lower latency, and high energy efficiency.

A larger neural network model usually results in a higher model accuracy. This means it is possible to tradeoff between the model accuracy and the hardware performance. Neural network researchers are designing more effcient network models from AlexNet [24] to ResNet [18], SqueezeNet [21] and MobileNet [20]. The main differences between these networks are the size of and the connections between each layer. The basic operations are the same and hardly affect the hardware design. Other methods try to achieve the tradeoff by optimizing existing NN models. Most of these methods are hardware oriented and will be discussed in detail in section 4.

The throughput of a neural network processing system can be expressed by equation 1. With a certain FPGA chip, the on-chip resource is limited. Increasing the peak performance means to reduce the size of each computation unit and increase the working frequency. Reducing the size of computation units can be achieved by simplifying the basic operations in neural network model, which may hurt the model accuracy and requires hardware-software co-design. Increasing working frequency, on the other hand is pure hardware design work. A high utilization ratio is kept by

reasonable parallelism implementation and efficient memory system. Most of this part is affected by hardware design. But model compression can also reduce the storage requirment of a neural network model and benefits the memory system.

$$throughput = \frac{peak\_performance \times utilization}{workload} \tag{1}$$

Energy efficiency is evaluated by the number of operations (multiplication or addition in this case) executed with unit energy cost. Given a certain network model, the energy efficiency of a neural network processing system is inversely proportional to the energy cost, which is expressed in equation 2. The energy cost comes from 2 parts: computation and memory access.

$$E_{total} = N_{effect\_op} \times E_{unit\_op} + N_{mem\_access} \times E_{unit\_mem\_access} \tag{2}$$

The first item in equation 2 is the energy cost for computation. This part is greatly affected by model compression. Model compression methods can reduce the actual number of operations (N_{effect_op}) to be executed on hardware and simplify the operations to reduce the unit energy cost of a single operation (E_{unit_op}). Given an FPGA chip, E_{unit_op} is also affected by its hardware implementation. The second item in equation 2 is the energy cost for memory access. The number of memory access N_{mem_access} is affected by the memory system and scheduling method. The energy for each memory access can be reduced by model compression methods by using a narrower data bit-width.

From the analysis of throughput and energy, we see that neural network accelerator involves software-hardware co-design. In the following sections, we will introduce previous work in software and hardware level respectively.

4 SOFTWARE DESIGN: MODEL COMPRESSION

As introduced in section 3, the design of energy efficient and high performacne neural network accelerator involves software and hardware co-design. In this section, we investigate the software level network model compression methods. Many researches on this topic have been proposed to reduce the number of weights or reduce the number of bits used for each neuron or weight, which helps reudce the computation and storage complexity. But these methods can also sacrifice the model accuracy. The trade-off between model compression and model accuracy loss is discussed in this section.

4.1 Data Quantization

One of the most commonly used method for model compression is the quantization on the weights and neurons. The neurons and weights of a neural network is usually represented by floating point data in common developing frameworks. Recent work try to replace this representation with low-bit fixed-point data or even a small set of trained values. On one hand, using less bits for each neuron or weight helps reduce the bandwidth and storage requirement of the neural network processing system. On the other hand, using a simplified representation reduce the hardware cost for each operation. The benefit on hardware will be discussed in detail in section 5. Two kinds of quantization methods are discussed in this section: linear quantization and non-linear quantization.

4.1.1 Linear Quantization. Linear quantization finds the nearest fixed-point representation of each weight and neuron. The problem of this method is that the dynamic range of floating point data greatly exceeds that for fixed point data. Most of the weights and neurons will suffer from overflow or underflow. Qiu, et al. [42] finds that the dynamic range of the weights and neurons in a single layer is much more limited and differs across different layers. Therefore they assign different fractional bit-widths to the weights and neurons in different layers. To decide the fractional

bit-width of a set of data, i.e. the neurons or weights of a layer, the data distribution is first analyzed. A set of possible fractional bit-widths are chosen as candidate solutions. Then the solution with the best model performance on training data set is chosen. In [42], the optimized solution of a network is chosen layer by layer to avoid an exponential design space exploration. Guo, et al. [13] further improves this method by fine tuning the model after the fraction bit-width of all the layers are fixed.

The method of choosing certain fractional bit-width equals to scale the data with a scaling factor of 2^k. Li, et al. [25] scales the weights with trained parameter W^l for each layer and quantize the weights with 2-bit data, representing W^l, 0 and $-W^l$. The neurons in this work is not quantized. So the the network still implements 32-bit floating point operations. Zhou, et al. [66] further quantize the weights of a layer with only 1 bit to $\pm s$, where $s = E(|w^l|)$ is the expectation of the absolute value of the weights of this layer. Linear quantization is also applied to the neurons in this work.

4.1.2 Non-linear Quantization. Compared with linear quantization, non-linear quantization independently assigns values to different binary codes. The translation from a non-linear quantized code to its corresponding value is thus a look-up table. This kind of methods helps further reduce the bit-width used for each neuron or weight. Chen, et al. [6] assign each of the weight to an item in the look-up table by a pre-defined hash function and train the values in look-up table. Han et al. [16] assigns the values in look-up table to the weights by clustering the weights of a trained model. Each look-up table value is set as the cluster center and further fine-tuned with training data set. This method is able to compress the weights of state-of-the-art CNN models to 4-bit without accuracy loss. Zhu, et al. [67] propose the ternary quantized network where all the weights of a layer are quantized to three values: W^n, 0, and W^p. Both the quantized value and the correspondance between weights and look-up table are trained. This method sacrifices less than 2% accuracy loss on ImageNet data set on state-of-the-art network models. The weight bit-width is reduced from 32-bit to 2-bit, which means about 16× model size compression.

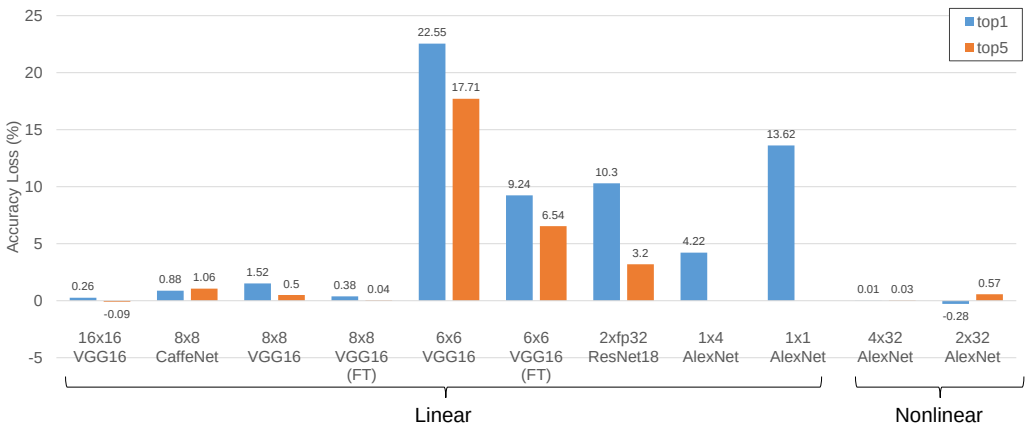

Fig. 3. Comparison between different quantization methods from [13, 16, 25, 42, 66, 67].

4.1.3 Comparison. We compare different quantization methods in Figure 3. The labels describe the experiments as $BW_{weight} \times BW_{neurons}$ and network name. The "(FT)" denotes that the network is fine-tuned after a linear quantization. Comparing different methods on different models is a little bit unfair. But it still gives some insights. For linear quantization, 8-bit is a clear bound to ensure

negligible accuracy loss. With 6 or less bits, using fine-tune or even training each weight from the beginning, will cause obvious accuracy degradation. If we require that 1% accuracy loss is within the acceptable range, we see that 8×8 linear quantization and 2×32 non-linear quantization is the lower bound of quantization. We will further discuss the performance gain of quantization in section 5.

4.2 Weight Reduction

Besides narrowing the bit-width of neurons and weights, another method for model compression is to reduce the number of weights. One kind of method is to approximate the weight matrix with a low rank representation. Qiu, et al. [42] compress the weight matrix W of an FC layer with singular value decomposition. An $m \times n$ weight matrix W is replaced by the multiplication of two matrices $A_{m \times p} B_{p \times n}$. For a sufficiently small p, the total number of weights is reduced. This work compress the largest FC layer of VGG network to 36% of its original size with 0.04% classification accuracy degradation. Zhang, et al. [64] use similar method for convolution layers and takes the effect of the following non-linear layer into the decomposition optimization process. The proposed method achieves 4× speed up on state-of-the-art CNN model targeting at ImageNet, with only 0.9% accuracy loss.

Pruning is another kind of method to reduce the number of weights. This kind of method directly remove the zeros in weights or remove those with small absolute values. The challenge in pruning is how to make more weights zero while keeping the model accuracy. One solution is the application of lasso object function during training. Liu, et al. [28] apply the spase group-lasso object function on the AlexNet [24] model. 90% weights are removed after training with less than 1% accuracy loss. Another solution is to prune the zero weights during training. Han, et al. [16] directly removes the values in network with zero or small absolute value. The left weights are then fine-tuned are training set to recover accuracy. Experimental result on AlexNet show that 89% weights can be removed while keeping the model accuracy.

The hardware gain from weight reduction is the reciprocal of the compression ratio. According to the above results, the improvment from weight reduction is upto 10×.

5 HARDWARE DESIGN: EFFICIENT ARCHITECTURE

In this section, we investigate the hardware level techniques used in state-of-the-art FPGA based neural network accelerator design to achieve high performance and high energy efficiency. We classify the techniques into 3 levels: computation unit level, loop unrolling level, and system level.

5.1 Computation Unit Designs

Computation unit level design affect the peak performance of the neural network accelerator. With a certain FPGA chip, the available resource is limited. A smaller computation unit design means more computation units and higher peak performance. If the computation units are simplified or the actual number of computation number is reduced, then the this also A carefully designed computation unit array can also increase the working frequency of the system and thus improve peak performance.

5.1.1 Low Bit-width Unit. Reduce the number of bit-width for computation is a direct way to reduce the size of computation units. The feasibility of using less bits comes from the quantization methods as introduced in section 4.1. Most of the state-of-the-art FPGA designs replace the 32-bit floating point units with fixed point units. Podili, et al. [40] implement 32-bit fixed point untis for the proposed system. 16-bit fixed point units are widely adopted in [10, 26, 42, 58, 59]. ESE [15] adopts 12-bit fixed-point weight and 16-bit fixed point neurons design. Guo, et al. [13] use 8-bit

units for their design on embedded FPGA. Recent work is also focusing on extremely narrow bit-width design. Prost-Boucle, et al. [41] implements 2-bit multiplication with 1 LUT for ternary network. Experiments in [39] shows that FPGA implementation of Binarized Neural Network (BNN) outperforms that on CPU and GPU. Though BNN suffers from accuracy loss, many designs explore the benefit of using 1 bit for computation [23, 27, 34, 36, 37, 51, 65].

The designs mentioned above are focused on computation unit for linear quantization. For non-linear quantization, translating the data back to full precision for computation is still of high cost. Samragh, et al. [44] proposes the factorized coefficients based dot product implementation. As the possible values of weights are quite limited for non-linear quantization, the proposed computation unit accumulates the multiplicator for each possible weight value and calculate the result as the weighted sum of the values in look-up table. In this way, the multiplication needed for one output neuron equals to the number of values in look-up table. The original multiplications are replaced by random addressed accumulations.

Most of the designs use a same bit-width through the process of a neural network. Qiu, et al. [42] finds that neurons and weights in FC layers can use less bits compared with CONV layers while the accuracy is maintained. Heterogeneous computation units are used in the designs of [12, 65].

The size of computation units of different bit-widths is compared in Figure 4. The resouce consumption is the synthesis result by Vivado 2017.2. All the IPs are required to not use DSP resources. Though we tend to use DSPs in real implementations, this result shows the actual hardware cost. It is also common to implement the computation units in a hybrid way like [42], where some of the operators are implemented by DSP and others by logic.

In neural network,the number of multiplications and addtions is approximately the same. So the sum of resource for multiplier and adder shows the overall cost. Operations with 32-bit fixed point data consumes similar resource as 32-bit floating point operations. For 16-bit operations, using fixed-point format saves about 30% resource. As introduced in section 4.1, 8-bit fixed point is the bound for linear quantization. Comapred with the 32-bit version, this allows 14× more hardware operators within the same area of logic. If further research can utilize 4-bit operations, this advantage becomes 54×. For 1-bit design, the improvement can be over 1000×.

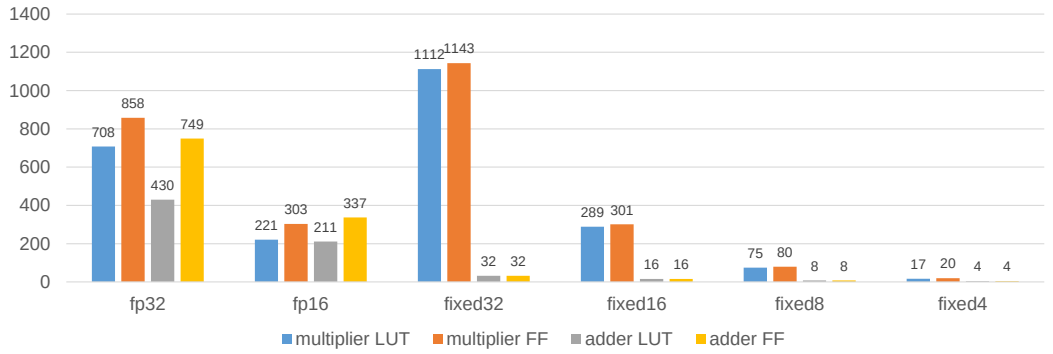

Fig. 4. FPGA resource consumption comparison for multiplier and adder with different types of data.

5.1.2 Fast Convolution Unit. For CONV layers, the convolution operations can be accelerated by special algorithms. Discrete Fourier Transformation (DFT) based fast convolution is widely adopted in digital signal processing. Zhang, et al. [61] propose a 2D DFT based hardware design for efficient CONV layer execution. For an $F \times F$ filter convolved with $K \times K$ filter, DFT converts the

$(F-K+1)^2K^2$ multiplications in sapce domain to F^2 complex multiplications in frequency domain. For a CONV layer with M input channel and N output channel, MN times of frequency domain multiplications are needed while only $(M+N)$ times DFT/IDFT are needed. The conversion of convolution kernels is once for all. So the domain conversion process are of low cost for CONV layers. This technique does not work for CONV layers with stride>1 or 1×1 convolution. Ding, et al. [8] suggests that a block-wise circular constraint can be applied on the weight matrix. In this way, the matrix vector multiplication in FC layers are converted to a set of 1D convolutions and can further accelerated in frequency domain. This method can also be applied to CONV layers by treating the $K \times K$ convolution kernels as $K \times K$ matrices and is not limited by K or stride.

Frequency domain methods require complex number multiplication. Another kind of fast convolution involves only real number multiplication [56]. The convolution of a 2D feature map F_{in} with a kernel K using Winograd algorithm is expressed by equation 3.

$$F_{out} = A^T[(GF_{in}G^T) \odot (BF_{in}B^T)]A \tag{3}$$

G, B and A are transformation matrix which only related to the sizes of kernel and feature map. $\odot$ denotes an element-wise multiplication of two matrices. For a 4×4 feature map convolved with 3×3 kernel, the transformation matrices are described as follows:

$$G = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 \end{bmatrix} \quad B = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix} \quad A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & -1 \\ 0 & -1 \end{bmatrix}$$

Winograd based methods are also limited by the kernel size and stride as DFT based methods. The most commonly used Winograd transformation is for 3×3 convolution in [30, 58].

The theoretical performance gain from fast convolution depends on the transformation kernel size. As DFT based method uses complex multiplication and costs more hardware, we estimate the theoretical gain according to the 6×6 Winograd used in [30], which is 4×.

5.1.3 DSP Optimization. Recent FPGAs implement hardened DSP units together with the reconfigurable logic to offer a high computation capacity. The basic function of a DSP unit is a multiplication accumulation (MAC). The bit-width for multiplication and addition is fixed. When the bit-width used in neural network does not match that of the DSP units, the FPGA is not fully utilized. The latest DSP units in Altera's FPGA implements 2 18×19 multipliers and can be configured into a 27×27 multiplier or a 32-bit floating point multiplier [1]. That of Xilinx's FPGA implements one 27×18 multiplier [2]. As mentioned in section 5.1.1, many designs adopt multiplication with less or equal than 16 bits, which can cause great DSP under utilization.

Nguyen, et al. [38] propose the design to implement two narrow bit-width fixed-point multiplication with a single wide bit-width fixed-point multiplier. In this design, AB and AC is executed with one multiplication $A(B << k + C)$. If k is sufficiently large, the bits for AB and AC does not overlap in the multiplication result and can be directly seperated. The design in [38] implements two 8-bit multiplications with one 25×18 multiplier, where k is 9. Similar method can be applied on other bit-width and DSPs. If this technique is used, the theoretical peak performance gain is 2×. But energy efficiency does not benefit from this technique because E_{unit_op} is not reduced.

5.1.4 Frequency Optimization Methods. All the above techniques introduced targets at increasing the number of computation units within a certain FPGA. Increasing the working frequency of the computation units also improves the peak performance.

To implement high parallelism, neural network accelerators usually implements matrix-vector multiplication or matrix-matrix multiplications rather than vector inner product as the basic operation. Different computation units share operators. Simply broadcast data to different computation units leads to large fan-out and high routing cost and thus reduce the working frequency. Wei, et al. [55] use the systolic array structure in their design. The shared data are transferred from one computation unit to the next in a chain mode. So the data is not broadcasted and only local connections between different computation units are needed. The drawback is the increase in latency. As the process of a neural network model is determined and the systolic structure is fully pipelined, the latency overhead can be fully covered.

Latest FPGAs support 700-900MHz DSP theoretical peak working frequency. But existing designs usually work at 100-300MHz [13, 32, 42, 59]. As claimed in [57], the working frequency is limited by the routing between on-chip SRAM and DSP units. The design in [57] use different working frequencies for DSP units and surrounding logic. Neighbour slices to each DSP unit are used as local RAMs to separate the clock domain. The prototype design in [57] achieves the peak DSP working frequency at 741MHz and 891MHz on FPGA chips of different speed grades. Yet this method is not adopted by a complete neural network accelerator design. Existing designs achieve 300MHz without this technique, so we esitimate the theoretical hardware performance gain as 2×.

5.2 Loop Unrolling Strategies

CONV layers and FC layers contribute to most of the computations and storage requirment of a neural network. As introduced in section 2. We express the CONV layer function in Figure 1(b) as nested loops in Algorithm 1. To make the code clear to read, we merge the loops along x and y directions for feature maps and 2-D convolution kernels respectively. An FC layer can be treated as a CONV layer with feature map and kernel both of size 1×1. Besides the loops in Algorithm 1, we also call the parallelism of the process of multiple inputs as a batch. This forms the batch loop.

Algorithm 1 Convolution Layer

Require: feature map F_{in} of size $M \times Y \times X$; convolution kernel Ker of size $N \times M \times K \times K$; bias vector b of size N

Ensure: feature map F_{out}

1: **function** ConvLayer(F_{in}, Ker)
2: Let $F_{out} \leftarrow$ zero array of size $N \times (Y - K + 1) \times (X - K + 1)$
3: **for** $n = 1; n < N; n++$ **do** ▷ Output channel loop
4: **for** $m = 1; m < M; m++$ **do** ▷ Input channel loop
5: **for** each (y, x) within $(Y - K + 1, X - K + 1)$ **do** ▷ Feature map loop
6: **for** each (ky, kx) within (K, K) **do** ▷ Kernel loop
7: $F_{out}[n][y][x] += F_{in}[m][y - ky + 1][x - kx + 1] * K[n][m][ky][kx]$
8: $F_{out}[n] += b[n]$
9: **return** F_{out}

To parallelize the execution of the loops, we unroll a certain part of the loops and map every operation in this part as a hardware computation unit. An inappropriate set of loop unroll parameter may lead to serious hardware underutilization. We take an example of three nested loops as shown in Figure 5. The big cube denotes all the operations within the loops. The length of each edge denotes the trip count of a loop. The small cube denotes the unrolled kernel, whose edges denote the unrolling parameter. A complete execution of the workload means to fullfill the big cube with

small cubes. Figure 5(a) shows an appropriate set of unroll parameters. But for Figure 5(b), the red part of some of the small cubes are out of the big cube, which means the hardware is wasted.

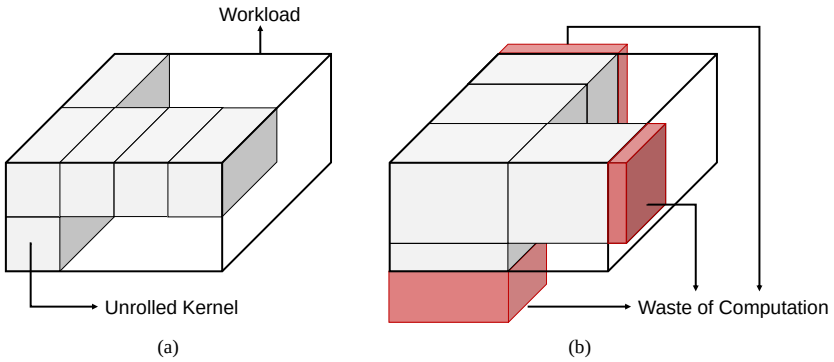

Fig. 5. Comparison between appropriate and inappropriate loop unroll parameters. (a) Appropriate parameters. (b) Inappropriate parameters.

It is obvious from Figure 5 that if the trip count of a loop is too small, the unroll parameter for this loop is limited. For a CNN model, the loop dimension varies greatly among different layeres. For a common network used on ImageNet classification like ResNet [18], the channel numbers vary from 3 to 2048, the feature map sizes vary from 224×224 to 7×7, the convolution kernel sizes vary from 7×7 to 1×1. Besides the under utilization problem, loop unrolling also affect the datapath and on-chip memory design. Thus loop unrolling strategy is a key feature for a neural network accelerator design.

Various work are proposed focusing on how the unroll parameter should be chosen. Zhang, et al. [60] propose the idea of unrolling the input channel and output channel loops and choose the optimized unroll parameter by design space exploration. Along these two loops, there is no input data cross dependency between neighbouring iterations. So no multiplexer is needed to rounte data from on-chip buffer to computation units. But the parallelism is limited as $7 \times 64 = 448$ multipliers. For larger parallelism, this solution is easy to suffer from the under utilization problem. Ma, et al. [32] further extends the design space by allowing parallelism on the feature map loop. The parallelism reaches $1 \times 16 \times 14 \times 14 = 3136$ multipliers. A shift register structure is used to route feature map pixels to the computation units.

The kernel loop is not chosen in the above work because kernel sizes vary greatly. Motamedi, et al [35] use kernel unrolling on AlexNet. Even with 3×3 unrolling for the 11×11 and 5×5 kenrels, the overall system performance still reaches 97.4% of its peak performance for the convolution layers. For certain networks like VGG [48], only 3×3 convolution kernels are used. Qiu, et al. [42] use the line-buffer structure to achieve 3×3 sliding window function and fully parallelize the kernel loop. Another reason to unroll kernel loop is to achieve acceleration with fast convolution algorithms. Design in [61] implements fully parallelized frequency domain multiplication on 4×4 feature map and 3×3 kernel. Lu, et al. [30] implement Winograd algorithm on FPGA with a dedicated pipeline for equation 3. The convolution of 3×3 kernel on 6×6 kernel is fully paralleized.

The above solutions are only for a single layer. But there is hardly a one-size-fits-all solution for a whole network, especially when we need high parallelism. Designs in [26, 29] propose fully pipelined structure with each layer a pipe stage. As each layer is executed with an independent part of hardware and each part is small, loop unrolling method can be easily chosen. This method is memory consuming because ping-pong buffers are needed between neighbouring layers for the

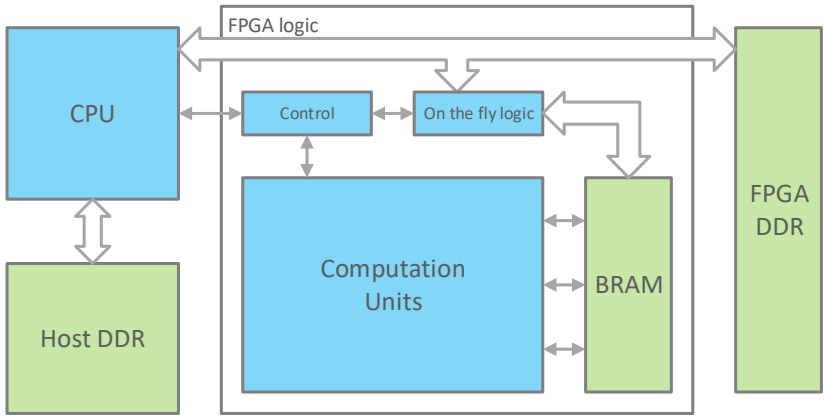

Fig. 6. Block graph of a typical FPGA based neural network accelerator system

feature maps. Design in [62] is similar but implemented on FPGA clusters to resolve the scalability problem. Shen, et al. [46] group the layers of a CNN by the loops' trip count and map each group onto one hardware module. Actuallly these solutions can be treated as unrolling the batch loop, because different inputs are processed in parallel on different layer pipe stages. The design in [30] implements parallelized batch both within a layer and among different layers. The drawback of batch parallel method is that the latency is higher compared with a $batch = 1$ design with a same parallelism.

Most of the current designs follow one of the above methods for loop unrolling. A special kind of design is for sparse neural network. Han, et al. [15] propose the ESE architecture for sparse LSTM network acceleration. Unlike processing a dense network, all the computation units will not work synchronously. This causes difficulty in sharing data between different computation units. ESE implements only the output channel (the output neurons of the FC layers in LSTM) loop unrolling within a layer to simplify hardware design and parallelize batch process.

5.3 System Design

A typical FPGA based neural network accelerator system is shown in Figure 6. The logic part of the whole system are denoted with the blue boxes. The host CPU issues workload or commands to the FPGA logic part and monitors its working status. On the FPGA logic part, a controller is usually implemented to communicate with host and generates control signals to all the other modules on FPGA. The controller can be an FSM or an instrcution decoder. The on the fly logic part is implemented for certain designs if the data loaded from external memory needs preprocess. This module can be data arrangement module, data shifter [42], FFT module [61], etc. The computation units are as discussed in section 5.1 and section 5.2.

The memory hiearchy of the system mainly contains three parts, denoted as the green boxes in Figure 6: Host DDR, FPGA DDR and on-chip block RAM. For state-of-the-art network, the number of weights can reach up to 100M. Using even 8-bit or less bit-width quantization will result in tens of MB storage requirement. Most of the current FPGAs implements less than 10MB on-chip memory and 1-8GB external DDR integrated on board. So for common designs, a two level memory hiearchy is used with DDR and on-chip memory.

From system level, the performance of a neural network accelerator is limited by two factors: the on-chip computation resource and the off-chip memory bandwidth. Various researches have been

proposed to achieve the best performance within a certain off-chip memory bandwidth. Zhang, et al. [60] introduce the roofline model in their work to analyze whether a design is memory bounded or computation bounded. An example of a roofline model is shown in Figure 7.

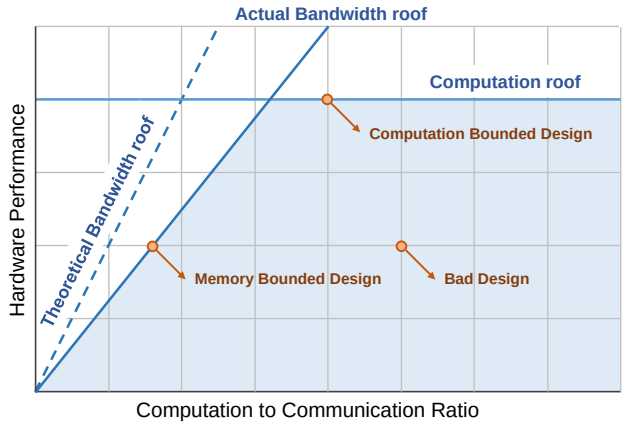

Fig. 7. An example of the roofline model. The shaded part denotes the valid design space given bandwidth and resource limitation.

The figure use the computation to communication (CTC) ratio as x-axis and hardware performance as y-axis. CTC is the number of operations that can be executed with a unit size of memory access. Each hardware design can be treated as a point in the figure. So y/x equals to the bandwidth requirement of the design. Given a certain platform, the availabel bandwidth is limited and can be described as the theoretical bandwidth roof in Figure 7. But the actual bandwidth roof is below the theoretical roof because for DDR access, the achievable bandwidth depends on the data access pattern. Sequential DDR access achieves much higher bandwidth than random access. The other roof is the computation roof, which is limited by the available resource on FPGA.

So a higher CTC ratio means the hardware is more likely to achieve the computation bound. Increasing the CTC ratio also reduce DDR access, which greatly reduce the energy cost according to [19]. In section 5.2, we have discussed the loop unrolling strategies to increase the parallelism while reducing the waste of computation for a certain network. When the loop unrolling strategy is decided, the scheduling of the rest part of the loops decides how the hardware can reuse data with on-chip buffer. This involves loop tiling and loop interchange strategy.

Loop tiling is a higher level of loop unrolling. All the input data of a loop tile will be stored on-chip and the loop unrolling hardware kernel works on these data. A larger loop tile size means that each tile will be loaded from external memory to on-chip memory less times. Loop interchange strategy decides the processing order of the loop tiles. External memory access happens when the hardware is moving from one tile to the next. Neighbouring tile may share a part of data. For example in a CONV layer, neighbouring tile can share input feature map or the weights. This is decided by the execution order of the loops.

In [32, 60], design space exploration is done on all the possible loop tiling sizes and loop orders. Many designs also explores the design space though some of the loop unrolling, tiling and loop order is already decided [35, 42]. Shen, et al. [47] also discuss the affect of batch parallelism over the CTC for different layers. This is a loop dimension not focused on in previous work.

All the above work give one optimized loop unrolling strategy and loop order for a whole network. Guo, et al. [13] implements flexible unrolling and loop order configuration for different

layers with an instruction interface. The data arrangement in on-chip buffers is controlled through instructions to fit with different feature map sizes. This means the hardware can always fully utilize the on-chip buffer to use the largest tiling size according to on-chip buffer size. This work also propose the "back and forth" loop execution order to avoid total on-chip data refresh when a innermost loop finishes.

Alwani, et al. [4] address the external memory access problem by fusing two neighbouring layers together to avoid the intermediate result transfer between the two layers. This strategy helps reduce 95% off-chip data transfer with extra 20% on-chip memory cost. Even software program gains 2× speedup with this scheduling strategy.

Besides increasing CTC, increasing the actual bandwidth roof helps improve the attainable performance with a certain CTC ratio. This is achieved by regularize the DDR access pattern. Common data format in external memory includes $NCHW$ or $CHWN$, where N means the batch dimension, C means the channel dimension, H and W means the feature map y and x dimension. Using any of these formats, a feature map tile may be cut into small data blocks stored in discontinuous addresses. Guan [10] suggest that a channel major storage format should be used for their design. This format avoids data duplication while long DDR access burst is ensured. Qiu et al. [42] propose a feature map storage format that arranges the $H \times W$ feature map into (HW/rc) tile blocks of size $r \times c$. So the write burst size can be increased from $c/2$ to $rc/2$.

6 EVALUTATION

In this section, we compare the performance of state-of-the-art neural network accelerator designs and try to evaluate the techniques mentiond in section 4 and section 5. The designs used for comparison are listed in Table 1. For data format, the "INT A/B" means that neurons are A-bit fixed point data and weights are B-bit fixed point data. We also investigate the resource utilization and draw advices to both accelerator designers and FPGA manufacturers.

Each of the designs in Table 1 drawn as a point in Figure 8, using $log_{10}(power)$ as x coordinate and $log_{10}(speed)$ as y axis. Therefore, $y - x = log_{10}(energy\_efficiency)$. Besides the FPGA based designs, we also plot the GPU experimental results used in [13, 15] as standards to measure the FPGA designs' performance.

6.0.1 Bit-width Reduction. Among all the designs, 1-2 bit based designs show outstanding speed and energy efficiency. This shows that extremely low bitwidth is a promising solution for high performance design. As introduced in section 4.1, linear quantized 1-2 bit network models suffer from great accuracy loss, so this technique is better applied on simple tasks now. The energy efficiency is improved by about 100× comapred with 32-bit floating point designs. But there is still a great gap towards the 1000× improvment estimation. This shows that memory access becomes the bottleneck for 1-2 bit designs, which only scales linearly with the bit-width. INT16/8, INT16 and INT8 are commonly adopted. But the difference between these designs are not obvious. This is due to the fact that current FPGAs implement wide multipliers in DSPs like 18×25 or 18×18. Using a less bits for computaion will not benefit the DSPs. The double MAC technique by [38] serves as a solution but is not adopted in the listed designs.

6.0.2 Fast Convolution Algorithm. Among all the 16-bit designs, [30] achieves the best energy efficiency and the highest speed with the help of the 6×6 Winograd fast convolution, which is 1.7× faster and 2.6× energy efficient than the 16-bit design in [63]. The design in [61] achieves 2× speedup and 3× energy efficiency compared with [60] where both designs use 32-bit floating point data. Overall, the improvement does not match the estimation but can still reach 2-3×.

Table 1. Performance and resource utilization of state-of-the-art neural network accelerator designs

	Data Format	Speed (GOP/s)	Power (W)	Efficiency (GOP/J)	Resource(%)			FPGA chip
					DSP	logic	BRAM	
[5]	FP16	1382	45	30.7	97	58	92	Arria 10 GX1150
[15]	INT16/12	2520	41	61.5	54.4	88.6	87.7	XCKU060
[52]	INT16	12.73	1.75	7.27	94.54	66.64	6.07	XC7Z020
[61]	FP32	123.5	13.18	9.37	87.5	85.4	64	Stratix V
[63]	INT16	1790	37.46	47.8	91	43	53	GX1150
	FP32	866	41.73	20.75	87	-	46	
[32]	INT16/8	645.25	21.2	30.43	100	38	70	GX1150
[42]	INT16	136.97	9.63	14.22	89.2	83.5	86.7	XC7Z045
[49]	INT16/8	117.8	19.1	6.2	12.5	22	65.2	5SGSD8
[60]	FP32	61.62	18.61	3.3	80	61.3	50	XC7VX485T
[10]	INT16	364.4	25	14.6	65	25	46	5SGSMD5
[30]	INT16	2940.7	23.6	124.6	-	-	-	ZCU102
[40]	INT32	229	8.04	28.5	100	83.7	17.6	Stratix V
[36]	1bit	329.47	2.3	143.2	0.5	34.4	11.4	Zynq XC7Z020
[23]	2bit	410.22	2.26	181.51	40.5	82.7	37.7	Zynq XC7Z020
[34]	1bit	40770	48	849.38	-	-	-	GX1155
[26]	INT16	565.94	30.2	22.15	59.56	63.21	65.07	XC7VX690T
[29]	INT16/8	222.1	24.8	8.96	39.9	26.6	39.7	XC7VX690T
[62]	INT16	1280.3	160	8	-	-	-	XC7Z020+ XC7VX690T×6
[13]	INT8	84.3	3.5	24.1	87	84	89	XC7Z020
[58]	INT16	229.5	9.4	24.42	91.9	71	83.2	XC7Z045
[11]	FP32	7.26	19.63	0.37	42	65.31	52.04	XC7VX485T
[59]	INT16	354	26	13.6	78	81	42	XC7VX690T

6.0.3 System Level Optimization. The overall system optimization is not well addressed in most of the work. As this is also related to the HDL design quality, we can just roughly evaluate the effect. Here we compare three designs[26, 29, 59] on the same XC7VX690T platform and try to evaluate the effect. All the three designs implement 16-bit fixed-point data format except that [29] uses 8-bit for weights. No fast convolution or sparsity is utilized in any of the work. Even though, [26] achieves 2.5× the energy efficiency of [29]. It shows that a system level optimization has a strong effect even comparable to the usage of fast convolution algorithm.

We also investigate the resource utilization of the designs in Table 1. Three kinds of resouce are considered, DSP, BRAM, and logic. We plot the designs in Figure 9 using two of the utilization ratio as x and y coordinate. We draw the diagonal line of each figure to show the designs' preference on hardware resource. The BRAM-DSP figure shows a obvious preference of hardware on DSP over BRAM. Similar preference is also between DSP and logic. This indicates that current FPGA designs are more likely computation bounded. FPGA manufacturers targeting neural network applications can adjust the resource allocation accordingly.

6.0.4 Comparision with GPU. In general, FPGA based designs have achieved comparable energy efficiency to GPU with 10-100GOP/J. But the speed of GPUs still surpass FPGAs.Scaling up the

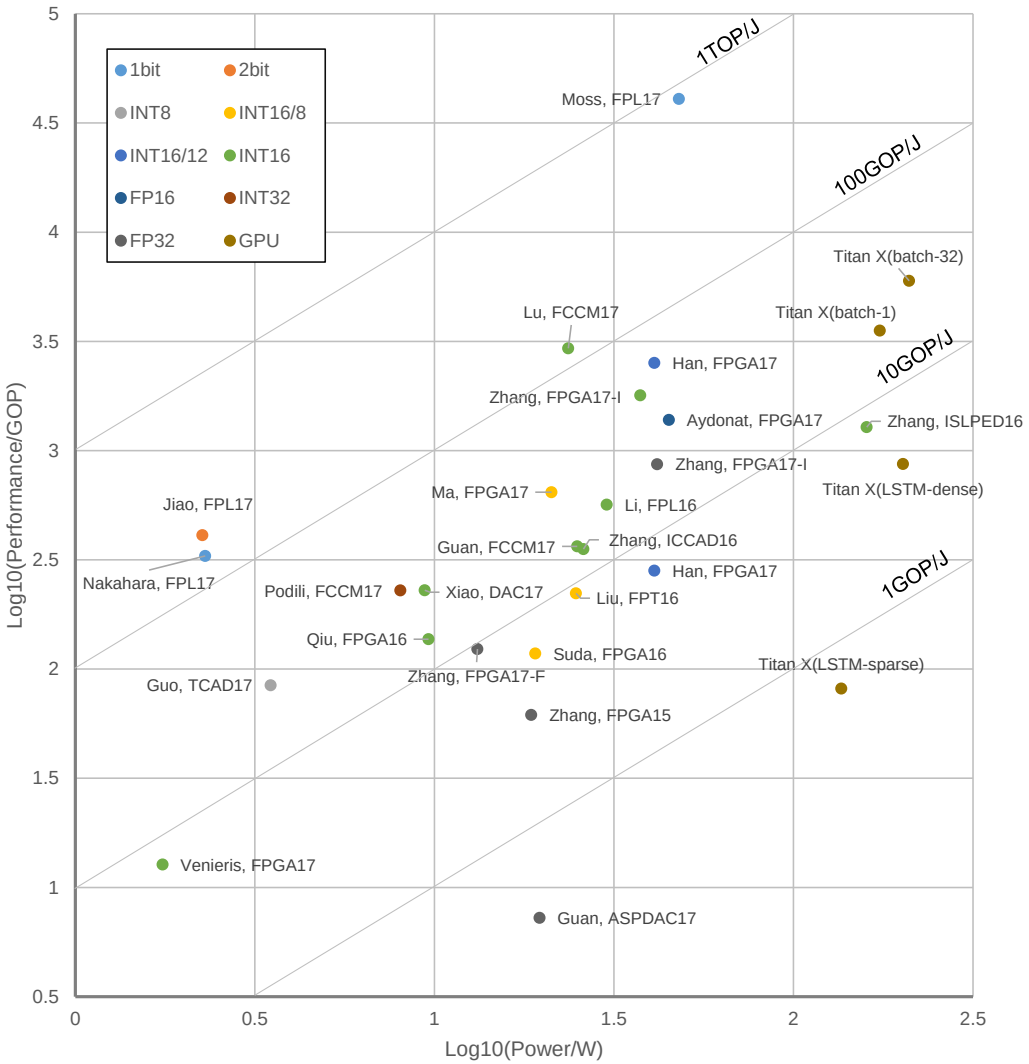

Fig. 8. A comparison between different designs on a logarithm coordinate of power and performance.

FPGA based design is still a problem. Zhang, et al. [62] propose the FPGA cluster based solution using 16-bit fixed point computation. But the energy efficiency is worse than the other 16-bit fixed-point designs.

Here we estimate the achievable performance of an ideal design. We use the 32-bit floating point design in [63] as a baseline. 8-bit linear quantization is used according to the analysis in section 4.1, which achieves 14× energy efficiency and speedup. Fast convolution and frequency optimization further improves the system by 4× and 2× respectively. Consider a 10× improvement from pruning data as in section 4.2, about 1000× speedup and 500× better energy efficiency. Even with a 100× conservative estimation, the system can achieve 80TOP/s with 80W power, reaching 1TOP/J, 1 magnitude over 32-bit floating point implementation on state-of-the-art GPU.

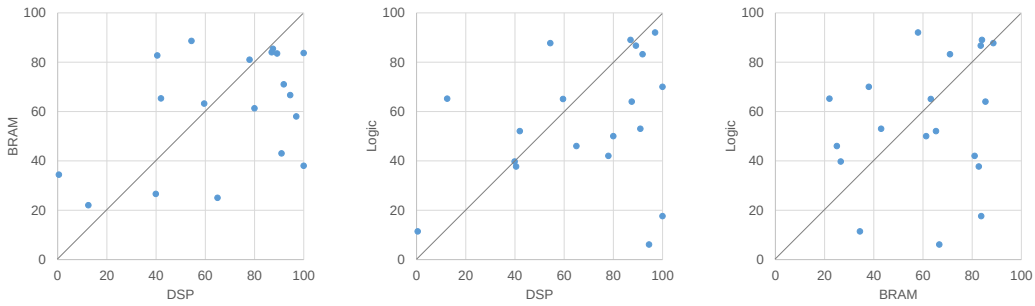

Fig. 9. Resource utilization ratio of different accelerator designs.

7 DESIGN FLEXIBILITY

Though we have discussed the techniques used to achieve high speed and high energy efficiency design, most of the designs are focusing on a certain network. In this section, we discuss the methods used to adapt a neural network design to different network models and corresponding development tools. Most of the development tools implement the interface at the network model level, for example using the prototext file in Caffe [22] as input. Towards the hardware flexibility, mainly two kinds of methods are used: HDL model based method and instruction based method.

7.1 HDL Model Based Method

HDL model based method is widely adopted in FPGA based accelerators [7, 31, 33, 45, 52–54]. These proposed techniques focus on automatically generate the HDL design based on the network parameter. Difference between these methods is the selection of an intermediate level description of the network to cover the gap between high level network description and low level hardware design.

A straight forward way is no intermediate description. The design flow in [31] search the optimized parameter for a hand coded verilog template with the input network description and platform constraint. This method is similar to the optimization methods mentioned in section 5. DiCecco, et al. [7] uses similar idea but based on OpenCL model. This enables that the development tool be integrated with Caffe and one network can be executed on different platforms.

Venireis, et al. [53] describes the network model as a DFG in their design tool. Then the network computaion process is translated to hardware design with DFG mapping method.

DnnWeaver [45] use an virtual instruction set to describe a network. The network model is first translated into instruction sequence. Then the sequence is mapped as hardware FSM states but not executed like traditional CPU instructions.

The HDL model based methods directly modifies the hardware design to each network. This means the hardware can always achieve the best performance on the target platform. This is suitable for FPGA because of its reconfigurability. It works in situations where network switching is not frequent and the reconfiguration overhead is not cared. For example, in the large scale cloud service, the change in network models can be covered by switching between different FPGA chips. So each of the FPGA do not need to be configured frequently.

7.2 Instruction Based Method

Instruction based methods try to run different networks on a same hardware design. The difference between these work is the granularity of instruction. At a lower level, Guo, et al. [13] propose the

instruction set with only three kinds of instructions: LOAD, CALC, and SAVE. The granularity of the LOAD and SAVE instructions are the same as the data tiling size. Each CONV executes a set of 2-D convolutions given the feature map size encoded in the instruction. The channel number is fixed as the hardware unrolling parameter. At this level, the software compiler is able to carry out static scheduling and dynamic data reuse strategy according the certain layer.

Zhang et al. [59] uses a layer level instruction. The control of a CNN layer is designed as a configurable hardware FSM. Compared with [13], this reduce the memory access for instruction access while increase the hardware cost on the configurable FSM.

Instruction based methods do not modify hardware and thus enables that the accelerator can switch between networks at run time. An example of the application senario is the real-time video processing system on a mobile platform. The process of a single frame can involve different networks if the task is complex enough. Reconfigure the hardware causes unacceptable overhead while instruction based methods can solve the problem if all the instructions of all the networks are prepared in memory.

7.3 Mixing Method

Wang, et al. [54] propose a method mixing the above two by both optimizing hardware design and compile software instructions. The hardware is first assembled with pre-defined HDL templates using the optimized hardware parameter. The data control flow of the computation process is controlled by software binaries, which is compiled according to the network description. It is possible that the hardware can be used for a new network by simply changing the software binaries.

8 CONCLUSION

In this paper, we review state-of-the-art neural network accelerator designs and summarize the techniques used. According to the evaluation result in section 6, with software hardware co-design, FPGA is able to achieve 13× better energy efficiency than state-of-the-art GPU while using 30% power with conservative estimation. This shows that FPGA is a promising candidate for neural network acceleration. We also review the methods used for flexible accelerator design, which shows that current development flow is able to achieve both high performance and run-time network switch.

But there is still gap between current designs and the estimation. Combining all the techniques requires software-hardware co-design. Using quantization and weight reduction together while maintaining the performance is challenging. Scaling up the design is another problem. Future work should focus on solving these challenges. There is still 10× potential performance gain from using 1/2 bits for neuron and weight representation. Future work should try to improve the model accuracy in these cases.

ACKNOWLEDGEMENT

This work was supported by National Natural Science Foundation of China (No. 61373026, 61622403, 61621091), National Key R&D Program of China 2017YFA0207600, and Joint fund of Equipment pre-Research and Ministry of Education (No. 6141A02022608).

REFERENCES

[1] [n. d.]. https://www.altera.com/products/fpga/stratix-series/stratix-10/features.html. ([n. d.]). Accessed Dec 7, 2017.
[2] [n. d.]. https://www.xilinx.com/support/documentation/user_guides/ug579-ultrascale-dsp.pdf. ([n. d.]). Accessed Dec 7, 2017.
[3] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (2016).

[4] Manoj Alwani, Han Chen, Michael Ferdman, and Peter Milder. 2016. Fused-layer CNN accelerators. In *Microarchitecture (MICRO), 2016 49th Annual IEEE/ACM International Symposium on*. IEEE, 1–12.

[5] Utku Aydonat, Shane O'Connell, Davor Capalija, Andrew C Ling, and Gordon R Chiu. 2017. An OpenCL (TM) Deep Learning Accelerator on Arria 10. *arXiv preprint arXiv:1701.03534* (2017).

[6] Wenlin Chen, James Wilson, Stephen Tyree, Kilian Weinberger, and Yixin Chen. 2015. Compressing neural networks with the hashing trick. In *International Conference on Machine Learning*. 2285–2294.

[7] Roberto DiCecco, Griffin Lacey, Jasmina Vasiljevic, Paul Chow, Graham Taylor, and Shawki Areibi. 2016. Caffeinated FPGAs: FPGA Framework For Convolutional Neural Networks. In *Field-Programmable Technology (FPT), 2016 International Conference on*. IEEE, 265–268.

[8] Caiwen Ding, Siyu Liao, Yanzhi Wang, Zhe Li, Ning Liu, Youwei Zhuo, Chao Wang, Xuehai Qian, Yu Bai, Geng Yuan, et al. 2017. CirCNN: accelerating and compressing deep neural networks using block-circulant weight matrices. In *Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture*. ACM, 395–408.

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 580–587.

[10] Yijin Guan, Hao Liang, Ningyi Xu, Wenqiang Wang, Shaoshuai Shi, Xi Chen, Guangyu Sun, Wei Zhang, and Jason Cong. 2017. FP-DNN: An Automated Framework for Mapping Deep Neural Networks onto FPGAs with RTL-HLS Hybrid Templates. (2017), 152–159.

[11] Yijin Guan, Zhihang Yuan, Guangyu Sun, and Jason Cong. 2017. FPGA-based accelerator for long short-term memory recurrent neural networks. In *Design Automation Conference (ASP-DAC), 2017 22nd Asia and South Pacific*. IEEE, 629–634.

[12] Jianxin Guo, Shouyi Yin, Peng Ouyang, Leibo Liu, and Shaojun Wei. 2017. Bit-Width Based Resource Partitioning for CNN Acceleration on FPGA. In *Field-Programmable Custom Computing Machines (FCCM), 2017 IEEE 25th Annual International Symposium on*. IEEE, 31–31.

[13] Kaiyuan Guo, Lingzhi Sui, Jiantao Qiu, Jincheng Yu, Junbin Wang, Song Yao, Song Han, Yu Wang, and Huazhong Yang. 2017. Angel-Eye: A Complete Design Flow for Mapping CNN onto Embedded FPGA. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* (2017).

[14] PK Gupta. 2016. Accelerating datacenter workloads. In *26th International Conference on Field Programmable Logic and Applications (FPL)*.

[15] Song Han, Junlong Kang, Huizi Mao, Yiming Hu, Xin Li, Yubin Li, Dongliang Xie, Hong Luo, Song Yao, Yu Wang, et al. 2017. ESE: Efficient Speech Recognition Engine with Sparse LSTM on FPGA.. In *FPGA*. 75–84.

[16] Song Han, Huizi Mao, and William J Dally. 2015. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149* (2015).

[17] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. 2014. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567* (2014).

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[19] M. Horowitz. [n. d.]. Energy table for 45nm process, Stanford VLSI wiki.[Online]. https://sites.google.com/site/seecproject. ([n. d.]).

[20] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. (2017).

[21] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. *arXiv preprint arXiv:1602.07360* (2016).

[22] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv preprint arXiv:1408.5093* (2014).

[23] Li Jiao, Cheng Luo, Wei Cao, Xuegong Zhou, and Lingli Wang. 2017. Accelerating low bit-width convolutional neural networks with embedded FPGA. In *Field Programmable Logic and Applications (FPL), 2017 27th International Conference on*. IEEE, 1–4.

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[25] Fengfu Li, Bo Zhang, and Bin Liu. 2016. Ternary weight networks. *arXiv preprint arXiv:1605.04711* (2016).

[26] Huimin Li, Xitian Fan, Li Jiao, Wei Cao, Xuegong Zhou, and Lingli Wang. 2016. A high performance FPGA-based accelerator for large-scale convolutional neural networks. In *Field Programmable Logic and Applications (FPL), 2016 26th International Conference on*. IEEE, 1–9.

[27] Yixing Li, Zichuan Liu, Kai Xu, Hao Yu, and Fengbo Ren. 2017. A 7.663-TOPS 8.2-W Energy-efficient FPGA Accelerator for Binary Convolutional Neural Networks. *arXiv preprint arXiv:1702.06392* (2017).

[28] Baoyuan Liu, Min Wang, Hassan Foroosh, Marshall Tappen, and Marianna Pensky. 2015. Sparse convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 806–814.

[29] Zhiqiang Liu, Yong Dou, Jingfei Jiang, and Jinwei Xu. 2016. Automatic code generation of convolutional neural networks in FPGA implementation. In *Field-Programmable Technology (FPT), 2016 International Conference on*. IEEE, 61–68.

[30] Liqiang Lu, Yun Liang, Qingcheng Xiao, and Shengen Yan. 2017. Evaluating fast algorithms for convolutional neural networks on fpgas. In *Field-Programmable Custom Computing Machines (FCCM), 2017 IEEE 25th Annual International Symposium on*. IEEE, 101–108.

[31] Yufei Ma, Yu Cao, Sarma Vrudhula, and Jae-sun Seo. 2017. An automatic RTL compiler for high-throughput FPGA implementation of diverse deep convolutional neural networks. In *Field Programmable Logic and Applications (FPL), 2017 27th International Conference on*. IEEE, 1–8.

[32] Yufei Ma, Yu Cao, Sarma Vrudhula, and Jae-sun Seo. 2017. Optimizing Loop Operation and Dataflow in FPGA Acceleration of Deep Convolutional Neural Networks. In *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*. ACM, 45–54.

[33] Raghid Morcel, Haitham Akkary, Hazem Hajj, Mazen Saghir, Anil Keshavamurthy, Rahul Khanna, and Hassan Artail. 2017. Minimalist Design for Accelerating Convolutional Neural Networks for Low-End FPGA Platforms. In *Field-Programmable Custom Computing Machines (FCCM), 2017 IEEE 25th Annual International Symposium on*. IEEE, 196–196.

[34] Duncan JM Moss, Eriko Nurvitadhi, Jaewoong Sim, Asit Mishra, Debbie Marr, Suchit Subhaschandra, and Philip HW Leong. 2017. High performance binary neural networks on the Xeon+ FPGAâĎć platform. In *Field Programmable Logic and Applications (FPL), 2017 27th International Conference on*. IEEE, 1–4.

[35] Mohammad Motamedi, Philipp Gysel, Venkatesh Akella, and Soheil Ghiasi. 2016. Design space exploration of fpga-based deep convolutional neural networks. In *Design Automation Conference (ASP-DAC), 2016 21st Asia and South Pacific*. IEEE, 575–580.

[36] Hiroki Nakahara, Tomoya Fujii, and Shimpei Sato. 2017. A fully connected layer elimination for a binarizec convolutional neural network on an FPGA. In *Field Programmable Logic and Applications (FPL), 2017 27th International Conference on*. IEEE, 1–4.

[37] Hiroki Nakahara, Haruyoshi Yonekawa, Hisashi Iwamoto, and Masato Motomura. 2017. A Batch Normalization Free Binarized Convolutional Deep Neural Network on an FPGA. In *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*. ACM, 290–290.

[38] Dong Nguyen, Daewoo Kim, and Jongeun Lee. 2017. Double MAC: Doubling the performance of convolutional neural networks on modern FPGAs. In *2017 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 890–893.

[39] Eriko Nurvitadhi, David Sheffield, Jaewoong Sim, Asit Mishra, Ganesh Venkatesh, and Debbie Marr. 2016. Accelerating Binarized Neural Networks: Comparison of FPGA, CPU, GPU, and ASIC. In *Field-Programmable Technology (FPT), 2016 International Conference on*. IEEE, 77–84.

[40] Abhinav Podili, Chi Zhang, and Viktor Prasanna. 2017. Fast and efficient implementation of Convolutional Neural Networks on FPGA. In *Application-specific Systems, Architectures and Processors (ASAP), 2017 IEEE 28th International Conference on*. IEEE, 11–18.

[41] Adrien Prost-Boucle, Alban Bourge, Frédéric Pétrot, Hande Alemdar, Nicholas Caldwell, and Vincent Leroy. 2017. Scalable high-performance architecture for convolutional ternary neural networks on FPGA. In *Field Programmable Logic and Applications (FPL), 2017 27th International Conference on*. IEEE, 1–7.

[42] Jiantao Qiu, Jie Wang, Song Yao, Kaiyuan Guo, Boxun Li, Erjin Zhou, Jincheng Yu, Tianqi Tang, Ningyi Xu, Sen Song, et al. 2016. Going deeper with embedded fpga platform for convolutional neural network. In *Proceedings of the 2016 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*. ACM, 26–35.

[43] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

[44] Mohammad Samragh, Mohammad Ghasemzadeh, and Farinaz Koushanfar. 2017. Customizing neural networks for efficient fpga implementation. In *Field-Programmable Custom Computing Machines (FCCM), 2017 IEEE 25th Annual International Symposium on*. IEEE, 85–92.

[45] Hardik Sharma, Jongse Park, Divya Mahajan, Emmanuel Amaro, Joon Kyung Kim, Chenkai Shao, Asit Mishra, and Hadi Esmaeilzadeh. 2016. From high-level deep neural models to FPGAs. In *Microarchitecture (MICRO), 2016 49th Annual IEEE/ACM International Symposium on*. IEEE, 1–12.

[46] Yongming Shen, Michael Ferdman, and Peter Milder. 2016. Overcoming resource underutilization in spatial CNN accelerators. In *Field Programmable Logic and Applications (FPL), 2016 26th International Conference on*. IEEE, 1–4.

[47] Yongming Shen, Michael Ferdman, and Peter Milder. 2017. Escher: A CNN Accelerator with Flexible Buffering to Minimize Off-Chip Transfer. In *Proceedings of the 25th IEEE International Symposium on Field-Programmable Custom Computing Machines (FCCMâĂŹ17). IEEE Computer Society, Los Alamitos, CA, USA*.

[48] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[49] Naveen Suda, Vikas Chandra, Ganesh Dasika, Abinash Mohanty, Yufei Ma, Sarma Vrudhula, Jae-sun Seo, and Yu Cao. 2016. Throughput-optimized OpenCL-based FPGA accelerator for large-scale convolutional neural networks. In *Proceedings of the 2016 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*. ACM, 16–25.

[50] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, et al. 2015. Going deeper with convolutions. Cvpr.

[51] Yaman Umuroglu, Nicholas J Fraser, Giulio Gambardella, Michaela Blott, Philip Leong, Magnus Jahre, and Kees Vissers. 2017. Finn: A framework for fast, scalable binarized neural network inference. In *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*. ACM, 65–74.

[52] Stylianos I Venieris and Christos-Savvas Bouganis. 2017. fpgaConvNet: Automated Mapping of Convolutional Neural Networks on FPGAs. In *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*. ACM, 291–292.

[53] Stylianos I Venieris and Christos-Savvas Bouganis. 2017. Latency-driven design for FPGA-based convolutional neural networks. In *Field Programmable Logic and Applications (FPL), 2017 27th International Conference on*. IEEE, 1–8.

[54] Ying Wang, Jie Xu, Yinhe Han, Huawei Li, and Xiaowei Li. 2016. DeepBurning: automatic generation of FPGA-based learning accelerators for the neural network family. In *Design Automation Conference (DAC), 2016 53nd ACM/EDAC/IEEE*. IEEE, 1–6.

[55] Xuechao Wei, Cody Hao Yu, Peng Zhang, Youxiang Chen, Yuxin Wang, Han Hu, Yun Liang, and Jason Cong. 2017. Automated Systolic Array Architecture Synthesis for High Throughput CNN Inference on FPGAs. In *Proceedings of the 54th Annual Design Automation Conference 2017*. ACM, 29.

[56] Shmuel Winograd. 1980. *Arithmetic complexity of computations*. Vol. 33. Siam.

[57] Ephrem Wu, Xiaoqian Zhang, David Berman, and Inkeun Cho. 2017. A high-throughput reconfigurable processing array for neural networks. In *Field Programmable Logic and Applications (FPL), 2017 27th International Conference on*. IEEE, 1–4.

[58] Qingcheng Xiao, Yun Liang, Liqiang Lu, Shengen Yan, and Yu-Wing Tai. 2017. Exploring Heterogeneous Algorithms for Accelerating Deep Convolutional Neural Networks on FPGAs. In *Proceedings of the 54th Annual Design Automation Conference 2017*. ACM, 62.

[59] Chen Zhang, Zhenman Fang, Peipei Zhou, Peichen Pan, and Jason Cong. 2016. Caffeine: Towards uniformed representation and acceleration for deep convolutional neural networks. In *Computer-Aided Design (ICCAD), 2016 IEEE/ACM International Conference on*. IEEE, 1–8.

[60] Chen Zhang, Peng Li, Guangyu Sun, Yijin Guan, Bingjun Xiao, and Jason Cong. 2015. Optimizing fpga-based accelerator design for deep convolutional neural networks. In *Proceedings of the 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*. ACM, 161–170.

[61] Chi Zhang and Viktor Prasanna. 2017. Frequency domain acceleration of convolutional neural networks on CPU-FPGA shared memory system. In *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*. ACM, 35–44.

[62] Chen Zhang, Di Wu, Jiayu Sun, Guangyu Sun, Guojie Luo, and Jason Cong. 2016. Energy-Efficient CNN Implementation on a Deeply Pipelined FPGA Cluster. In *Proceedings of the 2016 International Symposium on Low Power Electronics and Design*. ACM, 326–331.

[63] Jialiang Zhang and Jing Li. 2017. Improving the Performance of OpenCL-based FPGA Accelerator for Convolutional Neural Network.. In *FPGA*. 25–34.

[64] Xiangyu Zhang, Jianhua Zou, Xiang Ming, Kaiming He, and Jian Sun. 2015. Efficient and accurate approximations of nonlinear convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1984–1992.

[65] Ritchie Zhao, Weinan Song, Wentao Zhang, Tianwei Xing, Jeng-Hau Lin, Mani B Srivastava, Rajesh Gupta, and Zhiru Zhang. 2017. Accelerating Binarized Convolutional Neural Networks with Software-Programmable FPGAs.. In *FPGA*. 15–24.

[66] Shuchang Zhou, Yuxin Wu, Zekun Ni, Xinyu Zhou, He Wen, and Yuheng Zou. 2016. DoReFa-Net: Training low bitwidth convolutional neural networks with low bitwidth gradients. *arXiv preprint arXiv:1606.06160* (2016).

[67] Chenzhuo Zhu, Song Han, Huizi Mao, and William J Dally. 2016. Trained ternary quantization. *arXiv preprint arXiv:1612.01064* (2016).