



**HAL**  
open science

## A Holistic Approach for Optimizing DSP Block Utilization of a CNN implementation on FPGA

Kamel Eddine Abdelouahab, Cédric Bourrasset, Maxime Pelcat, François Berry, Jean-Charles Quinton, Jocelyn Sérot

► **To cite this version:**

Kamel Eddine Abdelouahab, Cédric Bourrasset, Maxime Pelcat, François Berry, Jean-Charles Quinton, et al.. A Holistic Approach for Optimizing DSP Block Utilization of a CNN implementation on FPGA. Proceedings of the 10th International Conference on Distributed Smart Cameras - ICDSC'16, Sep 2016, Paris, France. hal-01415955

**HAL Id: hal-01415955**

**<https://hal.science/hal-01415955>**

Submitted on 20 Dec 2017

**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.

# A Holistic Approach for Optimizing DSP Block Utilization of a CNN implementation on FPGA

K.Abdelouahab<sup>1</sup>, C.Bourrasset<sup>3</sup>, M.Pelcat<sup>1,2</sup>, F.Berry<sup>1</sup>, J.C.Quinton<sup>4</sup>, and J.Sérot<sup>1</sup>

<sup>1</sup>Institut Pascal, Université Clermont Auvergne, France

<sup>2</sup>IETR, INSA Rennes, France

<sup>3</sup>Atos/Bull, Montpellier, France

<sup>4</sup>Université Grenoble-Alpes, Grenoble, France

**Abstract**—Deep Neural Networks are becoming the de-facto standard models for image understanding, and more generally for computer vision tasks. As they involve highly parallelizable computations, Convolutional Neural Networks (CNNs) are well suited to current fine grain programmable logic devices. Thus, multiple CNN accelerators have been successfully implemented on FPGAs. Unfortunately, Field-Programmable Gate Array (FPGA) resources such as logic elements or Digital Signal Processing (DSP) units remain limited. This work presents a holistic method relying on approximate computing and design space exploration to optimize the DSP block utilization of a CNN implementation on FPGA. This method was tested when implementing a reconfigurable Optical Character Recognition (OCR) convolutional neural network on an Altera Stratix V device and varying both data representation and CNN topology in order to find the best combination in terms of DSP block utilization and classification accuracy. This exploration generated dataflow architectures of 76 CNN topologies with 5 different fixed point representation. Most efficient implementation performs 883 classifications/sec at  $256 \times 256$  resolution using 8 % of the available DSP blocks.

## I. INTRODUCTION

Convolutional Neural Network (CNN) techniques are taking part in an increasing number of computer vision applications. They have been successfully applied to image classification tasks [1], [2] and are wildly being used in image search engines or in data centers [3].

CNN algorithms involve hundreds of regular structures processing convolutions alongside non linear operations which allow CNNs to potentially benefit from a significant acceleration when running on fine grain parallel hardware. This acceleration makes FPGAs a well suited platform for CNN implementation. In addition, FPGAs provide a lower power consumption than most of the Graphics Processing Units (GPUs) traditionally used to implement CNNs. It also offers a better hardware flexibility and a reasonable computation power, as recent FPGAs embed numerous hard-wired DSP units. This match motivated multiple state-of-the-art approaches [4], as well as industry libraries [5] focusing on CNN implementation on FPGAs.

Nevertheless, efficient CNN implementations for FPGA are still difficult to obtain. CNNs remain computationally intensive while the resources of FPGAs (logic elements and DSP units) are limited, especially in low-end devices. In addition, CNNs

have a large diversity of parameters to tweak which makes exploration and optimization a difficult task.

This work focuses on DSP optimization and introduces a holistic design space exploration method that plays with both the CNN topology and its fixed point arithmetic while respecting some Quality of service (QoS) requirements. In summary, the contribution of this paper is threefold:

- A tool named Hardware Automated Description of CNNs (HADOC) is proposed. This utility generates rapidly dataflow hardware descriptions of trained networks.
- An iterative method to explore the design space is detailed. An optimized hardware architecture can be deduced after monitoring performance indicators such classification accuracy, hardware resources used or detection frame-rate.
- TPR/DSP metric is introduced. This ratio measures the quotient between classification performance of a CNN (its True Positive Rate (TPR)) and the number of DSPs required for its FPGA implementation. This gives a quantitative measure of an implementation efficiency.

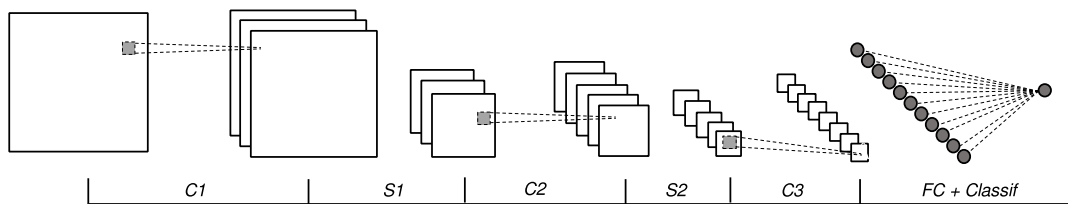
Thus, this paper is organized as follows: Section II summarises state-of-the art approaches for ConvNets implementations and optimization on FPGAs. Section III provides CNN background and links it to dataflow Model of Computation (MoC). Section VI introduces design space exploration for CNNs on FPGAs and our method for holistic optimizing. Section V details exploration and implementation results. Finally, section VI concludes the papers.

## II. RELATED WORK

Neural networks are nowadays wildly used for computer vision tasks. Hence, multiple CNN accelerators have been successfully implemented on FPGA. A non-exhaustive review of these can be found in [6].

First attempt was in 1996 with Virtual Image Processor (VIP) [7] : An FPGA based Single Instruction on Multiple Data (SIMD) processor for image processing and neural networks. However, since FPGAs in that times were very constrained in terms of resources and logic elements, VIP performance was quite limited.

Nowadays, FPGAs embed much more logic elements and hundreds of hardwired MAC operators (DSP Blocks). State-of-



**Fig. 1: An example a CNN topology, 3 convolutional layers interspersed with 2 sub-sampling layers and one fully connected stage**

the-art takes advantage of this improvement in order to implement an efficient feed-forward propagation of a CNN. Based on [6], and to our knowledge, best state-of-the-art performance for feed forward CNN acceleration on an FPGA was achieved by Ovtcharov in [3], with a reported classification throughput of 134 images /second on ImageNet 1K [1]. Such a system was implemented on an a Stratix V D5 device and outperformed most of state-of-the-art implementations such [8], [9], [10]. Most of these designs are FPGA based accelerators with a relatively similar architecture of parallel processing elements associated with soft-cores or embedded hardware processors running a software layer.

Regarding dataflow approaches for CNN implementations, the most notable contribution was neuFlow [4]: A runtime reconfigurable processor for real-time image classification. In this work, Farabet and al. introduced a grid of processing tiles that were configured on runtime to build a dataflow graph for CNN applications. It was associated to "luaFlow": a dataflow compiler that transforms a high-level flow-graph representation of an algorithm (in a Torch environment[11]) into machine code for neuFlow. Such architecture was implemented on a Virtex 6 VLX240T and provided a 12 fps categorization for 512x375 images.

In [12], an analytical design scheme using the roofline model and loop tiling is used to propose an implementation where the attainable computation roof of the FPGA is reached. This loop tiling optimization is performed on a C code then implemented in floating point on a Virtex 7 485T using Vivaldo HLS Tool. Our approach is different as it generates a purely dataflow architecture where topologies and fixed-point representations are explored.

### III. CNNs: BACKGROUND AND IMPLEMENTATION

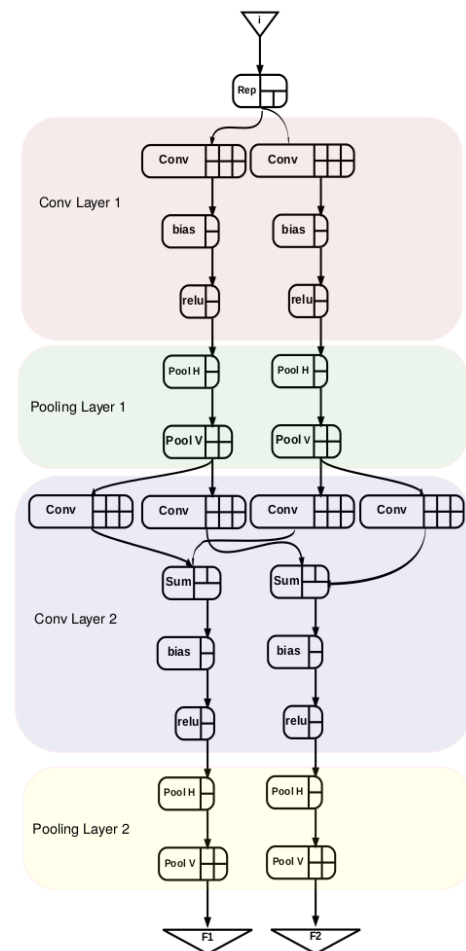
#### A. Convolutional networks topology

Convolutional Neural Networks, introduced in [2], have a feed-forward hierarchical structure consisting of a succession of convolution layers interspersed with sub-sampling operators. Each convolution layer includes a large number of neurons, each performing a sum of elementary image convolutions followed by a squashing non-linear function (Figure 1). A network topology can be described by its depth, the number of its neurons and their arrangement into layers. State-of-the-art CNNs for computer vision, such as [1], are usually deep networks with more than 5 hidden layers and with thousands of neurons.

Numerous machine learning libraries [11], [13], [14] can be used to design, train and test CNNs. Caffe [14] is a C++

package for deep learning developed by the Berkeley Vision and Learning Center (BVLC). This framework is leveraged on in this work as it benefits from a large community, contains Python and Matlab bindings, an OpenCL support and a "Model zoo repository", i.e. a set of popular pre-trained models to experiment on. Moreover, CNN topologies can easily be explored using Caffe.

Convolutional Neural Networks and more generally image stream processing algorithms can usually be expressed as sequences in an oriented graph of transformations. The CNN layout matches intuitively a dataflow model of computation.



**Fig. 2: Dataflow graph of an elementary CNN with 4 layers of 2 neurons**

## B. Dataflow MoC and CNNs

The foundations of dataflow MoC were formalized by [15] in order to create an architecture where multiple fragments of instructions can process simultaneously a stream of data. Programs respecting dataflow semantics are described as a *network* (graph) of fundamental processing units commonly called *Actors* and communicating abstract data messages called *tokens* on unidirectional First-In First-Out (FIFO) channels. As each neuron applies convolutions with known kernels on streams of feature maps, a dataflow processing model can be appropriate for CNNs [16]. Thus, a high parallelism degree can be introduced at each layer on the network and the successive layers can be fully pipelined.

As an example, Figure 2 shows the dataflow graph of a simple feature extractor composed of two convolutional layers of two neurons each interspersed with pooling layers. Each actor in this graph performs elementary operations of a CNN such as convolutions (`conv` actor), pooling (`poolV`, `poolH` actors), non linear activation (`ReLU` actor) and summation (`sum` actor). Such processing is done on a continuous stream  $i$  to extract two features  $F1$  and  $F2$  in the example.

## C. Implementation challenges on FPGAs

The implementation of feed-forward propagation of a deep neural network in FPGAs is constrained by the available computational resources. Using floating-point to represent network parameters makes the convolution arithmetic very costly and requires many logic elements and DSP blocks to be performed. Thus, CNN implementations on FPGAs usually build on fixed point representations of their data. Many studies have dealt with the aspects of deep learning with limited numerical precision [17], [18], [19]. The most common approach, such as in [10], [20], [4], consists of using a 16-bit fixed-point number representation which incurs little to no degradation in classification accuracy when used properly.

In this work, We explore different CNN topologies and data representations. These parameters are adjusted to give the best trade-off between performance and resource consumption.

## IV. DESIGN SPACE EXPLORATION

### A. The TPR to DSP Ratio (TDR)

Design Space exploration can be seen in our case as a method to deduce an efficient CNN implementation that optimizes either classification accuracy (TPR), hardware cost (with a focus in this paper on DSP utilization), or a trade-off between these two elements. To measure this "trade-off", the TPR to DSP metric (TDR in short) is introduced in equation 1. It computes the ratio between classification accuracy of the implementation (TPR) and the number of instantiated DSP blocks. TDR can be seen as the amount of classification accuracy that a DSP block contribute with. As an example, a TDR of 0.4 % means that every DSP block brings 0.4 % of classification accuracy. The higher the TDR is, the more efficient the implementation will be.

$$\text{TDR}(\%) = \frac{\text{TPR}(\%)}{\text{DSP}} \quad (1)$$

This work aims to maximize the TDR with a holistic approach that explores the CNN topology and the data representation of the learned weights and biases. These two sets of parameters can be expressed as follows:

- The number of bits required to quantify network parameters (weights and biases)

- $N_1, N_2, \dots, N_d$  : The number of neurons at each layer of the network.

Other parameters, such as  $d$  the depth of the network, or  $k$  the size of the convolution kernels can also be adjusted, but this can only be done at a price of an increased exploration complexity. Thus, these parameters are left to future work.

One way to explore the design space is by using an iterative method that generates the corresponding CNN hardware architecture for each network topology and data representation size. The proposed design space exploration method is described with algorithm 1. The CNN performances as well as the required hardware resources are monitored at each iteration. The implementation that maximizes the TDR can be considered the one balancing the most application requirements and hardware constraints.

```

forall possible network topologies do
  Train network
  forall possible fixed point representations do
    Generate hardware description
    Estimate classification accuracy
    Compute hardware utilization
  end
end

```

**Algorithm 1:** Design Space Exploration Procedure

### B. The HADDOC tool

The Hardware Automated Dataflow Deployment Of CNNs (HADDOC) utility is the tool proposed in this study for network exploration. Starting from a CNN description designed and learned using Caffe, HADDOC generates the corresponding dataflow graph (Actors and FIFOs) described as a Caph network. HADDOC also extracts the learned CNN parameters from a caffe trained model (represented in a 32-bit floating point format) and quantizes the data into the desired fixed point-representation.

Caph [21] is a dataflow language used here as an intermediate representation between the Caffe CNN network and its hardware description in VHDL. It is an image processing specific High-Level Synthesis (HLS) that interprets a desired algorithm to create a digital hardware description that implements its behaviour. Compared to other HLS tools, Caph main feature is to generate a purely dataflow and platform independent architecture from a graph network. This architecture, described in VHSIC Hardware Description Language (VHDL) can be implemented using a synthesis tool. These tools indicates the used hardware resources (logic elements, memory blocks and hard-wired DSPs) of the FPGA target. Moreover, Caph provides a systemC back-end to perform a functional simulation of the designed hardware. This was used to estimate classification accuracy for each network topology

and data representation scheme. Figure 3 summarizes the conception flow and tools involved in this work.

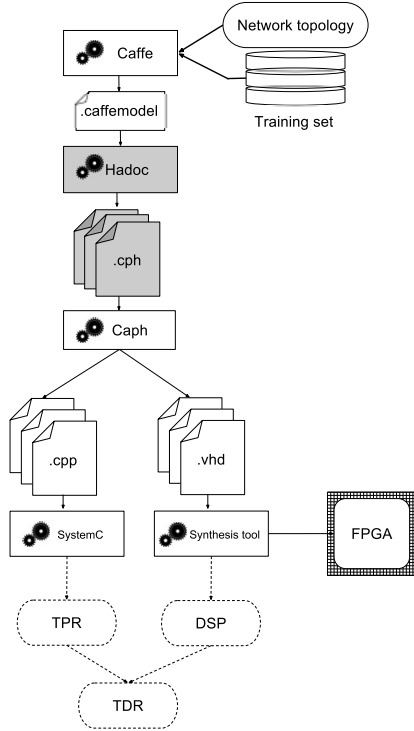


Fig. 3: Conception flow of design space exploration

## V. RESULTS

This section describes an example of design space exploration method. *Dreamnet*, a small convolutional neural network is explored, optimized and implemented on an Altera Stratix V 5SGSED8N3F4514 device.

*Dreamnet* is a light CNN designed for OCR applications. It is inspired from the LeNet5 [2] as it includes 3 convolutional layers of 3x3 kernels interspersed with 2 sub-sampling stages that perform max-pooling operations. The depth of the network is the only constant parameter, the network topology (number of neurons per layer  $N_1$ ,  $N_2$ , and  $N_3$ ), and parameter representation  $B$  being varied and explored. Table I describes its topology. *Dreamnet* is trained on 10000 images from the Mixed National Institute of Standards and Technology database (MNIST) handwritten digit database of which a few samples are displayed in Figure 4a. The classification accuracy of its implementation is then estimated on 1000 images of handwritten digits from the United States Postal Service (USPS) database. The USPS database contains digits difficult to identify, as shown by samples in Figure 4b.

In order to establish the optimal CNN topology, the space of possible configurations is explored. At each iteration, the tool chain of Figure 3 is leveraged on. It consists of the caffe tool for specifying the network and learning parameters, the HADDOC tool to generate Caph code, the Caph compiler to generate VHDL, and finally the Altera Quartus II synthesizer to evaluate the required hardware resources.

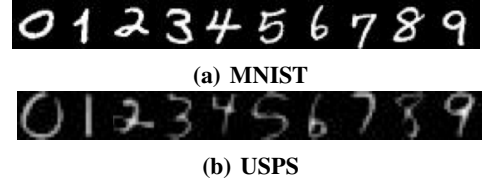


Fig. 4: Differences between MNIST and USPS handwritten digits databases

Layer	Size	Operation	Kernel
C1	$N_1$	Convolution	3x3
S1	$N_1$	Max sub-sampling	2x2
C2	$N_2$	Convolution	3x3
S2	$N_2$	Max sub-sampling	2x2
C2	$N_3$	Convolution	3x3
FC	10	Inner product	-
Classif	10	Softmax	-

TABLE I: Dreamnet topology

For each topology, *Dreamnet* is trained using caffe before generating the corresponding hardware. In this work, the choice is made to set, in the most resource-hungry case, a limited number of 5 neurons for the first layer, 10 for the second layer, and 14 for the third layer. This topology is sufficient enough to offer reasonable classification accuracy for OCR purpose on the MNIST database (99.7 % TPR on the MNIST test-set). As a CNN extracts features from an image hierarchically, the number of neurons in a layer  $N_i$  should be higher than the number of neurons in a previous layer  $N_{i-1}$ . Finally, a constant step  $step = 2$  is chosen to iteratively increment the topology parameters. This step can be reduced to 1 to have to have a more accurate optimization. The other explored parameter is the data representation size  $B$ . On *Dreamnet*, a data can have a maximum size of 7 bits. On one hand, this representation engenders a relatively low error rate compared to a floating-point reference. On the other hand, it prevents arithmetic over-flows to happen especially in the last stages of the network. In contrast, a minimum of 3 bits were used to represent the parameters which was the weakest precision usable to have acceptable classification rates.

The design space boundaries being defined (summarized in table II), Algorithm 1 can be reformulated as Algorithm 2. These boundaries will lead to explore a total of 76 networks with 5 different data type representations (A total of 380 combinations). In order to estimate classification accuracy, SystemC processed the 10000 images of the test set at a rate of 66.6 classifications/second while the synthesis tool takes an average of 6 minutes to compute the number of required DSP blocks. Thus, an architecture is explored every 8.5 minutes with an Intel i7-4770 CPU.

Parameter	Description	Min	Max
$N_1$	Number of neurons in C1	3	5
$N_2$	Number of neurons in C2	5	10
$N_3$	Number of neurons in C3	7	14
$B$	Representation size (in bits)	3	7

TABLE II: Design space boundaries: Dynamics of the explored parameters

```

for  $N_1$  in  $\min(N_1)$  to  $\max(N_1)$  do
  for  $N_2$  in  $N_1 + \text{step}$  to  $\max(N_2)$  do
    for  $N_3$  in  $N_2 + \text{step}$  to  $\max(N_3)$  do
      Caffe: Train network
      for  $B$  in  $\min(B)$  to  $\max(B)$ 
        Hadoc + Caph : Generate hardware
        SystemC: Simulate TPR
        Quartus: Compute DSP utilization
      end
    end
  end
end

```

**Algorithm 2:** Design space exploration on *Dreamnet*

A few remarkable implementations are detailed in tab III. The most efficient implementation considering TDR is I1: it presents the best trade-off between hardware cost and classification accuracy. Table IV gives post-fitting reports of I1. This architecture consumes low resources of the FPGA as it uses 161 from the 1963 available DSP blocks of the Stratix V device and 20 % of the available logic elements while maintaining a 64.8% classification rate on USPS at a rate of 57.93 MHz per pixel (which corresponds to 883 classifications per second with  $256 \times 256$  image resolution). Therefore, I1 could be implemented on a low-end device with less logic resources and DSP blocks. I2 is the implementation with the lowest number of neurons and data representation size, thus, this implementation has the weakest DSP usage. Finally, we found that the configuration with the greatest classification accuracy on USPS is I3. This implementation is among the ones with the highest number of neurons and fixed-point representation, considering the design space boundaries established above.

	$N_1$	$N_2$	$N_3$	$B$	TPR USPS	TPR MNIST	DSP	TDR
I1	4	6	8	5	64.8 %	98.3 %	161	0.40 %
I2	3	5	7	3	48.7 %	82.4 %	140	0.34 %
I3	4	8	12	7	73.2 %	99.7 %	428	0.17 %

**TABLE III:** Remarkable implementations

LOGIC UTILIZATION ( IN ALMS )	53,779 / 262,400 ( 20 % )
TOTAL RAM BLOCKS	109 / 2,567 ( 4 % )
TOTAL DSP BLOCKS	161/1963 ( 8 % )
FREQUENCY	57.93 MHz
CLASSIFICATION RATE ( AT $256 \times 256$ )	883 frame/s

**TABLE IV:** I1 implementation features on Stratix 5SGSED8N3F4514 device

#### A. Topology exploration

When only network topology is explored (fixed-point representation maintained constant at 3,4,...,7 bits), we find that both of classification accuracy and DSP utilization increase linearly with the number of neurons as shown in figures 5a and 5b. This causes implementations efficiency to decrease as the number of neurons grows, as represented in figure 5c. The more "sized" a CNN is, the less efficient its implementation will be.

#### B. Data-representation exploration

To see the approximate computing and numerical rounding effects on CNNs implementations, figures 6a,6b and 6c are plotted. They show the evolution of the average and standard deviation of network performances (in terms of TPR and DSP usage) for various data representations.

It can be seen that the mean classification accuracy (for all explored topologies) grows with numerical precision. Moreover, figure 6a shows how a 5 bit representation can be sufficient enough to maintain tolerable classification accuracy. In addition, figure 6b shows that DSP utilization grows quadratically when using different sizes of fixed point representations. Thus, the mean implementation efficiency (plotted in Figure 6c) have a maximum value (in our case 5 bits) that gives the best classification accuracy to DSP utilization trade-off.

#### C. Holistic Approach

Table V presents the architectures with more than 70 % classification accuracy on USPS. 5 of these implementations have a different network topology while 3 different data-representations are present.

If only a topology exploration with a 7 bits representation was performed, best reachable TDR would have been 0.171 i.e a relative loss of 41.9 % of efficiency compared to the optimum TDR of 0.298. On the other hand, if network topology was ignored and design space exploration focused only on fixed point data-representation, we show that there can be a loss of 6.8 % of classification accuracy between two implementations with same data-representation and different topologies (considering our design space size). This underlines how important a holistic approach is, where both topology and data-representation of a CNN implementation are explored.

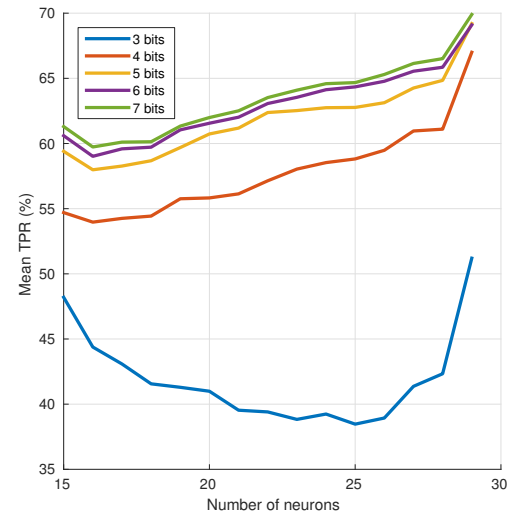
Figure 7 gives the result of such exploration as it shows that the most efficient implementations can be obtained after exploring various topologies and data representations. It also appears that a gradient descent optimization can be considered which could lead to a faster exploration process.

$N_1$	$N_2$	$N_3$	$B$	TPR <sub>usps</sub>	DSP	TDR
4	8	12	5	73	245	0.298
5	9	12	5	70.1	243	0.288
4	8	12	6	72.5	279	0.260
5	9	12	6	70.74	275	0.257
3	8	13	6	71.8	296	0.242
4	7	13	6	71	296	0.240
3	9	14	6	70.4	320	0.220
4	8	12	7	73.2	428	0.171
5	9	12	7	71.3	417	0.171
3	8	13	7	72.3	441	0.164
4	7	13	7	70.8	438	0.162
3	9	14	7	71.3	475	0.150

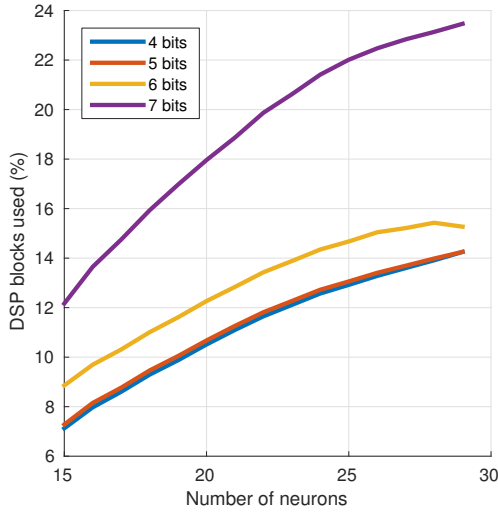
**TABLE V:** Top 10 efficient implementations

## VI. CONCLUSION

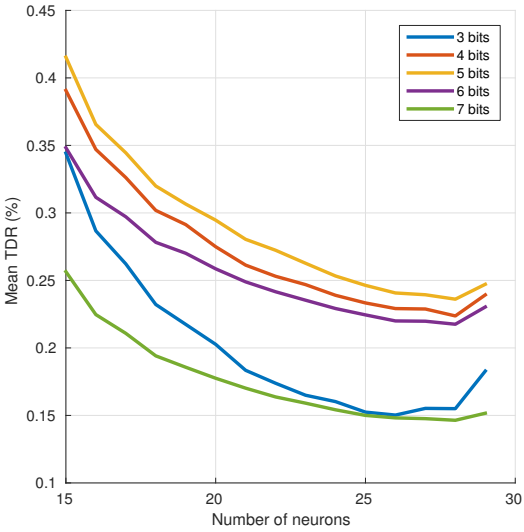
This work presented a method to optimize DSP utilization in FPGAs for CNN implementations. It relies on a holistic design space exploration on CNN topology and data-representation size to determine the most efficient architecture. As an example, this optimization was applied for OCR applications but



(a) Classification accuracy

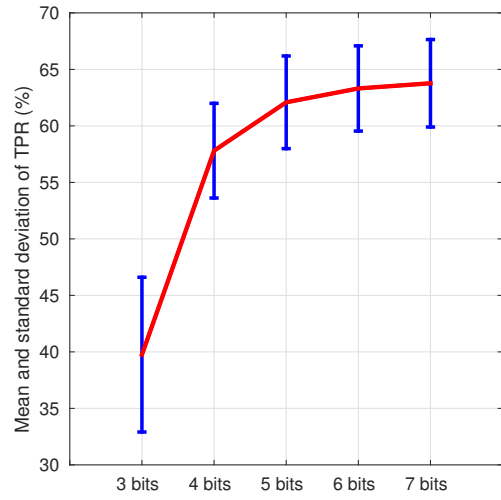


(b) DSP utilization

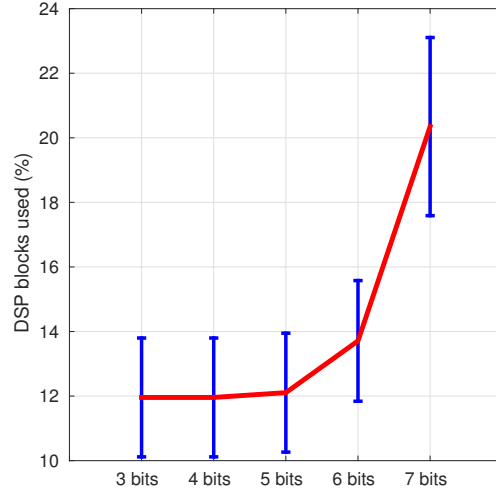


(c) Implementation efficiency

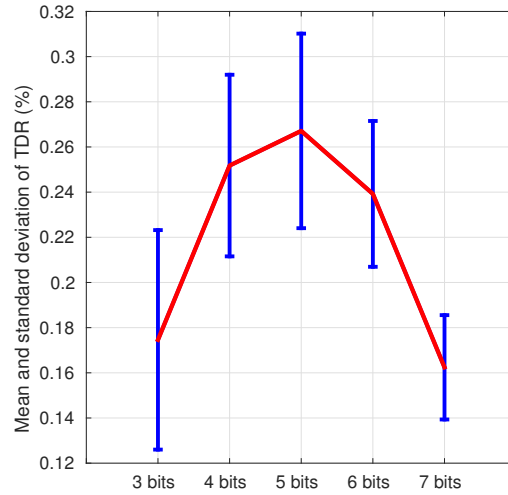
Fig. 5: Design space exploration for different CNN topologies



(a) Classification accuracy

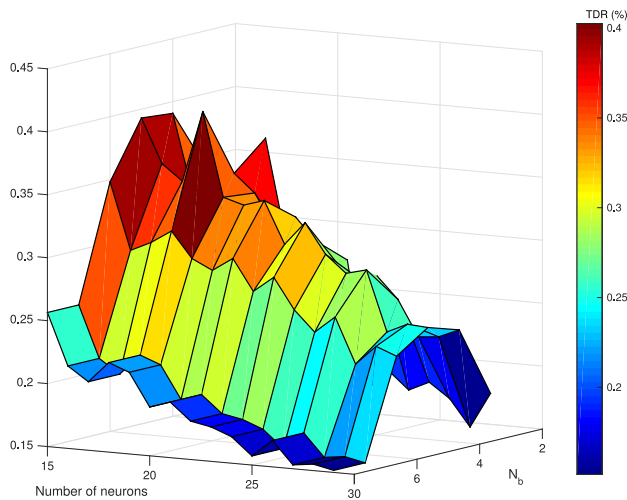


(b) DSP utilization



(c) Implementation efficiency

Fig. 6: Design space exploration for different representation size



**Fig. 7: Implementation efficiency of the CNN for multiple CNN sizes and fixed point representations**

can be transposed to other CNN classification tasks. It has been shown in this paper that a holistic approach is needed to optimize DSPs, as both fixed point arithmetic and topology network aspects should be explored.

The soft degradation in terms of quality when the number of bits is reduced or the topology simplified shows that CNNs are particularly well suited to approximate computing with a controlled rate of errors. Future work aim to improve optimization by augmenting the size of the explored design space and explore the CNN depth. Moreover, we expect that more efficient architectures can be implemented when bypassing the Caph HLS layer and generate directly the corresponding Hardware Description Language (HDL) of a neural network. Finally, it is planned to transpose design space exploration method to take more hardware constraints into account, such memory or logic elements utilization.

## VII. ACKNOWLEDGMENT

This work was funded by the french ministry of higher education MESR at Institut Pascal (UMR 6602). We thank them and all the collaborators for their support to this research.

## REFERENCES

- [1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. 2012.
- [2] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.
- [3] Kalin Ovtcharov, Olatunji Ruwase, Joo-Young Kim, Jeremy Fowers, Karin Strauss, and Eric S. Chung. Accelerating deep convolutional neural networks using specialized hardware. Microsoft Research, Feb 2015.
- [4] C. Farabet, B. Martini, B. Corda, P. Akselrod, E. Culurciello, and Y. LeCun. Neuflow: A runtime reconfigurable dataflow processor for vision. In *CVPRW'11, IEEE Computer Society Conference*.
- [5] Altera. FPGAs Achieve Compelling Performance-per-Watt in Cloud Data Center Acceleration Using CNN Algorithms, 2015.
- [6] G. Lacey, G. W. Taylor, and Areibi. Deep Learning on FPGAs: Past, Present, and Future. *ArXiv e-prints*, 2016.
- [7] J. Cloutier, E. Cosatto, and S. Pigeon. Vip: an fpga-based processor for image processing and neural networks. In *Microelectronics for Neural Network*, 1996.

- [8] Srimat Chakradhar, Murugan Sankaradas, Venkata Jakkula, and Srihari Cadambi. A dynamically configurable coprocessor for convolutional neural networks. *ACM- SIGARCH Comput. Archit. News*.
- [9] M. Peemen, A. Setio, B. Mesman, and H. Corporaal. Memory-centric accelerator design for convolutional neural networks. In *ICCD, 2013 IEEE*.
- [10] C. Farabet, C. Poulet, J. Y. Han, and Y. LeCun. Cnp: An fpga-based processor for convolutional networks. In *FPL International Conference on*, 2009.
- [11] R. Collobert. Torch. NIPS Workshop on Machine Learning Open Source Software, 2008.
- [12] Chen Zhang, Peng Li, Guangyu Sun, Yijin Guan, Bingjun Xiao, and Jason Cong. Optimizing fpga-based accelerator design for deep convolutional neural networks. *FPGA*, 2015.
- [13] Frédéric Bastien, Pascal Lamblin, and Goodfellow. Theano: new features and speed improvements. NIPS 2012 Workshop.
- [14] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, and Long. Caffe: Convolutional architecture for fast feature embedding. *arXiv*, 2014.
- [15] Jack B. Dennis and David P. Misunas. A preliminary architecture for a basic data-flow processor. *ISCA '75*. ACM.
- [16] Martin Abadi and al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.
- [17] Hong-Phuc Trinh & Marc Duranton & Michel Paidavoine. Efficient data encoding for convolutional neural network application. *ACM (TACO)*, 2015.
- [18] Anwar. S & Kyuyeon Hwang & Wonyong Sung. Fixed point optimization of deep convolutional neural networks for object recognition. *ICASSP, 2015 IEEE International Conference*, 2015.
- [19] Suyog Gupta, Ankur Agrawal, and Pritish Narayanan. Deep learning with limited numerical precision. *JMLR Workshop and Conference Proceedings*, 2015.
- [20] Vinayak Gokhale, Jonghoon Jin, Aysegül Dunder, Berin Martini, and Eugenio Culurciello. A 240 g-ops/s mobile coprocessor for deep neural networks. In *The IEEE (CVPR) Workshops*, June 2014.
- [21] J. Sérot and F. Berry. High-level dataflow programming for reconfigurable computing. In *Computer Architecture and High Performance Computing Workshop*, 2014.