Deep Neural Network on the Versat Reconfigurable Processor

João Pedro Costa Luís Cardoso

Thesis to obtain the Master of Science Degree in

Electrical and Computer Engineering

Supervisor(s): Prof. José João Henriques Teixeira de Sousa

September 2022

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Dedicated to my fiancée Matilde Moreira and soon to be born son, Francisco.

Acknowledgments

I want to thank my supervisor and Professor José Teixeira de Sousa for his eternal patience with me finishing this dissertation and the opportunity to work on the Versat CGRA. I would also like to acknowledge my friends and my parents who are always there for me and a special mention to my fiancée who has supported me all the way through my Bachelor and Masters and has pushed me to finally finish this document.

Resumo

Esta tese apresenta uma solução para simular o Deep Versat, uma CGRA que é acupulada de um processador RISC-V. Tambem é apresentada nesta tese ferramentas para o Deep Versat correr Redes Neuronais Convolucionais. Estas cargas de trabalho são usadas em diversos algoritmos de Inteligencia Arteficial como a deteção de objetos em imagens. As vantagens da ferramenta são várias. Primeiro, a escrita das configurações do Versat é preciso conhecimento da arquitetura a nivel detalhado e das suas APIs. Segundo, a escrita de algoritmos complexos para o Versat é preciso muitas horas de desenvolvimento e mais outras quantas para testar em hardware, ou seja o custo de usar o Versat baixa consideravelmente e a performance é otimizada á configuração do Versat escolhida podendo testar centenas de configurações de Hardware para otimizar a performance de uma rede em especifico.

Palavras-chave: CGRA, Versat, Darknet, Redes Neuronais

Abstract

Insert your abstract here with a maximum of 250 words, followed by 4 to 6 keywords...

Keywords: keyword1, keyword2,...

Contents

List of Tables

List of Figures

List of Acronyms

AGU Address Generation Unit

ALU Arithmetic Logic Unit

API Application Programming Interface

ASIC Application-Specific Integrated Circuit

ASIP Application Specific Instruction Set Processor

BRAM Block Random Access Memory

CGRA Coarse-Grain Reconfigurable Array

CM Configuration Module

CNN Convolutional Neural Network

CPI Cycles Per Instruction

CPU Central Processing Unit

DE Data Engine

DMA Direct Memory Access

DSP Digital Signal Processor

FPGA Field-Programmable Gate Array

FU Functional Unit

HDL Hardware Description Language

ISA Instruction Set Architecture

LED Light Emitting Diode

LUT Lookup Table

RTL Register-Transfer Level

UART Universal Asynchronous Receiver-Transmitter

UUT Unit Under Test

VCD Value Change Dump

VPI Verilog Procedural Interface

Chapter 1

Introduction

In this report, the problem of accelerating the execution of Deep Neural Networks (DNNs) using Coarse GRained Reconfigurable Arrays (CGRAs) is studied, with special emphasis on compiling a DNN description into code that runs on CPU/CGRA system. The Deep Versat Architecture [1] CGRA will be used as an implementation tool in this work.

1.1 Problem

Neural Networks have been an object of study since the 1940's, but until the beginning of this decade their applications were limited and did not play a major role in computer vision conferences. With its meteoric rise in research, several solutions to accelerate this algorithm have appeared, from Field Programmable Gate Arrays (FPGA) to Application Specific Integrated Circuits (ASIC) implementations.

Convolutional Neural Networks (CNNs) are a particular kind of DNN where the output values of the neurons in one layer are convolved with a kernel to produce the input values of the neurons of the next layer. This algorithm is compute bound, that is, its performance depends on how fast it can do certain calculations, and depend less on the memory access time. Namely the convolutional layers take approximately 90% of the computation time.

The acceleration of these workloads is a matter of importance for today's applications such as image processing for object recognition or simply to enhance certain images. Other uses like instant translation and virtual assistants are applications of neural networks and their acceleration is of vital importance to bring them into Internet of Things.

A suitable circuit to accelerate DNNs in hardware is the CGRA. A CGRA is a collection of Functional Units and memories with programmable interconnections in order to form computational datapaths. A CGRA can be implemented in both FPGAs and ASICs. CGRAs can be reconfigured much faster than FPGAs, as they have much less configuration bits. If reconfiguration is done at runtime, CGRAs add temporal scalability to the spacial scalability that characterize FPGAs. Moreover, partial reconfiguration is much easier to do in CGRAs compared to FPGAs which further speeds up reconfiguration time. Another advantage of CGRAs is the fact that they can be programmed entirely in software, contrasting with the large development time of customized Intellectual Property (IP) blocks. The Coarse Grain Reconfigurable Arrays (CGRA) is a midway acceleration solution between FPGAs, which are flexible but large, power hungry and difficult to reprogram, and ASICs, which are fast but generally not programmable.

However, mapping a specific DNN to a CGRA requires knowledge of its architecture, latencies and register configurations, which may become a lengthy process, especially if the user wants to explore the design space for several DNN configurations. An automatic compiler that can map a standard DNN description into CPU/CGRA code would dramatically decrease time to market of its users. Currently there are equivalent tools for CPUs and GPUs and even for FPGAS.

1.2 Solution

The proposed solution is a compiler that takes a configuration file from a neural network framework like Caffe or Darknet. This new tool inputs the parameters of Deep Versat, such as the number of layers and functional units, and produces the C code needed for the Versat runs. This code is run on the RISC-V picorv32 [2] CPU controller that has Deep Versat as a peripheral.

1.3 Report Outline

This report is composed of 4 more chapters. In the second chapter, the state-of-the-art of neural networks and the difficulties accelerating them is described. In the third chapter, the Deep Versat architecture and how to program it is explained. In the fourth chapter, CNN compiler techniques are explored. Finally, the last chapter contains the proposed solution and the plan for its execution.

Chapter 2

Background

Insert your chapter material here...

2.1 Theoretical Overview

Some overview of the underlying theory about the topic...

2.2 Theoretical Model 1

The research should be supported with a comprehensive list of references. These should appear whenever necessary, in the limit, from the first to the last chapter.

A reference can be cited in any of the following ways:

- Citation mode #1 - [**?**]
- Citation mode #2 - **?**]
- Citation mode #3 - [**?**]
- Citation mode #4 - **?**]
- Citation mode #5 - [**?**]
- Citation mode #6 - **?**
- Citation mode #7 - **?**
- Citation mode #8 - **?**
- Citation mode #9 - **?**
- Citation mode #10 - [**?**]

Several citations can be made simultaneously as [**? ?**].

This is often the default bibliography style adopted (numbers following the citation order), according to the options:
`\usepackage{natbib}` in file `Thesis_Preamble.tex`,
`\bibliographystyle{abbrvnat}` in file `Thesis.tex`.

Notice however that this style can be changed from numerical citation order to authors' last name with the options:
`\usepackage[numbers]{natbib}` in file `Thesis_Preamble.tex`,
`\bibliographystyle{abbrvunsrtnat}` in file `Thesis.tex`.

Multiple citations are compressed when using the `sort&compress` option when loading the `natbib` package as `\usepackage[numbers,sort&compress]{natbib}` in file `Thesis_Preamble.tex`, resulting in citations like [**? ? ? ?**].

2.3 Theoretical Model 2

Other models...

Chapter 3

Deep Neural Networks

A Neural Network (NN) is an interconnected group of nodes that follow a computational model that propagates data forward while processing. The earliest NNs were proposed by McCulloh and Pitts [3], in which a neuron has a linear part, based on aggregation of data and then a non-linear part called the activation function, which is applied to the aggregate sum. The issue with using only one neuron is that it is not able to be used in non-linear separable problems. By aggregating several neurons in layers and the input of each neuron as in figure 3.1 being based on the previous layers, that problem can be eliminated.

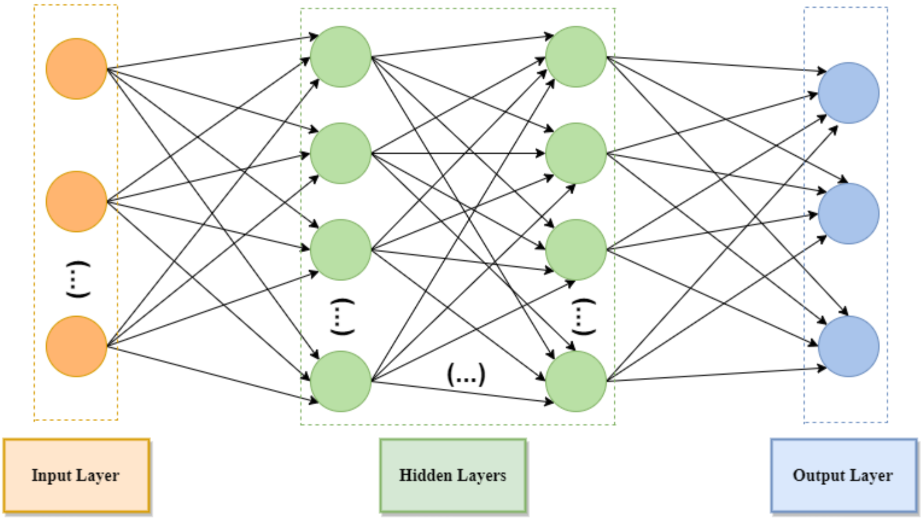


Figure 3.1: Deep Neural Network Structure

Each input to a neuron contributes differently to the output. The share is dependent on the weight value. These are obtained by training the network through various techniques, one of which is called Deep Supervised Learning [4] . For a certain input, there is an expected output and the real output of the NN. Then the loss function (the difference) is calculated and the weight values are iteratively modified for improving the outputs of the NN.

A Deep Neural Network (DNN) is a Neural Network that uses this approach for learning. It has multiple hidden layers and it can model complex non-linear relationships. If the activation function is non polynomial, it satisfies the Universal approximation problem [5].

One of the limitations of traditional NNs is the complexity of layer interconnections. Using as example the hand digit recognition problem and MNIST data set, composed of 28x28 grayscaled images [6], in a traditional fully connected NN, a neuron from the second layer would have 28x28 weights. That is 3.136 kiloBytes per neuron of weight values while using 32-bit floating-point numbers (FP32). When building a more complex network for image recognition, the computationally complexity grows quadratically with the number of neuros per layer.

3.1 Convolutional Neural Networks

Convolutional Neural Networks (CNN) are a class of DNNs used in Image and Video recognition due to their shift invariance characteristic. They were first proposed in the 1980s but it was not until 2012 with AlexNet [7] that CNNs really took off. Fundamentally, CNNs are a regularized version of Multilayer Perceptrons (MLP). These networks fix the complexity issue discussed as each neuron is only connected to a few neurons of the previous layer.

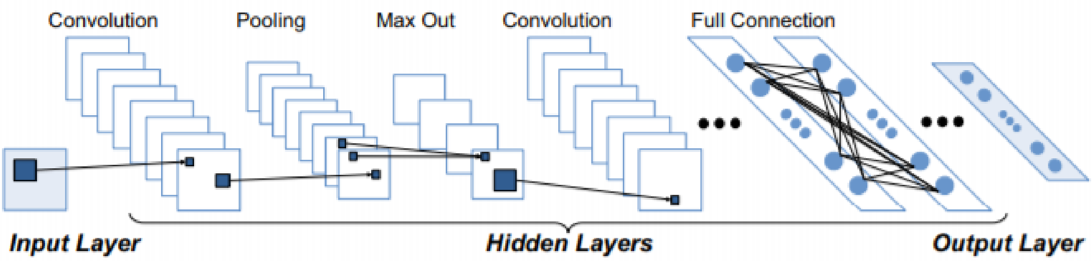


Figure 3.2: CNN architecture example, taken from [8]

3.1.1 Architecture Overview

Convolutional Layer

In a typical CNN not all layers are convolutional, but the convolutional layers are the most compute intensive ones. CNNs take input images with 3 dimensions (width, height and color space); for the following convolutional layers 3D arrays are used (width, height and number of channels). For the earlier example of the MNIST data set, the input would have dimensions 28x28x1 as it is a 2D image in grayscale.

To compute a neuron in the next layer we use the convolution equation 3.1 aided by Figure 3.3.

$$x_j^{l+1} = \delta(\sum_{i \epsilon M_j} x_i^l * k_{ij}^{l+1} + b_j^{l+1}) \tag{3.1}$$

where x_j^{l+1} is the output, δ is the activation function, which depends on the architecture, x_i^l is the input of the convolution layer, k_{ij}^{l+1} is the kernel of said layer which is obtained by training the network, and b_j^{l+1} is the bias.

Thus an output neuron depends only on a small region of the input which is called the local receptive field.

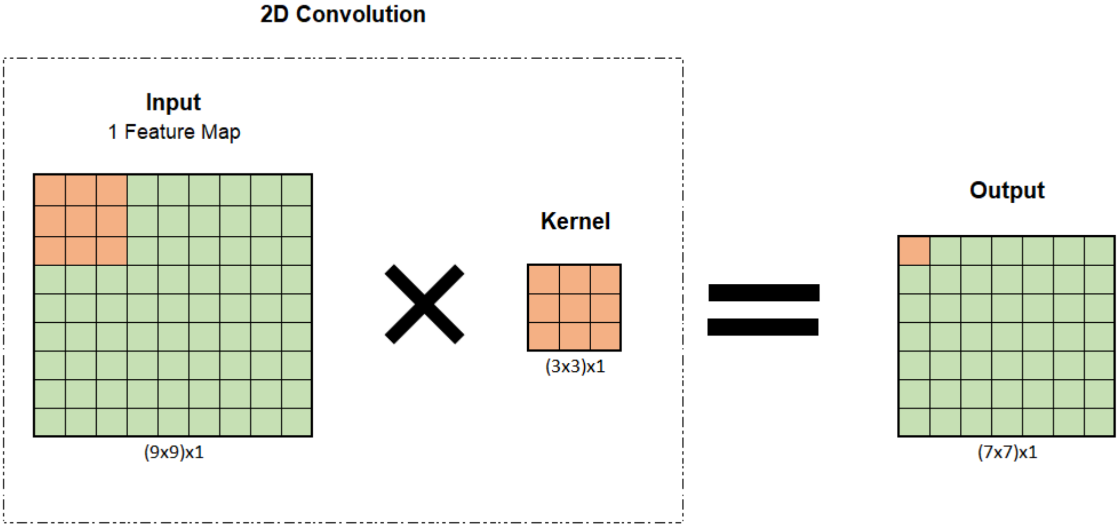


Figure 3.3: 2D convolution with stride = 1 and without zero padding

The output's dimensions depend on several parameters of the convolution such as zero-padding and stride. The former means to add zeros around the edges of the input matrix. The latter means the step used for the convolution, if the value is e.g 2, it will skip a pixel each iteration of the convolution. Equation 3.2 can be used to calculate the output size.

$$n^{l+1} = \frac{n^l - b^l + 2 \times p}{s} + 1 \tag{3.2}$$

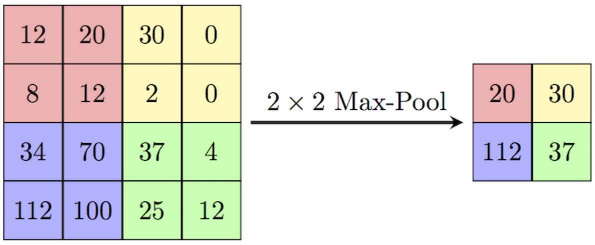
where n is the width/height of the input of layer l, b is the width/height of the kernel, p is zero-padding while s is the stride.

The number of channels of the output is equal to the number of filters in the convolutional layer.

Pooling Layer

The MaxPool or AvgPool are layers used in Convolutional Neural Networks to downsampling the feature maps to make the output maps less sensitive to the location of the features.

Maximum Pooling or MaxPool, like it is suggested in its name groups $n*n$ points and outputs the pixel with highest value. The output will have its size lowered by n times. The Average Pooling or AvgPool, instead takes all of the input points and calculates the average. Downsampling can also be achieved by using convolutions with stride 2 and padding equal to 1. Upsample layers can be also used that turn each pixel into n^2, where n is the amount of times the output will be bigger than the input.



12	20	30	0
8	12	2	0
34	70	37	4
112	100	25	12

20	30
112	37

Figure 3.4: Simple example of a maxpool layer, taken from [9]

Fully Connected Layer

The fully connected layer is mostly used for classification in the final layers of the NN. It associates the feature map to the respective labels. It takes the 3D vector and outputs a single vector thus it is also known as flatten. Equation 3.3 describes the operation.

$$x_j^{l+1} = \delta(\sum_i (x_i^l \times w_{ji}^{l+1}) + b_j^{l+1}) \tag{3.3}$$

where w_{ji}^{l+1} are the weights associated with a specific input for each output.

Route & Shortcut Layer

The Shortcut layer or skip connection was first introduced in Resnet [10]. It allows to connect the previous layer to another to allow the flow of information across layers. The Route layer, used in Yolov3 [11], concatenates 2 layers in depth (channel) or skips the layer forward. This is used after the detection layer in Yolov3 to extract other features.

Dropout Layer

This type of layer was conceived to avoid overfitting [12] by dropping the neurons with probability below the threshold. In Figure 3.5, there is a graphical representation.

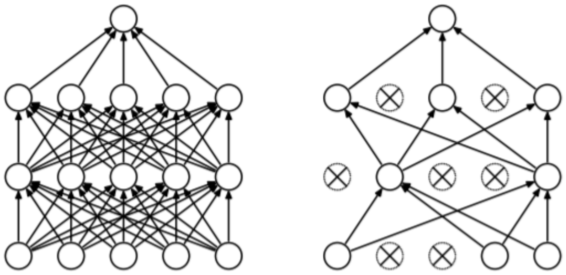


Figure 3.5: Dropout if applied to all layers, adapted from [12]

Activation Functions	Computation Equation
Sigmoid	$f(x) = \frac{1}{1 + e^{-x}}$
Tanh	$f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$
Softmax	$f(x_i) = \frac{x_i}{\sum_j e^{x_j}}$
ReLU	$f(x) = \begin{matrix} x & if & x \geq 0 \\ 0 & if & x < 0 \end{matrix}$
LReLU	$f(x) = \begin{matrix} x & if & x > 0 \\ \alpha x & if & x \leq 0 \end{matrix}$
ELU	$f(x) = \begin{matrix} x & if & x > 0 \\ \alpha e^{x} - 1 & if & x \leq 0 \end{matrix}$

Table 3.1: Popular activation functions

Activation Functions

Activation Functions (AF) are functions used in each layer of a NN to compute the weighted sum of input and biases, which is used to give a value to a neuron. Non-linear AFs are used to transform linear inputs to non-linear outputs. While training Deep Neural Networks, vanishing and exploding gradients are common issues, in other words, after successive multiplications of the loss gradient, the values tend to 0 or infinity and thus the gradient disappears. AFs help mitigate this issue by keeping the gradient within specific limits. The most popular activation functions can be found in table 3.1.

3.2 Frameworks for Neural Networks

To run a Neural Network model there are several popular frameworks like Tensorflow, PyTorch, Caffe and Darknet. Their purpose is to offer abstraction to software developers that want to run these networks. They also offer programming for different platforms like nVidia GPUs by using the CUDA API.

3.2.1 Darknet

Darknet [13] is an open source neural network framework written in C and CUDA. It is used as the backbone for Yolov3 [11] and supports several different network configurations such as AlexNet and Resnet. It utilizes a network configuration file (.cfg) and a weights file (.weights) as input for inference.

Listing 3.1: cfg code for a Convolutional Layer used in Yolov3 [11]

```
[convolutional]
batch_normalize=1
filters=32
size=3
stride=1
pad=1
```

```
activation=leaky
```

In Listing 3.1, there is a snippet of the file featuring a convolution layer with 32 kernels of size 3x3. It has stride 1 and zero padding of 1, meaning the output size equals the input size. The input size can be calculated by analyzing the previous layers and the network parameters. The network parameters in Listing 3.2 includes data to be used for training while only the first three parameters are needed for inference.

Listing 3.2: cfg code for the network parameters

```
[net]
width=608
height=608
channels=3

learning_rate=0.001
burn_in=1000
max_batches = 500200
policy=steps
steps=400000,450000
scales=.1,.1
```

3.2.2 Caffe

Convolutional Architecture for Fast Feature Embedding (Caffe) [14] is also an open source framework written in C++ with a Python interface. Caffe exports a neural network by serializing it using the Google Protocol Buffers (ProtoBuf) serialization library. Each network has 2 prototxt files:

- deploy.prototxt- File that describes the structure of the network that can be deployed for inference.
- train_val.prototxt- File that includes structure for training. it includes the extra layers used to aid the training and validation process.

The Python interface helps generate these files. For inference only the deploy file matters. In Listing 3.3, there is a snippet of a deploy file.

Listing 3.3: prototxt file for the input data and the first convolution layer of AlexNet [7]

```
name: "AlexNet"
layer {
  name: "data"
  type: "Input"
  top: "data"
  input_param { shape: { dim: 10 dim: 3 dim: 227 dim: 227 } }
}
layer {
  name: "conv1"
  type: "Convolution"
  bottom: "data"
  top: "conv1"
  param {
    lr_mult: 1
    decay_mult: 1
  }
  param {
    lr_mult: 2
    decay_mult: 0
  }
  convolution_param {
    num_output: 96
    kernel_size: 11
    stride: 4
  }
}
```

Chapter 4

Deep Versat

Versat is a Coarse Grained Reconfigurable Array (CGRA) Architecture. CGRAs are in-between Field Programmable Gate Arrays (FPGA) and general purpose processors (GPP). The former is fully reconfigurable and the highest performance for a workload can be achieved as the Architecture is tailored to the workload. GPPs on the other hand, are nor reconfigurable and thus slower but are more generic and can process different workloads. While FPGAs have the granularity at the gate level, CGRAs have the granularity at the functional unit level. They are configurable at run-time and the datapath can be changed in-between runs.

In this chapter, the base Versat Architecture will be explained and then the Deep Versat Architecture and its improvements.

4.1 Versat Architecture

The Versat Architecture [15–18] is depicted in Figure 4.1. Its composed by the following modules: DMA,Controller,Program Memory,Control File Registry,Data-Engine and Configuration module. The Controller accesses the modules through the control bus. The code made in assembly or C is loaded into the program Memory (RAM) where the user can write to the configuration module for the versat runs. Between runs of the Data Engine, the Controller can start doing the next run configuration and calculations.

4.1.1 Data Engine

The Data Engine which is represented in Figure 4.2 carries out the computation needed on the data arrays. Its a 32 bit Architecture with up to 11 Functional Units (FU): Arithmetic and Logic Unit(ALU), stripped down ALU (ALU-Lite), Multiplier and Accumulator (MAC) and Barrel Shifter. Depending on the project and calculations, a new type of FU or the existing ones can be altered to support the algorithm. The DE has a full mesh topology, that means that each FU can be the output to another, which leads to a decrease in operating frequency.

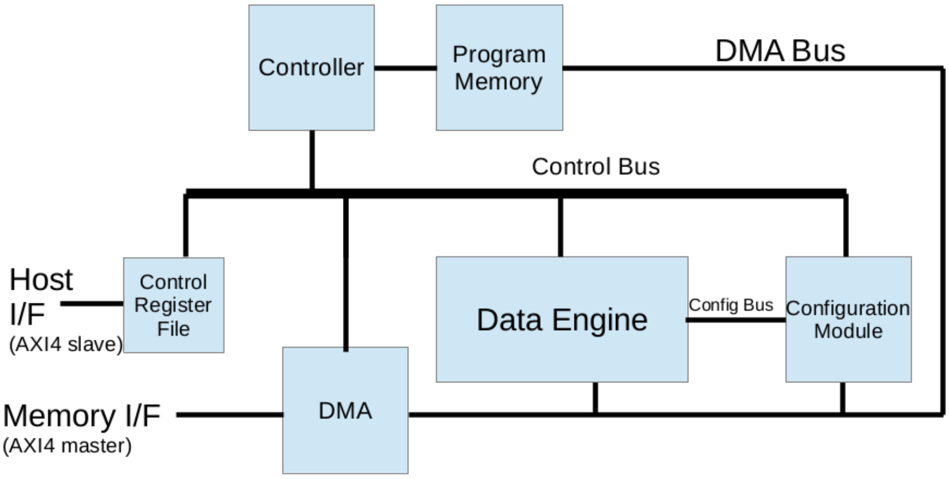


Figure 4.1: Versat Topology, taken from [16]

Each Input of a Functional Unit has a Mux with 19 entries, 8 of which are from the memories (2 from each Mem out of 4 total units) and the rest from the Functional Units (11).

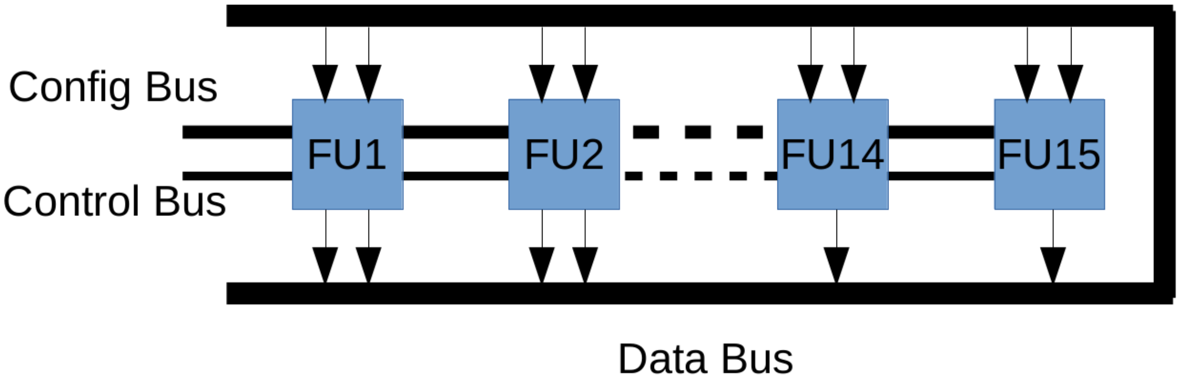


Figure 4.2: Versat Data Engine Topology, taken from [17]

The 4 Memories are dual port and for the input of both ports, there is an Address Generation Unit (AGU) that is able to reproduce two nested loops of memory indexes. The AGUs control which MEM data is the input of the FUs and where to store the results of the operation. Also, the AGUs support a delayed start to line up timings due to latencies.The memory module is represented in Fig 4.3.

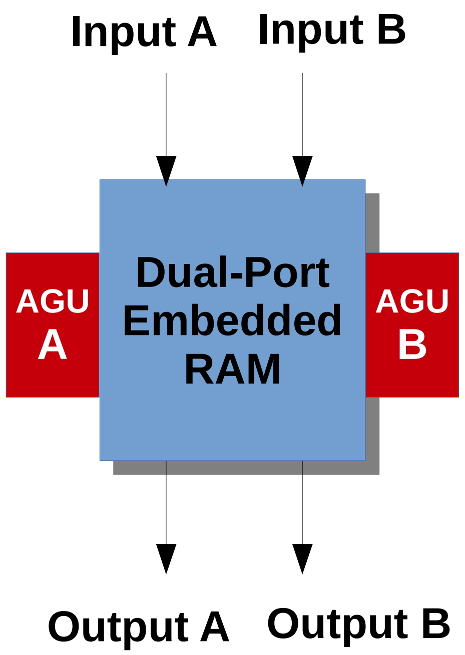


Figure 4.3: Versat Memory Unit with one AGU per port, taken from [19]

4.1.2 Configuration Module

Versat has several configuration spaces devised for each Functional Unit, with each space having multiple fields to define the operation of the Functional unit (e.g which op for the ALU). These are accessed before the run by the controller to define the datapath.

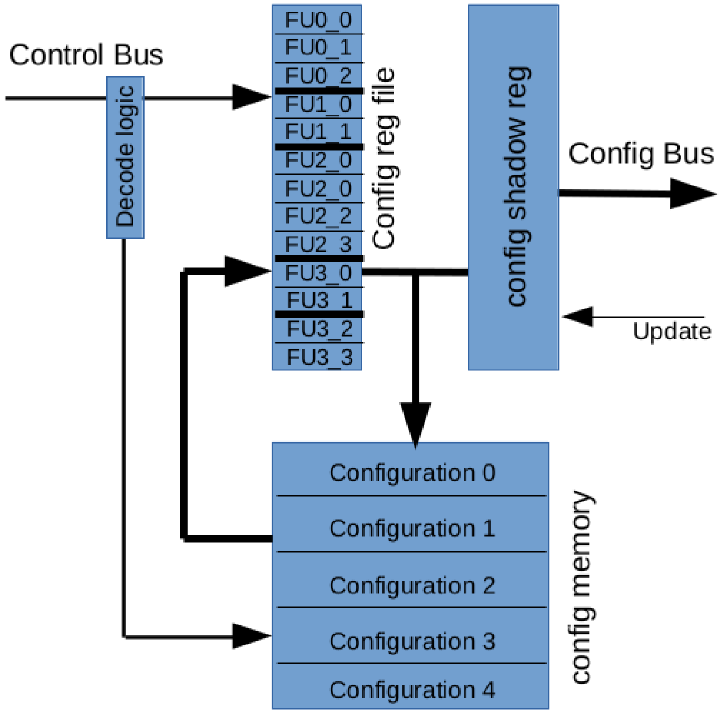


Figure 4.4: Configuration Module,taken from [16]

The Configuration Module (CM), depicted in figure 4.4, has three components:configuration memory, variable length configuration register file and configuration shadow register. The latter holds the current configuration so the controller can change the values of the configuration file in-between runs. The decode logic finds which component to write or read, if its the registers, it ignores read operations. Meanwhile, the configuration memory interprets both write and reads. When it receives a read, it writes into the register configuration data, when its a write, it stores the data instead.

4.2 Deep Versat Architecture

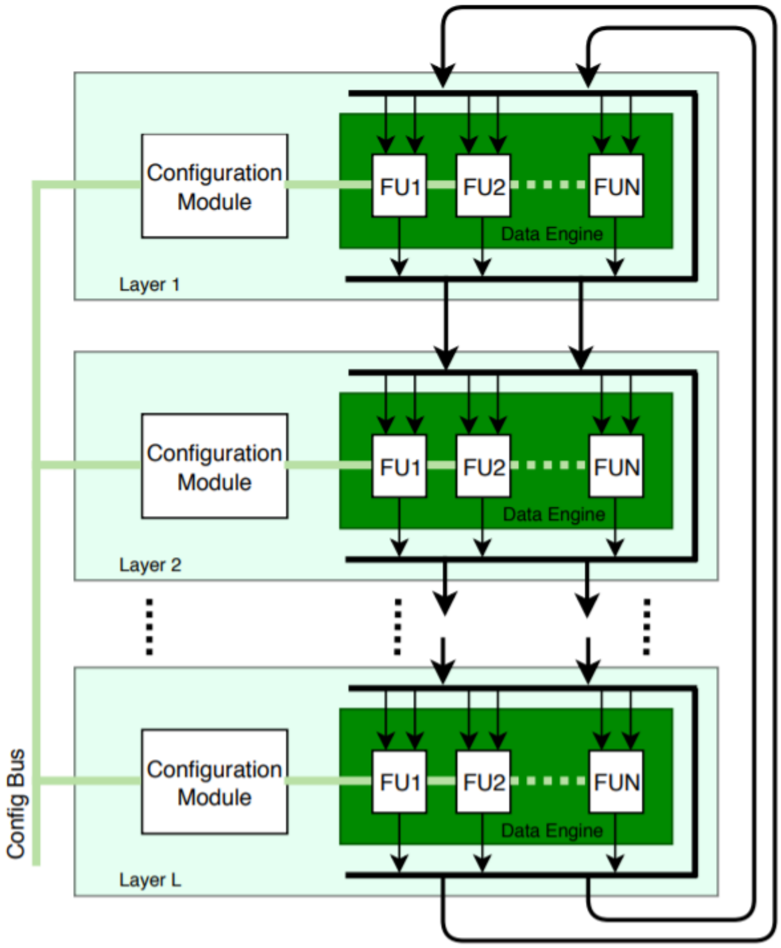


Figure 4.5: Deep Versat Architecture, taken from [1]

The Deep Versat Architecture [1] , in figure 4.5, decouples the Data Engine (DE) from all control and as such, it can be used with any CPU. It can be paired with hard cores in FPGA boards like the ZYNC board with its A9 ARM dual core CPUs or pair it with a soft core.

Its principle is to create the concept of a Versat Core: Configuration Module (CM) and its Functional Units (FU) connected with a control bus and a data bus. Instead of writing to a memory, there is the option to write for the next Versat Core to create more complex and more complete Datapaths, to avoid having to reconfigure the cores.

The number of Layers and FUs are reconfigurable pre-silicon with the only limitation that each layer is identical. To program Deep Versat, an API is generated from the Verilog .vh files.

4.2.1 Deep Versat System

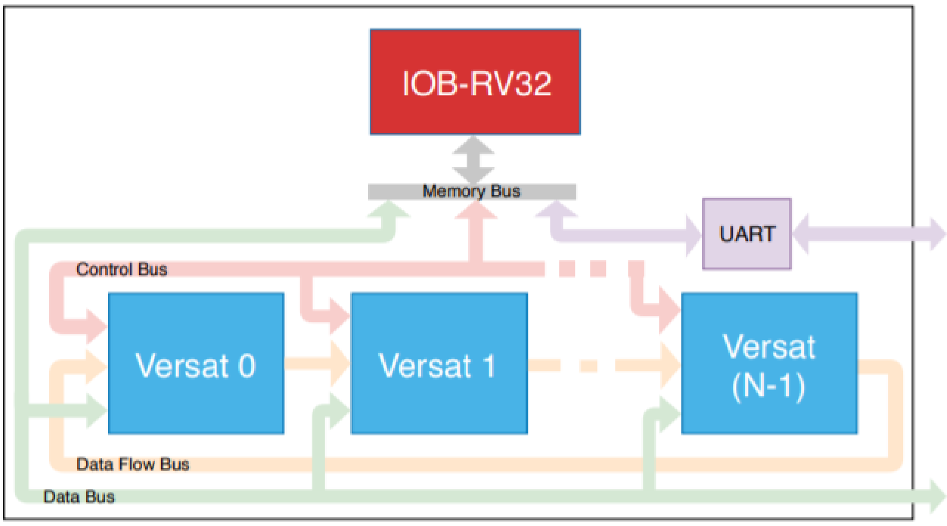


Figure 4.6: Deep Versat System using a RISC-V RV32IMC soft core, taken from [1]

To make a complete system, a new controller is needed with a more robust toolchain. In a recent dissertation [1], the IOB-RV32 processor was used which uses the RISC-V Instruction Set (ISA) with 32 bit Integer base alongside Multiplication and Division extension and Compact Instruction extension. The core is derived from the open source PicoRV32 CPU [2]. The IOB-RV32 uses its memory bus to access peripherals in which Deep Versat and the UART module are connected as such. The control bus is used to access the configuration modules of Deep Versat. The data bus is used to read and write large amount of data into Deep Versat. The data flow bus is reserved for inter Versat Core communication.

Peripheral	**Memory address**
UART module	12’h100xxxxx
Deep Versat control bus	8’h11xxxxxx
Deep Versat data bus	8’h12xxxxxx

Table 4.1: Deep Versat Memory Map

The memory map to address the peripherals, including deep versat, is in table 4.1. Each Versat has 15 bits of address while the CPU addresses the peripherals with 32 bits, with 8 of those occupied to chose the peripheral in question. That leaves 9 bits to address several Versat Cores which brings the theoretical maximum versat cores to 512. The IOB-RV32 is compatible with the GNU toolchain to offer better portability of code and alongside the C++ Versat API the difficulty to code for the System diminishes.

Chapter 5

CNN Compiling in FPGAs

This chapter presents an overview of toolflows that map convolutional neural networks into FPGA using the frameworks presented in Section 3.2. Next, the concepts for mapping CNNs into CGRAs are introduced.

5.1 Toolflows for Mapping CNNs in FPGAs

Several software frameworks have been developed to accelerate development and execution of CNNs. The neural networks frameworks discussed in section 3.2 provide high level APIs together with high performance execution on multi-core CPUs, GPUs, Digital Signal Processors (DSPs) and Neural Processing Units (NPUs) [20]. FPGAs provide an alternative to these architectures as they provide high-performance while also being low-power. FPGAs can meet several requirements like throughput and latency in diversity of applications. Thus, several toolflows that map CNN descriptions into hardware in order to perform inference have been created. In table 5.1, a list of notable ones is presented.

Toolflow Name	Interface	Year
fpgaConvNet	Caffe & Torch	May 2016
DeepBurning	Caffe	June 2016
Angel-Eye	Caffe	July 2016
ALAMO	Caffe	August 2016
Haddoc2	Caffe	September 2016
DNNWeaver	Caffe	October 2016
Caffeine	Caffe	November 2016
AutoCodeGen	Proprietary Input Format	December 2016
Finn	Theano	February 2017
FP-DNN	Tensorflow	May 2017
Snowflake	Torch	May 2017
SysArrayAccel	C	June 2017
FFTCodeGen	Proprietary Input Format	December 2017

Table 5.1: CNN to FPGA Toolflows, adapted from [21]

5.1.1 Supported Neural Network Models

These toolflows support the most common layers in CNNs, which are discussed in chapter 3. The acceleration target changes depending on the toolflow. For example, the fpgaConvNet [22] toolflow focuses more on feature extraction while offering non accelerated support for fully connected layers.

5.1.2 Architecture & Portability

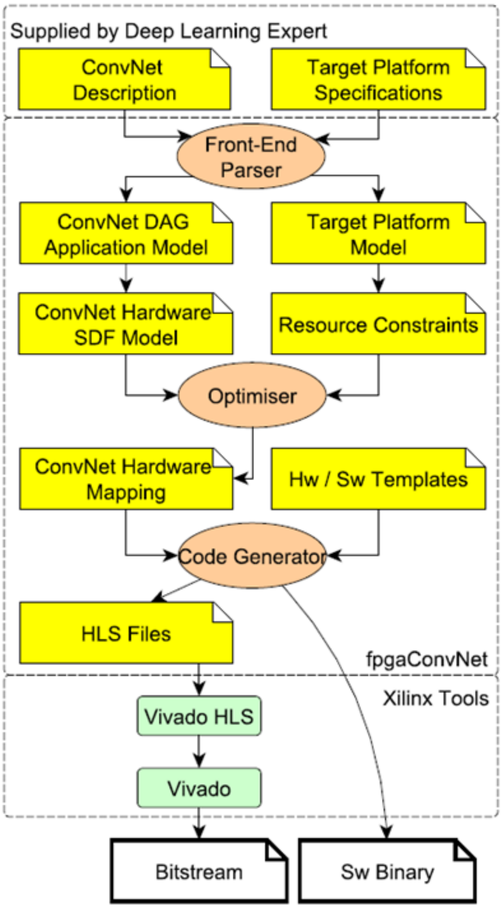


Figure 5.1: fpgaConvNet Architecture. Taken from [22]

As shown in figure 5.1, the fpgaConvNet architecture consists of a Front-End Parser that reads a (ConvNet) description of the network and a description of the target platform and produces, on the one hand a Directed Acyclic Graph (DAG), which is then converted to a Synchronous Data Flow (SDF) hardware model, and on the other hand, a model of the target platform from which resource constraints are derived. The hardware model thus obtained goes into an Optimiser procedure, which produces a hardware mapping. Using hardware and software templates, a Code Generator procedure, generates both the High Level Synthesis (HLS) input files and the software binaries that will run on the control CPU embedded in the FPGA. The HLS files go into the Xilinx (FPGA manufacturer) tools so that the configuration bitstream of the FPGA is produced.

Chapter 6

Proposed Work and Planning

The proposed work for the dissertation consists in the development of an automatic compiler of DNN description into Deep Versat / IOB-RV32 C++ code. The purpose of this work is to be able to run any state of the art CNN on the Deep Versat system with no effort on the user side, allowing the design and architectural exploration.

For the proof of concept stage, Darknet and Caffe will be the frameworks chosen to be supported by the compiler. Deep Versat can customized with the number of layers, numbers of each FU type and other options. Hence, the compiler must be able to take the configurations into account when producing computational datapaths for Deep Versat.

6.1 Flowchart

Figure 6.1 presents the flowchart of the system to be developed. The steps of the algorithm are explained in the next paragraphs.

The frameworks to be adopted for the Configuration files are Darknet [13] and Caffe [14].The former is to be adopted to support other dissertations that use Yolov3 [11] while the latter is one of the most used frameworks as seen in table 5.1.Also, unlike some other frameworks, Caffe is also based on Google's Protocol Buffers serialization library which is easier to work with. The goal after parsing is for the end result to be framework abstract, so other frameworks can be implemented into the compiler. When parsing the network, the compiler propagates the constants i.e input and output parameters of each layer.

With the network defined and Deep Versat deployed, each layer of the network can be mapped onto Versat (Layer Optimizer). Because the input size of each layer can be bigger than a Versat Memory, certain optimizations such parallelism exploration must be made to guarantee that each network layer is processed as efficiently as possible. How much performance can be extracted depends on the Deep Versat Configuration that is deployed, that is, performance per functional Unit, how many per Versat core and total number of cores.

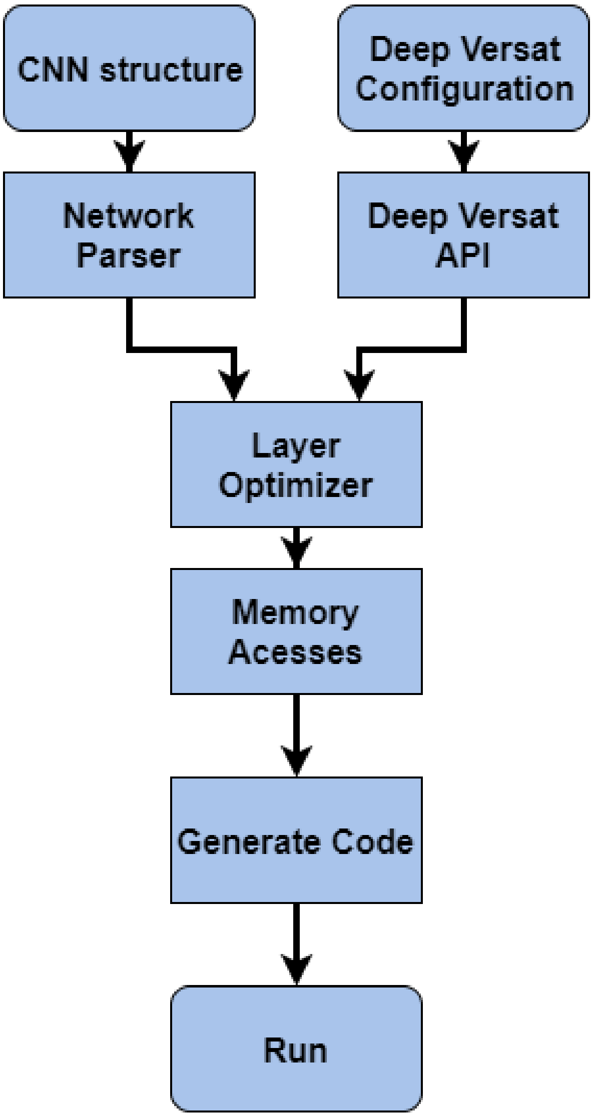


Figure 6.1: Flowchart of the software architecture

Memory accesses are all the operations to manage data from external memory to the RISC-V core and to the Deep Versat Cores and vice-versa. Finally the compiler generates the code for the CPU that controls Deep Versat.

6.2 Workplan

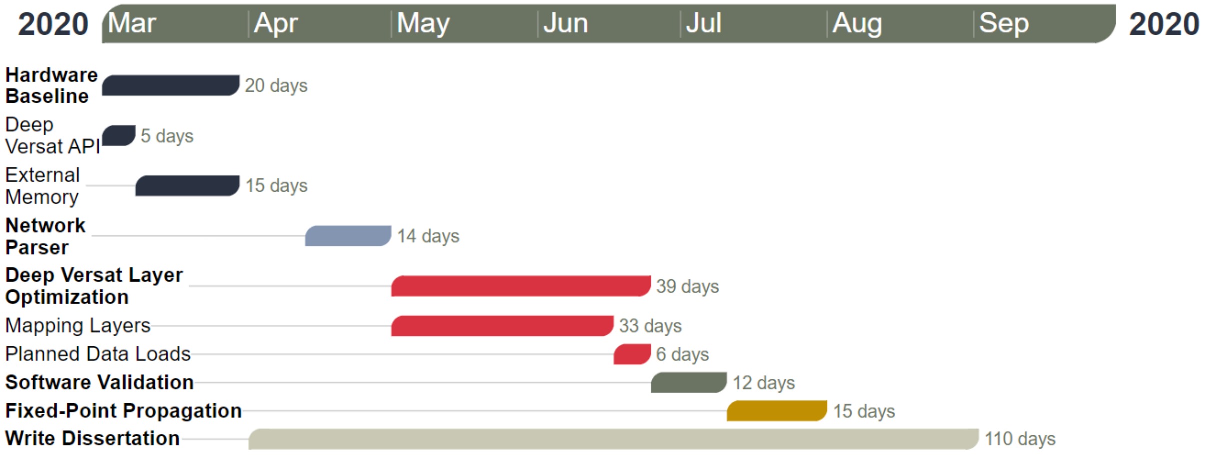


Figure 6.2: GANT chart of Proposed Work

In Fig 6.2, a GANT chart with the proposed schedule for the planned work is presented. In the GANT chart, 20 days will be used to deploy Deep Versat system and study external memory use and testing memory accesses with the RISC-V core and with Deep Versat. Then the core components in 6.1 are designed. Afterwards software validation and testing is done to ensure the compiler works as intended. Finally, because Deep Versat uses Integer functional units, fixed-point arithmetic is used instead. Fixed-Point formats are prone to overflow or underflow due to the fixed range it offers, thus 15 working days are allocated to make sure the outputs of any CNN network are valid and according to expected results.

Bibliography

[1] V. J. B. Mário. Deep versat: A deep coarse grain reconfigurable array. Master's thesis, Instituto Superior Técnico, November 2019.

[2] Picorv32- a size-optimized risc-v cpu. URL `https://github.com/cliffordwolf/picorv32`.

[3] G. Piccinini. The first computational theory of mind and brain: A close look at mcculloch and pitts's "logical calculus of ideas immanent in nervous activity". *Synthese*, 141, 08 2004. doi: 10.1023/B:SYNT.0000043018.52445.3e.

[4] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–44, 05 2015. doi: 10.1038/nature14539.

[5] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861 – 867, 1993. ISSN 0893-6080. doi: https://doi.org/10.1016/S0893-6080(05)80131-5. URL `http://www.sciencedirect.com/science/article/pii/S0893608005801315`.

[6] mnist database of hand-written digits. URL `http://yann.lecun.com/exdb/mnist/`.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. 2012.

[8] M. Tanomoto, S. Takamaeda-Yamazaki, J. Yao, and Y. Nakashima. A cgra-based approach for accelerating convolutional neural networks. pages 73–80, 09 2015. doi: 10.1109/MCSoC.2015.41.

[9] Max-pooling / pooling. URL `https://computersciencewiki.org/index.php/Max-pooling_/_Pooling`.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015.

[11] J. Redmon and A. Farhadi. Yolov3: An incremental improvement, 2018.

[12] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[13] J. Redmon. Darknet: Open source neural networks in c. `http://pjreddie.com/darknet/`, 2013–2016.

[14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *CoRR*, abs/1408.5093, 2014. URL http://arxiv.org/abs/1408.5093.

[15] R. Santiago, J. D. Lopes, and J. T. de Sousa. Compiler for the versat reconfigurable architecture. REC 2017, 2017.

[16] J. D. Lopes, R. Santiago, and J. T. de Sousa. Versat, a runtime partially reconfigurable coarse-grain reconfigurable array using a programmable controller. Jornadas Sarteco, 2016.

[17] J. D. Lopes and J. T. de Sousa. Fast fourier transform on the versat cgra. Jornadas Sarteco, 09 2017.

[18] J. D. Lopes and J. T. de Sousa. Versat, a minimal coarse-grain reconfigurable array. In D. I., C. R., B. J., and M. O., editors, *High Performance Computing for Computational Science – VECPAR 2016*, pages 174–187. Springer, 2016. doi:10.1007/978-3-319-61982-8_17.

[19] J. D. Lopes. Versat, a compile-friendly reconfigurable processor – architecture. Master's thesis, Instituto Superior Técnico, November 2017.

[20] A. Ignatov, R. Timofte, P. Szczepaniak, W. Chou, K. Wang, M. Wu, T. Hartley, and L. Van Gool. Ai benchmark: Running deep neural networks on android smartphones, 10 2018.

[21] S. I. Venieris, A. Kouris, and C.-S. Bouganis. Toolflows for mapping convolutional neural networks on fpgas: A survey and future directions, 2018.

[22] S. I. Venieris and C.-S. Bouganis. fpgaConvNet: A Framework for Mapping Convolutional Neural Networks on FPGAs. In *2016 IEEE 24th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*, pages 40–47. Institute of Electrical and Electronics Engineers (IEEE), May 2016. doi: 10.1109/FCCM.2016.22. URL http://dx.doi.org/10.1109/FCCM.2016.22.