EdgeNet: SqueezeNet like Convolution Neural Network on Embedded FPGA

Kathirgamaraja Pradeep
University of Moratuwa
deep0kathir@gmail.com

Kamalakkannan Kamalavasan
University of Moratuwa
kkvasan@live.com

Ratnasegar Natheesan
University of Moratuwa
rnatheesan@gmail.com

***Abstract*—In recent years, Convolution Neural Network (CNN) gained great success in many application specially in computer vision. Now adapting CNN inference on edge device have became the active research in embedded vision and hot topic in Edge AI. The major design hurdles for implementing CNN inference on embedded systems are limited computation resource, memory resource and power budget. In this study we are presenting a novel architecture for SqueezeNet like CNN models and this can be extended to support any CNN model as well. We address two approaches to mitigate resource constraints. First, we use custom floating point(12 bit for computation and 8bit for storing). Second is slicing the model into repetitive block called computation blocks. We have implemented SqueezeNet v1.1 for Image-Net for large scale classification which achieved around 9 FPS at 100MHz. Accuracy loss due to using custom float is measured to be less than 2%. Unlike other implementations which use FPGA boards with large amount of resources, our experiments are done in DE10 Nano, this mimics actual embedded system like environment.**

***Index Terms*—FPGA, SqueezeNet, Convolution Neural Network, Edge AI, Embedded FPGA**

I. Introduction

This document is a model and instructions for LaTeX. Please observe the conference page limits.

II. Background

A. Convolution neural Network

B. SqueezeNet v1.1

C. Related work

III. Design Statergy

Hardware resources(LUT, register, etc), on-chip memory or BRAM, bandwidth to access external memory and latency of systems are major constraints in embedded system. Typical CNN models used in computer vision task usually demand huge floating point arithmetic operations and large size of storage for parameters which cannot be stored in on-chip memory. So we cannot deploy entire model in hardware at once. At the same time it is not efficient to deploy single layer by layer, because each layer output should be stored and read back from memory and this will drastically degrade performance. We have carefully designed the architecture using two strategies, using custom floating point representation and computation blocks, such that any of the mentioned constraints will not badly affect the overall performance. We also use techniques like double buffering, prefetching and pipelining to further improve the performance.

A. Strategy I: Custom floating point number

Most of CNN inference are implemented in 32 bit float numbers. For some application 32 bit precision is too much and we can save resource and power by using low bit with representations. we use two type of floating point representation, 12 bit representation is for computation and 8 bit representation for storing input, parameters and intermediate results. 8 bit representation used 5 bit exponent and 2 bit mantissa and 12 bit uses 5 bit exponent and 6 bit mantissa. This floating point representation is chosen by an empirical study on SqueezeNet model with different format. We only sacrifice 2% of accuracy which gives top 1 accuracy of 70% on ImageNet dataset, enough for most of embedded vision applications. We also can easily adapt any floating format according to models requirement and resource availability. Using 8 bit for model parameter save 75% of memory required for model parameters. Using 12 bit operation save ???% of hardware resources compare to 32 bit operation

TABLE I
Floating point resource comparision

	12 bit		**32 bit**	
	ALM	***Register***	***ALM***	***Register***
Addition	131	109	???	???
Multiplication	43	45	???	???

1) Advantage of using custom floating point : Using 32 bit floating operation are not affordable in embedded systems. Studies [x,y,z] uses alternative representations like fixed point number, dynamic fixed point numbers(papers). When using fixed point intermediate output range may change drastically and designer have to be very careful about saturation values. In dynamic fixed point, its needed to adjust fractional part, which will consume additional resources. When using custom floating point numbers, user no need to worry much about the intermediate results saturation, because floating numbers can represent large range of numbers compare to range of fixed point representation of same size.

B. Strategy II: Computation block

Computation block is the key feature in our design. It is a collection of layers fused together as single block. It contains expand layers(3x3, 1x1 convolution) followed by maxpool, squeeze layer(1x1 convolution) and finally average pool layer.

With all these layers computation block will easily fit into any small FPGA like cyclone V. figure?? Show how squeezenet is sliced into 9 partition which can be executed in our computation block in sequential manner without reprogramming the hardware. Here is the list of main features of computation blocks

- Configurable size: size of input dimension, number of kernels can be configured on fly by setting configuration registers and computation block will operate accordingly. This make computation block can be reused for different size of input and number of kernel within provided range.
- Configurable layers: layers in side computation block can be enabled or disabled through control signal or by API. This make computation blocks to operate in different mode. For example at start of SqueezeNet model, only 3x3 convolution is enabled and at the end, only 3x3 and average pool are enabled.
- Low memory footprint: computation block won't begin with squeeze layer as in fire module of SqueezeNet. We purposely did this because, after squeeze layer, model shirks to fewer size, so slicing the model after squeeze layer and making it as entry and exit point of the computation block will reduce memory access. And the intermediate results after expand layer, which are larger in size, are processed internally without sending to memory. This cleaver decision have enormously reduced the memory bandwidth demand.
- Cascadable design: computation generate the out as in the same order of the input. This make computation block cascadable. For example in a larger FPGA, we can put two computation block one after another. This will increase the throughput by more than 2 times.
- Compact size: computation block contains few collection of layers though it is compact in size and can be accomodated small FPGA too. So computation block can be a general solution across different devices.

C. Architecture

D. Implementation

1) Functional flow: First we have to train the CNN model using TensorFlow or any framework. Its better to use SqueezeNet like models will achieve highest efficiency in our architecture. Then carefully select the floating point representation required for the trained model. Then convert trained weights and bias parameters to custom floating point values using provided APIs. This initial part can be done offshore in servers. Now converted parameters handed over to ARM processor in SoC FPGA. A host program also compiled and loaded in processor. Host program load the parameter in memory and send the address to FPGA, so it can access parameters. Then host program generates the configuration parameters required to execute particular CNN model and send it to configuration controller in FPGA. After initializing, it will start to load the input images into memory and pass the address to FPGA. In FPGA, configuration controller reads configurations and distributes to all other modules (input layer, output layer, kernel loader, computation block). Eventually, it will set all the required configuration registers in these modules. configuration controller issue a "start" signal to all blocks. Now weights and input data are read from memory and feeded to kernel controller and input layer modules. They will rearrange the order as required by computation block. Then computation block process the data according to configurations. Output from computation block send to output layer module which will write the output to memory. Again in next iteration input are read from memory and configuration block is executed with another configuration. This loop continues until it reaches the final layer. After the final iteration, configuration controller generate a interrupt to host programme denoting that inference have completed. Now host program can access the output of inference and use it to perform higher level task.

E. Some Common Mistakes

- The word "data" is plural, not singular.
- The subscript for the permeability of vacuum μ_0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter "o".
- In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an "inset", not an "insert". The word alternatively is preferred to the word "alternately" (unless you really mean something that alternates).
- Do not use the word "essentially" to mean "approximately" or "effectively".
- In your paper title, if the words "that uses" can accurately replace the word "using", capitalize the "u"; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones "affect" and "effect", "complement" and "compliment", "discreet" and "discrete", "principal" and "principle".
- Do not confuse "imply" and "infer".
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the "et" in the Latin abbreviation "et al.".
- The abbreviation "i.e." means "that is", and the abbreviation "e.g." means "for example".

An excellent style manual for science writers is [7].

F. Authors and Affiliations

The class file is designed for, but not limited to, six authors. A minimum of one author is required for all confer-

ence articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

G. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced.

H. Figures and Tables

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. **??**", even at the beginning of a sentence.

TABLE II
TABLE TYPE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a]Sample of a Table footnote.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".

ACKNOWLEDGMENT

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.