(SD2020-367) Bit-Parallel Vector Composability For Neural Acceleration (US patent rights)

Tech ID: 32070 / UC Case 2020-367-0

BACKGROUND

Conventional neural accelerators rely on isolated self-sufficient functional units that perform an atomic operation while communicating the results through an operand delivery-aggregation logic. Each single unit processes all the bits of their operands atomically and produce all the bits of the results in isolation.

Allowed US claims (will issue shortly) available:

https://patents.google.com/patent/US20230244484A1/

TECHNOLOGY DESCRIPTION

Engineers (Hadi Esmaeilzadeh and Soroush Ghodrati) from UC San Diego have designed and patented a neural accelerator that uses a new hardware implementation for performing vector dot-product operation. This innovative compute engine for vector dot-product also supports dynamic flexibility to support vector dot-product operation with flexible bit-widths. The compute engine then is integrated in a conventional architecture to accelerate deep neural networks (neural accelerator).The building block of the UCSD researchers neural accelerator is a Composable Vector Unit that is a collection of Narrower-Bitwidth Vector Engines, which are dynamically composed or decomposed at the bit granularity.

Traditionally, neural accelerators have relied on extracting Data-Level Parallelism (DLP) using isolated and self-sufficient compute units that process all the bits of operands. This innovation offers a different design style, bit-parallel vector-composability, that operates on operand bit-slices to interleave and combine the traditional data-level parallelism with bit-level parallelism. Across a range of deep models the results show that the innovative design style offers significant performance and efficiency compared to even bit-flexible accelerators.

CONTACT
Skip Cynar
scynar@ucsd.edu
tel: 858-822-2672.

OTHER INFORMATION

KEYWORDS

Hardware Acceleration, Deep Neural Networks, Artificial Intelligence, mobile devices, Robotic, convolutional neural nets, parallel architectures, recurrent neural nets, power aware computing, Nvidia's RTX 2080 TI GPU, high-bandwidth off-chip memory, LSTM deep networks, CNN, bit-level parallelism, neural acceleration, INT-4 execution, Bit-Parallel Vector Composability, Composable Vector Unit

CATEGORIZED AS

- **Computer**
  - Hardware

RELATED CASES

2020-367-0

(19) **United States**
(12) **Patent Application Publication**
Ghodrati et al.

(10) **Pub. No.: US 2023/0244484 A1**
(43) **Pub. Date: Aug. 3, 2023**

(54) **BIT-PARALLEL VECTOR COMPOSABILITY FOR NEURAL ACCELERATION**

(71) Applicant: **The Regents of the University of California**, Oakland, CA (US)

(72) Inventors: **Soroush Ghodrati**, La Jolla, CA (US); **Hadi Esmaeilzadeh**, San Diego, CA (US)

(21) Appl. No.: **18/004,802**

(22) PCT Filed: **Jul. 9, 2021**

(86) PCT No.: **PCT/US21/41167**

§ 371 (c)(1),
(2) Date: **Jan. 9, 2023**

Related U.S. Application Data

(60) Provisional application No. 63/049,982, filed on Jul. 9, 2020.

Publication Classification

(51) **Int. Cl.**
G06F 9/30 (2006.01)
G06F 9/38 (2006.01)

(52) **U.S. Cl.**
CPC ***G06F 9/30036*** (2013.01); ***G06F 9/30032*** (2013.01); ***G06F 9/3877*** (2013.01); ***G06F 9/3822*** (2013.01)

(57) **ABSTRACT**

Methods, apparatus and systems that relate to hardware accelerators of artificial neural network (ANN) performance that significantly reduce the energy and area costs associated with performing vector dot-product operations in the ANN training and inference tasks. Specifically, the methods, apparatus and systems reduce the cost of bit-level flexibility stemming from aggregation logic by amortizing related costs across vector elements and reducing complexity of the cooperating narrower bitwidth units.

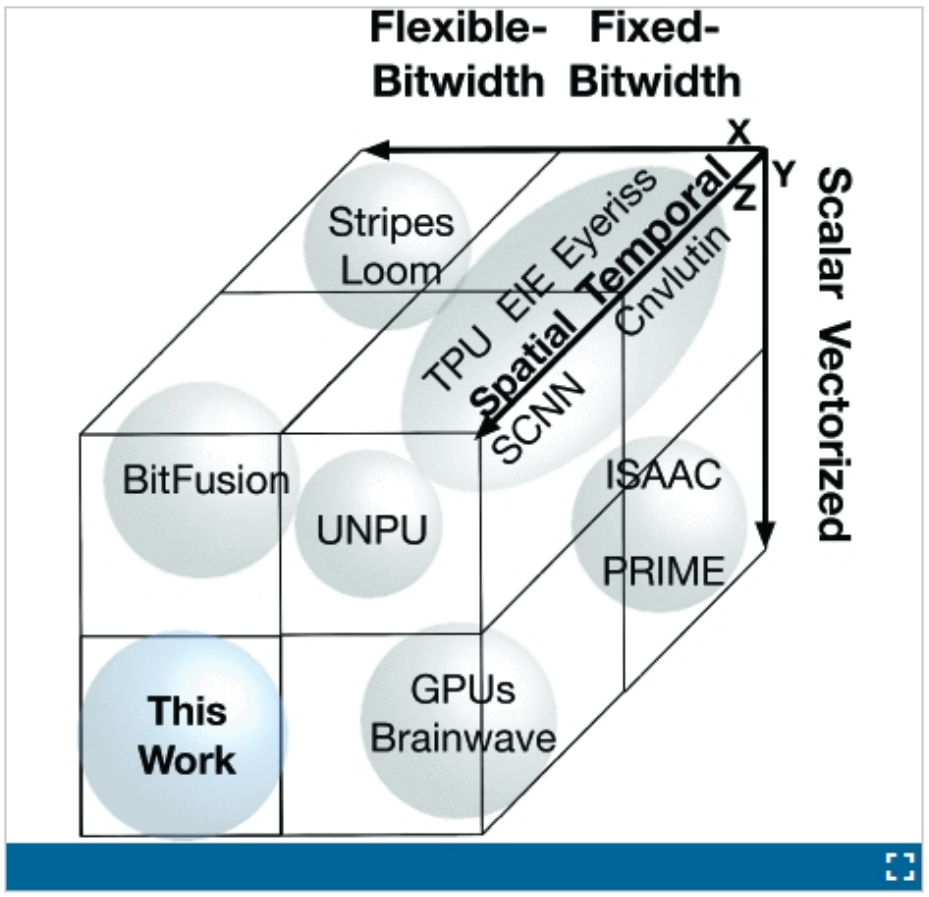

Fig. 1:
The landscape of DNN accelerators and how this work fits in the picture.

APPLICATIONS

This invention can be exploited in any neural processing unit for acceleration of neural networks. It can be also used in any other computer hardwares for application that require tensor operations such as linear algebra, digital signal processing, Artificial Intelligence, mobile devices, Hardware Acceleration, Deep Neural Networks, Robotics.

ADVANTAGES

This invention leverages an innovative design, where each unit is only responsible for a slice of the bit-level operations to interleave and combine the benefits of bit-level parallelism with the abundant data-level parallelism in deep neural networks. A dynamic collection of these

units cooperate at runtime to generate bits of the results, collectively. Such cooperation requires extracting new grouping between the bits, which is only possible if the operands and operations are vectorizable. The abundance of Data-Level Parallelism and mostly repeated execution patterns, provides a unique opportunity to define and leverage this new dimension of Bit-Parallel Vector Composability. This design intersperses bit parallelism within data-level parallelism and dynamically interweaves the two together. As such, the building block of our neural accelerator is a Composable Vector Unit that is a collection of Narrower-Bitwidth Vector Engines, which are dynamically composed or decomposed at the bit granularity.

STATE OF DEVELOPMENT

The compute engine in a systolic array architecture to evaluate the effectiveness of our methods on acceleration of deep neural networks. Using six diverse CNN and LSTM deep networks, the inventors evaluated this design style across four design points: with and without algorithmic bitwidth heterogeneity and with and without availability of a high-bandwidth off-chip memory. Across these four design points, Bit-Parallel Vector Composability brings (1.4x to 3.5x) speedup and (1.1x to 2.7x) energy reduction. They also comprehensively compare their design style to the Nvidia RTX 2080 TI GPU, which also supports INT-4 execution. The benefits range between 28.0x and 33.7x improvement in Performance-per-Watt.

INTELLECTUAL PROPERTY INFO

Please contact UCSD if you are interested in commercializing this patent-pending technology.

Published patent application available: https://patents.google.com/patent/US20230244484A1

Patent title: BIT-PARALLEL VECTOR COMPOSABILITY FOR NEURAL ACCELERATION

US 20230244484A1

(19) **United States**

(12) **Patent Application Publication**

(10) **Pub. No.: US 2023/0244484 A1**

(43) **Pub. Date: Aug. 3, 2023**

Ghodrati et al.

(54) **BIT-PARALLEL VECTOR COMPOSABILITY FOR NEURAL ACCELERATION**

(71) Applicant: **The Regents of the University of California**, Oakland, CA (US)

(72) Inventors: **Soroush Ghodrati**, La Jolla, CA (US); **Hadi Esmaeilzadeh**, San Diego, CA (US)

(21) Appl. No.: **18/004,802**

(22) PCT Filed: **Jul. 9, 2021**

(86) PCT No.: **PCT/US21/41167**

§ 371 (c)(1),
(2) Date: **Jan. 9, 2023**

Related U.S. Application Data

(60) Provisional application No. 63/049,982, filed on Jul. 9, 2020.

Publication Classification

(51) **Int. Cl.**
G06F 9/30 (2006.01)
G06F 9/38 (2006.01)

(52) **U.S. Cl.**
CPC ***G06F 9/30036*** (2013.01); ***G06F 9/30032*** (2013.01); ***G06F 9/3877*** (2013.01); ***G06F 9/3822*** (2013.01)

(57) **ABSTRACT**

Methods, apparatus and systems that relate to hardware accelerators of artificial neural network (ANN) performance that significantly reduce the energy and area costs associated with performing vector dot-product operations in the ANN training and inference tasks. Specifically, the methods, apparatus and systems reduce the cost of bit-level flexibility stemming from aggregation logic by amortizing related costs across vector elements and reducing complexity of the cooperating narrower bitwidth units.

1st patent-pending claim:

1. An apparatus for performing an energy-efficient dot-product operation between two input vectors, comprising:

a plurality of vector computation engines, wherein each vector computation engine from the plurality of vector computation engines comprises:

an array of multipliers connected through one or more add units and configured to generate an output of the vector computation engine based on a dot-product operation on a subset of bits of the two input vectors;

a plurality of shifters configured to shift the outputs of the vector computation engines; and

an aggregator coupled to the plurality of shifters and configured to generate a scalar output for the energy-efficient dot-product operation based on aggregating the shifted outputs.

2. The apparatus of claim **1**, wherein the plurality of vector computation engines is configured to operate in

RELATED MATERIALS

- S. Ghodrati, H. Sharma, C. Young, N. S. Kim and H. Esmaeilzadeh, "Bit-Parallel Vector Composability for Neural Acceleration," 2020 57th ACM/IEEE Design Automation Conference (DAC), 2020, pp. 1-6, - 07/20/2020

University of California, San Diego

Office of Innovation and Commercialization

9500 Gilman Drive, MC 0910, ,

La Jolla,CA 92093-0910

Tel: 858.534.5815

innovation@ucsd.edu

https://innovation.ucsd.edu

Fax: 858.534.7345