Bolhasani *et al. Journal of Big Data* (2023) 10:76
https://doi.org/10.1186/s40537-023-00775-8

Journal of Big Data

RESEARCH Open Access

DLA-E: a deep learning accelerator for endoscopic images classification

Hamidreza Bolhasani[1], Somayyeh Jafarali Jassbi[1*] and Arash Sharifi[1]

*Correspondence: s.jassbi@srbiau.ac.ir

[1] Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran

Abstract

The super power of deep learning in image classification problems have become very popular and applicable in many areas like medical sciences. Some of the medical applications are real-time and may be implemented in embedded devices. In these cases, achieving the highest level of accuracy is not the only concern. Computation runtime and power consumption are also considered as the most important performance indicators. These parameters are mainly evaluated in hardware design phase. In this research, an energy efficient deep learning accelerator for endoscopic images classification (DLA-E) is proposed. This accelerator can be implemented in the future endoscopic imaging equipments for helping medical specialists during endoscopy or colonoscopy in order of making faster and more accurate decisions. The proposed DLA-E consists of 256 processing elements with 1000 bps network on chip bandwidth. Based on the simulation results of this research, the best dataflow for this accelerator based on MobileNet v2 is kcp_ws from the weight stationary (WS) family. Total energy consumption and total runtime of this accelerator on the investigated dataset is 4.56×10^9 MAC (multiplier–accumulator) energy and 1.73×10^7 cycles respectively, which is the best result in comparison to other combinations of CNNs and dataflows.

Keywords: Deep learning accelerator, Dataflow, Deep neural networks, Convolutional neural networks, Medical Images, Endoscopy

Introduction

During the recent years, applications of deep learning have been grown very fast. A wide range of areas from agriculture to autonomous vehicles, aerospace industry and medicine are benefiting from the super power of deep neural networks for handling their tasks with higher speed and accuracy than humans [1, 2]. Computer vision and image recognition are considered as two of the most popular applications. Especially in medical and healthcare systems, there are a great quantity of use cases for image analysis such as generating radiology and histopathology reports [3–5]. Four main types of medical images that are now broadly used for training deep neural networks in order of diagnosis and classification are shown in Fig. 1. But there are also some other types of medical images that is not depicted in this figure like X-ray, angiographic images and Electroencephalography (EEG) signals [6].

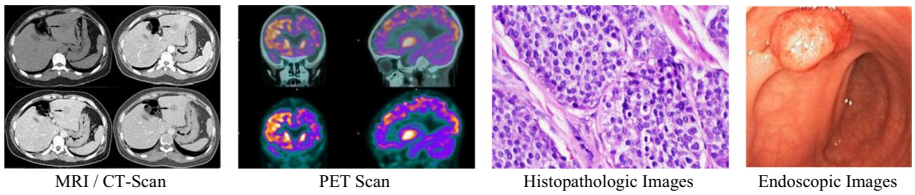

Fig. 1 Four main types of medical images

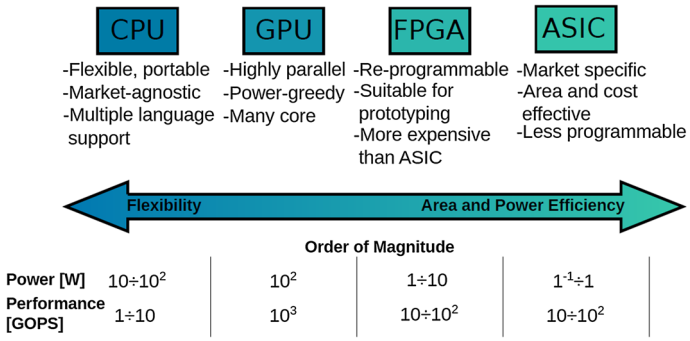

Fig. 2 Comparison of capabilities, performance and efficiency of well-known processors [11]

Convolutional Neural Networks (CNNs) or in some literatures, deep convolutional neural networks (DCNNs) have shown a significant success in medical images classification that is applicable for disease prediction, detection, diagnosis, grading and prognosis [7–9]. Some recent investigations have been proved that the accuracy and performance of these type of neural networks in several real time medical applications like endoscopy and colonoscopy are promising [10].

CNN models may require billions of computations, chiefly multiplication and accumulation (MAC) in both training and inference phase [11]. Undoubtedly, in case of achieving the desired results in a short time, these operations needs resources that have high level of energy consumption. As a result, the demand for cost-effective and energy efficient hardware for these specific purposes have been increased dramatically. In the recent decade, Graphic Processing Units (GPUs) have become very prevalent for speeding up the computations in variant use cases like image processing, cryptography, data mining, medical physics, biostatistics and many other applications. The key success factor of GPU was utilizing high level of parallelism for data and computing tasks [12, 13].

Desirable performance of GPUs led to emersion of some new efficient and powerful application-specific integrated circuits (ASIC) like Tensor Processing Unit (TPU). In 2015, a team of researchers and experts in Google, deployed TPU for accelerating neural networks operations in the inference phase in their data centers [14]. This accelerator consists of 65,536 MAC (8 bit) matrix which make it capable of maximum 92 Tera Operations Per Second (TOPS) throughput. Meanwhile, it is 15 to 30 times faster than its equivalent CPU or GPU and up to 200 times less energy consuming. A summary of capabilities, performance, efficiency, pros and cons of these processors have been depicted in Fig. 2. As it can be seen in this figure, ASICs are market

specific, area and cost effective but less programmable. Thus, these types of processors can be the best candidate for specific applications, particularly in the cases that major performance indicators like power consumption and area are very substantial.

Deep learning accelerators (DLAs) are considered as hardware that accelerate deep learning operations while taking care of energy consumption. In other words, efficient data processing methods should be applied for these accelerators. Specifically, in some use cases of deep learning, like real-time applications, embedded systems and mobile devices, the need for applying a proper DLA is inevitable.

In this research, a deep learning accelerator for endoscopic images classification is proposed that is called DLA-E. This ASIC accelerator is a Network on Chip (NoC) architecture made up of 240 processing elements (PEs) and 192 bps bandwidth. The proposed architecture for the selected CNN model (MobileNet v2) had the best performance result based on Row Stationary (RS) dataflow in comparison to other well-known dataflows.

The organization of this paper is offered in this order: "Related works" section covers a brief review on the related works. In "Fundamental concepts" section, some fundamental concepts in the context of this research are described. The proposed architecture of this research, DLA-E is discussed in "DLA-E (The proposed architecture)" section. And at the end, "Conclusion" section includes conclusion and open problems for the future.

Related works

Jin Hee Kim, et al. [15] introduced an FPGA accelerator for CNN inference. This accelerator is synthesized in C software and its mechanism is chiefly based on parallelism. The designed software translates the parallel threads to spatial hardware architecture. The proposed accelerator, have utilized some major computer architecture techniques like zero-weight skipping and reduced precision and could achieve up to 138 GOPS (Giga Operations Per Seconds) performance on VGG16.

Yongming Shen et al. [16] offered a CNN accelerator with flexible buffering to minimize off-chip transfer, named Escher. The authors of this research have focused on making balance between input and weight data transfer bandwidth. The bandwidth requirement of the studied accelerator has been reduced up to 1.7, 2.4 and 10.5 times for fully connected and convolutional layers in VGGNet-E, AlexNet and GoogLeNet respectively. Management and control of image batch size and buffering data on chip is also considered as a contribution in this architecture.

Yongming Shen, et al. [17] presented a new FPGA-type CNN accelerator that its main idea is principally based on resource partitioning for the different layers of the neural networks. Performance analysis showed that the proposed accelerator is significantly more efficient in comparison to its contemporary processor Xilinx Virtex-7 FPGA. The analysis results showed that the achieved throughput is 3.8 times higher for AlexNet and around 2 times faster for SqueezeNet and GoogLeNet.

Ximing Qiao, et al. [18] suggested AtomLayer which its emphasis is on computation improvement in atomic layer. This CNN accelerator is based on resistive random-access memory (ReRAM) to handle CNN operations both in training and inference phases. The mentioned architecture processes just one network layer per each time slot in order of preventing pipeline delays. This accelerator also benefits from some famous techniques

like data reuse and filter mapping. AtomLayer has 1.1, and 1.6-times higher power efficiency than the ISAAC [19] and PipeLayer [20] accelerators in inference and training phases respectively.

Lin Bai, et al. [21] revealed a high-performance depth wise CNN accelerator based on FPGA which is convolution separable. The authors of this research have compared the speed of this accelerator on an Arria 10 SoC FPGA with a CPU and found that it was 20 times faster for the MobileNet v2 network. The image classification time for a picture over the introduced architecture is 3.75 ms with a rate equal to 266.6 frames per second (fps). The low power consumption and high speed make this accelerator a satisfactory choice for the portable or mobile devices.

Chao-Tsung Huang et al. [22] introduced a block-based CNN accelerator with high parallelism capability named eCNN. Inference operation in CNNs consists of an enormous volume of computations and this is very challenging in the edge, since generally there are embedded, portable and mobile devices. Because in these devices, the problem is not just the speed, but the most important issue is area and power consumption. Inference in some applications like video processing involves massive data movement from off-chip memory and it has lots of costs. In order of reducing these costs, a hardware-oriented network called ERNet and also FBISA, a customized instruction set architecture (ISA) are proposed in this architecture.

Shiguang Zhang et al. [23] presented a reconfigurable FPGA based CNN accelerator for YOLO network. This accelerator is in the family of advanced RISC machine (ARM) architectures. The preferred configuration signals are passed through ARM and inference computation for the different layers of the CNN are handled accordingly over the time. In the suggested accelerator, operation of convolution and pooling layers are merged and consequently, the data movement from off-chip memory and in overall, the costs are reduced. This architecture has been implemented on the Xilinx ZCU102 FPGA for the inference of YOLO v2 and the achieved performance in peak was 289 GOPS.

In Table. 1, a summary of the discussed related works and some other researches are presented based on the reference, year and the main idea.

Fundamental concepts

There are five popular types of layers in deep neural networks: Convolutional layers (CONV), Fully connected layers (FC), Pooling, Nonlinearity and Normalization layers [28]. Multiply and accumulation are dominant computing operations in CONV and FC layers. CNNs are a type of deep neural networks with multiple CONV layers and mainly used for computer vision and image processing. As shown in Fig. 3, CONV layers include a 7-Dimension computing problem: $R \times S \times X \times Y \times C \times K \times N$. The position and role of each of these dimensions are depicted in this figure. R and S are respectively height and length of the filter (weight) sliding window. Length and height of the input data are represented by X and Y. C is the number of filter / input feature maps channels. K and N are number of channels of filter (weight) maps and output feature maps. Various arrangement and orchestration of these data on the processing elements will result in emergence of new different dataflows that each of them is efficient for one or some specific applications. Therefore, the need for customized dataflows is increasing drastically.

Table 1 A summary of the related works plus some other researches and their contribution

Refs.	Year	Architecture name	Main idea/Contribution
[24]	2016	ConvAU	A systolic array architecture based on Google's TPU for accelerating massive matrix multiplication and accumulations
[15]	2017	-	An FPGA accelerator for CNN inference utilizing parallelism
[16]	2017	Escher	+ CNN accelerator with flexible buffering to minimize off-chip transfer + Making balance between weight and input data transfer
[17]	2017	-	Resource partitioning for different layers of CNN
[25]	2018	CoNNA	+ A reconfigurable coarse-grained FPGA CNN accelerator + Accelerating all layer types
[18]	2018	AtomLayer	+ CNN accelerator equipped with ReRAM + Handle CNN operations both in training and inference phases + Computation improvement in atomic layer
[21]	2018	-	+ High-performance depth wise CNN accelerator based on FPGA + Low power consumption and high speed for mobile devices
[22]	2019	eCNN	+ CNN accelerator for edge inference + Block based inference flow for removing all DRAM bandwidth related to feature maps reading
[23]	2020	-	A reconfigurable FPGA-based CNN accelerator for YOLO network
[26]	2021	SWM	+ A CNN accelerator for inference phase + High performance accelerator using a dynamic scheduling strategy + Focusing on sparsity for reducing both memory and computation usage
[27]	2022	MVP	A systolic array architecture for efficient processing of memory and computation intensive operations in CNNs like EicientNet and MobileNet

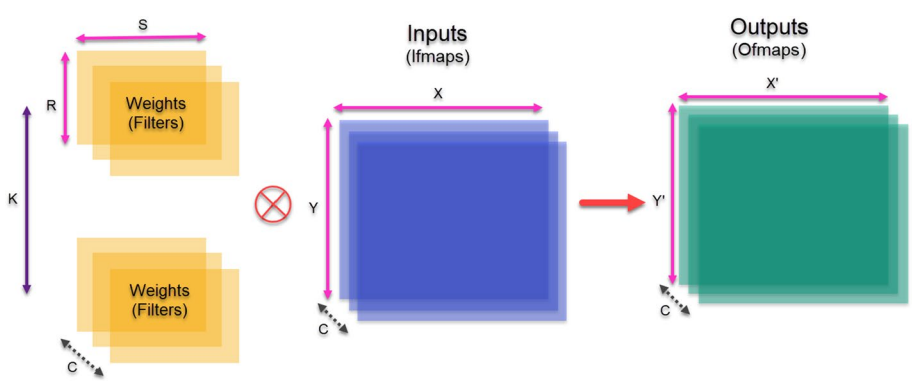

Fig. 3 Conceptual diagram of 7-Dimension computing problem: R × S × X × Y × C × K × N in a CNN

Yu-Hsin Chen, et al. [29] introduced Row Stationary (RS) dataflows for their proposed architecture, Eyeriss which is an energy-efficient reconfigurable accelerator for deep convolutional neural networks. In this dataflow, data reuse for filters and ifmpas have been utilized and as a consequence, data movement is minimized which is ideal for the assigned tasks related to CNN operations. In this research, it has been proved that RS had the highest efficiency and lowest power consumption in comparison to the other investigate dataflows like Input Stationary (IS), Output Stationary (OS), Weight Stationary (WS) and No Local Reuse (NLR).

DLA-E (The proposed architecture)

Object detection and image classification are two of the most spreading use cases of CNNs. Currently image classification using deep learning algorithms is becoming an inseparable part of various fields of study like medical science. This is not just limited to the theoretical part, and nowadays there are plenty of applications in the real world. Almost all of these applications are computing intensive or in the other word, computation hungry. Therefore, the need for energy-efficient hardware accelerators is inevitable and now, a large number of efforts have been done or in the progress.

In some medical operations like endoscopy or colonoscopy, an accurate and fast image classification during the operation can be very helpful for the medical specialists in order of making better decisions and more precise diagnosis [10]. Hence, several valuable datasets for this purpose have been prepared by the researchers from different countries and universities [31–33]. To the best of our knowledge, Kvasir [30], is the first and at the same time the most comprehensive dataset for this purpose.

This dataset contains 4,000 images from Gastrointestinal (GI) tract taken from real endoscopy, labeled by medical specialists into 8 categories. Amongst these 8 classes, three of them are labeled anatomically that are: (1) Pylorus, (2) Cecum and (3) Z-line and three of them are classified based on pathological findings: (4) Esophagitis, (5) Polyps and (6) Ulcerative Colitis. There are also two class dedicated for removal of polyps: (7) Dyed and Lifted Polyps and (8) Dyed Resection Margins. Each of these categories consists of 500 images. A sample image from each of these classes are presented in Fig. 4. The resolution of images are in the range of 720×576 up to 1920×1072.

This is a multi-class image classification problem. After some preprocessing operations on the images in the dataset, various state of the art deep neural networks architectures like ResNet, Inception, VGG19, Inception_ResNet and MobileNet has been investigated. The summary of obtained results including DNN model precision per each class is reported in Table. 2.

As it can be seen in Table. 2, Inception_ResNet V2 and MobileNet V3 achieved the highest average accuracy for this specific image classification problem amongst other neural network architectures. This is a satisfactory accuracy, but from the point of

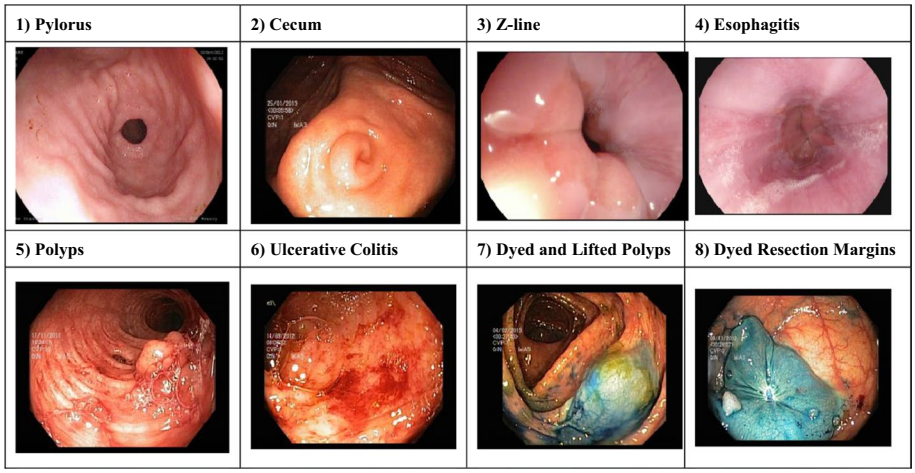

Fig. 4 Image sample of each class in Kvasir dataset

Table 2 A summary of obtained results related to endoscopic image dataset classification

No	Class / DNN Model Precision	Inception V3	ResNet V2	Inception_ ResNet V2	VGG19	MobileNet V3
1	Pylorus	0.91	0.92	0.93	0.91	0.95
2	Cecum	0.85	0.85	0.86	0.85	0.86
3	Z-line	0.87	0.84	0.93	0.96	0.95
4	Esophagitis	0.93	0.93	0.96	0.96	0.96
5	Polyps	0.97	0.99	0.96	0.88	0.94
6	Ulcerative Colitis	0.77	0.78	0.82	0.83	0.77
7	Dyed and Lifted Polyps	0.94	0.99	0.96	0.91	0.93
8	Dyed Resection Margins	0.97	0.96	0.98	0.96	0.97
Average Accuracy		0.90	0.91	**0.93**	0.91	**0.92**

Bold values indicate better results (Avarage Accuracy) in comparison to other results

Table 3 A summary of DLA-E architecture parameters

No	Architecture Parameter (Configured in MAESTRO)	Description	Value
1	num_pes	Number of Processing Elements (PEs)	256
2	l1_size_cstr	L1 Cache Size Constraint (KB)	100
3	l2_size_cstr	L2 Cache Size Constraint (KB)	3000
4	noc_bw_cstr	Network on Chip (NoC) Bandwidth Constraint (bps)	1000
5	offchip_bw_cstr	Off-Chip Bandwidth Constraint (GB/s)	50
6	dataflow	Specific Dataflow for the architecture	kcp_ws

computer architecture and design, it's not enough. As discussed earlier, these deep neural networks consist of huge volume of computations which are power consuming and as a result, considering some key performance indicators like total throughput (MACs / Cycle), total energy consumption (X MAC Energy) and total runtime (Cycles) is really important. Thus, there is a trade-off between accuracy and energy efficiency and the best choice is a combination of them based on the problem or assigned tasks.

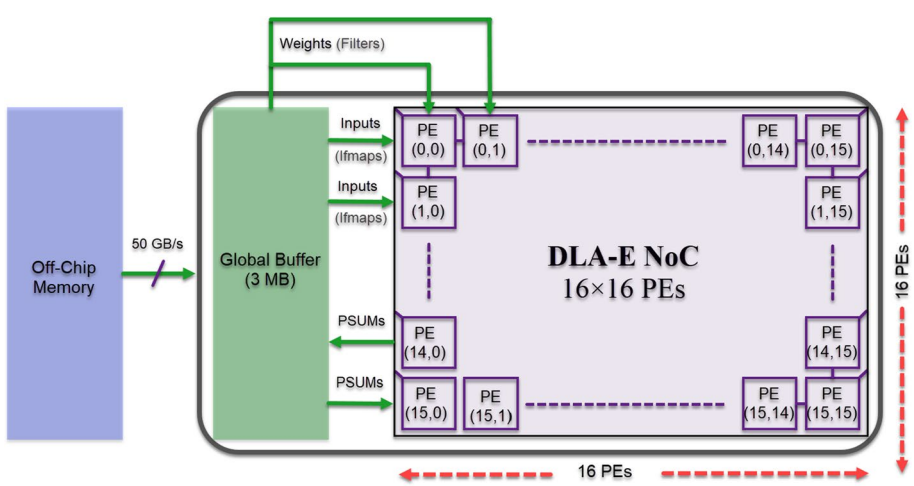

Fig. 5 Schematic design of DLA-E accelerator

The proposed architecture of this research is named DLA-E, which stands for a Deep Learning Accelerator for Endoscopic Image Classification. The main tools that were used for designing this accelerator was MAESTRO [34] that is an open-source infrastructure for modeling dataflows within deep learning accelerators. DLA-E has 256 processing elements (PEs) with 1000 bps NoC bandwidth, up to 100 KB L1 cache, 3 MB L2 cache and 50 GB/s off-chip memory bandwidth. A summary of this architecture specification and its schematic design is offered in Table. 3 and Fig. 5 respectively.

The most significant item that makes this accelerator efficient for the assigned tasks is its dataflow. There are lots of well-known dataflows that each of them may be optimum for some specific problems or tasks. Based on the simulation results with MAESTRO simulator that are presented in Table. 4, kcp_ws dataflow from the family of Weight Stationary (WS) for MobileNet V2 shows the highest efficiency amongst other combinations. Total Energy Consumption and Total Runtime for this arrangement are

Table 4 DLA-E Simulation results with MAESTRO on various DNN models and Dataflows

DNN	**MobileNet V2**			**ResNet**			**VGG19**		
Dataflow	**kcp_ws**	**xp_ws**	**rs**	**kcp_ws**	**xp_ws**	**rs**	**kcp_ws**	**xp_ws**	**rs**
Number of MACs	2.98×10^8	2.98×10^8	2.98×10^8	1.01×10^{10}	1.01×10^{10}	1.01×10^{10}	1.77×10^{10}	1.77×10^{10}	1.77×10^{10}
Avg L1 Size Requirement	7	7	22	8	8	37	15	15	59
Max L1 Size Requirement	18	18	96	98	98	224	18	18	96
Avg L2 Size Requirement	977	887	1,087	765	79	289	1,577	367	705
Max L2 Size Requirement	4,608	4,608	5,408	4,608	10,766	3,528	4,608	3,996	2,720
Avg Number of Utilized PEs	208	76	77	254	19	23	243	54	99
Max Number of Utilized PEs	256	240	240	256	110	191	256	222	222
Avg NoC Bandwidth	49.23	44.91	43.16	30.01	13.66	13.91	40.48	6.66	10.91
Max NoC Bandwidth	158.40	143.97	143.97	159	56	52.71	159.92	25	24.16
Avg Throughput (MACs / Cycle)	33.92	22.38	20.38	41.75	18.32	28.71	39.21	30.27	96.97
Max Throughput (MACs / Cycle)	48.00	56.00	146.67	42.67	56.00	197.73	42.67	54	214.47
Total Throughput (MACs / Cycle)	3,596.52	2,373	2,418	13,026	5,717	8,959	1,490	1,150	3,685
Total Energy Consumption (X MAC Energy)	4.56×10^9	1.31×10^{10}	8.40×10^9	1.21×10^{11}	3.18×10^{11}	2.10×10^{11}	1.75×10^{11}	2.26×10^{11}	1.93×10^{11}
Total Runtime (Cycles)	1.73×10^7	4.48×10^7	4.75×10^7	5.00×10^8	1.41×10^9	1.01×10^9	8.85×10^8	1.30×10^9	6.12×10^8

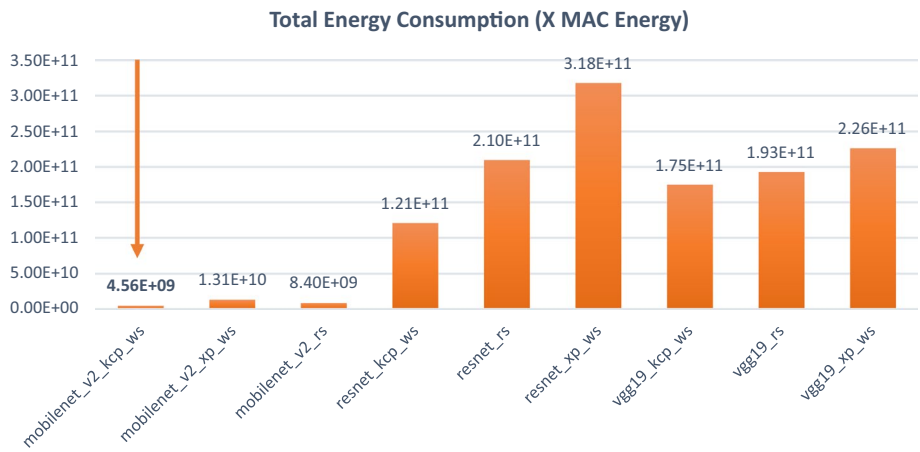

Fig. 6 Total Energy Consumption (X MAC Energy) in DLA-E for various DNN + Dataflow arrangements

respectively 4.56 × 109 X MAC Energy and 1.73 × 107 Cycles that are lowest values in comparison to the achieved values of other DNNs and dataflows combination. For more clarification, these two comparisons are represented in Figs. 6 and 7 sequentially.

Conclusion

In the recent years, applications of deep neural networks, especially convolutional neural networks in computer vision use cases like object detection and image classification is growing tremendously. Healthcare experts and medical science specialists are also benefiting from these applications in a very large scale. Nowadays, numerous deep learning algorithms are used for medical image analysis in order of detection, diagnosis, grading, treatment and prognosis. There are multiple medical images for these purposes like MRI, CT scan, EEG, Histopathologic and Endoscopic images. The state-of-the-art DNN models that are being used for these tasks mostly consists of thousands of layers

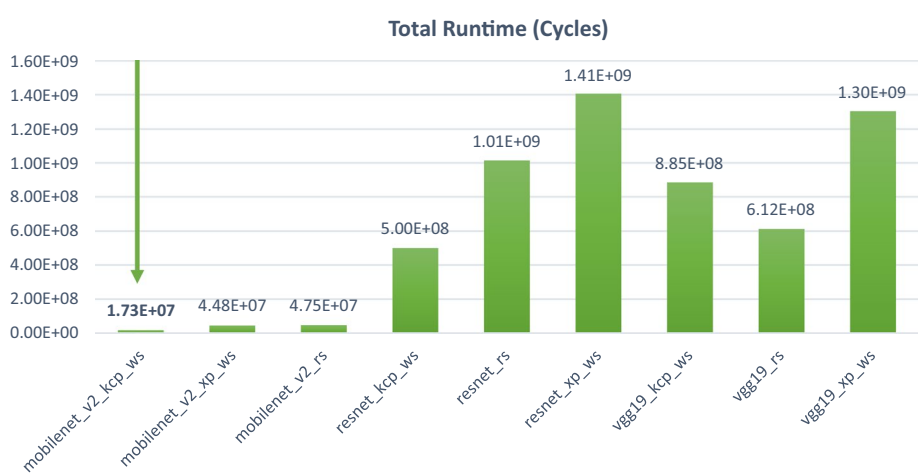

Fig. 7 Total Runtime (Cycles) of DLA-E for various DNN + Dataflow arrangements

including convolutional layers that are extremely power consuming. Thus, the need for energy-efficient deep learning accelerators is vital. In this research, an energy-efficient deep learning accelerator for endoscopic image classification, DLA-E is proposed. This accelerator is made up of 256 PEs, up to 100 KB L1 cache, 3 MB L2 cache, 1000 bps NoC bandwith and 50 GB/s off-chip bandwidth. The performance of this accelerator has been evaluated by MAESTRO simulator over various arrangements of DNN models and dataflows. Simulation results showed that kcp_ws dataflow from the family of weight stationary data flows on MobileNet V2 with 4.56×109 X MAC total energy consumption and 1.73×107 cycles total runtime (lowest values in comparison to the other ones) was the best arrangement for DLA-E.

Abbreviations

DLA-E	Deep Learning Accelerator for Endoscopic Image Classification
WS	Weigh Stationary
RS	Row Stationary
IS	Input Stationary
OS	Output Stationary
NLR	No Local Reuse
MAC	Multiplier = Accumulator
EEG	Electroencephalography
CNN	Convolutional Neural Network
DCNN	Deep Convolutional Neural Network
GPU	Graphic Processing Unit
TPU	Tensor Processing Unit
ASIC	Application-Specific Integrated Circuit
GOPS	Giga Operations Per Second
TOPS	Tera Operations Per Second
DLA	Deep Learning Accelerator
NoC	Network on Chip
PE	Processing Element
CONV	Convolutional Layer
FC	Fully Connected
FPS	Frames Per Second
RISC	Reduced Instruction Set Computing
ARM	Advanced RISC Machine
ISA	Instruction Set Architecture
GI	Gastrointestinal

Acknowledgements
Not Applicable.

Author contributions
Fine-Tunning a Deep Neural Network for Endoscopic Images (Gastrointestinal and Colorectal) Classification with high accuracy (93%). + Investigating stated of the art deep learning accelerators and summarizing them. + Evaluating various combinations of DNN/Dataflow in order of achieving the lowest power consumption and computation runtime. + Proposing DLA-E architecture as a Deep Learning Accelerator for Endoscopic Images Classification. + Finding the optimum arrangement of architecture/dataflow for DLA-E as deep learning accelerator for endoscopic images classification. All authors read and approved the final manuscript.

Funding
Not Applicable.

Availability of data and materials
Available.

Declarations

Ethics approval and consent to participate
Not Applicable.

Consent for publication
Not Applicable.

Competing interests

A new application-specific integrated circuit design in order of accelerating deep learning computations with low power consumption and computing runtime. All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version. This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue. The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript. The following authors have affiliations with organizations with direct or indirect financial interest in the subject matter discussed in the manuscript.

Received: 21 July 2022 Accepted: 17 May 2023
Published online: 25 May 2023

References

1. Das, A.; Rad, P. Opportunities and challenges in explainable artificial intelligence (xai): A survey. arXiv:2006.11371, arXiv 2020.
2. Lee SM, Seo JB, Yun J, Cho Y-H, Vogel-Claussen J, Schiebler ML, Gefter WB, Van Beek EJ, Goo JM, Lee KS, et al. Deep learning applications in chest radiography and computed tomography. J Thoracic Imaging. 2019. https://doi.org/10.1097/RTI.0000000000000387.
3. Monshi MMA, Poon J, Chung V. Deep learning in generating radiology reports: a survey. Artif Intell Med. 2020;106:101878.
4. Van der Laak J, Litjens G, Ciompi F. Deep learning in histopathology: the path to the clinic. Nat Med. 2021;27(5):775–84.
5. Wood DA, et al. Deep learning to automate the labelling of head MRI datasets for computer vision applications. Eur Radiol. 2021. https://doi.org/10.1007/s00330-021-08132-0.
6. Safayari A, Bolhasani H. Depression diagnosis by deep learning using EEG signals: a systematic review". Med Novel Technol Devices. 2021. https://doi.org/10.1016/j.medntd.2021.100102.
7. Yadav SS, Jadhav SM. Deep convolutional neural network based medical image classification for disease diagnosis. J Big Data. 2019;6(1):113.
8. Yari Y, Nguyen TV, Nguyen HT. Deep learning applied for histological diagnosis of breast cancer. IEEE Access. 2020;8:162432–48.
9. Lao J, Chen Y, Li ZC, et al. A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. Sci Rep. 2017;7(1):10353.
10. Debesh Jha, Sharib Ali, Håvard D. Johansen, Dag Johansen, Jens Rittscher, Michael A. Riegler, and Pål Halvorsen. 2020. Real-Time Polyp Detection, Localisation and Segmentation in Colonoscopy Using Deep Learning. arXiv preprint, arXiv:2006.11392 (2020).
11. Capra M, et al. Hardware and software optimizations for accelerating deep neural networks: survey of current trends, challenges, and the road ahead". IEEE Access. 2020. https://doi.org/10.1109/ACCESS.2020.3039858.
12. Cano A. A survey on graphic processing unit computing for large-scale data mining. Wiley Interdiscip. Rev Data Min Knowl Discov. 2018;8(1):1232.
13. Naz N, Malik AH, Khurshid AB, Aziz F, Alouffi B, Uddin MI, AlGhamdi A. Efficient processing of image processing applications on CPU/GPU. Math Probl Eng. 2020;2020:1–14.
14. Jouppi, N. P., Young, C., Patil, N. et al. In-datacenter performance analysis of a tensor processing unit. In Proceedings of the 44th Annual International Symposium on Computer Architecture, ISCA '17, 1–12 (ACM, 2017).
15. J. H. Kim, B. Grady, R. Lian, J. Brothers, and J. H. Anderson, "FPGA-Based CNN Inference Accelerator Synthesized from Multi-Threaded C Software," IEEE SOCC, 2017.
16. Y. Shen,M. Ferdman, and P.Milder, "Escher: A cnn accelerator with flexible buffering to minimize off-chip transfer," in FCCM, 2017.
17. Y. Shen, M. Ferdman, and P. Milder, "Maximizing CNN accelerator efficiency through resource partitioning," in 44th International Symposium on Computer Architecture (ISCA), 2017.
18. X. Qiao et al., "Atomlayer: a universal reram-based cnn accelerator with atomic layer computation," in DAC, 2018.
19. Ali Shafiee et al. ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars. In ISCA, 2016.
20. Linghao Song et al. PipeLayer: A pipelined ReRAM-based accelerator for deep learning. In HPCA, 2017.
21. Bai L, Zhao Y, Huang X. A CNN accelerator on FPGA using depthwise separable convolution. IEEE Trans Circuit Syst II Express Br. 2018. https://doi.org/10.1109/TCSII.2018.2865896.
22. C.-T. Huang et al., "ecnn: A block-based and highly-parallel cnn accelerator for edge inference," in MICRO, 2019.
23. Zhang, S.; Cao, J.; Zhang, Q.; Zhang, Q.; Zhang, Y.;Wang, Y. An fpga-based reconfigurable cnn accelerator for yolo. In Proceedings of the 2020 IEEE 3rd International Conference on Electronics Technology (ICET), Chengdu, China, 8–11 May; pp. 74–78.
24. K. Kiningham, M. Graczyk and A. Ramkumar, "Design and Analysis of a Hardware CNN Accelerator," Small, vol. 27, no. 6, Jun. 2016.
25. R. Struharik, B. Vukobratovi´c, A. Erdeljan, and D. Rakanovi´c, "Conna – compressed cnn hardware accelerator," in 2018 21st Euromicro Conference on Digital System Design (DSD). IEEE, 2018, pp. 365–372.
26. Wu Di, et al. SWM: A high-performance sparse-Winograd matrix multiplication CNN accelerator. IEEE Trans VLSI Syst. 2021;29(5):936–49.
27. Lee S, et al. MVP: An Efficient CNN Accelerator with Matrix, Vector, and Processing-Near-Memory Units.". ACM Trans Design Autom Electron Syst. 2022;27(5):1–25.
28. Sze V, et al. Efficient processing of deep neural networks. Synth Lect Comput Archit. 2020;15(2):1–341.

29. Chen Y-H, et al. Eyeriss: an energy-efficient reconfigurable accelerator for deep convolutional neural networks. IEEE J Solid-State Circuits. 2016;52(1):127–38.
30. Pogorelov, Konstantin, et al. "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection." Proceedings of the 8th ACM on Multimedia Systems Conference. 2017.
31. Jha, Debesh, et al. "Kvasir-seg: A segmented polyp dataset." International Conference on Multimedia Modeling. Springer, Cham, 2020.
32. Smedsrud PH, et al. Kvasir-Capsule, a video capsule endoscopy dataset. Sci Data. 2021;8(1):1–10.
33. Jha, Debesh, et al. "Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy." International Conference on Multimedia Modeling. Springer, Cham, 2021.
34. H. Kwon, M. Pellauer, T. Krishna, Maestro: an open-source infrastructure for modeling dataflows within deep learning accelerators (2018). arXiv preprint arXiv:1805.02566.

Publisher's Note