 doi:10.20944/preprints202107.0375.v1

Review

Low-power Ultra-small Edge AI Accelerators for Image Recognition with Convolution Neural Networks: Analysis and Future Directions

Weison Lin [1,*], Adewale Adetomi [1] and Tughrul Arslan [1]

[1] Institute for Integrated Micro and Nano Systems, University of Edinburgh, Edinburgh EH9 3FF, UK; adewale.adetomi@ed.ac.uk; tughrul.arslan@ed.ac.uk
* Correspondence: Weison.Lin@ed.ac.uk

Abstract: Edge AI accelerators have been emerging as a solution for near customers' applications in areas such as unmanned aerial vehicles (UAVs), image recognition sensors, wearable devices, robotics, and remote sensing satellites. These applications not only require meeting performance targets but also meeting strict reliability and resilience constraints due to operations in harsh and hostile environments. Numerous research articles have been proposed, but not all of these include full specifications. Most of these tend to compare their architecture with other existing CPUs, GPUs, or other reference research. This implies that the performance results of the articles are not comprehensive. Thus, this work lists the three key features in the specifications such as computation ability, power consumption, and the area size of prior art edge AI accelerators and the CGRA accelerators during the past few years to define and evaluate the low power ultra-small edge AI accelerators. We introduce the actual evaluation results showing the trend in edge AI accelerator design about key performance metrics to guide designers on the actual performance of existing edge AI accelerators' capability and provide future design directions and trends for other applications with challenging constraints.

Keywords: edge AI accelerator; CGRA; CNN

1. Introduction

Convolution neural network (CNN), which has been applied to image recognition, is a kind of machine learning algorithm. CNN is usually adopted by software programs that are supported by the Artificial intelligence (AI) framework, such as TensorFlow and Caffe. These programs are usually run by central processing units (CPUs) or graphics processing units (GPUs) to form the AI systems which construct the image recognition models. The models which are trained by massive data such as big data and infer the result by the given data have been commonly seen running on cloud-based systems.

Hardware platforms for running AI technology can be sorted into the following hierarchies: data center bound system, edge-cloud coordination system, and 'edge' AI devices. The three hierarchies of hardware platforms from the data center to edge devices require different hardware resources and are exploited by various applications according to their demands. The state-of-the-art applications for image recognition such as unmanned aerial vehicles (UAVs), image recognition sensors, wearable devices, robotics, remote sensing satellites belong to the third hierarchy and are called edge devices. Edge devices refer to the devices connecting to the internet but near the consumers or at the edge of the whole Internet of things (IoT) system. This indicates the size of the edge devices is limited. They are also called edge AI devices when they utilize AI algorithms. The targeted AI algorithm of the accelerators in this paper is CNN.

 doi:10.20944/preprints202107.0375.v1

The most important feature of these edge AI devices is the real-time computing ability for predicting or inferring the next decision by pre-trained data. For practicing CNN algorithms, CPUs and GPUs have been used a lot in the first two hierarchies of AI hardware platforms. Some edge AI systems are not power-sensitive such as surveillance systems for face recognition and unmanned shops. Although these kinds of applications do not care about power consumption, they tend to aware of data privacy more. As a result, they also avoid using the first and second hierarchy platforms. However, the scope of these power non-sensitive edge AI systems is not a target in this paper. This paper focuses on surveying power-sensitive edge AI devices based on batteries or limited power resources such as systems using solar panels. Due to the inflexibility of CPUs and the high-power consumption of GPUs, they are not suitable for power-sensitive edge AI devices. As a result, power-sensitive edge AI devices require a new customized and flexible AI hardware platform to implement arbitrary CNN algorithms for real-time computing with low power consumption.

Furthermore, as the edge devices are developing into various applications such as monitoring natural hazards by UAV, detecting radiation leakage for nuclear disaster by robotics, and remote sensing in space by satellites, these applied fields are more critical than usual. These critical environments such as radiation fields can cause a system failure. As a result, power consumption is not only the key but also the fault tolerance of edge AI devices for satisfying their compact and mobile feature with reliability. There are various research articles have been proposed targeting fault tolerance. [1] introduces a clipped activation technique to block the potentially faulty activations and maps them to zero on a CPU and two GPUs. [2] focuses on systolic array fault mitigation, which utilizes fault-aware pruning with/without retraining technique. With the retraining feature, it takes 12 minutes to finish the retraining at least, and the worst case is 1 hour for AlexNet. It is not suitable for edge AI. For permanent fault, [3] proposes a fault-aware mapping technique to minus the permanent fault in MAC units. For power-efficient technology, [4] proposes a computation re-use-aware neural network technique to reuse the weight by constructing a computational reuse table. [5] uses approximate a computing technique to retrain the network for getting the resilient neurons. It also shows that dynamic reconfiguration is the key feature for the flexibility to arranging the processing engines. These articles focus on fault tolerance technology specifically. Some of them such as [4,5] address the relationship between accuracy and power-efficient together but lack computation ability information. Besides these listed articles, there are still many published works targeting fault tolerance in recent years. This indicates that the edge AI with fault tolerance is the trend.

In summary, the significant issues of edge AI devices facing are power sensitivity with limited battery capacity, device size limitation, limited local-processing ability, and fault tolerance. To address these issues, examining the related works is necessary for providing future design directions. From the point of view of an edge AI system, the released specifications of the above fault tolerance articles are not comprehensive. This might be because they focus on the fault tolerance feature more than a whole edge AI system. Furthermore, most related edge AI survey works tend to focus on the specific topics and features of an individual structure without comparing the three features such as computation ability, power consumption, and area size. As a result, this paper focuses on evaluating the prior arts on the three key features.

To achieve the flexibility of the hardware structure and dealing with its compact size, one of the solutions for edge AI platforms is dynamic reconfiguration. The reconfigurable function realizes different types of CNN algorithms such that they can be loaded into an AI platform depending on the required edge computing. Moreover, the reconfigurable function also potentially provides fault tolerance to the system by reconfiguring the connections between processing elements (PEs). Therefore, this work not only focuses on the released commercial accelerators but also edge accelerators architecture based on coarse-grained reconfigurable array (CGRA) technology. Overall, this survey will benefit those who are looking up the low-power edge AI accelerators' specifications and setting up their

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 July 2021 doi:10.20944/preprints202107.0375.v1

designs. This paper helps designers in choosing or designing a suitable architecture by indicating reasonable parameters for their low-power edge AI accelerator.

The rest of this paper is organized as follows: Section 2 introduces the hardware types adopted by AI applications. Section 3 introduces the edge AI accelerators including prior edge AI accelerators, CGRAs accelerators, the units used in this paper for evaluating their three key features, and the suitable technologies for implementing the accelerators. Section 4 releases the analysis result and indicates the future direction. Conclusion and future works are summarized in section 5.

2. System Platform for AI Algorithms

To achieve the performance of AI algorithms, there are several design trends of a complete platform for AI systems, such as cloud training and inference, edge-cloud coordination, near-memory computing, and in-memory computing [6]. Currently, AI algorithms rely on the cloud or edge-cloud coordinating platforms, such as Nvidia's GPU-based chipsets, Xilinx's Versal platform, MediaTek's NeuroPilot platform, and Apple's A13 CPU [7]. The advantages and disadvantages of CPU and GPU for applying on edge devices are shown in Table 1 [8,9]. As shown in Table 1, CPU and GPU are more suitable for data-center-bound platforms because CPU's processing ability, GPU's power consumption, and their system size do not meet the demand of low-power edge devices, which are strictly power-limited and size sensitive [10].

When it comes to the edge-cloud coordination system, the second hierarchy, the network connectivity is necessary because those devices cannot run in the areas where is no network coverage. Data transfer through the network has significant latency, which is not acceptable for real-time AI applications such as security and emergency response [9]. Privacy is another concern when personal data is transferred through the internet. Low-power edge devices require hardware to support high-performance AI computation with minimal power consumption in real-time. As a result, designing a reconfigurable AI hardware platform allowing adopting arbitrary CNN algorithms for low-power edge devices under no internet coverage is the trend.

Table 1. Pros and cons of CPU, GPU, and Edge AI accelerator

Pros and cons	**Processor**		
	CPU	***GPU***	***Edge AI accelerator***
Advantage	• Easily to implement any algorithms	• Can process high throughput video data • High memory bandwidth • High parallel processing ability [11-14]	• Power and computation efficient • Compact size • Customizable design for the specific application
Disadvantage	• The sequential processing feature does not match the characteristic of CNN, requiring massively parallel computing.	• Requires massive power support • Restricts its application for power-sensitive edge devices • Images in a streaming video and some tracking algorithms are inputted sequentially but not parallel. [15]	• Customizable for the specific targeted application (inflexible for all type computations) • Computational power limited compared to data center CPU and GPU
Application platform	• More suitable for datacenter • Cooperate with AI accelerator	• More suitable for datacenter • Cooperate with AI accelerator	• Customized for specific edge devices • Can cooperate with CPU or GPU

3. Edge AI Accelerators

To achieve compact size, low power consumption, and computation ability of edge devices, there are several architectures and methods have been proposed. The following subsections, 3.2 and 3.3 introduce the released commercial edge AI accelerators and state-of-the-art edge accelerators based on CGRA, which are potentially suitable for low-power edge AI devices.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 July 2021 doi:10.20944/preprints202107.0375.v1

Most of the proposed review articles introduce accelerators feature-by-feature, and some of them miss mentioning the three key features. These articles tend to report the existing works only but not compare them. On the other hand, another edge AI accelerator articles contain all the three key features, but the result they release is hard to understand because they release comparison results with reference accelerators only. It turns out that the used units in these articles for comparison of the architecture area, power consumption, and computation ability are not expected by the edge AI designers' community. Instead of using millimeters squared (mm^2), watt (W), operations per second (OPs), the results show how many 'times' better their reference works. As a result, we decide to compare the edge AI accelerators and CGRA architectures by the units used by most AI accelerator designers.

3.1. Specification normalization and evaluation

The presented unit of the computation ability in the following section is OPs (operation per second); MOPs, GOPs, and TOPs represent Mega, Giga, and Tera OPs, respectively. The arithmetic of each accelerator varies in data representation, e.g., floating-point (FP) and fixed-point (Fixed). To compare the accelerators, using different arithmetic will lose impartiality. As a result, converting the units will be the following task.

The computation ability will be represented as c. If the arithmetic of the accelerator is FP, its c will be represented as cFP. On the other hand, $cFixed$ represents the computation ability under Fixed arithmetic. In the computation rows of Table 2-7, the initial c is the original data released by the reference works. It might vary in arithmetic type and precisions. Based on [16,17], the computation ability of FP (cFP) can be converted to the computation ability of $cFixed$ by scaling 3 times. As a result, (1) is introduced.

Converted computation ability to fixed point as:

$$cFixed = cFP \times 3 \tag{1}$$

However, because not all accelerators have the same data precision, $cFixed$ is not convincing while comparing the accelerators' ability. [18] indicates that if a structure is not being optimized for precisions like Nvidia does in their GPUs, the theoretical performance of half-precision follows the natural 2×/4× speedups against single/double precisions, respectively. As a result, the computation ability performance of accelerators needs to be normalized, and it is normalized to 16-bit as it is used in the majority without loss of generality. After normalization, the computation ability of each accelerator can be represented as $cFixed16$. To specify the accelerators' performance fairly, (2) is introduced. The lasted computation abilities shown in the computation rows in Table 2-7 are the computation ability in 16bit fixed-point format.

Converted computation ability to 16-bit fixed point as:

$$cFixed16 = cFixed \times (precision / 16) \tag{2}$$

To specify the accelerators' synergy performance, (3) is introduced for representing the evaluation value of accelerators. Since the edge devices require low power consumption and compact size, in (3), the denominator will be power consumption p (w) times chip size s (mm^2), and the fraction will be computation ability $cFixed16$ (GOPs).

The equation for evaluating accelerator's synergy performance:

$$Evaluation\ value\ (E) = cFixed16 / (p \times s) \tag{3}$$

3.2. Prior art edge AI accelerators

The following will show the edge AI accelerators [9,19-28], which focus on the demands of edge devices and are organized into Table 2-4 according to their precision and power consumption. Table 2 shows the accelerators with 16-bit precision and containing less than a watt power consumption. Table 3 shows the 16-bit precision accelerators

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 July 2021 doi:10.20944/preprints202107.0375.v1

contain relatively high-power consumption, higher than a watt. Table 4 shows the accelerators which do not belong to 16-bit precision.

After calculating their evaluation value *E*, accelerators [9] and [21] show similar abilities, *E*=80s. On the other hand, [19,22,26] share similar *E*, in the 20s. Although some of the accelerators have close evaluation value *E*, the *cFixed16*, *p*, *s* values of these accelerators still need to be examined respectively according to the different applying purposes such as targeting applications and environments. For example, [16] has the highest evaluation value *E*, but its size is 9.8 times [9], 3 times [21], and nearly 2 times [26]. Overall, the evaluation value *E* can tell us a general efficiency of an AI accelerator, which is computation ability per unit area and watt. In Table 2-4, there are several accelerators [27,28,30] lack details of the specifications since they only release the module-level data. As a result, the evaluation value *E* of them should be treated more conservatively. On the other hand, [25] is a completed system on an FPGA board and does not release its size on a single chip, so its evaluation value *E* is hard to measure. Nevertheless, its data is a good study material for the designers who intend to build their future projects on an FPGA board for prototyping. These AI accelerators have been proposed as commercial products. Commonly being seen, they have been made as developing boards or USB sticks.

Table 2. Prior Art Edge AI Accelerators

Three key features and the Evaluation value	Edge AI accelerators		
	Kneron 2018 [9]	Eyeriss (MIT) 2016 [19]	1.42TOPS/W 2016 [21]
Computation ability	152 GOPs	84 GOPs	64 GOPs
Precision	16-bit Fixed	16-bit Fixed	16-bit Fixed [9]
Power consumption	350mW	278 mW	45mW
Size	TSMC 65nm RF 1P6M Core area 2mmx2.5mm	TSMC 65nm LP 1P9M Chip size 4.0mmx4.0mm Core area 3.5mmx3.5mm	TSMC 65nm LP 1P8M Chip size 4.0mmx4.0mm
Evaluation value *E*	86.86(core)	18.88 24.66 (core)	88.88

Table 3. Prior Art Edge AI Accelerators

Three key features and the Evaluation value	Edge AI accelerators			
	Myriad x (Intel) 2017 [22]	NVIDIA Tegra X1 TM660M 2019 [23,24]	Rockchip RK1808 2018 [25]	Texas Instruments AM5729 2019 [28]
Computation ability	1 TFlops ≈3 TOPs	472 GFlops ≈1.42 TOPs	100 GFlops ≈300 GOPs	120GOP/s
Precision	16-bit FP	16-bit FP	16-bit FP 300 GOPs@ INT16	16-bit Fixed
Power consumption	<2 Watt	5-10 Watts (module-level)	~3.3W (module-level)	≈6.5W
Size	8.1mm × 8.8mm (package)	28-nm 23mm×23mm	22nm ≈13.9mm×13.9mm	28-nm 23 mm × 23 mm
Evaluation value *E*	21.55(package *s*)	5.34×10^{-4} (module-level *p*)	3.81 (module-level *p*)	3.5×10^{-5}

Table 4. Prior Art Edge AI Accelerators

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 July 2021 doi:10.20944/preprints202107.0375.v1

Three key features and the Evaluation value	Edge AI accelerators		
	GTI Lightspeeur SPR2801S 2019 [16]	Optimizing FPGA-based 2015 [20]	Google Edge TPU 2018 [26,27]
Computation ability	5.6 TFlops =9.45 TOPs	61.62 GFlops= 61.62 GOPS [20]	4 TOPs =2 TOPs
Precision	Input activations: 9-bit FP Weights: 14-bit FP	32-bit FP	INT8
Power consumption	600mW 2.8 TOPs@300mW	18.61Watt	2W (0.5W/TOPs)
Size	28nm 7.0x7.0mm	On Virtex7 VX485T	5.0mmx5.0mm
Evaluation value *E*	329.14	--	40.96

Some works such as [29] use analog components- memristors- to mimic neurons for CNN computing. However, none of the commercial proposed systems uses memristors. Several developers have researched memristor technology including HP, Knowm, Inc., Crossbar, SK Hynix, HRL Laboratories, and Rambus. HP built the first workable memristor in 2008, yet until now it still has a distance from prototype to commercial application.

Knowm, Inc. sold their fabricated memristor for experimentation purposes. Again, the memristor is not intended for application in commercial products [30]. Besides, it is worth mentioning that many CPUs in smartphones contain built-in neural processing units (NPU) or AI modules; for example, MediaTek Helio P90, Apple A13 Bionic, Samsung Exynos 990, Huawei Kirin 990, etc. However, individual NPU or so-called AI modules' detailed performance in these commercial CPUs is not public. As a result, these AI modules are hard to compare with the pure AI accelerators, but it is worth keeping an eye on these commercial products to prevent losing the latest information.

3.3. Coarse-grained cell array accelerators

Dynamically reconfigurable technology is the key feature of an edge AI hardware platform for flexibility and fault tolerance. The term 'dynamic' explains that during the runtime, reconfiguring the platform is still possible. There are several kinds of reconfigurable architectures, and they can be grouped into two major types, fine-grained reconfigurable architecture (FGRA) and coarse-grained reconfigurable architecture (CGRA). FGRA contains a large amount of silicon to be allocated for interconnecting the logic together. This implies that FGRA impacts the rate for reconfiguring devices in real-time due to the larger bitstreams of instructions that are needed. As a result, CGRA is a better solution for real-time computing.

[31] presents many CGRAs and categorizes them into different categories, such as early pioneering, modern, larges, deep learning. The article includes plentiful information, and the authors also collect the statistics to let readers know the developing trend of CGRA comparing to GPU. However, [31] does not focus on the three key features' comparison for the CGRAs. For understanding the performance between CGRAs and figuring out which architectures are the potential candidates for edge AI accelerators, this paper presents architectures [32-42] published in the recent few years for comparison. The units or reference standards are different in each architecture so that this paper consults the references and converts the various units to be standardized according to the revealed information of each architecture in Table 5-7. Table 5 shows the CGRAs use 32-bit precision while Table 6 shows the 16-bit precision CGRAs. Last but not the least, Table 7 presents the CGRAs, which do not use 32-bit, neither 16-bit precision.

Table 5. Coarse-grained Cell Array Accelerators

	Coarse-grained Cell Array Accelerators

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 July 2021 doi:10.20944/preprints202107.0375.v1

Three key features and the Evaluation value	ADRES 2017 [32]	VERSAT 2016 [33]	FPCA 2014 [36]	SURE based REDEFINE 2016 [37]
Computation ability	4.4 GFlops = 26.4GOPs (A9)	1.17 GFlops = 7.02 GOPs (A9)	9.1 GFlops = 54.6 GOPs (A9)	450 Faces/s ≈201.6GOPs (ref.)
Precision	32-bit FP (A9)	32-bit Fixed	32-bit FP (A9)	32-bit Fixed
Power consumption	115.6 mW	44 mW	12.6mW	1.22W
Size	0.64 mm^2	0.4mm^2	Xilinx Virtex6 XC6VLX240T	5.7mm^2
Evaluation value *E*	356.84	398.86	--	29.48

Some of the works do not show the computation ability in OPs, the expected reference unit for the AI accelerator designers and clients. Instead, a few of them compare their computation ability with the ARM Cortex A9 processor [32,33,36]. For example, [33] releases Versat's performance in operation cycles by running the benchmarks and shows the ARM Cortex A9 processor's ability on those benchmarks for comparison. However, the article does not show Versat's OPs, the expected reference unit for the AI accelerator designers and clients. According to the results, it shows that Versat is 2.4 times faster than the ARM Cortex A9 processor on average. Then, [43] shows the performance of the ARM Cortex A9 processor is 500 mega floating points per second (MFlops). After calculation, Versat's operation ability is equal to 1.17 Giga Flops (GFlops). However, GFlops is still not the preferred unit for edge AI devices' designers and clients. Based on (1), GFlops can be converted to GOPs by scaling 3 times. Finally, the performance of Versat is gotten and equal to 3.51 GOPs in 32-bit precision. We adopt (2) to get its cFixed16, 7.02 GOPs, as shown in Table 5 for easily comparing to other accelerators. Similar works are done for the rest of the architectures in Table 5-7. For exception, the area size of [34] is unable to be found out due to the lack of information. [36] does the work on an FPGA, so its core size is unable to be evaluated.

Table 6. Coarse-grained Cell Array Accelerators

Three key features and the Evaluation value	Coarse-grained Cell Array Accelerators		
	TRANSPIRE 2020 [34]	DT-CGRA 2016 [38,39]	Heterogenous PULP 2018 [42]
Computation ability	136MOPs (binary8 benchmark)	95 GOPs	170 MOPs
Precision	8/16-bit Fixed	16-bit Fixed	16-bits Fixed
Power consumption	0.57mW	1.79 W	0.44 mW
Size	N/A	3.79 mm^2	0.872 mm^2
Evaluation value *E*	--	14	443.08

Table 7. Coarse-grained Cell Array Accelerators

Three key features and the Evaluation value	Coarse-grained Cell Array Accelerators		
	SOFTBRAIN 2017 [35]	Lopes et al. 2017 [40]	Eyeriss v2 2016 [41]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 July 2021 doi:10.20944/preprints202107.0375.v1

Computation ability	452 GOPs (test under 16-bit mode)	1.03 GOPs =1.545 GOPs	153.6 GOPs
Precision	64-bit Fixed (DianNao)	24-bit Fixed	Weight/ iacts: 8-bit Fixed psum: 20-bit Fixed
Power consumption	954.4mW	1.996 mW	160 mW
Size	3.76mm^2	0.45mm^2 (not include the buffer, memory, and control systems)	24.5 mm^2 (2 times of v1 [19])
Evaluation value *E*	125.96	1720.1	39.18

The computation unit used by [37] is faces-per-second because it targets face recognition. This makes the computation unit's converting work even harder and having significant deviation when referencing the ability from other similar work. Table 5 shows the computation ability of [37] is 450 faces/second, roughly equal to 201.6 GOPs [44]. In [37], it achieves recognizing 30 faces in a frame while the frame rate is 15 per second which amounts to 450 recognitions to be performed per second. On the other hand, the reference work [44] recognizes up to 10 objects in a frame with a 60 per second frame rate. As a result, the converted computation ability of [37] remains for reference with a certain deviation. [40] has the highest evaluation value *E* of all the listed works. However, the revealed size of [40] is only part of the architecture, so edge AI designers should be more conservative in assessing specific architecture specifications. [42] does not release its computation ability. Since [42] shares the same architecture with [45], the operation/power ratio in [45] can be the reference. Furthermore, [42] contains double cores and extra heterogenous PEs comparing to [45]. As a result, the overall operation/power ratio of [42] would be higher than being evaluated.

Overall, the evaluation results show that [37,38,41] have tens' grade evaluation values *E*, between 10 to 40. [32,33,35,42] have hundreds' grade evaluation values *E*, between 100 to 400. According to CGRA's evaluation value *E*, CGRAs show the potential ability to execute edge AI applications, like the outstanding commercial edge AI accelerators in Table 2-4.

3.4. Implementation technology

As shown in Table 2-7, it can be noticed that FPGA and application-specific integrated circuits (ASIC) are the most used approaches to implement edge AI devices due to their customized ability and low-power consumption. The non-recurring engineering expense (NRE expense) and flexibility of ASIC are high and low, respectively, compared to FPGA. As a result, building a system by ASIC has a higher cost than FPGA when the amount of the products is small. Not only that but also the developing time of ASIC is longer than FPGA. At the beginning of the system developing process, FPGA-based platforms are the better solution due to their high throughput, reasonable price, low power consumption, and reconfigurability [46]. Accordingly, at the prototyping stage, building future AI platform design on a suitable FPGA platform at the system level [47] is suggested.

4. Architecture Analysis and Design Direction

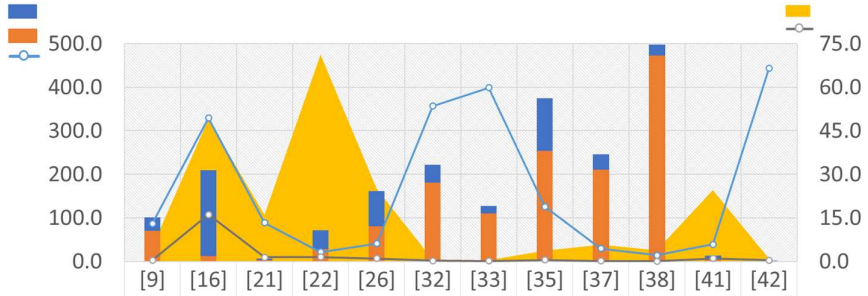
Fig. 1 organizes the accelerators whose evaluation value *E* is in the grade of tens or hundreds from Table 2-7. [20,34,36,40] are not included in Fig. 1 because they lack the area data at the chip level. The power consumption data of [23,25] is only released at module-level, so to be fair, they are not listed in Fig. 1, either. The yellow area in Fig. 1 represents the area size of each accelerator. Every bar in Fig. 1 is composed of two parts, up and down in blue and orange color, respectively. The upper part in blue represents the GOPs/mm^2; the lower part in orange represents the mW/mm^2. The two lines, the upper one and the lower one in Fig. 1, represent evaluation value *E* and the ratio of GOPs and mW,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 July 2021 doi:10.20944/preprints202107.0375.v1

respectively. To focus on accelerators targeting ultra-small areas, the accelerators whose area size is below 10 mm² are selected and their three key features are presented in Fig. 2 (a). In Fig. 2 (a), each accelerator contains two bars, the left orange one represents the power consumption in mW while the right blue one represents the computation ability, GOPs. According to Fig. 2 (a), the accelerators can be grouped into two categories by area size as units' grade and decimal grade. In the units' grade group are [9,35,37,38]. On the other hand, [32,33,42] belong to the decimal grade group. In the units' grade group, [9,35] share a similar ratio of GOPs/mW while [37] has a relatively lower ratio, and [38] has the lowest ratio. The result of the ratio can be analyzed as below. Although [37] has close computation ability about 1.3 times and 0.45 times of [9,35], respectively, it consumes too much power, near 3.5 times of [9] and 1.3 times of [35]. When it comes to [38], its computation ability almost reaches a hundred but with huge power consumption, even higher than [37]. As a result, [9,35] have better performance of computation ability and power consumption in the units mm² area grade. It is interesting to know in the decimal grade group, the area size and GOPs/mW ratio have a positive correlation. [42] has a relatively large area size compared to its computation ability. For more detail, the analysis will be introduced in the next paragraph.

Fig. 2 (b) shows the normalized three key features of the accelerators in Fig. 2 (a). The three key features of the accelerators are normalized to the same grade of computation ability by scaling up [32,33,38,42] and scaling down [35,37], linearly [48]. The result shows that except for [38,42], the remaining five accelerators have a similar trend in power consumption and area size. After normalization, the result emphasizes the insufficient performance of [38,42] for low-power edge AI devices. [38] consumes too much power while [42] has a too big area size compared to its computation ability. However, if the targeting application requires ultra-low power consumption and can accept hundred-grade MOPs, [42] is a good choice. Overall, a trading-off between computation ability and power consumption can be taken once the architecture size has been chosen. Designers can set accelerator's specifications according to its targeting application.

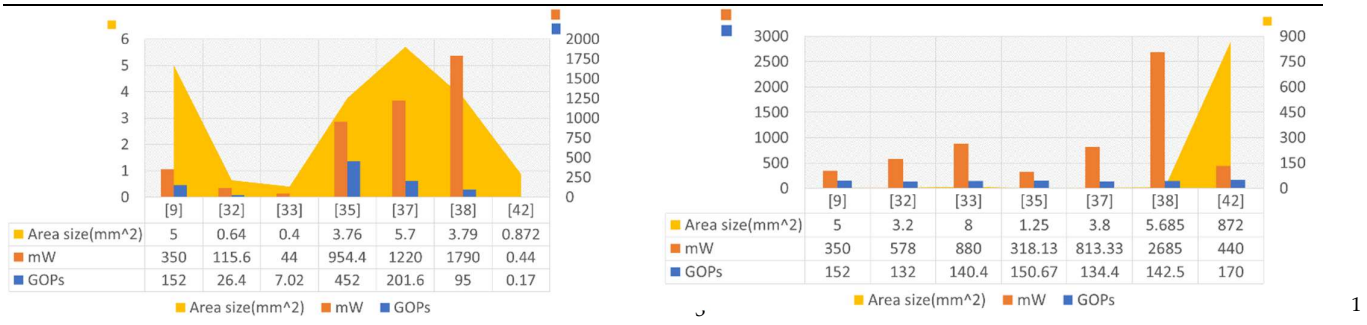
As a result, if designers want to design an architecture for an edge AI accelerator in an ultra-small area (units' mm² area size), the power consumption and operation ability should be in the order of hundreds of mWs and GOPs, respectively.



	[9]	[16]	[21]	[22]	[26]	[32]	[33]	[35]	[37]	[38]	[41]	[42]
Area size(mm^2)	5.0	49.0	16.0	71.3	25.0	0.6	0.4	3.8	5.7	3.8	24.5	0.9
GOPs/mm^2	30.4	197.5	4.0	43.1	81.9	41.3	17.6	120.2	35.4	25.1	6.3	0.2
mW/mm^2	70.0	12.2	2.8	28.1	80.0	180.6	110.0	253.8	210.5	472.3	6.5	0.5
evaluation value (E)	86.9	329.1	89.0	21.5	41.0	356.8	398.9	126.0	29.5	14.0	39.2	443.1
ratio (GOPs/mW)	0.4	16.1	1.4	1.5	1.0	0.2	0.2	0.5	0.2	0.1	1.0	0.4

Figure 1. Power consumption and operations per area statistics

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 July 2021 doi:10.20944/preprints202107.0375.v1



(a) (b)

Figure 2. (a) Three key features statistics (accelerators under 10 mm²) and (b) the statistics normalized to the same grade of GOPs

5. Conclusions and Future Works

This paper has presented a survey of up-to-date edge AI accelerators and CGRA accelerators that can apply to image recognition systems and introduced the evaluation value *E* for both edge AI accelerators and CGRAs. CGRA architectures meet the evaluation value *E* of the existing commercial edge AI accelerators, which implies the potential suitability of CGRA architectures for running edge AI applications. The result reveals the evaluation values *E* of commercial edge AI accelerators and CGRAs are between tens to four hundred, which indicates that the future design trend of edge AI accelerators should meet this grade. Overall, the analysis shows that the future design of ultra-small area (under 10 mm²) accelerators' power consumption and operation ability should be in the order of hundreds of mWs and GOPs, respectively.

As the edge devices are finding their way into various applications such as monitoring natural hazards by UAVs, detecting radiation leakage for nuclear disaster by robotics, and remote sensing in space by satellites, these applied fields are more critical than usual. Many research articles are targeting the fault tolerance feature for edge AI accelerators in recent years. This indicates that the trend of the edge AI accelerators is resilient in terms of reliability and high radiation field applicability.

Author Contributions: For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used "Conceptualization, W.L., A.A, and T.A.; methodology, W.L.; formal analysis, W.L.; investigation, W.L.; resources, W.L. and T.A.; data curation, W.L.; writing—original draft preparation, W.L.; writing—review and editing, W.L. and T.A.; visualization, W.L.; supervision, T.A. All authors have read and agreed to the published version of the manuscript." Please turn to the CRediT taxonomy for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

Funding: Please add: This research received no external funding.

Data Availability Statement: The original contributions presented in the study are included in the article, and further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hoang, L. -H.; Hanif, M. A.; Shafique, M. FT-ClipAct: Resilience Analysis of Deep Neural Networks and Improving their Fault Tolerance using Clipped Activation, 2020 Design, Automation & Test in Europe Conference & Exhibition (DATE), Grenoble, France, 9th -13th Mar. 2020.
2. Zhang, J. J.; Gu, T.; Basu, K.; Garg, S. Analyzing and mitigating the impact of permanent faults on a systolic array based neural network accelerator, 2018 IEEE 36th VLSI Test Symposium (VTS), San Francisco, CA, USA, 22nd -25th, Apr. 2018.
3. Hanif, M. A.; Shafique, M. Dependable Deep Learning: Towards Cost-Efficient Resilience of Deep Neural Network Accelerators against Soft Errors and Permanent Faults, 2020 IEEE 26th International Symposium on On-Line Testing and Robust System Design (IOLTS), Napoli, Italy, 13th-15th Jul. 2020.

4. Yasoubi, A.; Hojabr, R.; Modarressi, M. Power-Efficient Accelerator Design for Neural Networks Using Computation Reuse, *IEEE Computer Architecture Letters* **2017**, 16, 72-75.
5. Venkataramani, S.; Ranjan, A.; Roy, K.; Raghunathan, A. AxNN: Energy-efficient neuromorphic systems using approximate computing, 2014 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED), La Jolla, CA, USA, 11th-13th Aug. 2014.
6. You, Z.; Wei, S.; Wu, H.; Deng, N.; Chang, M.-F.; Chen, An et al. *White paper on ai chip technologies (2018)*; Tsinghua University and Beijing Innovation Centre for Future Chips: Beijing, China, 2008.
7. Montaqim, A. Top 25 ai chip companies: A macro step change inferred from the micro scale, *Robotics and Automation News* **2019**. Available online: https://roboticsandautomationnews.com/2019/05/24/top-25-ai-chip-companies-a-macro-step-change-on-the-micro-scale/22704/ (accessed on 4th May 2021)
8. Simonyan, K.; Zisserman A. Very deep convolution networks for large-scale image recognition, *arXiv* **2015**, arXiv:1409.1556.
9. Du, L.; Du, Y. Li, Y.; Su, J.; Kua, Y.-C.; Liu, C.-C. et al. A Reconfigurable Streaming Deep Convolutional Neural Network Accelerator for Internet of Things, *IEEE Transactions on Circuits and Systems* **2018**, 65, 198-208.
10. Clark, C.; Logan, R. Power budgets for mission success (2011), *Clyde Space Ltd* **2011**. Available online: http://mstl.atl.calpoly.edu/~workshop/archive/2011/Spring/Day%203/1610%20-%20Clark%20-%20Power%20Budgets%20for%20CubeSat%20Mission%20Success.pdf (accessed on 4th May 2021)
11. Yazdanbakhsh, A.; Park, J.; Sharma, H.; Lotfi-Kamran, P.; Esmaeilzadeh, H. Neural acceleration for GPU throughput processors, Proceedings of the 48th International Symposium on Microarchitecture, Waikiki, HI, USA, 5-9 Dec. 2015.
12. Hinton, G.; Deng, L.; Yu, D.; Dahl, G. E.; Mohamed, A.-R.; Jaitly, N. et al., Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, *IEEE Signal processing magazine* **2012**, 29, 82–97.
13. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R. et al. Caffe: Convolutional architecture for fast feature embedding, Proceedings of the 22nd ACM international conference on Multimedia, New York, NY, USA, Nov. 2014.
14. Vasudevan, A.; Anderson, A.; Gregg, D. Parallel multi channel convolution using general matrix multiplication, 2017 IEEE 28th International Conference on Application-specific Systems, Architectures and Processors (ASAP), Seattle, WA, USA, 10th Jul. 2017.
15. Guo, K.; Sui, L.; Qiu, J.; Yu, J.; Wang, J; Yao, S. et al. Angel-eye: A complete design flow for mapping CNN onto embedded FPGA, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* **2018**, 37, 35–47.
16. Gyrfalcon Technology Inc. (GTI), Lightspeeur® 2801S, Available online: https://www.gyrfalcontech.ai/solutions/2801s/ (accessed on 4th May 2021)
17. Farahini, N.; Li, S.; Tajammul, M. A.; Shami, M. A.; Chen, G.; Hemani, A. et al., 39.9 GOPs/watt multi-mode CGRA accelerator for a multi-standard basestation, 2013 IEEE International Symposium on Circuits and Systems (ISCAS), Beijing, China, 19th-23rd May 2013.
18. Abdelfattah, A.; Anzt, H.; Boman, E. G.; Carson, E.; Cojean, T.; Dongarra, J. et al. A survey of numerical methods utilizing mixed precision arithmetic, *arXiv* **2020**, arXiv:2007.06674.
19. Chen, Y.-H.; Krishna, T.; Emer, J. S.; Sze, V. Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks, *IEEE Journal of Solid-State Circuits* **2017**, 52, 262-263.
20. Zhang, C.; Li, P.; Sun, G.; Guan, Y.; Xiao, B.; Cong, J. Optimizing FPGA-based accelerator design for deep convolution neural networks, Proceedings of the 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, Monterey, California, USA, Feb. 2015.
21. Sim, J.; Park, J.-S.; Kim, M.; Bae, D.; Choi, Y.; Kim, L.-S.; A 1.42TOPS/W deep convolution neural network recognition processor for intelligent IoE systems, 2016 IEEE International Solid-State Circuits Conference, San Francisco, CA, USA, 31st Jan.-4th Feb. 2016.
22. Oh, N. Intel Announces Movidius Myriad X VPU, Featuring 'Neural Compute Engine', *AnandTech* **2017**, Available online: https://www.anandtech.com/show/11771/intel-announces-movidius-myriad-x-vpu (accessed on 4th May 2021)
23. NVIDIA, JETSON NANO, Available online: https://developer.nvidia.com/embedded/develop/hardware (accessed on 4th May 2021)
24. Wikipedia, Tegra, Available online: https://en.wikipedia.org/wiki/Tegra#cite_note-103 (accessed on 4th May 2021)
25. Toybrick, TB-RK1808M0, Available online: http://t.rock-chips.com/portal.php?mod=view&aid=33 (accessed on 5th May 2021)
26. Coral, USB Accelerator, Available online: https://coral.ai/products/accelerator/ (accessed on 5th May 2021)
27. DIY MAKER, Google Coral edge TPU, Available online: https://s.fanpiece.com/SmVAxcY (accessed on 5th May 2021)
28. Texas Instruments, AM5729 Sitara processor, Available online: https://www.ti.com/product/AM5729 (accessed on 5th May 2021)
29. Shafiee, A.; Nag, A.; Muralimanohar, N.; Balasubramonian, R.; Paul Strachan, J.; Hu, M. et al. ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars, 2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA), Seoul, South Korea, 18th-22nd Jun. 2016. (accessed on 5th May 2021)
30. Arensman, R. Despite HPs delays, memristors are now available, *Electronics 360* **2016**, Available online: https://electronics360.globalspec.com/article/6389/despite-hp-s-delays-memristors-are-now-available
31. Podobas, A.; Sano, K.; Matsuoka, S. A Survey on Coarse-Grained Reconfigurable Architectures From a Performance Perspective, *IEEE Access* **2020**, 8, 146719-146743.
32. Karunaratne, M.; Mohite, A. K.; Mitra, T.; Peh, L.-S. HyCUBE: A CGRA with Reconfigurable Single-cycle Multi-hop Interconnect, 2017 54th ACM/EDAC/IEEE Design Automation Conference (DAC), Austin, TX, USA, 18th-22nd Jun. 2017.

33. Lopes, J. D.; de Sousa, J. T.; Versat, a Minimal Coarse-Grain Reconfigurable Array, International Conference on Vector and Parallel Processing, Porto, Portugal, 28th-30th Jun. 2016.
34. Prasad, R.; Das, S.; Martin, K.; Tagliavini, G.; Coussy, P.; and Benini L. et al. TRANSPIRE: An energy-efficient TRANSprecision floatingpoint Programmable archItectuRE, 2020 Design, Automation & Test in Europe Conference & Exhibition (DATE), Grenoble (virtual), France, 9th-13rd Mar. 2020.
35. Nowatzki, T.; Gangadhar, V.; Ardalani, N.; and Sankaralingam, K. Stream-Dataflow Acceleration, 2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA), in IEEE Proc. ISCA, Toronto, ON, Canada, 24th-28th Jun. 2017.
36. Cong, J.; Huang, H.; Ma, C.; Xiao, B.; Zhou, P. A Fully Pipelined and Dynamically Composable Architecture of CGRA, 2014 IEEE 22nd Annual International Symposium on Field-Programmable Custom Computing Machines, Boston, USA, 11th-13th May 2014.
37. Mahale, G.; Mahale, H.; Nandy, S. K.; Narayan, R. REFRESH: REDEFINE for Face Recognition Using SURE Homogeneous Cores, *IEEE Transactions on Parallel and Distributed Systems* **2016**, 27, 3602-3616.
38. Fan, X.; Li, H.; Cao, W.; Wang, L.; DT-CGRA: Dual-Track Coarse Grained Reconfigurable Architecture for Stream Applications, 2016 26th International Conference on Field Programmable Logic and Applications (FPL), Lausanne, Switzerland, 29th Aug.-2nd Sep. 2016.
39. Fan, X.; Wu, D.; Cao, W.; Luk, W.; and Wang, L.; Stream Processing DualTrack CGRA for Object Inference, *IEEE Trans. VLSI Syst.* **2018**, 26, 1098–1111.
40. Lopes, J.; Sousa, D.; Ferreira, J. C.; Evaluation of CGRA architecture for real-time processing of biological signals on wearable devices, 2017 International Conference on ReConFigurable Computing and FPGAs (ReConFig), Cancun, Mexico, 4th-6th Dec. 2017.
41. Chen, Y.-H.; Yang, T.-J.; Emer, J.; and Sze, V. Eyeriss v2: A Flexible Accelerator for Emerging Deep Neural Networks on Mobile Devices, *IEEE Trans. Emerg. Sel. Topics Circuits Syst.* **2019**, 9, 292–308.
42. Das, S.; Martin, K. J.; Coussy, P.; Rossi, D. A Heterogeneous Cluster with Reconfigurable Accelerator for Energy Efficient Near-Sensor Data Analytics, 2018 IEEE International Symposium on Circuits and Systems (ISCAS), Florence, Italy, 27th-30th May 2018.
43. Nikolskiy, V.; and Stegailov, V. Floating-point performance of ARM cores and their efficiency in classical molecular dynamics, *Journal of Physics: Conference Series* **2016**, 681, 012049.
44. Kim, J.; Kim, Lee, M. S.; Oh, J.; Kim, K.; Yoo, H. A 201.4 GOPS 496 mW Real-Time Multi-Object Recognition Processor With Bio-Inspired Neural Perception Engine, *IEEE Journal of Solid-State Circuits* **2010**, 45, 32-45.
45. Gautschi, M.; Schiavone, P. D.; Traber, A.; Loi, I.; Pullini, A.; Rossi, D. et al., Near-Threshold RISC-V Core With DSP Extensions for Scalable IoT Endpoint Devices, *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* **2017**, 25, 2700 - 2713.
46. Shawahna, A.; Sait, S. M.; El-Maleh, A. FPGA-Based Accelerators of Deep Learning Networks for Learning and Classification: A Review, *IEEE Access* **2019**, 7, 7823-7859.
47. Lavagno L.; Sangiovanni-Vincentelli, A. System-level design models and implementation techniques, Proceedings 1998 International Conference on Application of Concurrency to System Design, Fukushima, Japan, 23rd -26th March 1998.
48. Takouna, I.; Dawoud, W.; Meinel, C.; Accurate Mutlicore Processor Power Models for Power-Aware Resource Management, 2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing, Sydney, NSW, Australia, 12th - 14th Dec. 2011.