Dean&Francis

ISSN 2959-6157

Emerging Technologies and Applications of AI Chips: Integrating Deep Learning Algorithms with Advanced Hardware Architectures

Yuting Zhang

Chongqing University of Posts and Telecommunications, Sichuan, China
*Corresponding author: 2161068@brunel.ac.uk

Abstract:

Due to the ongoing promotion of the latest technological revolution and industrial upgrading, as well as the widespread use of artificial intelligence and deep learning, the intelligent Internet of Things terminal has reached a more advanced stage of development. The combination of training data, model algorithms, and processing power is the basis for advancing artificial intelligence. The semiconductor business plays a crucial role in shaping the future of AI algorithms and the computing power industry. This study explores the progression of artificial intelligence, and essential technologies, and examines the categorization and structure of AI chips. The process of deep learning was then analyzed, demonstrating the application of AI in real life. Next, this study showcases distinct AI chip-related models, specifically designed for the automated implementation of Convolutional Neural Networks (CNNs) on Field-Programmable Gate Arrays (FPGAs). The study then proceeds to examine the unique procedure and the significance of the resulting data. The future of AI was examined in the domains of space exploration, healthcare, and autonomous driving, along with future forecasts.

Keywords: Deep Learning; Artificial Intelligence; AI Accelerator Chip.

1. Introduction

Artificial Intelligence is described as "A system's ability to accurately analyze external data, draw conclusions, and apply these findings to achieve specific objectives and tasks through adaptable change."[1] It is the capacity to create intelligent human machines via learning and reasoning. Artificial Intelligence was formally defined for the first time at the Dartmouth Conference in 1956, which also marked the official beginning of the AI revolution [1]. The cybernetics group gradually began to focus on neural networks [2], which are computer models that mimic the composition and operation of the human brain and can be thought of as approximating any functioning black box; in other words, this is an intelligent human-machine that focuses through learning and reasoning.

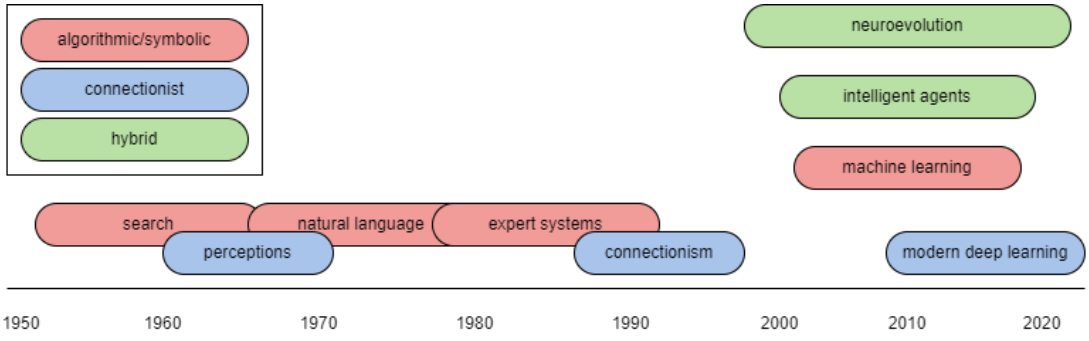

Fig. 1 A timeline of AI development over nearly 70 years from 1950 to 2020.

Fig. 1 illustrates how the connectionist and algorithmic/symbolic research scenarios have been increasingly popular in recent years, with the latter and their combination being utilized more frequently. Nowadays, there are four core areas of AI technology, namely Machine Learning, Deep Learning, Computer Vision, and Natural Language Processing (NLP).

One of the fundamental technologies of the AI age is artificial intelligence (AI) chips, which are typically chips made for AI algorithms. The three elements of al-

gorithms—computing power, huge data, and balance—need to be created in tandem with each other and the rapid growth of AI processors. In the field of Artificial Intelligence, the more current ones utilize three types of chips: general purpose chips (GPUs), semi-custom chips (FPGAs), and full custom chips (ASICs) [4]. At the current stage GPUs and CPUs are still the mainstay of AI chips. GPUs have a lot more Arithmetic Logic Units (ALUs) than CPUs, which helps them process dense data [5]. They also feature a parallel architecture that allows them to process several data sets with many computing units and ultra-long pipelines in a single instruction.

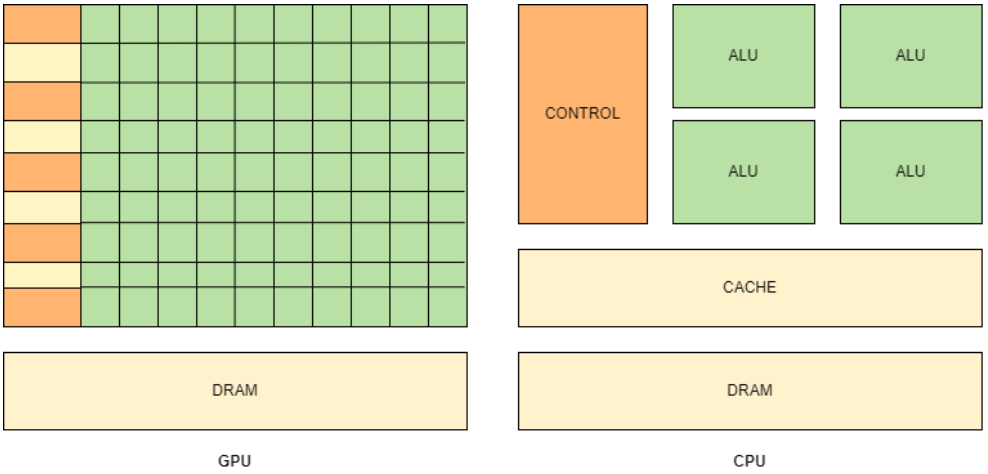

Fig. 2 The architectural differences between the GPU and CPU chips.

As can be seen in Fig. 2, the GPU compresses as well as splits the footprint of the caches and controllers to give more space to the ALU. That means more computational resources are provided to be able to have higher efficiency in image processing as well as complex algorithms.

The composition of this paper is divided into the following main sections. The literature review section shows the history of AI, application scenarios, and the advantages of AI chips. The deep learning procedure and the model flowchart are briefly illustrated in the model-specific part, which is followed by a detailed examination of the Cyclone V chip architecture. The application part primarily showcases the practical implementation of artificial intelligence (AI) in real-world settings. It primarily enumerates the various scenarios where AI is combined with Internet of Things (IoT) devices and mobile phones. The experimental section focuses on presenting a particular model for the automated deployment of Convolutional Neural Networks (CNNs) on Field-Programmable Gate Arrays (FPGAs). It includes an analysis of the results and data, as well as a performance comparison between various chips. Finally, there is a concluding section that summarizes the previous section and explains the future direction of AI.

2. Literature Review

In the early period of AI, in the 1960s, many celebrities appeared to issue predictions, and at this time research in AI was divided into two directions: algorithms and symbolic methods and artificial neural networks (ANNs) [6]. Deep learning made significant strides in the early 2010s with the advent of the big data revolution. In the early 2010s, the advent of the big data revolution marked a real breakthrough in deep learning, where researchers used computational resources as well as large amounts of data to train deep networks and made breakthroughs in image recognition, audio recognition, and more [7]. AI can be used in the field of nursing and can assist in complex nursing situations [8]. Similarly, AI enables remote interaction with patients, as well as the ability to track, monitor, and categorize their health status. Artificial Intelligence in industry has led to the development of smart manufacturing [9], combining technologies such as industrial internet, big data analytics, cloud computing, and networked physical systems [10]. Artificial Intelligence is also widely used in the manufacturing industry, mainly in the design of systems and products as well as in building complexes. In the field of computer science, artificial intelligence also holds a significant place. AI is capable of processing large amounts of data efficiently [12], which enables both network quality and system administration to be optimized.

The development of AI chips, in general, can be divided into two periods. One period is 2013-2015, this period is mainly the construction of basic method functions such as deep learning; the second period is after 2015, after which AI chips began to pursue high processing rates [13]. Older technology has caused traditional chips to function poorly, but Moore's Law has recently led to the rapid development of high-quality AI chips [14]. Conventional chips

used to function poorly because of antiquated technology, but Moore's Law has recently caused the development of high-quality AI processors to happen quickly [14]. Artificial intelligence (AI) aids in the comprehension and processing of data obtained from a variety of sensors. While trained neural chips can provide the AI with a high degree of accuracy, noise can also assault the AI and cause the accuracy to decline [15]. Both the training and prediction of neural networks require strong computational skills. While ordinary CPU chips have continuous computational abilities, they are unable to provide the intensive parallel computation that dedicated AI processors offer for deep neural networks (DNNs) [16].

3. Methodological and Technical Modelling Basis

3.1 Deep learning models

3.1.1 Machine learning

Machine learning is divided into a total of three levels, presenting an inclusive relationship from the largest to the smallest, which are i) machine learning algorithms, ii) artificial neural networks, and iii) deep neural networks [17].

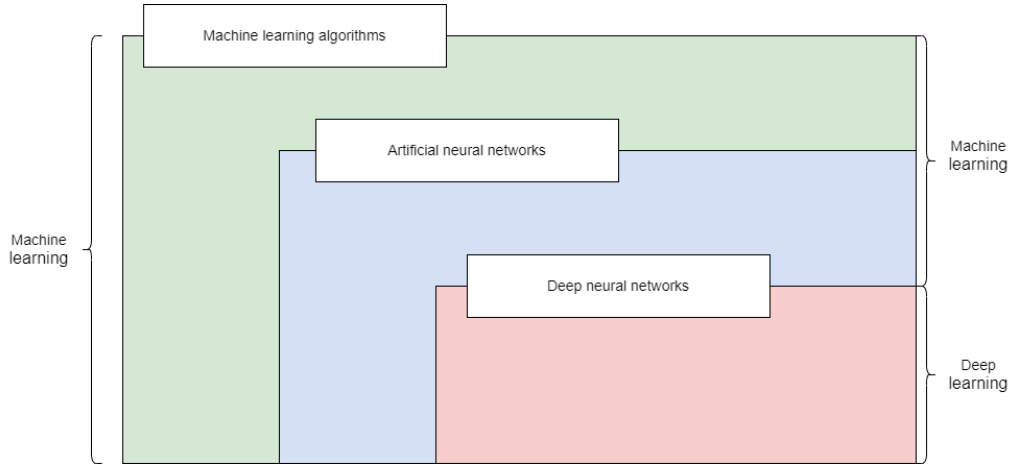

Fig. 3 The relationship between the three levels of machine learning and what is included.

From Fig. 3, it can be concluded that both machine learning algorithms and artificial neural networks belong to shallow machine learning, while the deep neural network part belongs to deep learning, which has more complex model structures and data types. Instead of following given instructions, machine learning can automatically learn and solve problems based on input data. Artificial intelligence (AI) can tackle complex issues [18].

3.1.2 Deep learning flowchart

Over nearly 70 years, the architecture of deep learning (DL) has changed, with neural network architectures gradually replacing artificial neural networks.

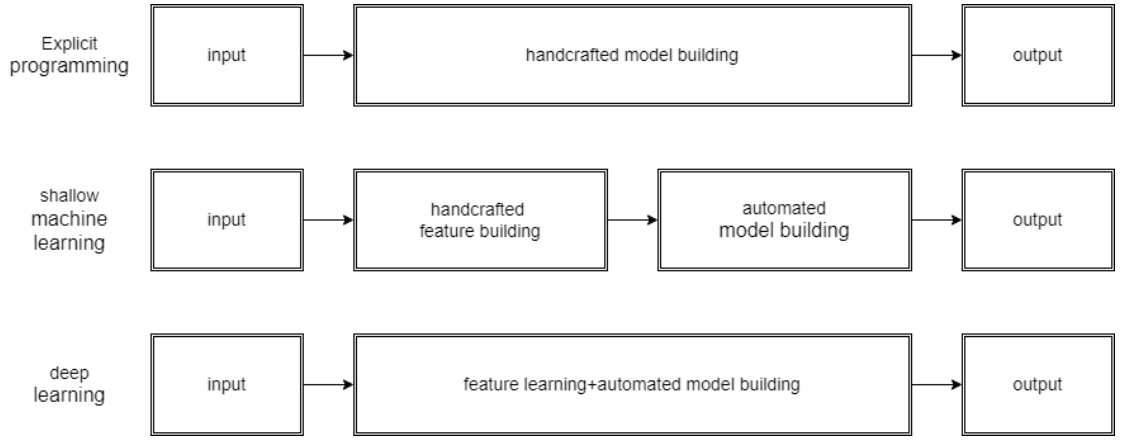

Fig. 4 The model diagrams corresponding to explicit programming, shallow ML, and DL [18].

As can be seen in Fig. 4, the process modules corresponding to different models are not the same. Shallow machine learning (SLIFT) necessitates the manual generation of the feature project, with the model being automatically constructed based on it. In contrast, explicit programming entails the manual construction of the model. Ultimately,

deep learning enables the complete automation of feature extraction and model creation, removing the requirement for manual procedures.

For the analysis of the overall process, the input of data is usually from various websites, social media platforms, and applications. Cross-modal learning, or the combining of several content forms with one another, and the processing of diverse data kinds are both made possible by deep learning. Whereas the feature extraction part describes the derived attributes of the original data and then expresses them in a suitable formal language [18]. Deep learning directly manipulates the high-dimensional raw input data to finish the automatic model construction process, which has the capacity for automated feature learning. The last step is model evaluation, which usually includes several factors, including computational resources. In a similar vein, deep learning puts high demands on hardware resources due to its need to handle millions of model parameters [20].

3.2 Architecture of artificial intelligence chips

As for the architecture of the AI chip, the FPGA chip Cyclone V FPGA Chips is used as an example, as shown below.

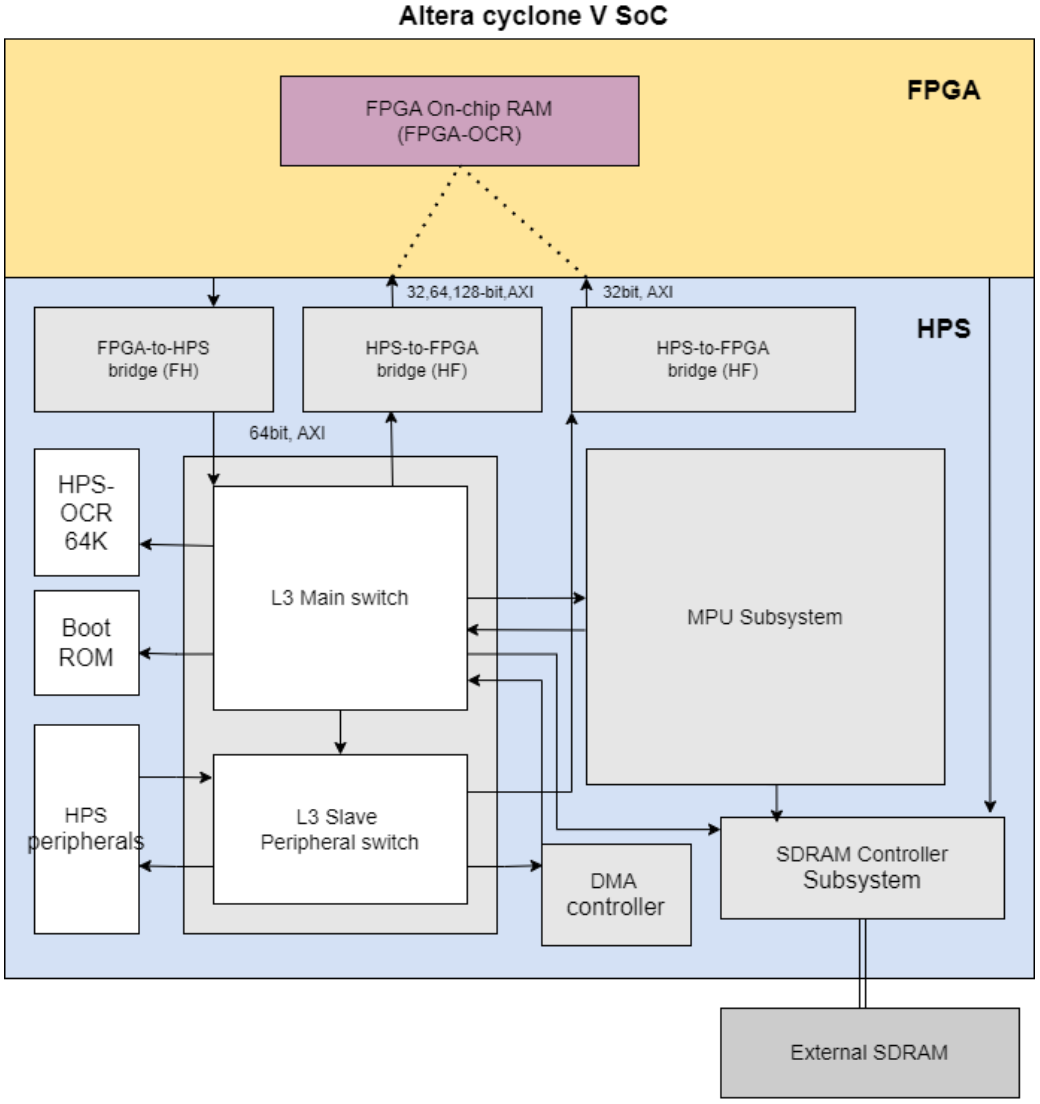

Fig. 5 A simplified diagram of the Cyclone V architecture

As can be seen in Fig. 5, the Adaptation Logic Module (ALM) constitutes about 300K equivalent logic elements (LEs). A hard processor system (HPS), HPS peripherals, on-chip RAM, boot ROM, and a microprocessor unit (MPU) made up of an MPU, an SDRAM controller (SDRAMC), and a DMA controller (DMAC) are all present in the Cyclone V FPGA [21]. The External Memory Interface Controller uses a high bandwidth connection backbone in the FPGA fabric, and the FPGA is located outside the HPS, mostly in the On-chip RAM module, and

connects to the HPS via HF as well as LW.

4. Applications

4.1 AI+Healthcare

The flowchart of design decisions in AI model development is depicted in the following image. AI-based models are frequently utilized in medicine [19]. These scenarios result in various solutions in deep learning models and standard machine learning methods.

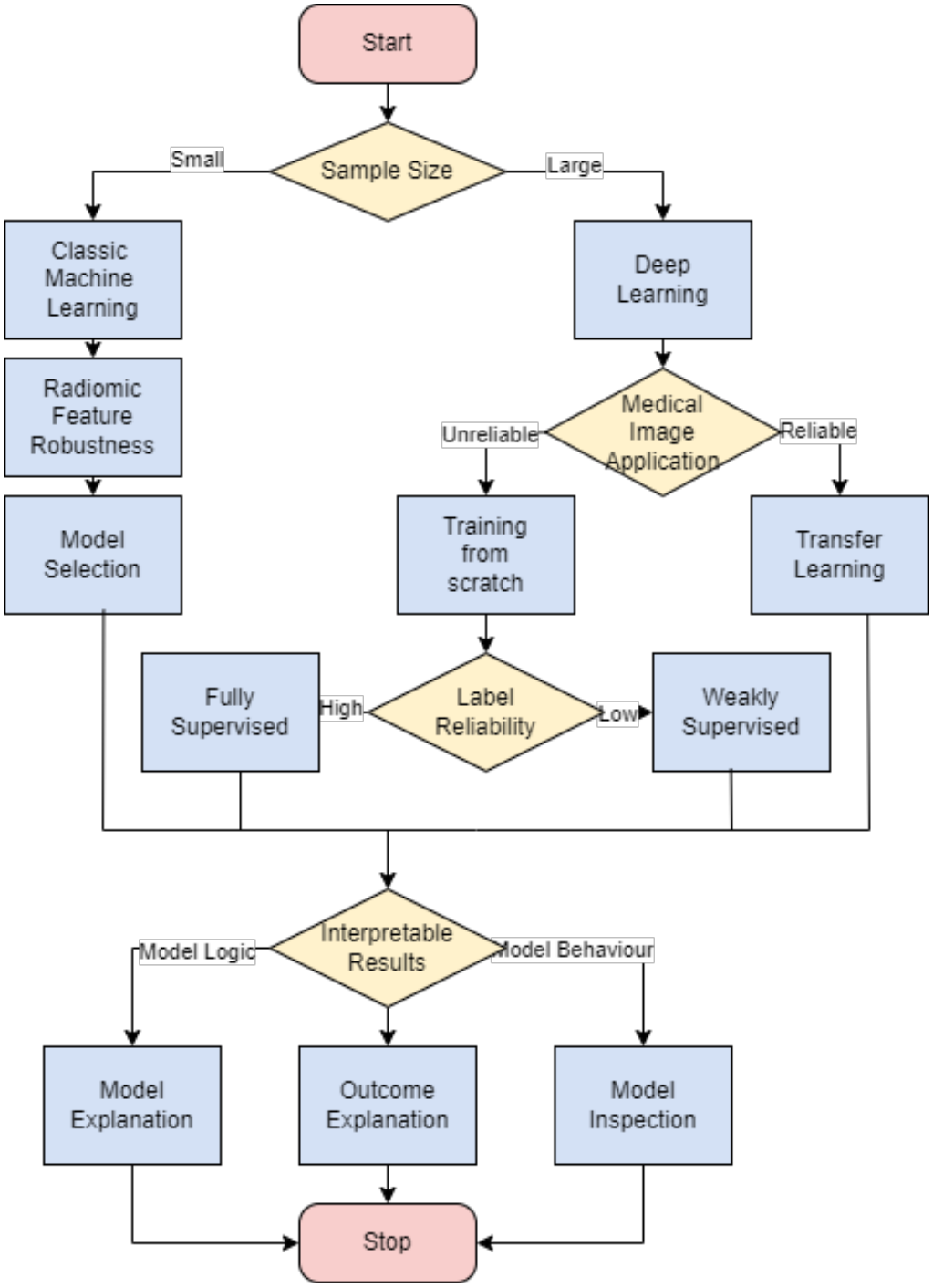

Fig. 6 Machine learning and deep learning models in healthcare systems

According to the expertise of the medical imaging sector, training faulty sample data from scratch is the first step in the deep learning line model process. This is followed by labeling and classification.

4.2 AI + IoT devices

The number of IoT devices has increased dramatically in recent years, and it is predicted that by 2025 there will be more than 75 billion devices connected to the internet. As a result, large amounts of data need to be computed as locally as possible and AI applied to end devices to reduce cloud traffic.

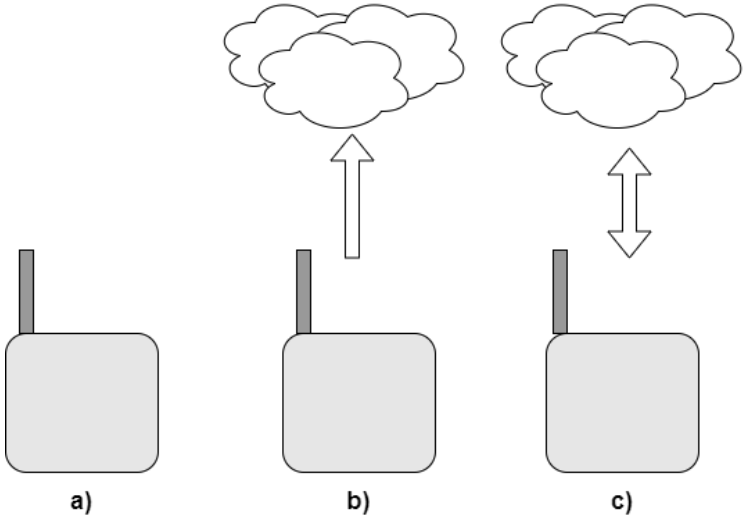

Fig. 7 Three types of computation: (1) DNNs are computed at the end-device end; (2) data is sent from the end-device to the edge servers for computation; and (3) joint computation rows are processed in the cloud.

As can be seen from Fig. 7. The easiest approach to using edge servers is to load all processing from the device end to the edge server. In this scenario, the end device transmits data to a nearby edge server, which processes it, and then receives the processed results. While edge servers can expedite DNN processing, edge devices don't necessarily need to execute DNNs on edge servers.

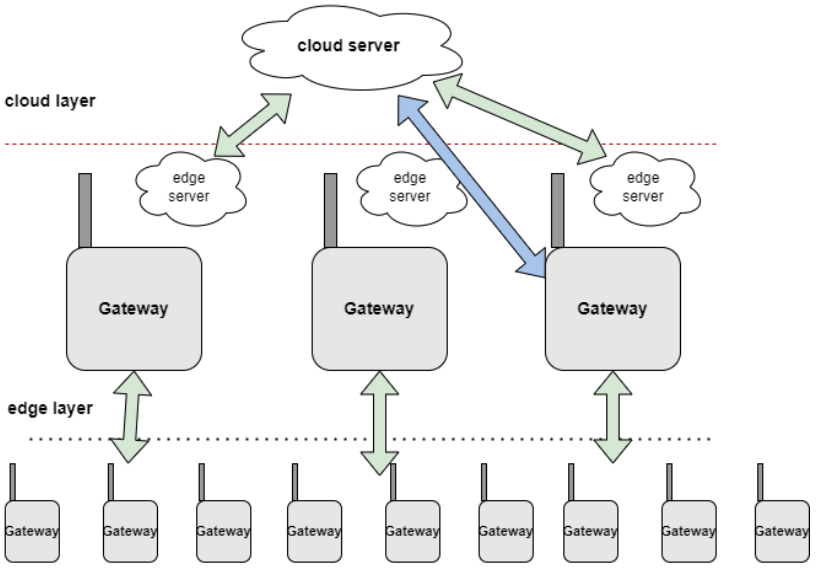

Fig. 8 Three offloading scenarios: (1) partial offloading of partitioned DNNs, (2) a layered architecture where offloading is performed in a combination of edge devices, cloud, and edge servers, and (3) a distributed device computing approach.

4.3 AI+ Mobile phones

Inside the application of AI combined with mobile phones, chips in the form of SoCs dominate smartphone chips [23]. Regarding a mobile device AI chip technology is catego-

rized into two types: adding new functions to an existing module or implementing AI features on top of a separate module. The former involves combining a software development kit (SDK) with a standard hardware module.

Table 1. Comparison of the advantages of the two technology directions for each metric [23].

Technical Route	Power and efficiency	Cost	R&D difficulty
Independent AI unit	Higher efficiency	Lower	Software and hardware development
General hardware unit +software scheduling	DSP: Higher efficiency CPU: Lower than DSP	lowest power consumption	Software development

Table 1. presents a comparison of the two chip technology types: Higher arithmetic efficiency and reduced cost are benefits of chips with autonomous AI unit technology; in contrast, a typical hardware unit outfitted with AI software uses less energy in this configuration, and software development will mostly drive future advancements.

5. Experimental and Modelling Evaluation

5.1 Experimental design

The accuracy of deep learning using Convolutional Neural Networks (CNN) in computer vision tasks [24]. Both Convolutional Neural Networks (CNN) and Deep Neural Networks (DNN) are types of deep learning algorithms, which are widely used in a variety of real-world application areas including autonomous driving. Accelerators are necessary for deep learning and can be broadly classified into three categories: The greatest reconfigurability is offered by FPGAs, whereas GPUs can process large amounts of data quickly but at a significant energy cost. The most computationally efficient shoes are Asics, but their development cycles are lengthy [25].

In recent years, Convolutional Neural Networks (CNNs) have achieved excellent results in several emerging classification tasks, but predictive implementations are usually accompanied by intensive workloads. The large amount of data makes the computation slow and there is a method used to automatically deploy CNNs on FPGAs which makes the arithmetic more powerful [26].

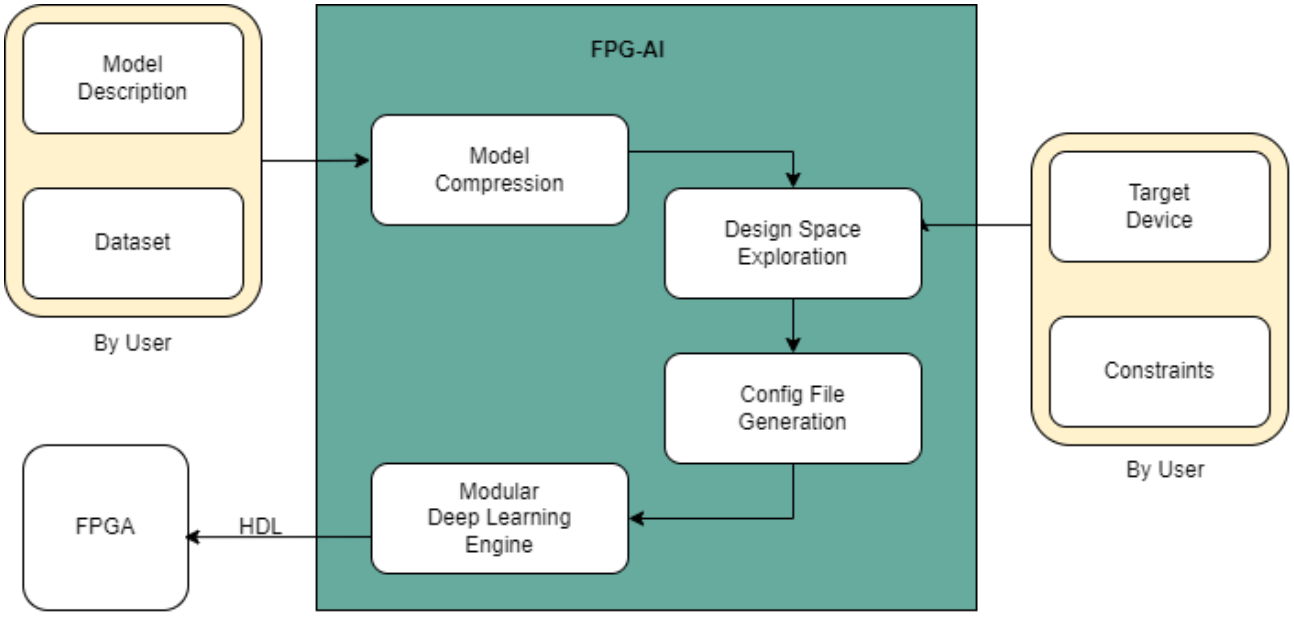

Fig. 9 The FPG-AI tool flow.

5.2 Experimental results

The graphs presented here depict the relationship between the number of bits in various chips and the accuracy of their performance under different bit configurations.

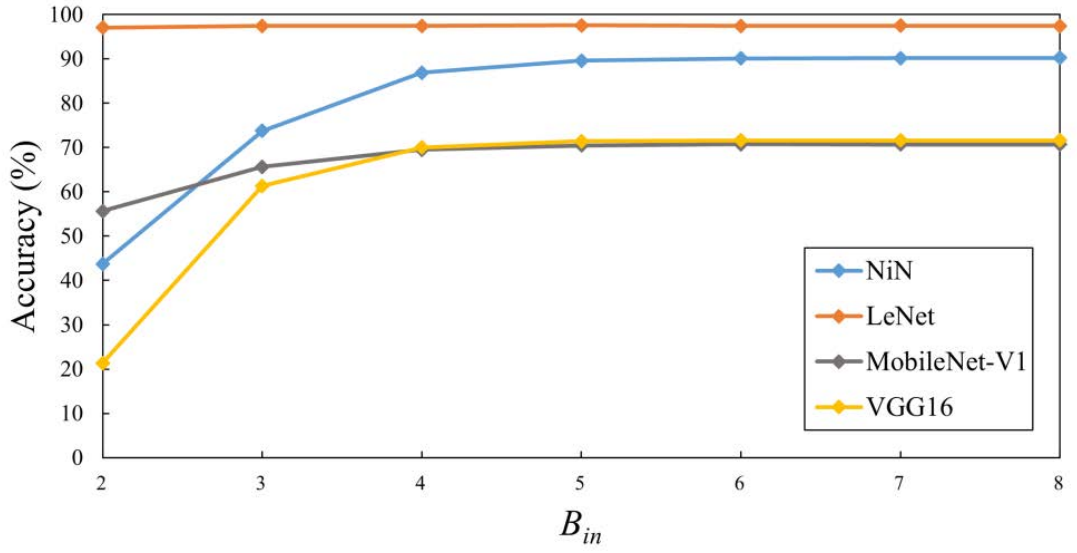

Fig. 10 The results of pre-analyzing the input dataset for the CNN [26].

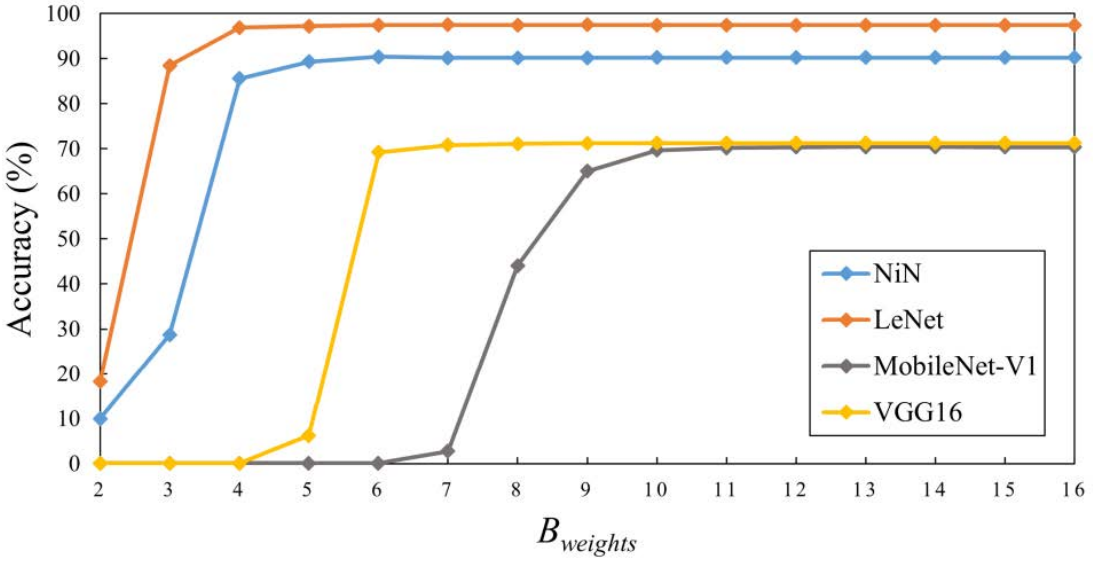

Fig. 11 The results of the weight pre-analysis of the CNN [26].

5.3 Analysis of Results

When it comes to accuracy, the utilization of a 4-bit FXP value for the input data does not have a substantial impact (Figs. 10, 11). VGGI6 is able to maintain great accuracy even with 6-bit or more-bit weight representations, in contrast to MobileNet-V1, which requires at least 9 bits in order to avoid having an impact on the classification algorithm. VGG16 demonstrates a high degree of adaptability to quantization effects in the data as a result of the redundancy of its parameters, whereas MobileNet-V1 is designed to have fewer parameters and, as a consequence, compresses the useful information in the weights in a more efficient manner [25].

For AI chips, different chips have different strengths and weaknesses bias, Table 2 exemplifies the comparison of different chips.

Table 2. The performance comparison of different brain-inspired chips [26]

	BrainScaleS	Neurogrid	TrueNorth	Loihi
Implementation	Analog	Analog	Digital	Digital
Time	Discretized	Real-Time	Discretized	Discretized
Neuron Update		Real-Time	Time MUX	Time MUX
Synapse Resolution	4b	13b shared	1b	1 to 64b
Bio-Mimicry	Not configurable	Not configurable	Limited to LIF	Configurable
On-Chip Learning	STDP only	No	No	Yes
NoC	Hierarchical	Tree multicast	2D mesh unicast	2D mesh unicast
Neurons per Core	8 to 512	65e3	256	max. 1024
Synapses per Core	~130k	100e6	65k x 1b	16000×64b
Core per Chips	352 (wafer scale)	1	4096	128
Chips area (mm^2)	50 (single core)	168	430	60
Technology (nm)	180	180	28	14 (FinFET)
Energy/SOP (pJ)	174	941	27	Min.105.3

6. Summaries

To tackle these concerns, this paper initially presents a chronological account of AI chips, differentiating between their central processing units (CPUs) and graphics processing units (GPUs), and provides a concise overview of the factors that led to the emergence of this technology. The benefits and application areas of AI chips are discussed next. Flowcharts and model diagrams are included in the third section, which is the deep learning application model specifically. The use of AI in mobile phones and Internet of Things devices is covered in the fourth section. The final section is the experimental section, which combines an FPGA and CNN model.

Space technology is the first to evolve in the direction of AI in the future. AI is crucial to the planning of celestial resources and the use of automated mining technology as human activities in space continue to grow. In a similar vein, there are great opportunities for the colonization of space and perhaps Earth's resources in the future [27]. Then there is the field of medical diagnostics, where artificial intelligence (AI) has increased precision in identifying medical issues and tracking the course of illnesses. In the future, even more sophisticated AI technology will be developed, known as quantum artificial intelligence (QAI). Large volumes of medical data may be analyzed by it, and it can enhance patient outcomes and build more effective healthcare systems [28]. Last but not least, AI has advanced in three areas: perception, object identification, and planning for autonomous vehicles. This will allow these vehicles to make real-time decisions with operational safety that complies with regulations in the future [29].

References

[1] A. Grzybowski, K. Pawlikowska–Łagód and W. C. Lambert, "A History of Artificial Intelligence," Clinics in Dermatology, vol. 42, (3), pp. 221-229, 2024.

[2] Yuan Xiaofeng et al, 'Deep learning in process data modeling for process industries,' Journal of Intelligent Science and Technology, vol. 2, (2), pp. 107-115, 2020.

[3] C. Adami, "A Brief History of Artificial Intelligence Research," Artificial Life, vol. 27, (2), pp. 131-137, 2021.

[4] YAO Peng, SONG Changming, HU Yang, CAI Jian, YIN Shouyi, WU Huaqiang. Future Technical Development Approach for High Computing Power Chips[J]. Science and Technology Foresight, 2022, 1(3): 115-129.

[5] B. Li, J. Gu and W. Jiang, "Artificial intelligence (AI) chip technology review," in 2019, DOI: 10.1109/MLBDBI48998.2019.00028.

[6] A. Joshi, Artificial Intelligence and Human Evolution: Contextualizing AI in Human History. (1st ed.) 2023. DOI: 10.1007/978-1-4842-9807-7.

[7] J. M. Górriz et al, "Artificial intelligence within the interplay between natural and artificial computation: Advances in data science, trends, and applications," Neurocomputing (Amsterdam), vol. 410, pp. 237-270, 2020.

[8] K. Seibert et al, "Application Scenarios for Artificial Intelligence in Nursing Care: Rapid Review," Journal of Medical Internet Research, vol. 23, (11), pp. e26522-e26522, 2021.

[9] M. Haenlein and A. Kaplan, "A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence," California Management Review, vol. 61, (4), pp.

5-14, 2019.
[10] I. K. Nti et al, “Applications of artificial intelligence in engineering and manufacturing: a systematic review,” Journal of Intelligent Manufacturing, vol. 33, (6), pp. 1581-1601, 2022.
[11] J. P. Nelson, J. B. Biddle, and P. Shapira, “Applications and societal implications of artificial intelligence in manufacturing: A systematic review,” Cornell University Library, arXiv.org, Ithaca, 2023. DOI: 10.48550/arxiv.2308.02025.
[12] Y. Li, “The application analysis of artificial intelligence in computer network technology,” in 2021, . DOI: 10.1109/IPEC51340.2021.9421146.
[13] H. Momose, T. Kaneko, and T. Asai, “Systems and circuits for AI chips and their trends,” Japanese Journal of Applied Physics, vol. 59, (5), pp. 50502, 2020.
[14] A. P. James, “Towards strong AI with analog neural chips,” in 2020, . DOI: 10.1109/ISCAS45731.2020.9180545.
[15] Q. Jiang and J. Zhan, “Traditional architecture artificial intelligence chip technology,” in 2021, . DOI: 10.1109/ICCSMT54525.2021.00086.
[16] G. Pang and G. Pang, “The AI Chip Race,” IEEE Intelligent Systems, vol. 37, (2), pp. 111-112, 2022.
[17] C. Legaard et al, “Constructing Neural Network Based Models for Simulating Dynamical Systems,” ACM Computing Surveys, vol. 55, (11), pp. 1-34, 2023.
[18] Janiesch, Christian, Patrick Zschech, and Kai Heinrich. “Machine learning and deep learning.” Electronic Markets 31.3 (2021): 685-695.
[19] I. Castiglioni et al, “AI applications to medical images: From machine learning to deep learning,” Physica Medica, vol. 83, pp. 9-24, 2021.
[20] Q. Jiang and J. Zhan, “Traditional architecture artificial intelligence chip technology,” in 2021, DOI: 10.1109/ICCSMT54525.2021.00086.
[21] L. Costas et al, “Characterization of FPGA-master ARM communication delays in zynq devices,” in 2017, DOI: 10.1109/ICIT.2017.7915487.
[22] Merenda, Massimo, Carlo Porcaro, and Demetrio Iero. “Edge machine learning for ai-enabled IoT devices: A review.” Sensors 20.9 (2020): 2533.
[23] Viswanathan S M. AI Chips: New Semiconductor Era[J]. International Journal of Advanced Research in Science, Engineering and Technology, 2020, 7(8): 14687-14694.
[24] Y. Xin et al, “Machine Learning and Deep Learning Methods for Cybersecurity,” IEEE Access, vol. 6, pp. 35365-35381, 2018.
[25] R. Gao and M. Song, “Performance comparative analysis of artificial intelligence chip technology,” in 2021. DOI: 10.1109/ICCEIC54227.2021.00037.
[26] T. Pacini, E. Rapuano, and L. Fanucci, “FPG-AI: A Technology-Independent Framework for the Automation of CNN Deployment on FPGAs,” IEEE Access, vol. 11, pp. 32759-32775, 2023.
[27] J. Garcia-del-Real and M. Alcaráz, “Unlocking the future of space resource management through satellite remote sensing and AI integration,” Resources Policy, vol. 91, pp. 104947, 2024.
[28] M. A. Al-Antari, “Artificial Intelligence for Medical Diagnostics-Existing and Future AI Technology,” Diagnostics (Basel), vol. 13, (4), pp. 688, 2023.
[29] Atakishiyev S, Salameh M, Yao H, et al. Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions[J]. IEEE Access, 2024.Fangfang. Research on power load forecasting based on Improved BP neural network. Harbin Institute of Technology, 2011.