Revolutionizing machine learning: A comprehensive analysis of ASIC hardware accelerators and their applications

Weiming Lei

Department of Electronic Information and Communication, Huazhong University of Science and Technology, Wuhan, 430074, China

u202013897@hust.edu.cn

Abstract. The growth of the web over the past few years has led to tremendous data growth, which has provided a powerful impetus for artificial intelligence and machine learning. Machine learning algorithms are widely used in various classification and prediction problems. However, with the rich data types and needs, the traditional software computing method that relies on CPU can no longer meet the application scenarios under different requirements. Machine Learning (ML) hardware accelerators, especially Application Specific Integrated Circuits (ASICs), have become trendy to meet a variety of needs. This paper reviews the research on ML ASIC in the past few years, reviews the development of ML and ASIC design, and summarizes the characteristics of their use. It is followed by examples of various scenarios for which it can be used, such as medical diagnostics and internet of thing (IOT) terminals. Finally, the author analyses the existing problems and limitations, and gives the improvement methods of hardware and algorithm to deal with the related obstacles. It is hoped that this paper can provide some help and convenience for the subsequent related research.

Keywords: ASIC; Machine Learning; Hardware Accelerators.

1. Introduction

In the past decade, artificial intelligence has ushered in the third wave of development, and has made breakthroughs in many fields such as computer vision, speech signal processing, and natural language processing, reaching or even exceeding the human level. The machine learning technology as the core of Artificial Intelligence (AI) has to face larger data and more complex problems. CPU and GPU cannot meet the machine learning tasks in the pure hardware environment, while FPGA design leads to a large area and energy consumption, and the design of Machine Learning (ML) ASIC has become the current mainstream development direction [1].

This article reviews some of the research designs related to ML ASICs over the past few years. The hardware features of ML ASIC are analysed from the aspects of efficiency, performance, and cost. The main application fields of this kind of hardware system, such as medical disease diagnosis and terminal control and function of the Internet of things, are summarized and studied in the past research, and the main role of ML ASIC chip is obtained. At the same time, compared with the existing solutions to the corresponding problems, the advantages of the designed ML ASIC in dealing with this problem are given. By analysing some optimization studies on ML ASIC systems, this paper presents some limitations of such hardware systems at present, such as the low flexibility of ASIC itself, and the high overhead reflected by the high complexity of ML algorithms on ASIC, and gives corresponding optimization methods in terms of hardware design and algorithms. In this paper, the author hope to review the past research on ML ASICs, give a summary, and provide convenience for the future design and optimization of ML ASICs.

2. Basic Theories

2.1. Machine Learning(ML)

In 1997, Tom m. Mitchell provided a widely cited and more formal definition: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with E" [2]. In other words, machine learning is a technology that allows computers to improve themselves through experience and data. In machine learning, computers can automatically discover patterns and models in the training data by analysis and learning, and then make predictions or decisions based on these results. The goal of machine learning is to endow computers with intelligence similar to humans, enabling them to autonomously learn and adapt to new tasks and environments.

In 1950s, Alan Turing came up with the famous Turing Test, which explored the question of whether a machine could exhibit human intelligence. Claude Shannon first mentioned the concept of machine learning in his paper "A Mathematical Theory of Transient Errors" [3].

In the following decade, artificial intelligence and machine learning began to become a hot topic, and the first machine learning algorithm, the Perceptron, appeared. Bernard Baars and Frederick Hebb proposed the concept of "the strength of the connection between neurons", which laid the foundation for the development of neural networks. In 1970s, machine learning was beginning to move from a single perceptron model to more complex models such as multi-layer perceptrons (MLPS) and support vector machines (SVMS).

There was a downturn in neural network research, which did not rise again until the 1980s. Machine learning then began to be applied to computer vision and speech recognition, and many new algorithms and models appeared, such as decision trees, random forests, naive Bayes and so on. The research of neural networks had been revived with the emergence of new models and algorithms such as backpropagation (BP) algorithm and convolutional neural network (CNN).

In 1990s, machine learning was applied in the field of data mining, and new algorithms such as clustering, association rule mining and anomaly detection appeared [4]. In 2000s, Machine learning had gradually become an important part of the field of data science, with the emergence of a large number of data analysis and ML tools and platforms. Deep learning began to rise, with the emergence of many new models and algorithms, such as recurrent neural networks (RNN), long and short term memory networks (LSTM).

2.2. Application Specific Integrated Circuit (ASIC)

ASIC offers numerous advantages over general integrated circuits, including compact size, reduced power consumption, enhanced reliability, superior performance, heightened confidentiality, and decreased cost. In the early stages, ASICs originated as circuits that utilized gate array technology. The introduction of bipolar diode-transistor logic (DTL) and transistor-transistor logic (TTL) gate arrays occurred in 1967 through Fairchild Semiconductor's Micromatrix family.

ASICs began to increase in size after CMOS (complementary metal oxide semiconductor) technology became available, and in 1974 Robert Lipp developed the first CMOS gate array for International Microcircuits. Fairchild and Motorola standardized MOS technology in the 1970s, when Micromosaic and Polycell standard units were created [5]. VLSI Technology's commercialization of this was only successful in 1979, and LSI Logic's commercialization of it started in 1981.

In 1981, the ZX81 8-bit chip introduced successful commercially viable applications for mass market users, and in 1982, the ZX Spectrum personal computer introduced successful commercially viable applications. During this period, gate arrays consisted of several thousand gates, now known as mediumscale integrations. Customization is accomplished by changing the metal and/or polysilicon interconnect mask [6].

The ASIC development process that involves a full custom design is typically the most expensive due to the design's need to start at the semiconductor level and use HDL to describe each layer of the ASIC. This approach is used by processor designers like Intel, AMD, and Nvidia to create more optimized ASICs. The Comparison among chips is shown in figure 1.

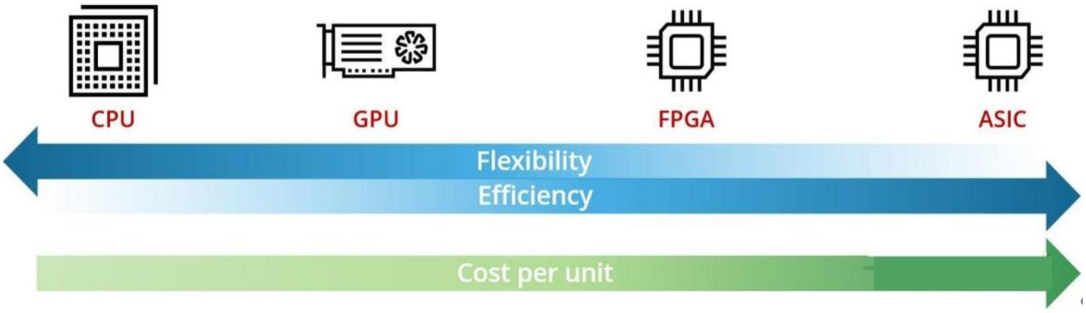

Figure 1. Comparison among chips (Photo/Picture credit: Original)

2.3. Features of ASIC in ML area

2.3.1. Efficiency

Compared to traditional chips designed for ML such as CPUs and GPUs, one of the ASIC's advantages is efficiency. According to the review from Talib, ASICs always yield less power consumption than FPGAs for the same system and ML algorithm [7]. Also, Akhoon mentioned that several ASIC system can reach 100GPOS to 1TPOS and the speed can be 60x to 460x faster than some existing GPU [8]. In Abubakar's research, the ASIC they used only required 1.08mm2 and 746nW which showed great efficiency both in area and energy [9]. Similarly, in other specific researches about chips using ML algorithms, ASIC designs all showed same characteristics: minimal area and extremely low power consumption compared to other design, which can reach the order of millimetre and nanowatt. This means ASICs can reach great efficiency in ML area.

2.3.2. Accuracy

When designing ASIC, the circuit is fixed to realize the specific algorithm with highest efficiency. So unlike FPGA design, ASIC has more compact circuit and higher utilization rate of its computing power, without the noise from inactive elements, which means that ASIC can effectively avoid hardware interference to reach better accuracy. Sudha designed a ASIC system for clinical decision support which performed better than any existing approaches and the accuracy could reach around 95% [10]. Tastan also achieved at least 90% top-1 accuracy on trained models and test data by new designed IoT ASIC [11].

3. ML-ASIC's applications

3.1. Medical disease detection

In medicine, analyzing images and waveforms is an important way to diagnose diseases, and this is an area where machine learning is good at. At the same time, in order to reduce the damage of all kinds of electronic testing equipment to the human body, and to obtain more special and accurate signals or images, all kinds of equipment have strict area and energy consumption requirements. For specific conditions, users want to be able to respond quickly to changes in the condition. Therefore, the low delay and high area and power efficiency of ML ASIC chips perfectly meet these requirements.

Lin focused on the Arrhythmia, a kind of heart disease which is gusty and accidental, and designed an ASIC circuit including machine learning engine to detect the occurrence of the disease by analysing the collected electrocardiogram (ECG) with trained models to give timely reports to patient or doctor [12]. He finally used a combination of SVM and linear classification algorithms, reached a high accuracy at 98%, which was both more efficient and more accurate than existing designs with 41.7% reduction in total energy consumption on average.

Sun also paid attention to the artificial neural networks (ANN) ASIC in classifying ECG signals and recognizing risks of cardiovascular diseases (CVDs). The researchers propose a new reusable neuron architecture (RNA) implemented with a three-layer ANN to perform the aiming functions. By implementing RNA design, the mobile device can train autonomously to update network’s weights and biases and give ECG classification [13]. The RNA has a 98.7% area save, a 99.1% dynamic power save and a 30-times overabundance of time delay compared to the Flat design. RNA had remarkable improvements in speed and energy consumption compared to software-based implementation on a smartphone for approximately 5000x and 4000x.

Abubakar made a further improvement in ECG classifying area. They used a three-layer TNN-based ECG processor ASIC which was wearable or implantable device. After using a moving average filter, the system would detect the binary image and extract features [9]. TNN part then gave its classification of 13 different types of ECG rhythms. The proposed ASIC was finally implemented using 65nm COMS technology and had 764 nW power consumption which is smaller than other works. The 99.3% accuracy made it stand out from the peer designs.

Sudha provided a more general aspect about the ML ASIC utilization in clinical decision support or disease prediction. The researchers gathered clinical data sets upload to a medical diagnose system attaching importance on heart disease, breast cancer, and fertility concerns [10]. Then, they created an ML system and an ASIC framework and compared their accuracy to that of the pre-existing classifiers. Finally, compared to other existing classifiers, the suggested system, ASIC-BPNN, performed at least 95% accurately in three challenges.

3.2. Internet of Things

In the Internet of Things, many terminal devices may work in a specific environment to perform the corresponding task, and some requirements, the use of fpga may cause a great waste of resources. For CPU, GPU, IOT terminals often lack a software operating environment, so ASIC can effectively reduce resource waste and human workload.

Fan indicated that the developments in Internet of Things (IoT) and ML are leading to the trend of combination of these two areas [14]. The researcher aimed to design an ANN-based ASIC -system node IP to optimize a time-varying algorithm for swarm robotics in the IoT. Through algorithm analysis and module design, the final ANN node chip can successfully communicate with Xilinx FPGA. Then the researcher simulated the artificial potential field model and used ANN to replace the original algorithm to realize the control of swarm robotics.

Liu focused on the speech human-computer interactions in the IoT environment. The existing high-accuracy deep learning (DL) algorithm leaded to high processing power consumption and large hardware overhead [15]. So Liu develop a DSCNN-based ASIC to lower the power consumption in speech keyword spotting problems with remained high accuracy about 90%. Finally on ASIC design, it only required 0.31mm2 and 8.05uW to reach similar performance compared with existing designs which required 52 to288uW. Other than that, the time latency per work was lowered to 14.4ms when working in 1MHz.

3.3. Other applications

Saric proposes a customized solar particle events (SPE) hourly predictor based on logistic regression (LR) to detect SPE timely to protect circuits for space applications from high radiation during SPE [16]. The researchers collected data from historical SPE selection and sram-based radiation monitor. Then they compared 8 ML model and finally selected LR because of its better performance during offline training and validation. The ultimate ASIC chip used 130 nm industrial technology which had the accuracy of 96% in detecting the occurrence of SPE.

Miryala proposed a possibility to deploy NN models on front-end ASICs to promote signal processing and quantization. The researchers compared the performances of MLP and CNN in three different sizes and the small size network showed better error in recover peak value from waveforms. Moreover, pruning and quantization can reduce the model size by 90% with little increment in error [17].

Borrego-Carrazo also reviewed the past researches about the utilization of embedded devices including ASIC in Advanced Driver Assistance System (ADAS) [18]. In ADAS area, there are 5 major task relevant for ML and embedded solutions. Most previous researches used SVM-based algorithms for those tasks because they are lightweight models and do not need any modification while other models like CNN are also used but demand more computation resources.

4. Restrictions of ASIC in ML area and further improvements

4.1. Low Flexibility

Because the ASIC design cannot be changed after manufacturing, it is unsuitable for applications that require updates after installation. ASICs have little flexibility and greater design costs and longer time cycles. Every improvement in algorithms or adjustments of device errors means totally redesigning the chip.

For most design process, despite the fact that the eventual goal is ASIC implementation, FPGAs are frequently utilized for prototyping and validation. For example, the researches mentioned above all choose simulation on the FPGA chips using Vivado or VHDL to assure the logic circuit and make correct the time sequence. Then they used CAD or other technis to propose the final ASIC design. Through this method, when the designer improves the ASIC design, he does not need to re-wire the system, but reduces the waste of circuit components and materials through the simulation design on the software, and also facilitates the designer to find the loopholes and errors in the design in time so as to correct them earlier.

4.2. Complexity of ML algorithms

The computational complexity of neural network technology is high, resulting in high processing power consumption and large hardware overhead. The better performance the designer required, the more area or energy may be needed, which may lead to great reduction in one advantage of ASIC, efficiency. Also, ML algorithms can hardly be transformed to hardware language, which restrict the development of relative designs.

Therefore, many researchers aim to develop ML algorithms to make it more concise and suitable for hardware design.

At present, the vast majority of ASIC designs use neural network-related algorithms, and more neurons and neural layers mean more computation and energy consumption, which is fatal to ASIC design. Therefore, we can reduce the corresponding loss by reducing the number of two parameters. This can lead to a drop in possible accuracy or performance, so we need to constantly adjust the parameters to get an acceptable effect.

On the other hand, many ASIC designs now use 32-bit floating-point numbers as the base data type for relevant ML calculations, which greatly increases the overall computation and memory area required on the chip. However, according to the existing literature research, the performance improvement effect of high-precision data computation on ASIC and ML algorithm is not obvious. Therefore, an effective way to improve the energy efficiency of ASIC is to replace the existing floating-point data with low-precision Int data of 4 or 8 bits, which will bring exponential efficiency improvement and space saving. This is extremely effective for ML ASIC design optimization.

5. Conclusion

By reviewing the ML ASIC related research in the past few years, this paper reviews the development history of ASIC design and ML algorithms and presents the characteristics of high energy efficiency, high area efficiency, high accuracy but high implementation cost of ML ASIC hardware system. In the third part, the author makes a generalization of its application areas and achievements. In medical disease diagnosis, especially heart rhythm analysis, the ASIC system performed well in time latency which were required in alarming doctors and patients, and the low energy requirements which enabled the device installed inside the body to get precise measurements. In IoT terminal devices and other fields, the low power consumption made the systems suitable for terminal design and extreme environments. In the last part, in order to solve the problem of low flexibility of existing ML ASICs and high-power consumption caused by high complexity of ML algorithm, software /FPGA pre-design and optimization methods of reducing precision and parameter size are presented. It is hoped that this paper can provide convenience for future related research.

References

[1] Fradkov A L. Early history of machine learning. IFAC-PapersOnLine, 2020, 53(2): 1385-1390.

[2] Zhang L, Tan J, Han D, et al. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. Drug discovery today, 2017, 22(11): 1680-1685.

[3] Plasek A. On the cruelty of really writing a history of machine learning. IEEE Annals of the History of Computing, 2016, 38(4): 6-8.

[4] Çelik Ö. A research on machine learning methods and its applications. Journal of Educational Technology and Online Learning, 2018, 1(3): 25-40.

[5] Basiladze S G. Application specific integrated circuits for ionizing-radiation detectors (review, part 1). Instruments and Experimental Techniques, 2016, 59: 1-52.

[6] Piso D, Veiga-Crespo P, Vecino E. Modern monitoring intraocular pressure sensing devices based on application specific integrated circuits. 2012.

[7] Talib M A, Majzoub S, Nasir Q, et al. A systematic literature review on hardware implementation of artificial intelligence algorithms. The Journal of Supercomputing, 2021, 77: 1897-1938.

[8] Akhoon M S, Suandi S A, Alshahrani A, et al. High performance accelerators for deep neural networks: A review. Expert Systems, 2022, 39(1): e12831.

[9] Abubakar S M, Yin Y, Tan S, et al. A 746 nW ECG Processor ASIC Based on Ternary Neural Network. IEEE Transactions on Biomedical Circuits and Systems, 2022, 16(4): 703-713.

[10] Sudha M. Evolutionary and neural computing based decision support system for disease diagnosis from clinical data sets in medical practice. Journal of medical systems, 2017, 41(11): 178.

[11] Taştan İ, Karaca M, Yurdakul A. Approximate CPU design for IoT end-devices with learning capabilities. Electronics, 2020, 9(1): 125.

[12] Lin Kaiwen. An Low power ECG Processor with Weak-Strong Hybrid Classifier for Arrhythmia Detection. MS thesis. Zhejiang University. 2018.

[13] Sun Y, Cheng A C. Machine learning on-a-chip: A high-performance low-power reusable neuron architecture for artificial neural networks in ECG classifications. Computers in biology and medicine, 2012, 42(7): 751-757.

[14] Fan Fangwen. Study on Application of reusable ANN node IP in IoT. MS thesis. Beijing University of Technology. 2018.

[15] Liu Qingsong. Neural Network based Low Power Keyword Spotting Hardware Design. MS thesis. University of Electronic Science and Technology of China. 2022.

[16] Sarić R, Chen J, Čustović E, et al. Design of ASIC and FPGA system with Supervised Machine Learning Algorithms for Solar Particle Event Hourly Prediction. IFAC-PapersOnLine, 2022, 55(4): 230-235.

[17] Miryala S, Mittal S, Ren Y, et al. Waveform processing using neural network algorithms on the front-end electronics. Journal of Instrumentation, 2022, 17(01): C01039.

[18] Borrego-Carazo J, Castells-Rufas D, Biempica E, et al. Resource-constrained machine learning for ADAS: A systematic review. IEEE Access, 2020, 8: 40573-40598.