Hardware Implementation for Convolutional Neural Networks in Artificial Intelligence

Hongxi Chen

School of Integrated Circuit Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China

hxchen@std.uestc.edu.cn

Abstract. Artificial Intelligence (AI) has brought great convenience and help to human society by improving efficiency, increasing productivity and reducing cost. As a part of deep learning, convolutional neural networks (CNNs) have been widely concerned by researchers in recent years. In this paper, five parts of the CNN structure, including input layer, convolutional layer, pooling layer, fully connected layer, activation function and output layer are going to be elaborated. Besides, four different kinds of hardware, which can be used in AI implementation, including graphics processing unit (GPU), field programmable gate array (FPGA), application-specific integrated circuit (ASIC) and brain-like chips will be discussed in this paper. Based on the different characteristics of these four groups of hardware, this paper will analyze their feasibility to implement artificial intelligence algorithm. After contrasting their cost, flexibility and power consumption, it is concluded that different hardware has different advantages to implement AI under different circumstances. GPU performs better to handle with parallel operations or construct complex network models of AI. FPGA is able to achieve flexible AI model programming. On the other hand, ASIC is preferred considering its low power consumption and cost to implement AI. Although brain-like chip is not as well developed as the other three chips, it is promising to implement AI in the future.

Keywords: CNN, GPU, FPGA, ASIC, Brain-like Chips.

1. Introduction

In recent years, convolutional neural networks (CNNs) have been widely used for their exceptional accuracy in various fields such as computer vision [1-4], natural language processing [5-7], information retrieval [8, 9] and speech recognition [10, 11]. The CNN structure comprises an input layer, convolutional layer, pooling layer, fully connected layer, activation function and output layer. The input layer processes the neural network's input by de-meaning and normalizing raw image data. The convolutional layer extracts feature from the input data, while the pooling layer shrinks the feature map to preserve important data and produce a smaller subgraph that reflects the original map. The fully connected layer is responsible for the final classification result, and the activation function measures the error between the predicted value and the real sample label.

Regarding hardware implementations, deep learning focused processors are classified into four groups: universal chips like the GPU, semi-customized chips like the FPGA, fully customized ASIC chips like Google's Tensor Processing Unit (TPU), and brain-like chips. GPUs are image processing-oriented microprocessors that have become the first choice for artificial intelligence hardware due to their high flexibility, many threads, registers, and single instruction multiple data (SIMD) units. On the other hand, FPGAs are semi-custom circuits that allow for flexible and highly effective programming, while ASIC chips are fully custom-designed solely for AI applications and have better performance considering energy consumption and other aspects. However, ASICs are inflexible and costly to produce once customized. The brain-like chip, which is able to learn with high degree of automation, is a major topic of future research.

2. Overview of CNN Structure

Fig. 1 shows a fundamental CNN structure, which includes the input layer, convolutional layer, pooling layer, fully connected layer, activation function and output layer.

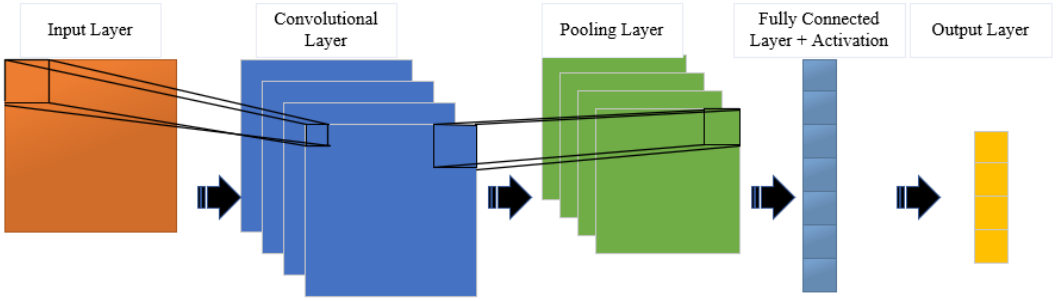


Figure 1. Fundamental CNN structure

2.1. Input Layer

The entire neural network's input is represented in the input layer. This layer's processing primarily consists of de-mean and normalization for the raw image data. A one-dimensional convolutional neural network's input layer often receives a one- or two-dimensional array, whereas a two-dimensional convolutional neural network's input layer typically receives a two- or three-dimensional array.

2.2. Convoluntional Layer

Based on the image recognition process of human brain, local pixels are closely related, while distant pixels are weakly related. Therefore, it is not necessary for each neuron to perceive the global image but to perceive the local information and synthesize the local information at a higher level to obtain the global image. The convolutional layer's job is to take the input data and extract features from it. There are numerous convolutional cores in it. Similar to the neuron in a feedforward neural network, each component of the convolutional kernel corresponds to a weight coefficient and a bias vector. There are several connections between each neuron in the convolutional layer and nearby neurons.

2.3. Pooling Layer

The pooling layer exists in the middle of the continuous convolution layers. After feature extraction in the convolutional layers, the output feature map will be transmitted to the pooling layer for feature selection and information filtering. The pooling layer's purpose is to shrink the feature map, preserve important data, and produce a smaller subgraph to reflect the original map. In essence, pooling is a process of down sampling the image. The pooling layer is calculated via a filter-like structure that moves up, down, and left to right, much like the convolution process. There are two Pooling methods: Max Pooling and Average Pooling.

2.4. Fully Connected Layer

The fully connected layer in convolutional neural networks is equivalent to the hidden layer in traditional feedforward neural networks. After the operation of multiple layers of convolutional layer and pooling layer, there will generally be 1-2 fully connected layers to give the final classification result. The fully connected layer is the last part of the hidden layer of the convolutional neural network and only transmits signals to other fully connected layers. The fully connected layer plays the role of "classifier" in the whole convolutional neural network but it is not expected to have the feature extraction ability. In practice, the fully connected layer can be realized by the convolution operation. In some convolutional neural networks, the function of full connection layer can be replaced by global average pooling.

2.5. Activation Function

The activation function is used to measure the error between the predicted value and the real sample label. The sigmoid function and the hyperbolic tangent function are two common functions. And the rectified linear unit (ReLU) [12] is becoming popular in recent years.

3. Hardware Implementation

Deep learning focused processors can be divided into four groups from the perspective of technological architecture: GPU, FPGA, ASIC and brain-like chips.

3.1. GPU

GPU is a kind of image processing oriented microprocessor. It has less on-chip storage like cache and local memory than CPU. But it has more threads, registers and single instruction multiple data (SIMD) units than CPU. Therefore, for deep learning algorithms containing a large number of parallel operations or complex network models, GPU has become the first choice of artificial intelligence hardware, which is currently the most widely applied AI hardware with the highest flexibility and is widely used in the early training of network models. In August 2018, NVIDIA unveiled its next-generation GPU architecture named Turing at SIGGRAPH. The Quadro RTX GPU and the GeForce RTX GPU are already powered by NVIDIA's latest Turing GPU architecture, featuring the new RT Core which speeds up ray tracing, and the new Tensor Core which is for AI-oriented reasoning. Although the Tensor Core is new to the GPU, it is not much different from the conventional arithmetic logic units (ALUs). Tensor Core is designed to deal with large matrix calculations, which means it will be better able to cope with the large amount of data required for deep learning training.

3.2. FPGA

FPGAs are semi-custom circuits that allow for flexible and highly effective programming. And a large number of basic gate logics and memories integrated in FPGA have been fixed at the factory. Under different programming conditions, the utilization rate of the circuit will be different, even there will be a large number of redundancies, which is caused by the universality of FPGA. While complicated algorithms can be realized on GPU and used for data training, this is much more challenging on FPGA.

3.3. ASIC

An alternative to semi-custom FPGAs is fully custom ASIC chips designed solely for AI applications. Compared with GPU and FPGA, ASIC chip shows better performance considering energy consumption and other aspects. The disadvantage of ASIC chip is that once customized, it cannot be changed and the development cycle is long and less flexible compared with GPU and FPGA. In recent years, the algorithm of deep learning is constantly improved, different models change quickly, and the depth of the network is increasingly large, which is a great challenge to customize a suitable ASIC chip. However, in the long-term plan, it is the mainstream choice to design an ASIC chip specifically for AI applications because cost, energy consumption and massive chip fabrication are key considerations that must be taken into account as artificial intelligence continues to advance rapidly. The TPU, which Google initially unveiled in 2016, was created exclusively for TensorFlow, their open source machine learning software framework.

The TPU's core consists of an 8-bit matrix multiply unit and 28MB of software-managed memory on chip with a maximum computing capacity of 92 TeraOp/s (TOPS). The second generation of the TPU was launched in 2017. Compared to the first generation, the TPU2.0 contains four chips that can be used not only for reasoning, but also for training neural networks, requiring no additional resources for training networks. By May 2018, Google had released TPU3.0, a chip capable of more than 100 petaflops at peak performance.

3.4. Brain-like Chip

As can be seen from the name of the chip, the ultimate goal of this kind of chip is to achieve a function similar to the human brain, which is the most complex computing system in nature, with hundreds of billions of neurons, capable of independent learning, memory, processing of a series of information, output corresponding behavioral instructions, and control the behavior of the body. The most remarkable feature of human brain neuron system is that some neurons are dormant when not in use, and the overall energy consumption is very low. The purpose of the brain-like chip is to enable the chip to learn autonomously, be able to calculate and recognize, and achieve strong cognitive ability. In 2014, IBM designed a brain-like integrated chip called TrueNorth using the idea of multi-core processing, based on the workings of human neural networks. The total number of programmable synapses on this chip is over 268 million, and the basic transistor count is 5.4 billion. Since memory, computing, and communication are all handled in 4,096 synaptic nuclei, TrueNorth bypassed the bottleneck of von Neumann's architecture and is very energy efficient. The future of brain-like chips is promising, but there are also many difficulties to be faced. Brain-like chips mainly realize the function of human brain on the physical level, but human brain has extremely low power consumption, high fault tolerance rate and can learn independently. These problems are needed to be solved in the current research and there is still a certain distance from the application stage of brain-like chips.

4. Conclusion

Through the analysis of the development and application of the above four types of AI chips, it can be found that GPU has powerful functions. FPGAs have relatively low power consumption and high flexibility in development, but it is difficult to implement complex algorithms; Dedicated ASICs have a high investment cost in the early stage of development, but they have low power consumption and moderate development difficulty. In mass production, their average cost is superior to GPU and FPGA, which makes ASIC chips have a great advantage in terminal AI chips. Because of its slow development, the brain chip has not yet formed a complete system, which is not dominant in large-scale application.

References

[1] Sivaranjani A, Senthilrani S, Ashokumar B, et al. An Improvised Algorithm for Computer Vision Based Cashew Grading System Using Deep CNN. 2018 IEEE International Conference on System, Computation, Automation and Networking (ICSCA). IEEE, 2018: 1-5.

[2] Avanzato R, Beritelli F, Puglisi V F. Dairy Cow Behavior Recognition Using Computer Vision Techniques and CNN Networks. 2022 IEEE International Conference on Internet of Things and Intelligence Systems (IoTaIS). IEEE, 2022: 122-128.

[3] Talukder M M H, Ria T A, Hasan M M, et al. A Computer Vision and Deep CNN Modeling for Spices Recognition. 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT). IEEE, 2022: 1-5.

[4] Gao Y, Shi S, Sun Z, et al. The combination of transformer and CNN in computer vision. 2022 IEEE 4th International Conference on Civil Aviation Safety and Information Technology (ICCASIT). IEEE, 2022: 321-325.

[5] Wang W, Gang J. Application of Convolutional Neural Network in Natural Language Processing. 2018 International Conference on Information Systems and Computer Aided Education (ICISCAE). IEEE, 2018: 64-70.

[6] Yang S. Natural Language Processing Based on Convolutional Neural Network and Semi Supervised Algorithm in Deep Learning. 2022 International Conference on Artificial Intelligence in Everything (AIE). IEEE, 2022: 174-178.

[7] Li T. Research on Related Technologies of Natural Language Processing Based on Convolutional Neural Network. 2021 International Conference on Electronic Information Technology and Smart Agriculture (ICEITSA). IEEE, 2021: 435-440.

[8] Patnaik S K, Babu C N. Information Retrieval from web with Faster R-CNN Deep Learning Networks: A New Perspective. 2021 6th International Conference on Computer Science and Engineering (UBMK). IEEE, 2021: 61-66.

[9] Wang L, Qian X, Zhang Y, et al. Enhancing Sketch-Based Image Retrieval by CNN Semantic Re-ranking. IEEE Transactions on Cybernetics, 2020, 50(7): 3330-3342.

[10] Avula H, R R, Pillai A S. CNN based Recognition of Emotion and Speech from Gestures and Facial Expressions. 2022 6th International Conference on Electronics, Communication and Aerospace Technology. IEEE, 2022: 1360-1365.

[11] Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Mansor, H., Kartiwi, M., & Ismail, N. Speech Emotion Recognition using Convolution Neural Networks and Deep Stride Convolutional Neural Networks. 2020 6th International Conference on Wireless and Telematics (ICWT) 2020: 1-6.

[12] Nair, V., & Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), 2010.