Dean&Francis

ISSN 2959-6157

Research and Analysis of the Application of Artificial Intelligence Chips in the Field of Intelligent Recognition

Yuanjing Li

Nanshan Foreign Language School, Shenzhen, China

*Corresponding author: yj.li729@outlook.com

Abstract:

With the rapid development of deep learning technology, intelligent recognition technology, centered around convolutional neural networks (CNNs), has been widely applied in scenarios such as healthcare, image recognition, and autonomous driving. However, the training and inference processes of intelligent recognition models, especially deep CNNs, require massive computational and memory access efforts, posing severe performance and energy efficiency challenges to traditional computing architectures (such as CPUs and GPUs). As intelligent recognition technology becomes increasingly practical and widespread across society, performance requirements for CNNs are steadily increasing. This has significantly driven research into specialized artificial intelligence chips, enabling architectural innovation and optimization of intelligent recognition models.This article will review the early development and current status of intelligent recognition technology, the composition and evolution of convolutional neural networks (CNNs), which play a core role in recognition technology, and how AI chips optimize and enhance CNN performance.In summary, AI chips are specialized accelerators for intelligent recognition computing, and their development is mutually reinforcing with advances in deep learning algorithms. As model complexity continues to grow, AI chips designed for emerging architectures, supporting higher energy efficiency and hardware-software co-design, will continue to be a core driving force behind the advancement of intelligent recognition technology.

Keywords: Artificial intelligence chips, convolutional neural networks, intelligent recognition

ISSN 2959-6157

1. Introduction

Intelligent recognition is a crucial and indispensable technology in contemporary society. Facial recognition has become deeply integrated into every aspect of life, from financial payment at shopping malls to access control systems in residential communities to security checks at train stations and airports. By 2023, my country's facial recognition market was projected to reach 8.5 billion yuan, and is projected to exceed 10 billion yuan.Among them, the security sector has the highest application demand, accounting for 54% of the market size, followed by finance (16%) and entertainment (10%). Intelligent driving technology is also gradually becoming more widespread throughout society. Many well-known brands such as Tesla, Mercedes-Benz, Audi, and Volvo, as well as key technology companies such as Huawei, Baidu (Apollo), and Google (Waymo), are promoting the development and application of intelligent driving technology through their own research and development or through collaboration with automakers. In 2022, global shipments of autonomous vehicles are expected to reach 17.7 million, an increase of 6.4 million units and a year-on-year growth of 56.6%.These phenomena have led to explosive growth in intelligent recognition technology. At the same time, the application of intelligent recognition technology in academic fields has gradually become widespread. For example, in seismic interpretation, by continuously learning the shallow and deep characteristics of faults in seismic profiles, a fault recognition model is finally trained and predicted to achieve the effect of fault recognition [1]. Combining it with neuroscience, computer science, cognitive science, and medicine can more effectively conduct research on emotion recognition [2]. The application of intelligent recognition in multiple fields has placed higher demands on its own real-time performance and recognition accuracy.General-purpose chips are not powerful enough to meet the needs of machine learning. Traditional general-purpose chips (GPUs) have limitations in terms of high energy consumption and high cost when deploying convolutional neural network models (CNNs). Convolution operations require a high parallel budget, while the serial architecture of the CPU is inefficient and is about 10-100 times slower than the GPU. AI chips, on the other hand, have additional neural processing units. They can deliver AI performance faster and have longer battery life. They provide more efficient energy efficiency and better power consumption. AI chips occupy four to five times the bandwidth of standard chips because AI applications require greater bandwidth between processors to run properly and efficiently [3].This is because they require parallel processing. Artificial intelligence can execute highly advanced algorithms, providing exceptional parallel processing capabilities.

Therefore, we need to improve object recognition performance by incorporating AI chips into intelligent recognition applications through architectural innovation, enabling the further maturation of intelligent recognition technology in all aspects of life.

This article will review the early development and current status of intelligent recognition technology, the composition and evolution of convolutional neural networks (CNNs), which play a core role in recognition technology, and how AI chips are being applied in the field of intelligent recognition technology.

This article first reviews the key developments in intelligent recognition technology and convolutional neural networks, analyzing their development trends and performance evolution, from the early LeNet to modern architectures such as the Deep Residual Network (ResNet). It then focuses on how AI chips optimize intelligent recognition technology. A comparison of traditional chips and dedicated AI chips highlights the significant performance improvements AI chips offer for CNNs. Examples of their application are also provided.

2. Early Development and Current Status of Intelligent Recognition Technology:

2.1 Early Development

In the 1970s, pattern recognition became a key area of computer vision. Researchers developed methods for identifying shapes, textures, and objects in images, paving the way for more complex visual tasks.

Template matching, one of the early pattern recognition methods, compares an image to a set of templates to find the best match. This method is limited by its sensitivity to changes in scale, rotation, and noise. Early computer vision systems were hampered by the limited computing power of the time. Computers in the 1960s and 1970s were bulky, expensive, and had limited processing power.

2.2 Current Status

Relying on AI chips and convolutional neural networks. Compared to the past, it was extremely dependent on human experts to manually design algorithms to extract important features of objects. In the era of CNN deep learning, humans only need to provide a large amount of data and define the network structure, allowing the model to discover its internal laws and features. Artificial intelligence also learns from data by training machines and adjusts its behavior based on experience, making it a more flexible and versatile problem-solving method. Artificial intelligence systems can analyze large amounts of data, identify patterns and correlations, and use this information to make predictions or take actions without explicit programming. [4]

3. Overview and Development of Convolutional Neural Network Technology

Convolutional neural networks (CNNs) are at the core of intelligent recognition, serving a function similar to the cerebral cortex. They enable computers to understand the content of images and extract valuable information from them, just as humans do. Through unique mechanisms such as local connections, weight sharing, and pooling, CNNs automatically, efficiently, and hierarchically learn image feature representations. This revolutionizes traditional methods that rely on handcrafted features, propelling image recognition technologies to unprecedented heights and becoming the absolute core of modern AI vision.

The basic structure of a CNN consists of an input layer, a convolutional layer, a sampling layer, a fully connected layer, and an output layer. Several convolutional and sampling layers are typically used, alternating between them. Each neuron in the output feature plane of a convolutional layer is locally connected to its input. The corresponding connection weights are then weighted and summed with the local input, plus a bias value, to produce the neuron's input value. This process is equivalent to a convolution process.

A CNN extracts relevant features from an input image, which are then used by one or more fully connected layers to make predictions. To use a CNN for image recognition, it must first be trained on a large dataset of labeled images containing the target object. During training, the CNN learns to associate the extracted features with the correct labels through a process of backpropagation and optimization. Once the CNN is trained, it can be used to make predictions on new, unseen images by feeding the image into the network and selecting the label with the highest predicted probability. [5]

Convolutional layers: These filters process structured, grid-like data by automatically learning hierarchical patterns. These filters detect various features such as edges, textures, and colors by sliding over the image and calculating dot products.

Activation functions: After convolution, activation functions such as ReLU (rectified linear unit) are used to output positive values if the input is positive and zero values if it is negative, helping the neural network to efficiently learn nonlinear relationships in the data.

Pooling layers: Pooling layers provide a downsampling operation that reduces the dimensionality of the feature map, helping to extract the most relevant features.

Fully connected layers: These layers interpret the features extracted by the convolutional and pooling layers to make predictions. These layers are similar to layers in traditional neural networks.[6]

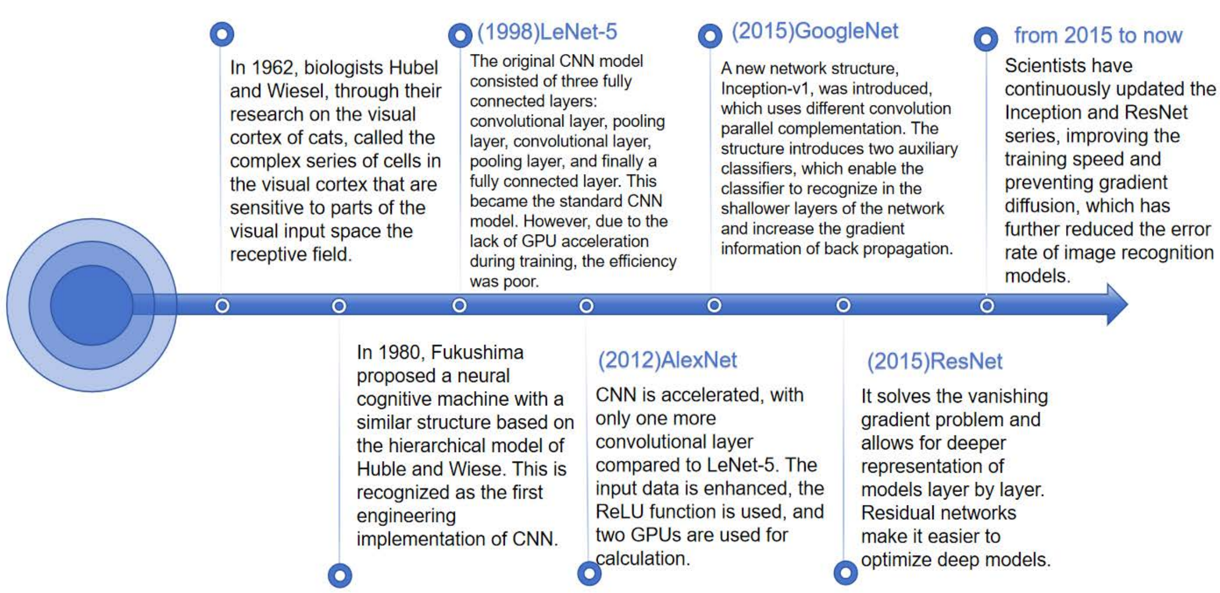

Fig. 1 The evolution of CNN models

Figure 1[7] shows the initial proposal, continuous optimization, and evolution of the CNN model over the past few decades. This illustrates the general trends in CNN development over these decades:

ISSN 2959-6157

The development of convolutional networks has shown a trend toward deeper network structures. For example, AlexNet (2012) boasted 8 convolutional layers, there are five convolutional layers and three fully connected layers,debuted ReLU and Dropout, and reduced the ImageNet error rate from 28.2% to 16.4%. VGG (2014) had 16-19 layers, all stacked with small 3×3 convolution kernels, resulting in fewer parameters and a receptive field equivalent to a larger kernel. ResNet (2015) boasted 152 layers, introducing residual connections (shortcuts) to address the vanishing gradient problem and support ultra-deep training. CNN computational efficiency has also significantly improved. MobileNetV1 boasts a 40% increase in energy efficiency compared to VGG, and LSNet, released in 2024, boasts approximately 300 times greater CPU efficiency.

4. Artificial intelligence chip optimizes convolutional neural network architecture

Artificial intelligence chips can be categorized by architecture into CPUs (general-purpose processors), GPUs, DSPs (digital signal processors), FPGAs, ASICs, and brain-inspired chips. Traditional CPUs and GPUs can both execute AI algorithms, but CPUs have relatively weak floating-point computing capabilities, especially when processing large-scale matrix operations, making them inefficient and unable to meet the computational demands of complex AI models. Furthermore, CPUs have relatively low memory bandwidth, which can lead to slow model parameter reading and affect overall computational efficiency.High-performance GPUs are expensive and require a cooling system, which increases hardware costs. Developing a CPU requires mastering specialized programming frameworks such as CUDA, which has a steep learning curve and is not suitable for novice developers. GPUs from different manufacturers have different compatibility, and software needs to be optimized for specific hardware. CPUs are primarily used in small-scale tasks to take advantage of their multi-core parallel processing capabilities. GPUs are more recommended for large-scale inference tasks, as their high parallel computing capabilities can significantly improve efficiency. [8]

ASIC is the abbreviation for Application-Specific Integrated Circuit, which refers to chips custom-designed for specific tasks. Compared with GPUs and CPUs, it has the characteristics of being specialized, high-performance, and low-power. Its high performance lies in its ability to eliminate redundant logic, and the computational density of ASICs far exceeds that of CPUs and GPUs. The TPUv4's matrix multiplication speed is 30 times faster than that of similar-generation GPUs. Its low power consumption stems from its streamlined instructions, resulting in a power efficiency ratio (TOPS/W) that is 5-10 times that of GPUs.

Table 1: Differences between CPUs, GPUs, and ASICs

Dimensions	CPU	GPU	ASIC
Core Objectives	General logic control	Highly parallel computing	Specific acceleration tasks
Computing Units	A few complex cores	Thousands of simple cores	Custom circuits
Flexibility:	Very high (arbitrary program)	High (supports AI frameworks),	Very low (hardened logic)
Energy Efficiency	0.1-1TOPS/W	10-30TOPS/W	50-1000TOPS/W

In fields like intelligent recognition, deep learning applications like object recognition require extremely fast computational response times. However, while ensuring high performance and efficiency, AI chips must also maintain low power consumption to preserve device battery life. Therefore, GPUs are not the optimal AI chip choice for intelligent recognition technology. Therefore, the development of ASICs (Integrated Design Units) has become inevitable. As shown in Table 2, ASICs are now widely used in high-efficiency intelligent recognition chips from leading chip companies.

Table 2: ASIC Application Examples

Chip Model	CNN Model Support	Performance Comparison (vs. GPU)	Intelligent Recognition Application Scenario
Google TPUv4	EfficientNet-B7	Triple Inference Speed, 60% Energy Consumption Reduction	Google Image Classification
Huawei Ascend 910	YOLOv4	2.1x Increase in Video Analysis FPS	Urban crowd statistics

5. Challenges and Outlook

With the continuous evolution and advancement of intelligent recognition technology and convolutional neural networks, along with the development of AI chips, has played a significant role in continuously optimizing computing performance, efficiency, and energy consumption in the intelligent recognition field. While the computing power of AI chips is rapidly increasing, the development of memory bandwidth and storage speed has lagged behind. These trends overlook an emerging challenge in training and serving AI models: the „memory wall" problem and communication bottlenecks. This is similar to the „data duplication" problem in convolution operations. Convolution itself has a high „computational data reuse" feature, but limited on-chip storage resources and huge external memory access latency and power consumption prevent the chip from fully utilizing this feature, and valuable bandwidth is wasted on moving duplicate data. The computing unit spends most of its time waiting for new data blocks to be loaded[9]. The bandwidth utilization rate of chips like Huawei's Ascend chip is only 38.2%[10]. Challenges still exist in terms of security and privacy risks caused by the lack of hardware protection[11].There are still challenges in various aspects. The training of generative AI requires a large amount of data, which includes processing a large amount of sensitive biometric information and scene data such as faces and fingerprints. [12] Therefore, as the underlying hardware, the security of AI chips is of vital importance. The popularization of generative AI may bring new challenges to privacy and personal data. Future development needs to innovate and modify the model architecture around existing problems. In the future, chips can adopt a more radical integrated memory and computing architecture, embed the computing unit into the memory, and perform calculations directly, which fundamentally reduces data transfer and improves energy efficiency.For multimodal fusion, processors that support hybrid computing can be selected to break through the „memory wall" problem. In terms of security, a data security module can be built into the chip to encrypt the data. During training, the chip can be used to process data locally and only upload gradient updates. Federated learning communication can be optimized to reduce energy consumption. [13], encryption computing technology and other methods can be used to deal with security risks.

The future of artificial intelligence chips will shift from „pure pursuit of computing power" to the coordinated development of „efficiency, intelligence, and security." Only through breakthroughs in algorithms, architecture, circuits, and security technologies can we overcome existing bottlenecks and truly empower the next generation of trusted and reliable intelligent recognition applications. The ultimate goal is to build a scenario-driven vertical optimization system: not only in intelligent recognition, but also in key areas such as medical diagnosis and autonomous driving, through the closed loop of „silicon-based capabilities + algorithm insights", to achieve the leap from „recognition intelligence" to „decision-making intelligence".

6.Conclusion

This article systematically reviews the application of artificial intelligence chips in driving the development of intelligent recognition technology and the development history of related technologies. Research shows that although intelligent recognition algorithms represented by convolutional neural networks (CNNs) have achieved a leap in accuracy, their massive number of parameters and complex computational operations pose a severe performance challenge to traditional computing architectures. This contradiction has become the fundamental driving force behind the development of artificial intelligence chips. The core argument of this article is that AI chips have fundamentally reshaped the computing paradigm of intelligent recognition through the coordinated optimization of hardware and algorithms. Graphics processing units (GPUs) have laid the foundation for modern deep learning with their massively parallel architecture.Application-specific integrated circuits (ASICs) optimize core operations like convolution. These chips utilize parallel computing, low-precision quantization, memory hierarchy optimization, and other key technologies to significantly improve the efficiency of intelligent recognition tasks.

However, this study also reveals key challenges for future development: First, the "memory wall" problem remains the fundamental bottleneck restricting the release of computing power, and the energy consumption and latency of data transfer have become the biggest obstacles to improving energy efficiency; second, as intelligent recognition penetrates into various industries, security and privacy risks are becoming increasingly prominent, and adversarial attacks and side-channel attacks at the hardware level pose new threats.

Looking ahead, the evolution of AI chips will move beyond the single dimension of computing power, toward the coordinated development of "efficiency, intelligence, and security." Breaking through the "memory wall" requires integrated storage and computing architectures or multimodal integration; addressing security challenges necessitates developing hardware-inherent security modules and trusted execution environments. Ultimately, through the deep synergy of algorithms, architectures, circuits, and security technologies, AI chips will continue to serve as core infrastructure, enabling the next generation of ac-

ISSN 2959-6157

cessible, reliable, and autonomous intelligent recognition systems.

References

[1]Yang J, et al. “Research progress on intelligent identification of earthquake faults based on deep learning.” Progress in Geophysics 37.1 2022: 298-311.

[2]Vempati, R, and Lakhan Dev S. “A systematic review on automated human emotion recognition using electroencephalogram signals and artificial intelligence.” Results in Engineering 18 2023: 101027.

[3]Darney, P. Ebby. “A Review on Artificial Intelligence Chip.” Recent Research Reviews Journal 1.1 2022: 99-109.

[4] Li, Z, et al. “A survey of convolutional neural networks: analysis, applications, and prospects.” IEEE transactions on neural networks and learning systems 33.12 2021: 6999-7019.

[5] Arsenov, Andrey, et al. “Evolution of Convolutional Neural Network Architecture in Image Classification Problems.” ITS. 2018.

6] Lu Yy, Sun Xy, Ji Wl, et al. Convolutional neural network accelerator based on “storage and computing integration” Journal of terahertz science and electronic information technology, 2025, 23(2): 170-174.

[7] Zhou Fy, Jin Lp, Dong Jun. A review of convolutional neural network research. Chinese Journal of Computers, 2017, 40(6): 1229-1251.

[8] Shi Yx. A review of artificial intelligence chip technology system research Telecommunications Science, 2019, 35(4): 114-119. DOI: 10.11959/j.issn.1000-0801.2019063.

[9]Gholami, Amir, et al. “Ai and memory wall.” IEEE Micro 44.3 (2024): 33-39.

[10]Kim et al., “Memory-Centric AI System Design”, *ISSCC 2023*

[11]L. Smith, M. Johnson, and A. Gupta, “Secure optical computing for privacy-preserving machine learning,” *Nature Photonics*, vol. 18, pp. 102–115, Mar. 2024, doi: [10.1038/s41566-024-01412-3]

[12] Chen Yw. “Beyond ChatGPT: Opportunities, Risks and Challenges of Generative AI.” Journal of Shandong University (Philosophy and Social Sciences) 3 2023: 127-143.

[13]Qualcomm, “AI 100 Federated Learning Benchmark”, Hot Chips 2023