DOI: 10.54254/2755-2721/86/20241551

A method to shorten the time of high-dimensional matrix operations and convolution operations

Hongfei Zhang

School of Computers and Artificial Intelligence, Wuhan Textile University, Wuhan, 430000, China

2104240202@mail.wtu.edu.cn

Abstract. Serverless computing, or Function-as-a-Service (FaaS), is a cloud computing execution model where the cloud provider dynamically manages the allocation and provisioning of servers which is a new way to solve the problem that the need of client.By serverless computing clients can easily solve their problems anytime,anywhere.However, the complex convolution operations and high-dimensional matrix operations can still consume a lot of resources such as computation time and memory occupancy in serverless computing. In this paper, firstly, the author selects two frequently used models in serverless computing, and then divide them into different numbers of partitions respectively to study the impact of partitioning methods on the computation time and memory occupancy. The author show that different partitioning methods can make an obviously impact on the computation time and memory occupancy,whitch shows that the proper partitioning of high-dimensional matrices and tensors can make a significantly reduction of calculation time and memory occupancy.

Keywords: Serverless, high dimensional matrix, cloud, matrix multiplication Convolution

1. Introduction

In recent years with the development of Big Data, Artificial Intelligence and other industries, serverless computing or Function-as-a-Service (FaaS) has received a lot of attention as an emerging computing model [1]. This emerging computing model is characterized by pay-as-you-go, event-driven lean deployment, and pay-per-use [2].

This study explores the application of cloud computing in various computer-related fields, focusing on high-dimensional matrix operations and convolution operations. The research involves implementing algorithms on the Alibaba cloud platform to assess the time and memory efficiency of partitioned computations. By dividing matrices and tensors into different parts, the study demonstrates how partitioned computing can enhance performance, reduce memory usage, and improve resource utilization, crucial for large-scale data processing and deep learning tasks. The findings highlight the potential of cloud computing to optimize complex computations in fields like machine learning, image processing, and industrial applications.

2. Application scenario analysis

Cloud computing can be applicated in a lot of computer-related fields. Cloud computing offers reliable storage services, enabling users to store and manage large amounts of data, as well as perform data backup and recovery [3]. Virtual server environments can be created through cloud computing platforms, allowing for flexible resource allocation and efficient utilization [4]. It also can be created through cloud computing platforms, allowing for flexible resource allocation and efficient utilization. The Cloud computing provides powerful computing capabilities to support the processing and analysis of large-scale data. Developers can utilize the cloud computing environment for collaborative development, version control, and testing, enhancing development efficiency. Cloud computing provides computing resources for artificial intelligence and machine learning algorithms, accelerating the training and inference processes. Enterprises can deploy applications to the cloud platform to achieve elastic scalability and high availability. Cloud computing supports data analysis and visualization, helping users better understand and gain insights from data. Cloud computing is combined with the IoT to enable the connection, data collection, and remote control of devices and it also offers graphics processing capabilities to support the operation of virtual reality and augmented reality applications. Cloud computing is used to build CDNs, accelerating the distribution and delivery of content.

Cloud computing also has a wide range of applications in industrial development [5]. The first is Industrial enterprises generate a large amount of data, including production data, monitoring data, quality data, etc. Cloud computing provides large-scale data storage and efficient data processing capabilities, which can help enterprises store and analyze this data to support decision-making and business optimization. Second, Cloud computing is combined with the IoT to achieve remote monitoring, control, and management of industrial equipment and systems. By connecting devices to the cloud platform, device data can be collected in real time, and predictive maintenance, fault diagnosis, and performance optimization can be performed. Third, Cloud computing can support collaborative management of the supply chain, enabling information sharing and business collaboration among suppliers, producers, and customers. Through the cloud platform, enterprises can better coordinate all aspects of the supply chain and improve its flexibility and efficiency. The fourth is Cloud computing provides computing resources and data analysis capabilities for intelligent manufacturing, supporting automated production, intelligent scheduling, and optimized production scheduling in factories [6]. At the same time, cloud computing can also promote collaborative innovation between manufacturing enterprises and the design, R & D, and other links. The fifth is Virtual simulation and modeling are often required in the process of industrial design and R & D. Cloud computing can provide high-performance computing resources to accelerate the process of simulation and modeling, and improve the quality and efficiency of product design. Sixth, Cloud computing makes large-scale big data analysis possible [7]. Industrial enterprises can use big data analysis tools on the cloud platform to discover potential business opportunities, optimize production processes, and improve product quality. Seventh Many industrial software and service providers deliver their products through cloud computing platforms, and users can subscribe and use these software and services on-demand, avoiding the complexity of local installation and maintenance. Eighth, Cloud computing provides data security and backup solutions to ensure the confidentiality, integrity, and availability of industrial data. Cloud service providers usually have powerful security measures and data backup mechanisms, reducing the security risks of enterprises. However, complex convolution operations and high-dimensional matrix operations still consume a lot of time in cloud computing. When the model is large, such as in deep learning, the number of convolution operations and high-dimensional matrix operations are very large. So, it is very necessary to reduce the requirement of time and memory.

High-dimensional matrix operation is used for image recognition, classification, and detection tasks, relying on efficient convolution operations in Convolutional Neural Networks (CNNs), they can also be used for processing sequential data, such as natural language processing and speech recognition in Recurrent Neural Networks (RNNs) and the Large-scale Data Training that requires high-dimensional matrix operations on large datasets, with GPU and TPU accelerators significantly improving computation efficiency [8, 9].

Convolution operations are fundamental in various fields such as signal processing, image processing, machine learning, and deep learning [10]. Convolution is a mathematical way that produces a third function by combining two functions, illustrating how one function alters the shape of the other.

Convolution operations are frequently used to Filtering, it is used to apply filters to signals, removing noise or enhancing specific features. In signal transformation, it helps in transforming signals from one domain to another, like from time to frequency domain. In Edge Detection Convolution with edge detection kernels helps identify edges in images. In Blurring and Smoothing Applying convolution with specific kernels can blur or smooth images, useful in reducing noise. In Sharpening Convolution can enhance the edges and fine details in images. In Convolutional Neural Networks (CNNs) Convolution layers in CNNs extract hierarchical features from input data, making them highly effective for tasks like image and speech recognition. In Feature Extraction Convolution operations help in extracting essential features from raw data, facilitating better learning and prediction.

3. System Requirements

3.1. Model Selection

To observe the time and memory usage required for calculations, author use python3.9 to complete the coding and upload it to Alibaba cloud platform. First, the author select two models that frequently be used, one is High-dimensional matrix operation which is fundamental to various advanced applications in computer science and engineering. These operations involve complex manipulations of large matrices, often requiring substantial computational resources, the other is Convolution operations, convolution often be used in image identify, it is a mathematical way that produces a third function by combining two functions. After that, the author designs the algorithm and present it, then the author implements it.

3.1.1. High Dimensional Matrix

High-dimensional matrix is a mathematical object that extends the concept of a matrix in traditional linear algebra to higher dimensions. In a high-dimensional matrix, each element is indexed by multiple indices instead of just two in a standard matrix. These additional indices allow for the representation and manipulation of data with more complexity and structure. High-dimensional matrices are commonly used in various fields such as data analytics, machine learning, computer vision, scientific computing. They can represent data with multiple features or variables, enabling more sophisticated analysis and modeling. Working with high-dimensional matrices often requires specialized algorithms and techniques to handle the increased dimensionality and complexity. Common operations on high-dimensional matrices include matrix multiplication, addition, and decomposition. These operations can be used to extract valuable information, perform classification, regression, or clustering tasks, and build predictive models. In summary, high-dimensional matrices provide a powerful tool for handling and understanding data in complex systems and are essential in many modern applications.

3.1.2. Convolution

Convolution is a fundamental operation in mathematics and has various applications in big data deep learning and other fields. In essence, convolution takes one function, known as the kernel or filter, and slides it over another function. At each point, the two functions are multiplied and then integrated or summed up. The result of this operation is a new function that represents the combined effect of the two input functions. In the context of image processing, convolution is commonly used for tasks such as filtering, feature extraction, and pattern recognition. For example, applying a Gaussian filter can smooth an image by blurring it, while using edge detection filters can highlight the boundaries in an image. In machine learning, convolution is a key component of convolutional neural networks (CNNs), which are widely used for image classification, object detection, and other tasks. The convolution operation helps CNNs learn local patterns and features in the input data, enabling efficient processing and accurate predictions. Overall, convolution is a powerful tool that allows for the extraction of meaningful information from functions and plays a crucial role in many areas of science, engineering, and technology.

3.1.3. Model requirements

Due to the high time complexity of high dimensional matrix operations and the exponential growth of computation as the matrix dimension increases, this can lead to a significant increase in computation time. The author explores the effect of different number of partitions on the computation time of high-dimensional matrices, exploring the time and memory size occupied by different partitioning methods.

3.2. Convolutional operations

Convolutional operations involve a large number of multiplication and accumulation operations, especially in the field of deep learning, where convolutional kernels usually have a large size and depth, which leads to a high time complexity of convolutional operations. The authors have explored the impact of different partitions on convolutional operation time and memory footprint.

4. System Design and Analysis of Experimental Results

4.1. System Design

For high dimensional matrix multiplication as example, the author divides it into 4, 9, 16, 25 partitions and Strassen's algorithm to explore the time required to run on cloud server for the same high dimensional matrix respectively. The author also uses the Strassen’s algorithm for multiplication, Strassen’s algorithm was developed by Volker Strassen in 1969, which is an efficient way to calculate matrix multiplication. It significantly reduces the complexity of multiplying two matrices compared to the conventional algorithm.

For convolutional operations, similarly the author divided it into 4, 6, 8 workers. exploring the time required to run on the server for different workers for the same tensor and convolutional box respectively. Each computation goes through the same steps: Create a high dimensional Matrice or convolution. The author use python to create a High-dimensional matrix(1000x1000) or a convolution. Then upload it to Alibaba cloud. In order to explore the impact of different partitioning methods on high-dimensional matrix operation time, the author design a algorithm to divide them into different parts. Next the author makes them to perform designated calculation like multiplication or addition and then write results back to storage. After that the author respectively observe the time and memory usage required for calculations.

4.2. Analysis of Experimental Results

4.2.1. Data-Analysis

The experiment was divided into multiple sets of different data, when divided into different workers, the final calculation time and memory usage are different. The time and memory usage is shown in figure 1.

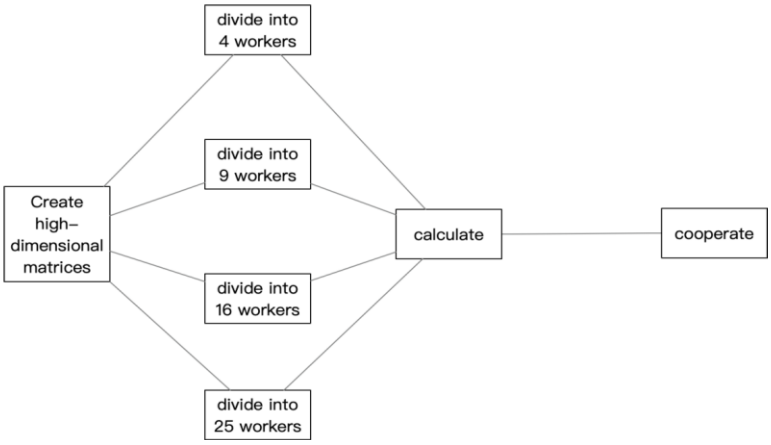

Figure 1. The time and memory usage

When the author divides the high dimensional matrix into 4 parts, the execution time is 1708ms and the memory usage is 136.82MB. When the author divided the high dimensional matrix into 9 parts, the execution time was 1396ms and the memory footprint was 134.93MB. When the author divides the high-dimensional matrix into 16 parts, the execution time is 1255ms and the memory footprint is 137.36MB. When the author partitions the high-dimensional matrix into 25 parts, the execution time is 1268ms and the memory footprint is 139.36MB. The detailed is shown in table 1.

Table 1. Create a high-dimensional matrix and divide it into different numbers of parts

Numbers of part	Time	Memory
4	1708ms	136.82MB
9	1396ms	134.93MB
16	1255ms	137.96MB
25	1268ms	139.36MB

When the author does not partition and instead compute the high-dimensional matrix using Strassen's algorithm, the execution time is 746ms and the memory footprint is 72.02MB. The author found that when the number of workers increases, the consumption of time tends to decrease and the usage of memory tend to decrease, the lowest time consumption occurs when using 16 workers and the lowest memory usage occurs when using 9 workers.

4.2.2. Data-Analysis

Create a tensor and divide it into different umbers of workers and convolute it with kernel is shown in figure 2.

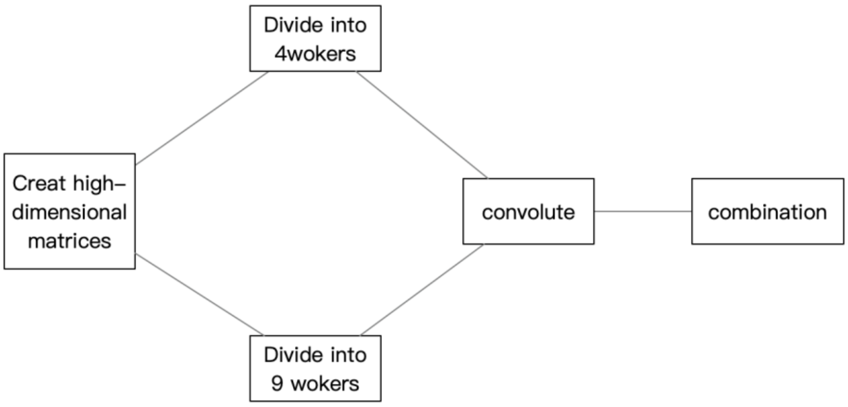

Figure 2. Create a tensor and divide it into different umbers of workers and convolute it with kernel

When the author divides the tensor into 4 parts, the execution time is 430ms and the memory footprint is 47.06MB. When the author divides the tensor into 6 parts, the execution time is 398ms and the memory footprint is 47.6MB. The detailed is shown in table 2.

Table 2. The time and memory usage

Numbers of part	Time	Memory
4	430ms	47.06MB
6	398ms	47.6MB

5. Conclusion

Partitioned computing can be used to speed up computation, which is important for tasks such as large-scale data processing and deep learning. In addition, splitting a matrix or tensor into multiple small blocks for calculation can enable computing tasks to be performed on multiple computers or computing units, improving computing efficiency and processing capabilities. In addition, splitting large matrices or tensors can into small matrices or tensors can reduce memory usage and effectively improve resource utilization, which is very important for resource-constrained devices and systems. Matrix partition operations can not only be used in the fields of linear algebra and numerical calculations, but also in various fields such as image processing and natural language processing to provide support and optimization for more application scenarios.

References

[1] Copik M, Kwasniewski G, Besta M, et al. Sebs: A serverless benchmark suite for function-as-a-service computing. Proceedings of the 22nd International Middleware Conference. 2021: 64-78.

[2] Zhao S, Miao J, Zhao J, et al. A comprehensive and systematic review of the banking systems based on pay-as-you-go payment fashion and cloud computing in the pandemic era. Information Systems and e-Business Management, 2023: 1-29.

[3] Chouhan V, Peddoju S K, Buyya R. dualDup: A secure and reliable cloud storage framework to deduplicate the encrypted data and key. Journal of Information Security and Applications, 2022, 69: 103265.

[4] Chen M, Wang T, Ota K, et al. Intelligent resource allocation management for vehicles network: An A3C learning approach. Computer Communications, 2020, 151: 485-494.

[5] Bello S A, Oyedele L O, Akinade O O, et al. Cloud computing in construction industry: Use cases, benefits and challenges. Automation in Construction, 2021, 122: 103441.

[6] Wang J, Xu C, Zhang J, et al. Big data analytics for intelligent manufacturing systems: A review. Journal of Manufacturing Systems, 2022, 62: 738-752.

[7] Stergiou C L, Plageras A P, Psannis K E, et al. Secure machine learning scenario from big data in cloud computing via internet of things network. Handbook of Computer Networks and Cyber Security: Principles and Paradigms, 2020: 525-554.

[8] Thudumu S, Branch P, Jin J, et al. A comprehensive survey of anomaly detection techniques for high dimensional big data. Journal of Big Data, 2020, 7: 1-30.

[9] Goldberg Y. Neural network methods for natural language processing. Springer Nature, 2022.

[10] Li Z, Liu F, Yang W, et al. A survey of convolutional neural networks: analysis, applications, and prospects. IEEE transactions on neural networks and learning systems, 2021, 33(12): 6999-7019.