Deployment and Optimization Strategy of Machine Learning Model Based on Cloud Computing

Yimeng Lyu

China Coal Technology&Engineering Group, CCTEG Coal Mining Research Institute,100000 Beijing,China

lvyimeng@live.com

***Abstract*—This article aims to explore the deployment strategy and optimization method of ML (Machine Learning) model based on cloud computing, so as to improve the performance of the model in the cloud computing environment, reduce the cost and ensure the security. Aiming at this goal, firstly, the challenges faced by the current ML model deployment are analyzed, including poor scalability, low resource utilization, high cost and security risks. Based on this, a complete set of deployment strategy and optimization system is proposed. The system covers micro-service architecture, distributed training, model quantification and pruning, hardware acceleration, cloud service selection and resource allocation. The results show that micro-service architecture has obvious advantages in scalability and management convenience compared with single architecture. Distributed training technology can significantly shorten the model training time; Reasonable cloud service selection and resource allocation strategies reduce operating costs. The research results verify the effectiveness of the proposed strategy and optimization method, and provide strong technical support for practical application.**

Keywords*—Cloud computing; Machine Learning model; Deployment strategy; Optimization method

I. Introduction

In the digital age, as the infrastructure of information technology, cloud computing is changing all walks of life at an unprecedented speed [1]. Its powerful computing power, flexible resource allocation and high scalability have brought revolutionary changes to the fields of big data processing and high-performance computing [2]. At the same time, the rapid development of ML technology makes it possible to extract valuable information from massive data and make intelligent decisions [3]. ML is a core branch of artificial intelligence. It enables computers to learn and improve from data without explicit programming. According to different learning styles, ML can be divided into supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning [4]. The selection and optimization of ML algorithm is very important for the performance and accuracy of the model, and it is also the basis of subsequent deployment and optimization [5]. However, it is a big challenge to effectively deploy the ML model into the production environment and ensure its efficient and stable operation [6].

Cloud computing is an important support of information technology [7-8]. Cloud computing, with its characteristics of on-demand service, resource pooling, extensive network access and rapid flexibility, has completely changed the traditional way of providing IT services. [9] With its unique advantages, cloud computing platform has become the first choice for ML model deployment. This research is carried out under this background. The research aims to explore how to make full use of the characteristics of cloud computing, realize the efficient deployment and optimization of ML model, and promote the wide application of artificial intelligence technology.

The significance of this study is to put forward a set of systematic deployment and optimization strategies by deeply analyzing the internal mechanism of the combination of cloud computing and ML. This can significantly improve the efficiency and accuracy of the ML model, at the same time effectively reduce the deployment cost and improve the utilization of resources. This is of great significance to promote the industrial application of ML technology and accelerate the intelligent transformation. This research will also help to build a more secure, reliable and extensible ML service system, provide support for data analysis and intelligent decision-making in the cloud computing environment, and promote the high-quality development of the digital economy. The core goal of this study is to design and implement an efficient ML model deployment and optimization strategy based on cloud computing platform, covering the whole process from model selection, preprocessing, deployment architecture design to performance optimization and cost control.

II. Deployment strategy of ML model based on cloud computing

To deploy the ML model in the cloud computing environment, it is necessary to comprehensively consider the training, reasoning, storage and access of the model [10]. Cloud service providers provide rich ML services, such as SageMaker of AWS, AI Platform of GCP and Machine Learning of Azure. These services simplify the process of model deployment and provide all-round support from data preparation, model training to deployment management. With the introduction of containerization technologies such as Docker and Kubernetes, the model can be packaged and deployed in a standardized way, which improves the flexibility and portability of deployment. Micro-service architecture serves the model, which is convenient for expansion and management and meets the needs of different application scenarios.

In the initial stage of ML model deployment based on cloud computing, model selection is very important [11].

This requires comprehensive consideration of the accuracy of the model, training time, reasoning speed and resource consumption according to the needs of specific application scenarios. CNN (Convective Neural Network) stands out in this project because of its excellent performance and wide applicability, and becomes our first choice model. CNN is a deep learning model especially suitable for processing grid data [12]. Through convolution layer, pooling layer and fully connected layer, it can automatically extract features from data and gradually construct a high-level abstract representation. In this project, the framework and parameters of the model are determined according to the requirements of specific application scenarios. By adjusting the number, size and step size of convolution layers, we can effectively control the complexity and calculation of the model while maintaining high accuracy. CNN's convolution layer formula is as follows:

$$o_{i,j}^{l} = \sigma\left(\sum_{m}\sum_{n} I_{i+m,j+n}^{l-1} \cdot K_{m,n}^{l} + b^{l}\right) \quad (1)$$

The output of the pool layer is expressed as:

$$P = pooling(O) \quad (2)$$

Where $o_{i,j}^{l}$ is the output characteristic diagram of the convolution layer of the l layer at position (i, j). $I_{i,j}^{l-1}$ is the value of the input feature map of $l-1$ layer at position (i, j). $K_{m,n}^{l}$ is the weight of the l layer convolution kernel at position (m,n). b^{l} is the bias of the l layer. σ is the activation function. m,n is the index of convolution kernel. $pooling$ is a pooling operation. The hop connection in CNN is shown in Figure 1:

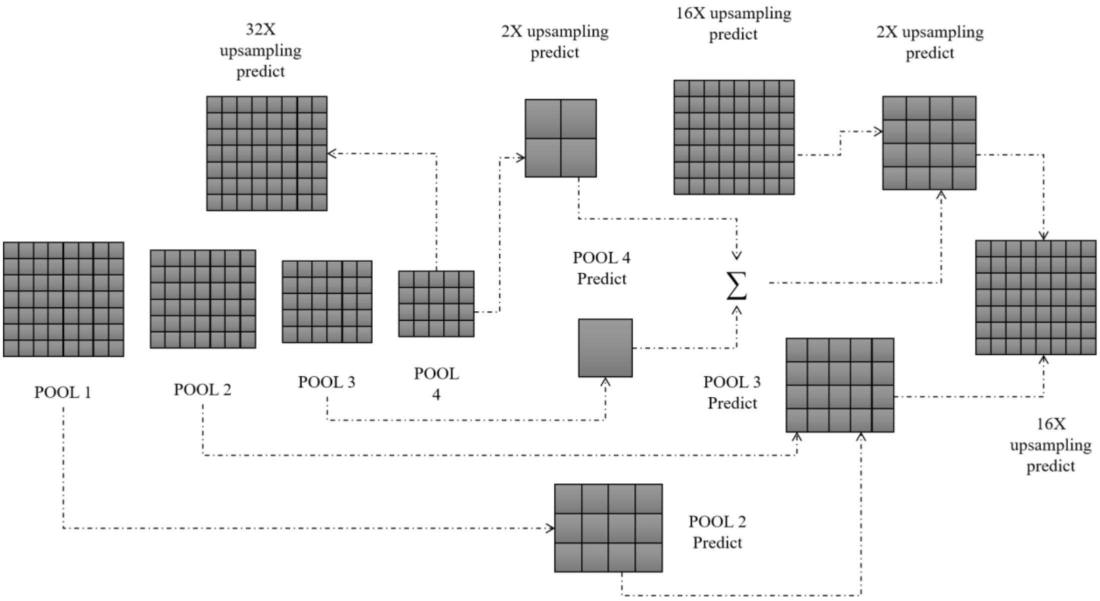

Fig. 1. Jumping connection in CNN

In this article, advanced training techniques are used to speed up the training process of the model and prevent over-fitting. The model training process is as follows:

$$\theta^{*} = \arg\min_{\theta} \sum_{i=1}^{k} Loss\left(D_{val}^{i}; \theta\right) \quad (3)$$

Where θ^{*} is the optimal model parameter found by grid search and cross-validation. k is the number of folds for cross-validation. D_{val}^{i} is the verification set of the i fold.

The volume and reasoning time of the model are further reduced by compressing the model such as quantization and pruning. In order to meet the strict requirements of resource consumption and response speed in the cloud computing environment. The quantization operation is expressed as:

$$Q = round\left(\frac{W}{S} + Z\right) \cdot S \quad (4)$$

Where: Q is the quantized weight. W is the original weight. S is the quantization step size. Z is the quantization zero. $round$ is a rounding operation. Pruning operation is expressed as:

$$W' = W \cdot M \quad (5)$$

Where: W' is the weight after pruning. W is the original weight. M is the pruning mask.

Deployment architecture design is the core of ML model deployment based on cloud computing. A reasonable deployment architecture should be able to support the rapid deployment, flexible expansion and efficient operation of the model [13]. In this article, micro-service architecture is adopted to serve the model, which is convenient for management and expansion. Each model service can be deployed and upgraded independently, which reduces the coupling between systems. Cloud service selection and resource allocation are the key steps of ML model deployment based on cloud computing. When selecting cloud services, factors such as service quality, price, availability, security and technical support should be considered comprehensively. At the same time, according to the specific needs of the model, reasonable allocation of computing resources, storage resources and network resources.

III. Optimization strategy of ML model

A. *Distributed training optimization*

Distributed training optimization is a key means to improve the training efficiency of ML model. In the cloud computing environment, this article uses distributed computing resources to process large-scale data sets in parallel. This can significantly shorten the model training time. To realize distributed training optimization, it is necessary to solve the problems of data parallelism and model parallelism. Data parallelism refers to dividing the data set into multiple subsets, training on different computing nodes respectively, and then summarizing and updating the model parameters. Model parallelism is to divide the model into multiple parts, each part is trained on different nodes, and finally integrated into a complete model. In order to further improve the efficiency of distributed training, this article adopts gradient compression, parameter server and other technologies to reduce communication overhead and improve training speed. At the same time, this article also pays attention to the consistency of data and the synchronization of models to ensure the accuracy and stability of distributed training.

B. *Reasoning performance optimization*

Reasoning performance optimization is the key to improve the response speed and processing ability of ML model in practical application. Reasoning performance optimization mainly includes model quantization, pruning and hardware acceleration. Model quantization is to convert model parameters from high precision (such as 32-bit floating-point numbers) to low precision (such as 8-bit integers). In this way, the model size and calculation amount are reduced and the reasoning speed is improved. Model pruning reduces the complexity of the model and improves the reasoning efficiency by removing unimportant connections or neurons in the model. Hardware acceleration is to use specialized hardware (such as GPU and FPGA) to accelerate the reasoning process of the model. In this article, in the cloud computing environment, the reasoning performance of the model is further improved by combining the reasoning acceleration service provided by the cloud service provider (Inferentia of AWS).

C. *Cost optimization*

Cost optimization is an important part in the deployment of ML model based on cloud computing. Cloud computing services are charged according to usage. Therefore, how to reasonably control the use of resources and reduce operating costs is the main goal of cost optimization. In this article, the cost optimization is realized through several aspects in Table 1:

Table 1: Detailed Cost Optimization Strategies for Machine Learning Model Deployment in the Cloud

Cost Optimization Strategy	Specific Implementation Methods	Expected Outcomes
Resource Allocation and Release on Demand	Dynamically adjust resources based on model load to avoid waste	Reduce unnecessary resource expenditure and lower costs
Utilize Auto-Scaling Features	Set auto-scaling rules to automatically increase or decrease resources based on demand	Improve resource utilization and reduce costs from idle resources
Select Appropriate Cloud Service Types and Configurations	Choose cost-effective cloud service types and configurations based on model requirements	Balance performance and cost, avoiding over-configuration
Optimize Model Training and Inference Processes	Adopt efficient algorithms and frameworks to reduce computation time and resource consumption	Shorten training and inference times, lowering computational costs
Use Data Compression and Preprocessing Techniques	Compress and preprocess data to reduce data transmission and storage costs	Lower data processing costs and improve processing efficiency
Implement Resource Monitoring and Optimization	Real-time monitor resource usage and adjust resource configurations promptly	Ensure efficient resource utilization and avoid waste
Leverage Reserved Instances and Savings Plans	Purchase reserved instances in advance or participate in savings plans to enjoy discounted prices	Reduce long-term usage costs and improve cost-effectiveness
Multi-Cloud Strategy and Supplier Optimization	Choose multiple cloud service providers based on needs to achieve resource complementarity and cost optimization	Obtain more options and reduce risks associated with single suppliers
Adopt Serverless Architecture or Functions-as-a-Service	Utilize serverless architecture or functions-as-a-service to pay only for what is used, avoiding resource idleness	Lower fixed costs and increase flexibility
Regularly Review and Optimize Resource Usage	Periodically check resource usage, identify and optimize inefficient or redundant resource configurations	Continuously optimize costs and improve resource utilization efficiency

D. *Security and privacy protection*

In the deployment of ML model based on cloud computing, security and privacy protection are very important. The data and models in the cloud computing environment may involve users' private information or business secrets, which will have serious consequences if they are leaked or maliciously used. Therefore, it is necessary to take multi-level security measures to protect the security of data and models. This includes data encryption, storage and transmission, access control, identity authentication and other technical means, as well as regular security audit and vulnerability scanning and other management measures. At the same time, we also need to pay attention to the privacy protection of the model to ensure that the user's privacy information is not leaked during the model training and reasoning process. By strengthening security and privacy protection, we can ensure the safe and reliable operation of ML model in cloud computing environment and win the trust and support of users.

IV. Experimental Design and Result Analysis

In order to verify the effectiveness of the ML model deployment strategy and optimization method based on cloud computing, this section designs related experimental schemes. The experiment aims to evaluate the advantages and disadvantages of different deployment strategies and optimization methods by comparing their performance, cost and security. Experimental evaluation indicators include training time, reasoning delay, resource consumption, cost and safety indicators of the model. This section builds an experimental environment, selects the mainstream cloud service providers, and configures the corresponding computing resources, storage resources and network resources. At the same time, we prepared a data set for training and testing, and selected a suitable ML model and framework for the experiment.

In the deployment experiment, according to the deployment strategy proposed above, the ML model is deployed to the cloud computing environment. Single architecture and micro-service architecture are adopted for deployment respectively, and their deployment efficiency, scalability and management convenience are compared. The results are shown in Table 2 and Table 3:

Table 2: Comparison of Deployment Efficiency

Architecture Type	Deployment Time (minutes)	Resource Utilization (%)	Deployment Success Rate (%)
Monolithic	45.8	60.8	90.6
Microservices	30.7	75.4	95.4

Table 3: Comparison of Scalability and Management Convenience

Architecture Type	Scalability Rating (1-10)	Management Convenience Rating (1-10)
Monolithic	6.4	7.8
Microservices	9.1	8.3

Scalability score: Micro-service architecture has a high scalability score of 9 because of its modularity and service characteristics. Due to tight coupling, the single architecture is relatively weak in scalability, with a score of 6.

Management convenience score: Although micro-service architecture has brought higher complexity, management convenience is still better than single architecture through good service governance and monitoring tools, with a score of 8. Because of its simplicity, the single structure is relatively easy to manage, but with the increase of scale, the management difficulty will also increase, with a score of 7.

This article also tests the deployment services of different cloud service providers and evaluates their service quality, price and technical support. The results are shown in Table 4:

Table 4: Comparison of Service Quality, Price, and Technical Support

Cloud Provider	Service Quality Rating (1-10)	Price (CNY/month)	Technical Support Response Time (hours)	Technical Support Satisfaction Rating (1-10)
Microsoft Azure	8.5	500	2	9
Tianyiyun	7.8	400	4	8
Amazon web service	9.0	600	1	9.5

In the optimization experiment, the training, reasoning and cost of ML model are optimized comprehensively. In the aspect of training optimization, distributed training technology is adopted, and the training efficiency under different data parallel and model parallel strategies is compared, as shown in Figure 2:

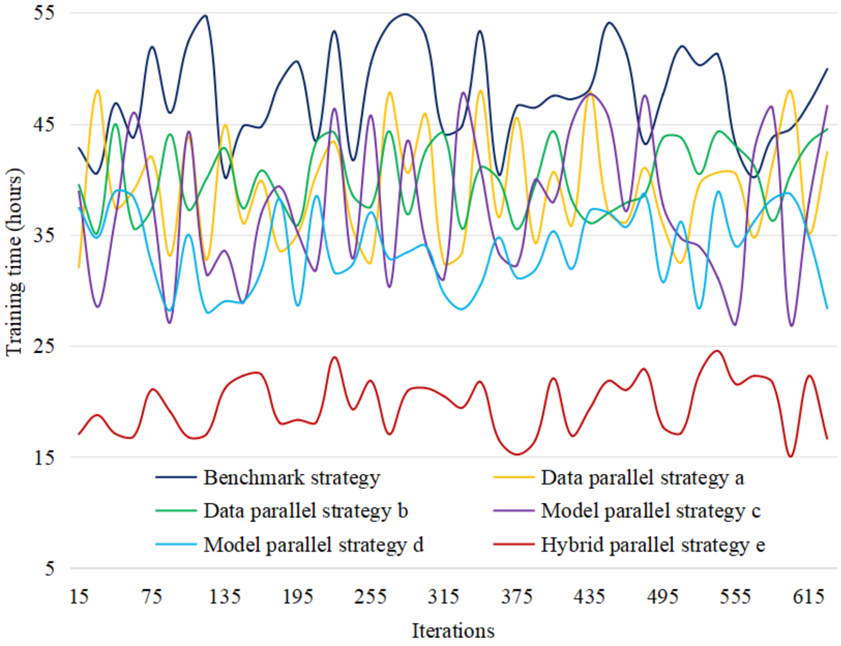

Fig. 2. Training efficiency test

In the aspect of cost optimization, the operating costs under different charging modes are compared, and the control effects of automatic expansion and resource management strategies on costs are evaluated. The results are shown in Table 5:

Table 5: Operational Costs under Different Billing Modes and Strategies

Billing Mode/Strategy	Fixed Cost (CNY/month)	Variable Cost (CNY/hour)	Total Cost (CNY/month)	Cost Savings Ratio (%)
On-Demand Billing (Baseline)	5000	1000 (assumed)	35000	-
Reserved Instance Billing Mode	15000	0	15000	57.14
Auto-Scaling Strategy A	5000	600 (optimized)	23000	34.29
Resource Management Strategy B	5000	700 (optimized)	26000	25.71
Auto-Scaling + Resource Management C	5000	450 (comprehensively optimized)	20500	41.43

The results show that micro-service architecture has obvious advantages in scalability and management convenience compared with single architecture; Distributed training technology can significantly shorten the training time of the model; Reasonable cloud service selection and resource allocation can reduce operating costs. This experiment verifies the effectiveness of the proposed strategy and provides practical guidance for the deployment of ML model based on cloud computing.

V. Conclusions

This study focuses on the deployment strategy and optimization method of ML model based on cloud computing. Through the combination of theoretical analysis and experimental verification, this article puts forward a complete deployment strategy and optimization system, aiming at improving the performance of ML model in cloud computing environment, reducing costs and ensuring security. Deployment strategy: This article compares the advantages and disadvantages of single architecture and micro-service architecture in ML model deployment, and finds that micro-service architecture has significant advantages in scalability, management convenience and service flexibility. At the same time, the paper puts forward the deployment process and specification of ML model based on cloud computing, which provides guidance for practical application.

The research results have significant contribution and influence on practical application. The proposed deployment strategy and optimization method provide technical support for the efficient deployment of ML model in cloud computing environment, and help to improve the operational efficiency and service quality of enterprises. By reducing operating costs and improving performance, this study is helpful to promote the wide application of ML technology and the rapid development of artificial intelligence industry. This study also provides suggestions for enterprises on how to reasonably select and configure cloud services, helping enterprises to make better use of cloud computing resources and realize digital transformation and upgrading. Through continuous research and practice, we expect to provide a more perfect, efficient and safe solution for ML model deployment based on cloud computing.

References

[1] Lu Hongming, Liu Xianfeng, Zhou Zhou, et al. Research on Energy Consumption Model of Cloud Data Center Based on Machine Learning Methods. Journal of Chinese Computer Systems, vol. 44, no. 9, pp. 1966-1973, 2023.

[2] Yu Qun, Shen Zhiheng, Sun Feifei, et al. Differential Privacy Protection Method for Electricity Load Data in Cloud Computing Applications. Electric Power Automation Equipment, vol. 42, no. 7, pp. 68-75, 2022.

[3] Zhang Lei, Cui Yong, Liu Jing, et al. Application of Machine Learning in Cyberspace Security Research. Chinese Journal of Computers, vol. 41, no. 9, pp. 1943-1975, 2018.

[4] Yu Shaofeng, Zhong Jianxu, Xi Lingzhi, et al. Design of Sensor Data Storage and Analysis System Based on Cloud Computing and Big Data Technology. Electronic Design Engineering, vol. 32, no. 18, pp. 105-109, 2024.

[5] Yu Shaofeng, She Jun, Zhong Jianxu, et al. Simulation of Cloud Monitoring Data Analysis Integrating Machine Learning. Automation Instrumentation, vol. 43, no. 3, pp. 75-78, 2022.

[6] Chen Si. Research on Naive Bayes Secure Classification Outsourcing Scheme in Cloud Computing Environment. Computer Applications and Software, vol. 37, no. 7, pp. 275-280, 2020.

[7] Deng Yi. Design and Implementation of Multimedia Teaching Platform Based on Cloud Computing. Electronic Design Engineering, vol. 26, no. 8, pp. 162-167, 2018.

[8] Liu Junxian, Wu Xinlong. Research on Multi-user Network Isolation Security Testing in Cloud Computing Centers. Microcomputer Applications, vol. 36, no. 9, pp. 156-159, 2020.

[9] Li Haixia, Song Danlei, Kong Jianing, et al. Evaluation of Hyperparameter Optimization Techniques for Traditional Machine Learning Models. Computer Science, vol. 51, no. 8, pp. 242-255, 2024.

[10] Jian Chengfeng, Kuang Xiang, Zhang Meiyu. An Improved Learning Model for Cloud Computing Group Optimization with Enhanced Timeliness. Computer Science, vol. 46, no. 5, pp. 290-297, 2019.

[11] Shan Zidan, Zou Ying, Li Yunzhu. Process Optimization and Decision Model for Service-Oriented Manufacturing Networks Based on Cloud Computing. Computer Integrated Manufacturing Systems, vol. 25, no. 12, pp. 3139-3148, 2019.

[12] Wang Bo, Xu Jing, Sun Xueying. Cloud Computing Resource Scheduling Model Based on Ant Colony Optimization Algorithm. Computer & Digital Engineering, vol. 48, no. 5, pp. 1009-1012, 2020.

[13] Zhai Ling, Shen Si, Cheng Shixing. Simulation of Optimal Balanced Allocation of Electronic Information Resources on Cloud Computing Platforms. Computer Simulation, vol. 36, no. 7, pp. 397-400, 2019.