Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/57/20241348

Dynamic resource allocation for virtual machine migration optimization using machine learning

Yulu Gong[1], Jiaxin Huang[2], Bo Liu[3,*], Jingyu Xu[4], Binbin Wu[5], Yifan Zhang[6]

[1]Computer & Information Technology, Northern Arizona University, Flagstaff, AZ, USA
[2]Information Studies, Trine University, Phoenix USA
[3]Software Engineering, Zhejiang University, HangZhou China
[4]Computer Information Technology, Independent Researcher, Flagstaff, AZ, USA
[5]Computer Network Engineering, Cisco Systems, Beijing, China
[6]Executive Master of Business Administration, Amazon Connect Technology Services (Beijing) Co. Ltd., Xi'an Shaanxi China

*Corresponding author: yg486@nau.edu

Abstract. This article delves into the importance of applying machine learning and deep reinforcement learning techniques in cloud resource management and virtual machine migration optimization, highlighting the role of these advanced technologies in dealing with the dynamic changes and complexities of cloud computing environments. Through environment modeling, policy learning, and adaptive enhancement, machine learning methods, especially deep reinforcement learning, provide effective solutions for dynamic resource allocation and virtual intelligence migration. These technologies can help cloud service providers improve resource utilization, reduce energy consumption, and improve service reliability and performance. Effective strategies include simplifying state space and action space, reward shaping, model lightweight and acceleration, and accelerating the learning process through transfer learning and meta-learning techniques. With the continuous progress of machine learning and deep reinforcement learning technologies, combined with the rapid development of cloud computing technology, it is expected that the application of these technologies in cloud resource management and virtual machine migration optimization will be more extensive and in-depth. Researchers will continue to explore more efficient algorithms and models to further improve the accuracy and efficiency of decision making. In addition, with the integration of edge computing, Internet of Things and other technologies, cloud computing resource management will face more new challenges and opportunities, and the application scope and depth of machine learning and deep reinforcement learning technology will also expand, opening new possibilities for building a more intelligent, efficient and reliable cloud computing service system.

Keywords: Cloud computing migration technology, Virtualization, Machine learning-based optimization, Dynamic resource allocation.

1. Introduction

In today's digital age, cloud computing technology has become indispensable for enterprises, providing them with flexibility, efficiency, and scalability. However, the complexity of managing virtual machines in cloud environments presents significant challenges. Traditional static methods for resource allocation and virtual machine migration are inadequate in dynamically changing environments, resulting in resource wastage and degraded performance. Therefore, leveraging advanced techniques such as machine learning is crucial for achieving dynamic resource allocation and optimizing virtual machine migration in cloud computing[1]. Virtual Machine migration plays a vital role in cloud computing, facilitating dynamic resource adjustments, load balancing, and fault tolerance without disrupting user services. Despite its benefits, VM migration encounters challenges such as performance loss, lengthy migration times, and data consistency issues. Machine learning offers a potential solution by analyzing historical data to adaptively allocate resources and intelligently select migration solutions based on real-time monitoring, mitigating performance losses and data consistency issues. [2] In terms of virtual machine migration, machine learning technology can use real-time monitoring data and predictive models to intelligently select the best migration solution, reducing performance losses and data consistency issues during migration. Therefore, dynamic resource allocation and virtual machine migration optimization using machine learning technology can better adapt to the dynamic changes of the cloud environment and improve the performance and reliability of the system.

2. Background and related work

2.1. VM Migration in the Cloud Computing Environment

Virtual machine migration, facilitated by virtualization technology, entails the seamless transfer of running VM instances from one physical server to another within the cloud environment. This technique optimizes resource utilization, enhances fault tolerance, and facilitates load balancing across servers. The article "Performance Framework for Virtual Machine Migration in Cloud Computing" by Tahir Alyas et al. (2022) discusses the significance of virtual machine (VM) migration in cloud computing environments and proposes a performance framework to address related challenges[3]. VM migration plays a crucial role in enabling dynamic resource adjustment, load balancing, and fault tolerance in cloud environments. The framework aims to optimize VM migration processes by considering factors such as performance loss, migration time, and data consistency issues. By leveraging machine learning techniques and real-time monitoring data, the framework intelligently selects migration solutions to enhance system performance and reliability. Through this research, Alyas et al. contribute to the advancement of dynamic resource allocation and optimization strategies in cloud computing environments, addressing the growing complexity of managing virtualized infrastructures. VM migration in the cloud computing environment is a fundamental technique for optimizing resource utilization, enhancing fault tolerance, and ensuring scalability. By understanding and addressing the challenges associated with VM migration, cloud providers can effectively leverage this technique to deliver seamless and efficient services to their users.

2.2. Resource allocation in the cloud computing environment

In cloud computing, efficient resource allocation is essential. Users make sequential requests for cloud resources, which are limited in the cloud computing center. Tran et al. (2022) explore resource allocation strategies, particularly for multi-tier applications, using the Q-learning algorithm. Their research aims to optimize virtual machine migration policies, enhancing performance. Dynamic resource allocation, facilitated by machine learning techniques like Q-learning, improves system efficiency, scalability, and cost-effectiveness.[4] This study highlights the significance of adapting resource allocation to dynamic workload changes, optimizing cloud infrastructure performance for modern applications.

For high availability, each VM must reside in a distinct fault domain to prevent economic losses stemming from cloud computing center failures. This study introduces a cloud resource allocation approach based on deep reinforcement learning, considering user requirements. It prioritizes assigning users to nearby servers to enhance service quality[5]. However, such proximity-based allocation might congest servers, leading to prolonged waiting times. Thus, the study proposes a cloud resource allocation method based on deep reinforcement learning. Deep reinforcement learning dynamically allocates resources based on the system's current state, maximizing cloud resource utilization efficiency and reducing user waiting times.

2.3. VM Migration Optimization and machine learning in Resource Management

VM migration optimization in resource management is a complex problem, which involves many aspects such as resource allocation, load balancing, and power consumption optimization. Machine learning (ML) methods can help optimize the virtual machine migration decision process by predicting loads, resource requirements, and so on to develop migration strategies. Here, we can explore a simplified model that uses machine learning to optimize virtual machine migration.

(1) Problem definition

Let's say our goal is to minimize the energy consumption of the entire data center while ensuring that the performance of all virtual machines (VMs) is not affected. We can achieve this by intelligently migrating VMs to different physical machines (PMs)[6].

(2) Feature selection

In order to train a machine learning model, we need to choose the right features. These characteristics may include:

Table 1. Table of optimized resource management indicators for VM migration

Feature	Description
CPU usage of the VM	The percentage of CPU resources being used by the VM.
Memory usage of the VM	The amount of memory resources being used by the VM.
Storage I/O usage of the VM	The rate at which data is being read from and written to storage by the VM.
Network I/O usage of the VM	The rate at which data is being sent and received over the network by the VM.
Remaining CPU capacity of the PM	The amount of CPU resources that are not currently being used by any VM on the physical machine (PM).
Remaining memory capacity of the PM	The amount of memory resources that are not currently being used by any VM on the PM.
Energy consumption of PM	The amount of electrical power being consumed by the PM.

Table 1 lists the key characteristics and metrics used for resource management and decision making during VM migration optimization. These characteristics and metrics are critical to understanding and evaluating virtual machine (VM) and physical machine (PM) performance, resource utilization, and power consumption[7]. By analyzing this data, you can help optimize resource allocation in the data center, improve energy efficiency, and ensure that the performance of the service is not affected.

(3) Model selection

To solve this problem, we can consider using reinforcement learning (RL), especially Q-Learning or Deep reinforcement learning (DRL), such as Deep Q-Networks (DQN). These models can learn optimal migration strategies by interacting with the environment.

(4) Practical application

In a practical application, we need a more detailed environment model to simulate the dynamic interaction of VM and PM. In addition, the training of the model may require a large amount of data and computational resources. Depending on the specific needs, we may also need to consider using other ML models, such as decision trees, support vector machines, or neural networks, to predict the VM's resource requirements. Therefore, through continuous iteration and optimization, machine learning models can help us manage resources more efficiently and optimize the migration strategy of virtual machines to achieve the goal of reducing energy consumption and maintaining service performance.

3. Vm resource optimization methodology

The essence of the online migration technology is that VMS can be migrated from one physical machine to another without stopping. Create VMS with the same configuration on the target physical machine, migrate all kinds of data, and quickly switch to a new VM on the target physical machine. During the migration process, user VMS can run normally for most of the time, and the last phase of the switchover process is very short, which does not interrupt user services and has little impact on user services running on VMS. This paper divides the online migration process into three stages: preparation stage, migration stage and switching stage. In the field of virtual machine migration optimization, the application of machine learning is gradually becoming a hot research spot, because it can help automate the optimization of resource allocation and migration decisions to improve the efficiency and performance of cloud computing environments. Here are some of the more innovative use cases that demonstrate the use of machine learning techniques in optimizing virtual machine migration:

3.1. Optimization of resource demand based on prediction

In this application case, the machine learning model is used to predict the future resource requirements (such as CPU, memory, etc.) of the virtual machine. By forecasting, the system can migrate VMS in advance to ensure effective resource allocation and avoid resource excess or insufficiency. This predictive approach can significantly reduce performance degradation due to resource reallocation and improve the overall efficiency of the cloud environment. Typically, a time series prediction model, such as a long short-term memory network (LSTM) or gated cycle unit (GRU), can be used to predict a virtual machine's future resource requirements. The model diagram is as follows:

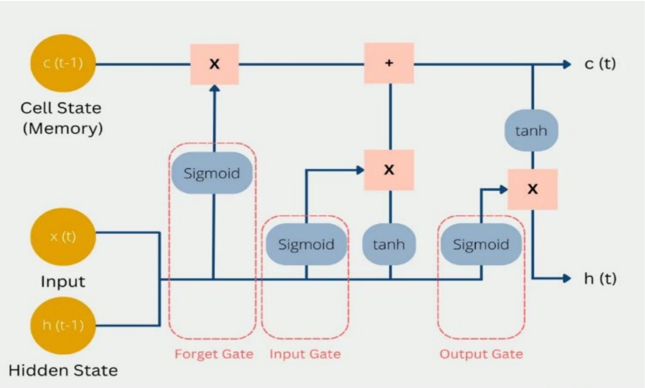

Figure 1. LSTM cloud computing resource optimization model

The LSTM (Long Short Term Memory) units address the gradient disappearance or explosion issue in traditional recurrent neural networks (RNNs) when handling long sequences. They maintain long-term memory through specialized structural units, including Forget, Input, and Output Gates, along with a Cell State. These equations outline the internal mechanics of LSTM units, pivotal in learning patterns and dependencies within sequential data like historical CPU and memory usage. In virtual machine resource demand forecasting, LSTM units are essential for accurately predicting future resource requirements.Let Xt be the input feature vector of time tt (for example, historical CPU and memory usage) and Yt be the predicted output of time tt (i.e. future resource requirements). LSTM units can be described by the following formula:

$$ft=\sigma(Wf\cdot[ht-1,Xt]+bf)ft=\sigma(Wf\cdot[ht-1,Xt]+bf) \tag{1}$$

$$it=\sigma(Wi\cdot[ht-1,Xt]+bi)it=\sigma(Wi\cdot[ht-1,Xt]+bi) \tag{2}$$

$$ot=\sigma(Wo\cdot[ht-1,Xt]+bo)ot=\sigma(Wo\cdot[ht-1,Xt]+bo) \tag{3}$$

$$C\sim t=\tanh\boxed{f_0}(WC\cdot[ht-1,Xt]+bC)C\sim t=\tanh(WC\cdot[ht-1,Xt]+bC) \tag{4}$$

$$Ct=ft*Ct-1+it*C\sim tCt=ft*Ct-1+it*C\sim t \tag{5}$$

$$ht=ot*\tanh(Ct)ht=ot*\tanh(Ct) \tag{6}$$

Here's how each component of the LSTM unit contributes to the overall function:

Forget Gate (ft): Determines which information from the previous cell state should be retained or forgotten, based on the current input and the previous hidden state. It helps the LSTM unit to selectively remember or forget information from previous time steps.

Input Gate (it): Decides which new information to store in the cell state based on the current input and the previous hidden state. It regulates the flow of new information into the cell state.

Output Gate (ot): Controls the information to be output from the cell state based on the current input and the previous hidden state. It helps the LSTM unit to decide what information to pass on to the next time step.

Candidate Cell State (C~t): Computes the candidate values that could be added to the cell state based on the current input and the previous hidden state. It calculates a potential update to the cell state based on the current input.

Cell State (Ct): Updates the cell state by combining information from the forget gate, input gate, and candidate cell state. It retains long-term dependencies by selectively updating and maintaining information over multiple time steps.

Hidden State (ht): Produces the output of the LSTM unit based on the current cell state passed through the output gate. It captures the relevant information from the cell state to generate predictions or representations of the input sequence.

Together, these components enable the LSTM unit to learn and remember long-term dependencies in sequential data, making it particularly effective for tasks such as time series forecasting, including virtual machine resource demand forecasting in cloud computing environments.

3.2. Virtual machine resource optimization experiment

Machine learning has the following advantages in optimizing virtual machine resources:

1) Machine learning models can take advantage of the elastic resources provided by virtual machines, such as automatically scaling and scaling computing resources to meet the demands of different workloads. This flexibility can help optimize resource utilization and improve system performance and efficiency.

2) In addition, machine learning models can take advantage of the diverse hardware resources provided by virtual machines, such as GPU and TPU accelerators, to accelerate the model training and inference process. By leveraging these dedicated hardware resources, the performance and throughput of machine learning models can be significantly improved, enabling faster training tasks and real-time inference services.

In this paper, the function and advantages of machine learning in virtual machine resource optimization are deeply discussed through the experiment of resource optimization on virtual machine. These experimental results will help us better understand the application of machine learning in the optimization of virtual machine resources, and provide guidance and enlightenment for the future development of virtualization technology and machine learning algorithms.

3.3. Data set preprocessing

This dataset contains questions, answer options, and relevant contextual information, and is commonly used for training and validation of machine learning models to solve question-answering tasks. These data sources may be collected from the Web or constructed by humans so that the model learns to choose the right answer options based on the question and context.

Table 2. Virtual machine question-answer dataset

	rompt	context	A	B	C	D	E	
0	What is genetic admixture?	Genetic admixture may have an important role f...	Genetic admixture is the study of the relation...	Genetic admixture is the study of the correlat...	Genetic admixture is a method used to correlat...	Genetic admixture is the process of mixing gen...	Genetic admixture is the study of populations ...	C
1	How does Kamen Rider Den-O defeat the Piggies ...	After being knocked away by Zeronos, the Piggi...	Kamen Rider Den-O transforms into Climax Form ...	Kamen Rider Den-O uses his Extreme Slash attac...	Kamen Rider Den-O gains control of the DenLine...	Kamen Rider Den-O transforms into Sword Form a...	Kamen Rider Den-O receives assistance from Kam...	B
2	Who was Mina Minovici's mentor during his fore...	Mina Minovici (; April 30, 1858 – April 25, 19...	Paul Brouardel	Ștefan Minovici	Nicolae Minovici	Carol Davila	Eforie Civilian Hospitals	A
3	What is the primary function of the blood-brai...	In its neuroprotective role, the blood–brain b...	To facilitate the delivery of therapeutic agen...	To enhance the delivery of therapeutic agents ...	To selectively allow the passage of therapeuti...	To prevent the passage of any diagnostic and t...	To restrict the flow of potentially important ...	C
4	What is the definition of scalability in an ec...	In industrial engineering and manufacturing, s...	The property of a system to handle a growing a...	The ability of a business to increase sales wi...	The characteristic of computers and networks t...	The closure under scalar multiplication in mat...	The capacity of a process, system, or organiza...	B

For virtual machine resource optimization experiments, the advantage of this dataset is that it provides a specific task scenario that can simulate a real question and answer scenario in a virtual machine environment. By conducting experiments on such data sets, we can evaluate the performance of machine learning models on question answering tasks and explore how to optimize virtual machine resources to improve the model's training efficiency and reasoning speed. In addition, this data set also provides a wealth of question and context information, which helps the model to better understand the background and context of the question, so as to improve the accuracy and reliability of the answer selection.

Preprocessing is the conversion of data into a format acceptable to the model. Because it uses a pre-trained tokenizer to convert text and answers into token IDs, truncated or populated to meet the input requirements of the model. Second, it generates input-output pairs where the input is tokenized text and the output is tokenized answers. Finally, it removes columns other than input and output from the data to optimize memory usage.

Next, the code is optimized with virtual machine resources and distributed training. It uses PyTorch/XLA for model training using Google Cloud TPUs. On each TPU core, it initializes the device and sets up data loaders, samplers, and data processing functions to enable distributed training between[8] TPU cores. It loads the pre-trained model and moves it onto the TPU device, then freezes most of the layers and sets up the optimizer and learning rate scheduler. Finally, it conducts the specified number of epochs and steps model training, and outputs the loss and learning rate during the training process.

3.4. Result analysis and interpretation

This paper presents the importance of virtual machine resource optimization and discusses the key role of online migration technology in this field. With online migration technology, VMS can be migrated from one physical machine to another without stopping, realizing dynamic resource adjustment and troubleshooting. This paper divides the process of online migration into preparation stage, migration stage and switching stage, and points out the importance and characteristics of each stage. [9] The principle of long short term memory network (LSTM) in resource demand prediction and the application and advantages of machine learning in virtual machine resource optimization experiment are introduced in detail. Finally, the paper emphasizes that the experimental results can better understand and evaluate the role of machine learning in the optimization of virtual machine resources, and provide guidance and enlightenment for the future development of virtualization technology and machine learning algorithms.

4. Conclusion

This article delves into the utilization of machine learning and deep reinforcement learning (DRL) techniques in cloud resource management and virtual machine migration optimization, emphasizing their significance in navigating the dynamic changes and complexities of cloud computing environments. Through environment modeling, policy learning, and adaptive capability enhancement, machine learning methods, particularly DRL, offer effective solutions for dynamic resource allocation and intelligent virtual machine migration. These technologies aid cloud service providers in enhancing resource utilization, reducing power consumption, and improving service reliability and performance [10-12]. It is anticipated that the application of these technologies in cloud resource management and virtual machine migration optimization will become more widespread and profound. Researchers will continue to explore more efficient algorithms and models to enhance decision-making accuracy and efficiency further. Additionally, with the integration of edge computing, the Internet of Things, and other technologies, cloud computing resource management will encounter new challenges and opportunities. The application scope and depth of machine learning and DRL technology will expand accordingly, paving the way for the establishment of a more intelligent, efficient, and reliable cloud computing service ecosystem.

References

[1] Arianyan E, Taheri H, Sharifian S. Novel heuristies for consolidation of virtual machines in cloud datacenters using multi-criteria resource management solutions!J. The Journal of Supercomputing, 2016, 72(2).688-717

[2] Zhang X, Zhao Y, Guo S, et al. Performance-Aware Energy-efficient Virtual Machine Placement in CloudData Center[C]. 2017 lEEE International Conference on Communications, 2017: 1-7.

[3] Liu, Bo, et al. "Integration and Performance Analysis of Artificial Intelligence and Computer Vision Based on Deep Learning Algorithms." arXiv preprint arXiv:2312.12872 (2023).

[4] Ye K, Wu Z, Wang C, et al. Profiling-Based Workload Consolidation and Migration in Virtualized DataCenters[J]. lEEE Transactions on Parallel and Distributed Systems, 2015, 26(3): 878-890.

[5] Liu, Bo, et al. "Integration and Performance Analysis of Artificial Intelligence and Computer Vision Based on Deep Learning Algorithms." arXiv preprint arXiv:2312.12872 (2023).

[6] Alharbi F, Tain Y C, Tang M, et al. Profile-based static virtual machine placement for energy-efficient datacenter[C]. 2016 lEEE 18th International Conference on High Performance Computing and Communications.2016:1045-1052.

[7] Liu, B., Zhao, X., Hu, H., Lin, Q., & Huang, J. (2023). Detection of Esophageal Cancer Lesions Based on CBAM Faster R-CNN. Journal of Theory and Practice of Engineering Science, 3(12), 36–42. https://doi.org/10.53469/jtpes.2023.03(12).06

[8] Uddin M, Shah A, Alsagour R, et al. Measuring Efficiency of Tier Level Data Centers to lmplement GreenEnergy Efficient Data Centers[J]. Middle East Journal of Scientific Research, 2013, 15(2): 200-207

[9] Hajam Shahid Sultan,and Sofi Shabir Ahmad. "Resource management in fog computing using greedy and semi-greedy spider monkey optimization." Soft Computing 27. 24 (2023):

[10] Qilin Zhou Lili Wang,and Shoulin Wu. "Resource management optimisation for federated learning-enabled multi-access edge computing in internet of vehicles." International Journal of Sensor Networks 42. 1 (2023):

[11] Chen Yifan, et al. "DRJOA: intelligent resource management optimization through deep reinforcement learning approach in edge computing." Cluster Computing 26. 5 (2022):