

SOME PROBLEMS IN MANAGING SERVER COMPUTING RESOURCES USING DEEP MACHINE LEARNING TOOLS

O.Khoshaba
Oleksandr.Khoshaba@gmail.com
Vinnitsia National Technical University

Annotation. The paper examines some problems managing server computing resources using deep machine learning tools and shows the relevance of studying the problems and possible solutions. Particular attention is paid to the consideration of deep machine learning tools.

Formulation of the problem. The main goal of this work is to consider some problems of managing server computing resources using deep machine learning tools, to determine the relevance of studying them, and to suggest possible ways to solve them. Special attention should be paid to the consideration of deep machine learning tools.

Introduction. Managing server computing resources has become increasingly important due to growing data volumes, computational complexity, and the need for automation in infrastructure management. Organizations are seeking ways to improve the efficiency of their computing resources while adapting to changing conditions. This includes predicting process execution, improving the quality of service, and staying competitive through innovation and learning new technologies. Deep machine learning tools have emerged as a powerful solution to these challenges.

The relevance of solving server computing resource management problems using deep machine learning tools has been increasing due to growing data volumes, computational complexity, and the need to increase computing resource efficiency. By automating infrastructure management in organizations, systems can be adapted to changing conditions, improve the quality of service, and predict process execution. Learning new technologies can also foster innovation and competitiveness.

Deep learning tools. As data volumes and computational complexity increase, the need for more efficient resource allocation arises. Deep learning helps predict workloads and optimize resource allocation in real-time. Deep learning tools analyze large data sets, identify resource inefficiencies, and provide suggestions to optimize them. This can reduce costs and improve server efficiency.

Deep learning can automate various infrastructure management tasks, such as monitoring, scaling, and disaster recovery, thereby reducing administrators' burden and improving system reliability. Deep learning systems can adapt to changes in workload and application requirements, ensuring optimal resource allocation despite uncertainty.

Deep learning improves service availability and performance by using predictive capabilities and automation to impact the end-user experience positively. Innovative

companies adopt deep learning for computing resource management to drive technological progress and increase competitiveness.

Considering these factors, using deep machine learning in managing server computing resources is a promising area with significant potential for development and implementation in various industries and fields of activity.

Problems of managing server computing resources using deep machine learning tools. While managing server computing resources using deep machine learning tools offers promising opportunities for optimization and automation, it also comes with challenges (Figure 1). Let us briefly describe some of them.

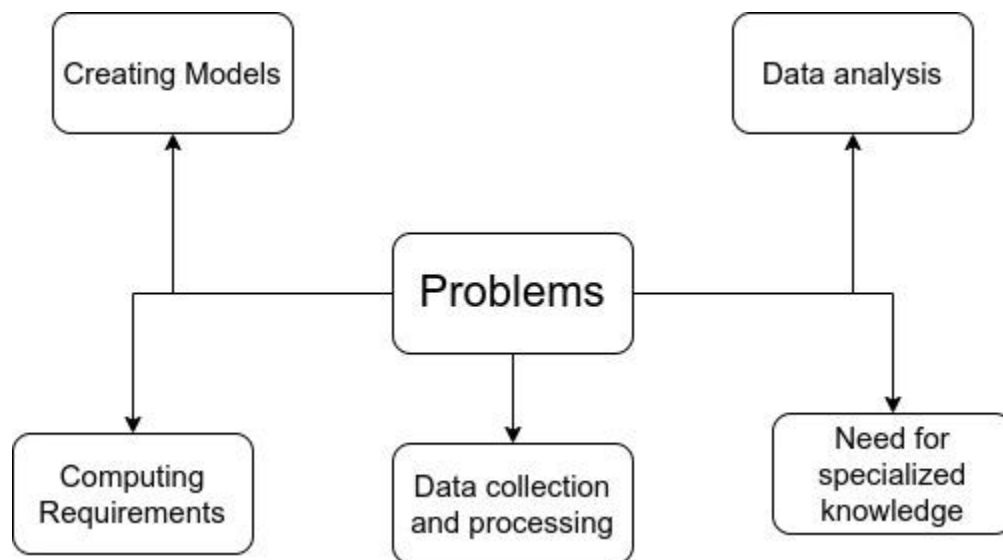


Figure 1. Some challenges of managing server computing resources using deep machine learning tools

Deep learning requires significant computing power to train models, which can result in high costs and the need for specialized hardware. Practical training of deep machine learning models requires large volumes of high-quality data. Collecting, cleaning, and preparing this data can be time-consuming and expensive.

Deep learning can be unpredictable, and models sometimes make incorrect predictions due to overfitting, undertraining, or biased data. This can lead to misallocation of resources. Deep neural networks are often criticized for their "black box" nature, where it is difficult to understand how a particular decision was made.

This makes diagnosing and resolving problems in the resource management system difficult. Models can be trained on data that does not fully reflect real-world operating conditions, reducing their effectiveness in dynamic environments. Developing, implementing, and maintaining deep learning-based systems requires specialized machine learning knowledge and skills, which can be a barrier for some organizations.

Deep Learning is a subfield of machine learning that uses neural networks with many layers (deep neural networks) to model complex abstractions in data. This allows you to process and interpret large volumes of data, find patterns, and make predictions

or decisions based on them. Deep machine learning has applications in various areas, from automatic speech and image recognition to autonomous control systems.

Various tools and libraries offer features and optimizations for training and developing deep learning models. These tools and libraries are used in research projects that require creating services based on artificial intelligence. The choice of a specific tool depends on the specifics of the task, the required performance, ease of use, and other factors.

The concept of server computing resources. Server computing resources are hardware and software components that provide the necessary power to run programs, process data, and service network requests. They include the following key elements.

The central processing unit executes software instructions, processing basic operations such as arithmetic, logic, control, and input/output. CPU power determines the server's ability to process tasks. Some specialized servers, such as those for machine learning or video processing tasks, use GPUs specialized for parallel computing. GPUs significantly impact the performance of deep machine learning tools. Some GPUs are explicitly developed for artificial intelligence and machine learning tasks. For example, NVIDIA has developed a Tensor Core architecture optimized to speed up neural network computing. These specialized architectures allow for even faster training and model inference. GPUs have high memory bandwidth, which allows large amounts of data to be transferred quickly between memory and the processor. This reduces processing latency and improves overall performance when training and using deep learning models. Although GPUs consume significant power while running, they can be more energy efficient than CPUs when performing deep learning tasks due to their higher performance and ability to process more operations per unit time.

Random access memory (RAM), which is used for temporary storage of data and programs, is very important for servers. More RAM allows you to handle more tasks simultaneously without accessing slower drives.

Hard disk storage is designed for long-term storage of data and programs. Solid-state drives (SSDs) provide faster read and write speeds than traditional hard drives. Network adapters enable the server to connect to the network and exchange data with other devices and users via local and global networks. Effective management of these resources is the basis for ensuring high availability, reliability, and performance of servers and the services they provide. Deep machine learning can play a significant role in optimizing the use of these resources by predicting workloads and adapting systems to changing operating conditions.

Conclusions. Despite these challenges, the benefits of using deep machine learning to manage server computing resources far outweigh its challenges as technology continually advances and methods for training and optimizing models improve. Various solutions and strategies have been developed to address the challenges of managing server computing resources using deep machine learning tools. Unfortunately, determining which are the most effective is beyond the scope of this work.