

РЕФЕРАТ

Актуальність теми. Актуальність теми "Оптимізація використання обчислювальних ресурсів у операціях машинного навчання" не може бути переоцінена у сучасному світі швидкого розвитку технологій. У контексті стрімкого зростання обсягів даних та постійного ускладнення моделей машинного навчання, виникає нагальна потреба в ефективному використанні обчислювальних ресурсів. Оптимізація цих ресурсів є ключовим елементом для досягнення високої продуктивності, точності та економічної ефективності в процесах машинного навчання.

З огляду на високі вимоги до обчислювальної потужності, необхідної для тренування та використання моделей машинного навчання, особливо актуальним стає питання оптимізації використання ресурсів для зниження вартості та збільшення ефективності обчислювальних процесів. Ця тема має велике значення для розробників, науковців, а також для комерційних організацій, які шукають способи максимально ефективного використання обчислювальних ресурсів.

Враховуючи швидке зростання кількості даних та їх складності, а також постійне оновлення і вдосконалення технологій машинного навчання, оптимізація обчислювальних ресурсів є важливим напрямком у дослідженнях і практичних застосуваннях. Це не тільки сприяє підвищенню ефективності використання технологій, але й відкриває нові можливості для інноваційних рішень у галузі штучного інтелекту та машинного навчання.

Об'єктом дослідження є операції машинного навчання.

Предметом дослідження є оптимізація використання ресурсів, таких як обсяг сховища, утилізація GPU, кількість мережевого трафіку та швидкодія у операціях машинного навчання

Мета роботи: аналіз існуючих програмних та інфраструктурних рішень до навчання моделей; виявлення їх переваг і недоліків; пошук вузьких місць; аналіз використання ресурсів; створення засобів для

усунення вад існуючих рішень та покращення продуктивності; дослідження створених засобів у близькому до реального середовищі.

Наукова новизна полягає в наступному:

1. Запропоновано стратегію оптимізації управління великими даними в операціях машинного навчання.
2. Розроблено спосіб оптимізації швидкості отримання напівструктурованих даних в операціях машинного навчання.

Практична цінність виявляється в кількох ключових аспектах:

1. **Покращення ефективності обробки даних:** За допомогою оптимізованого управління великими даними забезпечується ефективніша обробка та аналіз даних, що є важливим для швидкого та точного отримання результатів у машинному навчанні. Це особливо важливо для проектів, які обробляють великі обсяги даних.
2. **Зменшення часу навчання моделей:** Оптимізація швидкості отримання напівструктурованих даних сприяє скороченню часу, необхідного для навчання моделей. Швидке завантаження та обробка даних веде до зниження загального часу витраченого на тренування та тестування моделей.
3. **Зниження витрат на обчислювальні ресурси:** Ефективне управління великими даними може призвести до зменшення навантаження на обчислювальні системи, що в свою чергу веде до оптимізації витрат. Це включає зменшення потреби у високопродуктивних обчислювальних ресурсах та зниження вартості зберігання даних.
4. **Підвищення гнучкості та масштабованості:** Запропоновані стратегії дозволяють підвищити гнучкість та масштабованість процесів машинного навчання, забезпечуючи можливість ефективно працювати з різними видами даних та в різних масштабах.

5. **Поліпшення якості даних:** Оптимізоване управління даними також сприяє підвищенню якості вхідних даних, що є критично важливим для точності моделей машинного навчання. Видалення шумів, непотрібних або дублюючих даних є ключовим для підвищення ефективності тренування моделей.

Ці стратегії є надзвичайно важливими для розвитку машинного навчання, оскільки вони сприяють створенню більш потужних, точних та ефективних моделей, які можуть вирішувати складніші завдання з меншими витратами ресурсів.

Апробація роботи. Основні положення і результати роботи були представлені та обговорювались на науковій конференції магістрантів та аспірантів «Прикладна математика та комп'ютинг» ПМК-2023 (Київ, 28-30 листопада 2023 р.).

Структура та обсяг роботи. Магістерська дисертація складається з вступу, трьох розділів та висновків.

У *вступному* розділі магістерської дисертації ретельно розглядається актуальність та важливість оптимізації використання обчислювальних ресурсів у сучасних процесах машинного навчання. Вступ підкреслює, що зі зростанням обсягів даних та складності моделей машинного навчання, ефективне управління обчислювальними ресурсами стає ключовим фактором для досягнення більшої ефективності, точності та економічної вигоди в цій галузі. У вступі також наголошується на зростаючому інтересі до цієї теми у науковій спільноті та необхідності розробки нових підходів та технік для вирішення актуальних викликів.

У *першому розділі* проводиться глибокий аналіз проблематики, пов'язаної з використанням ресурсів у машинному навчанні та їх наслідків. Розділ починається з огляду основних понять, що стосуються операцій машинного навчання, включаючи їх складові, сфери застосування, та роль у

економічному розвитку та суспільстві. Далі розглядаються конкретні проблеми, які виникають у процесі використання ресурсів, включаючи інфраструктурні складнощі та виклики, пов'язані з роботою з даними. Особлива увага приділяється аналізу використання ресурсів під час тренування моделей машинного навчання, а також засобам оцінювання та моніторингу цих процесів. Розділ завершується оглядом існуючих рішень для оптимізації використання обчислювальних ресурсів, включаючи апаратні, мережеві та програмні рішення, та розглядає негативні наслідки відсутності оптимізації.

У *другому розділі* детально розглядаються засоби оптимізації, спрямовані на удосконалення процесів управління великими даними та підвищення ефективності отримання напівструктурованих даних. Розділ включає комплексний аналіз стратегій, які можуть бути застосовані для ефективного збереження, обробки та використання великих обсягів даних у машинному навчанні. Особливий акцент робиться на аспектах, які впливають на продуктивність обчислювальних систем, включаючи методи зменшення комунікаційних затримок та підходи до структуризації неструктурованих даних. Також у цьому розділі розглядаються інноваційні стратегії для створення масштабованої інфраструктури та ефективного використання ресурсів під час тренування моделей машинного навчання, що включає аналіз різних технічних та методологічних підходів.

У *третьому розділі* здійснюється аналіз результатів, отриманих від застосування розроблених методів оптимізації. Розділ починається з опису проведених експериментів та методології вимірювання метрик, які використовуються для оцінки ефективності запропонованих рішень. Далі подається детальний опис компонентів та розроблених рішень, які були використані в ході експериментів. У цьому розділі також проводиться глибокий аналіз отриманих результатів, включаючи оцінку ефективності різних стратегій оптимізації. Особлива увага приділяється порівнянню запропонованих підходів із існуючими рішеннями, щоб визначити їх

переваги та недоліки. Це включає аналіз впливу оптимізаційних стратегій на швидкість обробки даних, комунікаційні затримки, витрати на обчислювальні ресурси та інші критично важливі параметри.

У висновках підсумована важливість та внесок проведеного дослідження в області оптимізації обчислювальних ресурсів машинного навчання.

Робота представлена на 87 аркушах, містить посилання на список використаних літературних джерел.

Ключові слова: операції машинного навчання, MLOps, великі дані, Big Data.

ABSTRACT

Relevance of the topic. The relevance of the topic "Optimizing the use of computing resources in machine learning operations" cannot be overestimated in today's world of rapid technological development. In the context of the rapid growth of data volumes and the ever-increasing complexity of machine learning models, there is an urgent need for efficient use of computing resources. Optimizing these resources is a key element in achieving high performance, accuracy and cost-effectiveness in machine learning processes.

Given the high demands on computing power required to train and use machine learning models, the issue of optimizing the use of resources to reduce the cost and increase the efficiency of computing processes is becoming particularly relevant. This topic is of great importance for developers, researchers, and commercial organizations looking for ways to maximize the efficient use of computing resources.

Given the rapid growth of data and its complexity, as well as the constant updating and improvement of machine learning technologies, optimisation of computing resources is an important area of research and practical applications. This not only helps to increase the efficiency of technology use, but also opens up new opportunities for innovative solutions in the field of artificial intelligence and machine learning.

The object of the study is machine learning operations.

The subject of the study is the optimization of usage of resource such as storage capacity, GPU utilization, network traffic, and performance in machine learning operations.

The aim of the work is to analyze existing software and infrastructure solutions for model training; identify their advantages and disadvantages; find bottlenecks; analyze resource usage; create tools to eliminate the shortcomings of existing solutions and improve performance; and test the created tools in a close-to-real environment.

The scientific novelty is as follows:

1. A strategy for optimizing big data management in machine learning operations is proposed.
2. A method for optimizing the speed of obtaining semi-structured data in machine learning operations is developed.

The practical value is manifested in several key aspects:

1. **Improved data processing efficiency:** Optimized big data management enables more efficient data processing and analysis, which is essential for fast and accurate machine learning results. This is especially important for projects that process large amounts of data.
2. **Reduced model training time:** Optimizing the speed of semi-structured data acquisition helps reduce the time required to train models. Fast data loading and processing results in a reduction in the overall time spent on model training and testing.
3. **Reduce the cost of computing resources:** Effective big data management can lead to a reduction in the load on computing systems, which in turn leads to cost optimisation. This includes reducing the need for high-performance computing resources and lowering the cost of data storage.
4. **Increase flexibility and scalability:** The proposed strategies increase the flexibility and scalability of machine learning processes, providing the ability to work effectively with different types of data and at different scales.
5. **Improved data quality:** Optimized data management also helps to improve the quality of input data, which is critical to the accuracy of machine learning models. Removing noise, unnecessary or duplicate data is key to improving the efficiency of model training.

These strategies are crucial for the development of machine learning, as they contribute to the creation of more powerful, accurate, and efficient models that can solve more complex problems with fewer resources.

Work approbation. The main provisions and results of the work were presented and discussed at the scientific conference of undergraduate and postgraduate students "Applied Mathematics and Computing" PMC-2023 (Kyiv, 28-30 November 2023).

Structure and scope of the work. The master's thesis consists of an introduction, three chapters and conclusions.

The *introductory* chapter of the master's thesis thoroughly discusses the relevance and importance of optimizing the use of computing resources in modern machine learning processes. The introduction emphasizes that with the increasing amount of data and complexity of machine learning models, efficient management of computing resources is becoming a key factor in achieving greater efficiency, accuracy and economic benefits in this field. The introduction also emphasizes the growing interest in this topic in the scientific community and the need to develop new approaches and techniques to address current challenges.

The *first chapter* provides an *in-depth* analysis of the issues surrounding the use of resources in machine learning and their implications. The chapter begins with an overview of the basic concepts related to machine learning operations, including their components, applications, and role in economic development and society. It goes on to discuss the specific challenges that arise in the process of resource utilization, including infrastructure and data challenges. Particular attention is paid to the analysis of resource utilization during machine learning model training, as well as to the means of evaluating and monitoring these processes. The chapter concludes with an overview of existing solutions for optimizing computing resource utilization, including hardware, networking, and software solutions, and discusses the negative consequences of a lack of optimization.

The *second section* discusses in detail optimisation tools aimed at improving big data management processes and increasing the efficiency of semi-structured data. The chapter includes a comprehensive analysis of strategies that can be applied to efficiently store, process, and use large amounts of data in machine learning. Particular emphasis is placed on aspects that affect the performance of computing systems, including methods for reducing communication latency and approaches to structuring unstructured data. This chapter also discusses innovative strategies for creating scalable infrastructure and efficient resource utilization when training machine learning models, including an analysis of various technical and methodological approaches.

The *third section* analyzes the results obtained from the application of the developed optimisation methods. The section begins with a description of the experiments conducted and the methodology for measuring the metrics used to evaluate the effectiveness of the proposed solutions. This is followed by a detailed description of the components and developed solutions used in the experiments. This section also provides an in-depth analysis of the results obtained, including an assessment of the effectiveness of various optimisation strategies. Particular attention is paid to comparing the proposed approaches with existing solutions to identify their advantages and disadvantages. This includes an analysis of the impact of optimisation strategies on data processing speed, communication latency, computing resource costs, and other critical parameters.

The conclusions summarize the importance and contribution of the study in the field of machine learning computing resource optimisation.

The paper is presented on 87 pages and contains references to the list of used literature.

Keywords: machine learning operations, MLOps, big data, Big Data.