Mesopotamian journal of Big Data
Vol. **(2023**) 2023, **pp**. 12–15
DOI: https://doi.org/10.58496/MJBD/2023/002 ISSN: 2958-6453
https://mesopotamian.press/journals/index.php/BigData

Research Article

Parallel Machine Learning Algorithms

Saba Abdulbaqi Salman*[1], Saad Ahmed Dheyab [2], Qusay Medhat Salih[3], Waleed A. Hammood [4]

[1] *Department of Computer, College of Education, AL-Iraqia University, Iraq*

[2] *College of Engineering University of Information Technology and Communications, Baghdad, Iraq*

[3]*Medical Equipment Department, Alsalam University College, Iraq*

[4] *Faculty of Engineering, Gilgamesh University, Iraq*

ARTICLE INFO

Article History

Received 09 Nov 2022
Accepted 10 Jan 2023
Published 22 Jan 2023

Keywords

Big data

Parallel Processing
Deep Learning

Machine Learning

ABSTRACT

To expedite the learning process, a group of algorithms known as parallel machine learning algorithms can be executed simultaneously on several computers or processors. As data grows in both size and complexity, and as businesses seek efficient ways to mine that data for insights, algorithms like these will become increasingly crucial. Data parallelism, model parallelism, and hybrid techniques are just some of the methods described in this article for speeding up machine learning algorithms. We also cover the benefits and threats associated with parallel machine learning, such as data splitting, communication, and scalability. We compare how well various methods perform on a variety of machine learning tasks and datasets, and we talk about the advantages and disadvantages of these methods. Finally, we offer our thoughts on where this field of study is headed and where further research is needed. The importance of parallel machine learning for businesses that want to glean insights from massive datasets is emphasised, and the paper provides a thorough introduction of the discipline.

1. INTRODUCTION

There is a class of machine learning algorithms[1, 2] designed to speed up training times by running on numerous processors or computers simultaneously. In today's data-driven environment, these algorithms are crucial because they speed up the training of machine learning [3] models on big datasets, allowing businesses to quickly glean actionable insights. Several factors highlight the significance of parallel machine learning algorithms[4]: Huge amounts of data: The rising volume and complexity of data being generated and collected nowadays is one of the primary drivers of the demand for concurrent machine learning techniques. When dealing with huge datasets, training a machine learning model[5] can be computationally intensive, making the use of a single processor impracticable. The ability to train machine learning models in parallel across several processors or computers is made possible by parallel machine learning techniques, allowing businesses to more effectively mine massive datasets for useful information.

A company's success often hinges on how quickly its machine learning models can be trained and put into production. The time it takes to train machine learning models can be drastically cut with the use of parallel machine learning algorithms, allowing businesses to get their wares to market much sooner. The computational expense of training machine learning models on huge datasets means that many businesses must resort to the usage of pricey cloud computing services or specialised hardware. By enabling businesses to train machine learning models on fewer, more powerful processors or computers, or by taking use of the economies of scale provided by cloud computing platforms, parallel machine learning algorithms can help minimise the cost of training machine learning models.

Training on larger datasets or employing more complicated models that would be impractical to train on a single processor are two ways in which parallel machine learning methods can boost the accuracy of machine learning models. In sum, parallel machine learning algorithms are a crucial resource for businesses that want to efficiently mine massive data sets for actionable intelligence.

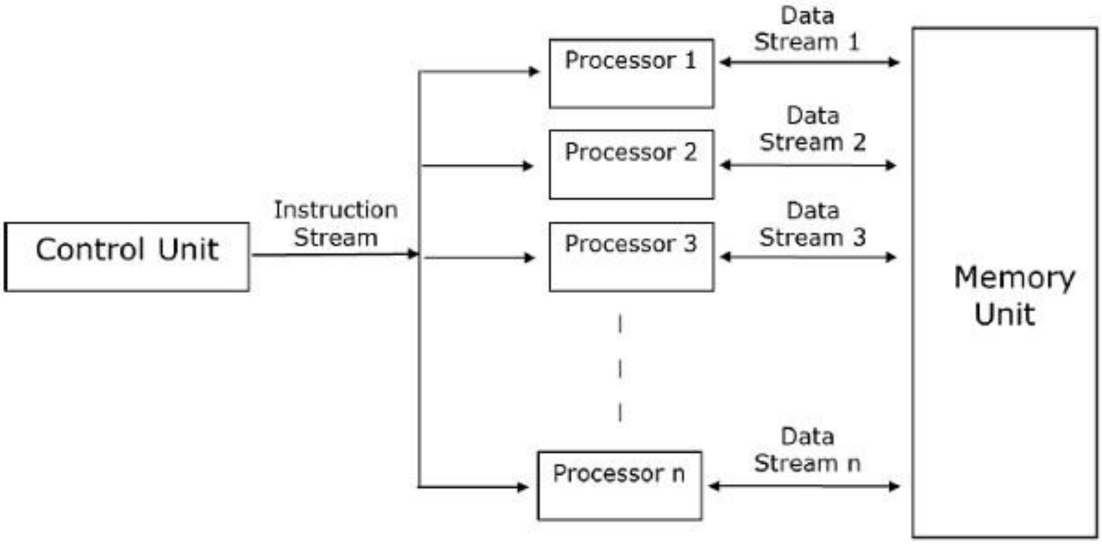

Fig. 1 Parallel data processing.

2. RELATED WORKS

The publication "Parallel stochastic gradient descent for deep learning: A survey" by J. Dean et al. (2012)[6] provides an overview of the numerous methods proposed for accelerating the implementation of the widely used optimisation technique for deep neural network training: stochastic gradient descent (SGD). The authors analyse the effectiveness of several ways to parallelizing SGD across a variety of datasets and architectures and explore the associated trade-offs. According to "Scalable and efficient parallel machine learning with MapReduce" by Y. Low et al.[7]This paper presents MapReduce, a programming approach for handling massive datasets on computer clusters, as a foundation for parallel machine learning. The authors show that their method works well for various machine learning problems and evaluate it against other parallel machine learning strategies.

Article by Y. LeCun et al. (2015) titled "Parallel distributed deep learning: A survey"Data parallelism, model parallelism, and hybrid approaches are all discussed, as well as other methods that have been proposed for parallelizing deep learning [8]. The authors also address the benefits and drawbacks of parallel deep learning and offer suggestions on where the field should go from here in terms of research.

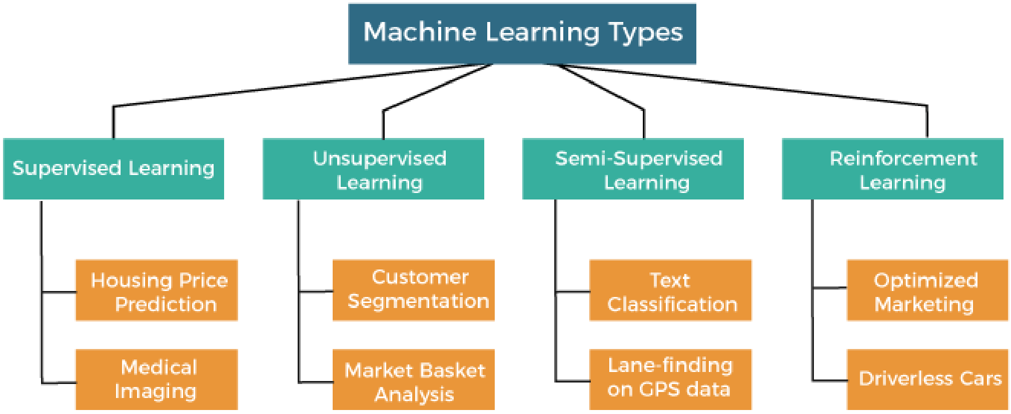

Fig. 2 Types of machine learning.

3. PARALLEL MACHINE LEARNING CHALLENGES

The term "big data" is used to describe the massive amounts of data, both structured and unstructured, that are created by corporations. Data segmentation, connectivity, and scalability are just a few of the obstacles and opportunities that come with parallel machine learning. Partitioning the data into smaller subsets, or "partitions," that can be independently handled by various processors or computers, is a significant challenge in parallel machine learning. Partitioning the data so that each processor may operate independently while yet being able to aggregate their results to arrive at a global solution is a difficult problem to solve.

Parallel machine learning faces additional difficulty when trying to exchange information between multiple computers or processors. The training process can be slowed down by communication bottlenecks if there is an excessive volume of data being transmitted between computers. Thirdly, scalability, or the ability to deal with larger datasets or more complicated models without significantly additional computational resources, is a barrier for parallel machine learning. It is crucial to discover parallel machine learning algorithms that scale well with the number and complexity of the data and models being utilised, and there are some that do so better than others. Despite these difficulties, parallel machine learning does provide some promising possibilities. Case in point: Rapidity is a key advantage of parallel machine learning. Companies can drastically cut the time it takes to train machine learning models on huge datasets by leveraging many processors or computers to do so simultaneously. Using fewer, more powerful processors or computers, or taking use of the economies of scale afforded by cloud computing platforms, can assist minimise the cost of training machine learning models, as was discussed earlier. Training on larger datasets or employing more complicated models that would be impractical to train on a single processor are two ways in which parallel machine learning methods can boost the accuracy of machine learning models. Finding the sweet spot between the obstacles and benefits of parallel machine learning is crucial for the success of any parallel machine learning method.

4. PROPOSED APPROACHES OF PARALLEL MACHINE LEARNING

Different strategies, such as data parallelism, model parallelism, and hybrid strategies, have been proposed to speed up machine learning algorithms. The concept of "data parallelism" refers to the practise of dividing a large dataset into smaller portions that can then be processed independently by separate computers or processors. Each processor's output is added to the others to provide a global answer. Machine learning models, such neural networks, that may be trained independently on distinct subsets of the data are prime candidates for data parallelism.

Model parallelism involves breaking down the machine learning model into smaller pieces and sending them to their own dedicated processors. This method is generally employed when training models with a complicated structure that can be partitioned into smaller portions, or when training models that are too huge to fit on a single processor. Hybrid methods: The benefits of both data parallelism and model parallelism can be taken advantage of by combining the two in a hybrid method. Partitioning a model into smaller pieces that may then be trained in parallel on separate subsets of the data is one example. All of these methods have their advantages and disadvantages, and choosing the most appropriate one for a given dataset and machine learning model is a highly contextualized decision. Finding the proper method for a particular challenge, such as which parallel machine learning algorithm is best for which dataset or model, might be crucial to attaining good performance.

5. LIMITATION OF PARALLELISM

Several approaches to parallel machine learning have their own set of advantages and disadvantages. Partitioning the data: breaking it up into manageable chunks for parallel processing isn't easy, but it's crucial for getting a high performance. Suboptimal performance can occur if the processors cannot function independently or cannot aggregate their outputs efficiently due to improper data partitioning. The volume of data being exchanged between processors can be substantial, making communication a bottleneck in parallel machine learning. This can reduce the system's scalability, resulting in longer training times.

Some concurrent machine learning methods may have difficulty scaling to larger datasets or more sophisticated models, reducing their utility for solving more involved issues. Hardware: Parallel machine learning techniques' efficiency may also be affected by the hardware configuration used. While graphics processing units (GPUs) and other specialised hardware can increase the speed at which some algorithms execute, this may not always be a viable or cost-effective option. Different methods of parallel machine learning have varying degrees of success depending on the tasks and datasets being used. Some methods may work better than others on specific tasks or data sets, while others may be flexible enough to handle a wide variety of applications. One way to narrow down potential solutions is to compare how well they perform across a range of activities and datasets.

6. CONCLUSION AND FURTHER WORK

In conclusion, businesses who want to efficiently mine meaningful insights from massive datasets should make use of parallel machine learning methods. Data parallelism, model parallelism, and hybrid techniques are all types of parallel machine learning, each with its own advantages and disadvantages. The properties of the dataset and the machine learning model in use might affect the efficiency of multiple approaches, making it difficult to determine which one is best for a given situation. Issues with data partitioning, communication, and scalability are only a few of the problems and trade-offs to think about when employing concurrent machine learning methods. Despite these difficulties, parallel machine learning is a vital area of study because of its potential for improved speed, efficiency, and accuracy. There is still a lot of room for development in the world of parallel machine learning. There is a need for more efficient algorithms and methods that can scale well with the amount and complexity of the datasets and models being used. More study is required to identify the constraints and trade-offs of various approaches, and more efficient tools and frameworks are required for creating and deploying parallel machine learning applications. The importance of parallel machine learning is expected to increase over the coming years as the volume and complexity of data continue to grow.

Conflicts of Interest

The author's paper explicitly states that there are no conflicts of interest to be disclosed.

Acknowledgment

The authors would like to express their gratitude to the AL-Iraqia University, the Informatics Institute for Postgraduate Studies, and the Al Salam University College for their moral support. Please accept my sincere gratitude for the useful recommendations and constructive remarks provided by the anonymous reviewers.

Funding

The author's paper clearly indicates that the research was conducted without any funding from external sources.

References

[1] C. Xu, G. Neglia, and N. Sebastianelli, "Dynamic backup workers for parallel machine learning," *Computer Networks,* vol. 188, p. 107846, 2021.

[2] D. Wu, C. Jennings, J. Terpenny, S. Kumara, and R. X. Gao, "Cloud-based parallel machine learning for tool wear prediction," *Journal of Manufacturing Science and Engineering*, vol. 140, no. 4, 2018.

[3] G. Xian, "Parallel machine learning algorithm using fine-grained-mode spark on a mesos big data cloud computing software framework for mobile robotic intelligent fault recognition," *IEEE Access,* vol. 8, pp. 131885-131900, 2020.

[4] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, "A survey of machine learning for big data processing," *EURASIP Journal on Advances in Signal Processing,* vol. 2016, pp. 1-16, 2016.

[5] Y. Low, J. E. Gonzalez, A. Kyrola, D. Bickson, C. E. Guestrin, and J. Hellerstein, "Graphlab: A new framework for parallel machine learning," *arXiv preprint arXiv:1408.2041,* 2014.

[6] G. Heigold *et al.*, "Multilingual acoustic models using distributed deep neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*, 2013, pp. 8619-8623: IEEE.

[7] B. Dong, J. Qiu, Q. Zheng, X. Zhong, J. Li, and Y. Li, "A novel approach to improving the efficiency of storing and accessing small files on hadoop: a case study by powerpoint files," in *2010 IEEE International Conference on Services Computing*, 2010, pp. 65-72: IEEE.

[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature,* vol. 521, no. 7553, pp. 436-444, 2015.