Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/102/20241145

Comparison Distributed and Parallel Machine Learning: Evidence from Models, Principles and Application Scenarios

Zhefan Zhang

Department of Communication Engineering, Wuhan University of Science and Technology, Wuhan, China

zhefanzhang@wust.edu.cn

Abstract. Contemporarily, the demand for processing large-scale data has been rapidly increasing, prompting continuous advancements in the field of machine learning. This study examines the state of parallel and distributed machine learning at present, both of which aim to enhance computational efficiency when dealing with large datasets and demonstrate promising applications across various domains. As the scale of data escalates, traditional single-machine learning methods are becoming increasingly inadequate, which lead to the emergence of parallel and distributed machine learning. These approaches enable substantial computations to be performed more efficiently through the collaboration of multi-core CPUs or multiple computing nodes. This research conducts an in-depth analysis of the inherent challenges associated with these methods, including data transmission latency, synchronization requirements, and user privacy concerns. Ultimately, this research emphasizes the tremendous potential of both methods for future applications, a potential that is bolstered by ongoing advancements in hardware and algorithm optimization. These results provide valuable insights for practitioners in the field and offers guidance for future research directions.

Keywords: Parallel and distributed machine learning, data scalability, hardware and algorithm optimization.

1. Introduction

Machine learning aims to use computer systems to automatically improve their performances by learning previous knowledge (data). In 1957, Frank Rosenblatt organized a team to create a machine called “perceptron” in Cornell University, which associated with the origin of early machine learning [1]. During the early period, researches in machine learning focused primarily on the developments of algorithms and theories. Recently, the significant advances in techniques and in-depth theoretical researches have led the growth of machine learning, which captured extensive interest from multiple fields. The recent success and rapid commercialization of deep learning have driven technological advancements across various industries, including computer vision, speech recognition, gaming, and machine translation [2]. In addition, machine learning is also employed as a powerful tool for the filtration of spam from textual documents, such as emails, thereby enhancing effectiveness and exactness of communication systems. Furthermore, machine learning techniques are leveraged to extract critical information from extensive datasets through sophisticated data mining methodologies, thereby facilitating the identification of significant trends and insights. Based on learning from meticulously

DOI: 10.54254/2755-2721/102/20241145

curated training data, machine learning algorithms enable the identification of intricate patterns, the construction of predictive models, and the formulation of informed forecasts. This pattern recognition capability is paramount for various applications, as it allows practitioners to draw meaningful conclusions from complex information. Moreover, the algorithms inherent to machine learning are extensively applied across a multitude of disciplines, including, but not limited to, biology and genomics, where they assist researchers in unraveling complex biological processes and understanding genetic variations. This interdisciplinary application of machine learning underscores its versatility and impact on advancing knowledge in contemporary scientific research [3].

Parallel machine learning can reduce training time and improve algorithm efficiency by running contemporaneously on multiple computers. Due to the dramatically increasing magnitude and intricacy of data generation and collection, there is an urgent need to advance parallel machine learning technologies [4]. A article illustrated a series–parallel machine learning strategic framework to describe the ice and liquid water uniformly and accurately, helping to understand the relevant process molecular simulation [5]. This study presented a functional molecular model for understanding phase transitions and crystallization phenomena in nanoconfined environments, along with a strategic framework for constructing a complicated molecular model in harsh environmental conditions.

Another approach to machine learning is distributed machine learning. Comparing with the conventional (centralized) machine learning, it can process the huge training data which exceed the hardware computing capabilities by distributing the workload across multiple machines [6]. For instance, the 6G technology in the future must face the problems mentioned before. Therefore, the demand space for 6G applications will encompass multiple dimensions. Thus, in the field of algorithms, the ability to train discontinuous and highly fragmented data, improve training efficiency in dynamic environments, and oversee collaborative learning transfer agents to enhance anomaly detection and management will become integral components of future solutions. Others shows the significant research of the distributed learning technologies in terms of 6G world, predicting the distributed machine learning can offer solutions to probably networking problems due to the complexity and high dimensionality of data [7].

Both the parallel machine learning and distributed machine learning are developed to resolve the current and upcoming problems. However, there are various differences in terms of the methods and algorithms, principles as well as the features. Hence, this research aims to compare the two class of machine learning. This work hope to gain useful insights, understand the current limitations, and make projections for the future through these comparisons. First of all, the article will introduce the common models of parallel machine learning and distributed machine learning, and then there is a detailed introduction to parallel machine learning and distributed machine learning. Finally, the study conducted comparisons multiple times to gain insight into the model's applicability and validity, as well as its current limitations and future prospects for optimization.

2. Machine learning scenarios

Machine learning models are algorithms that have the ability to identify patterns or make predictions on datasets that have never been seen before. These models can evolve over time as new data enters the system, unlike rule-based programs, which must be explicitly coded. There are various common machine learning models below. The process involves providing the machine with input and desired output and labels are used to categorize input and output data. The predicted target variable is based on the given set of predictors. Logistic regression, decision trees, support vector machines (SVM), and neural networks are some of the algorithms that can be utilized. Classification methods for discrete outputs and regression methods for continuous outputs are both included. Datasets with input data without labelled responses are used for this type of learning to draw inferences, as if there are no supervisors to guide. Models like k-means clustering and PCA are used to analyze unlabeled data and identify patterns and structures without predefined labels.

The models interact with the machine's environment to learn. Correct actions will be rewarded by the supervisor and wrong actions will be punished. The machine is trained to be autonomous and make its own decisions here. Reinforcement learning is based on learning from past experiences. Accurate

DOI: 10.54254/2755-2721/102/20241145

decisions can be achieved by capturing the best possible knowledge. For instance, there are models like Q-learning and deep Q-networks [8].

Combining small amounts of labeled data with large amounts of unlabeled data is done to enhance classification accuracy and efficiency. Semi-Supervised Support Vector Machines is an extension of SVM that uses both labeled and unlabeled data to enhance classification accuracy. Selecting the most suitable method for a specific problem requires understanding common machine learning models. The performance of foundational models can be improved and solutions developed can be more effective by optimizing them, and the continuous advancements in these models will continue to unlock new possibilities across various domains. For instance, some scholars suggesting a novel attack detection model that uses the distributed machine learning in order to detect attacks at network edge devices [9]. The result shows that the model which employs an advanced voting algorithm for fundamental logic systems between the server and the worker nodes exhibits excellent performance in edge computing. In conclusion, it is crucial to select appropriate models based on different occasions and implement reasonable optimizations tailored to different issues.

3. Description of parallel machine learning

In detail, parallel machine learning refers to the concurrent execution of machine learning tasks across multiple processing units, such as CPU and GPU, to enhance computational efficiency and reduce training time. Therefore, parallel machine learning is a fundamental aspect of modern machine learning practices because it divides the workload and allows for faster processing of large datasets and complex models. In parallel computing, communication and synchronization between the processing units is crucial for ensuring data consistency and model accuracy, which includes sharing model parameters, aggregating computation results, and more. Parallel machine learning aims to maximize the utilization of computational resources by dispersing the workload, thus easing the strain on individual processing units and speeding up the training process, especially when dealing with large-scale datasets and complex models. Therefore, the efficiency of model training and data processing is enhanced by the collective support of these principles in parallel machine learning implementation.

Furthermore, in parallel machine learning, prediction involves training models using synthesized fabricated data to achieve improved generalization, or training intelligent agents in a virtual environment to obtain improved policies, avoiding the use of fabricated data and fictitious scenarios for assessing the efficacy of machine learning outcomes. The pipeline of computational experiments consists of training and testing processes [10]. The domain gap between the virtual and real worlds necessitates multiple trials to repeat the phases of description and prediction, which leads to computational overhead.

There are a variety of approaches for achieving parallel machine learning [11]. In data parallelism, multiple processing units are used to split the dataset into smaller batches. Despite sharing the same model parameters, each unit performs computations independently on its subset of data. The global model is updated by aggregating the gradients (or updates) computed on the different subsets after processing, for example, through averaging. This strategy for training learning models on large datasets is particularly effective. When it comes to the model parallelism, different processing units are assigned to different sections or layers within the model. In cases where the model is too large to fit into one unit's memory, this approach is advantageous. As required, the processing unit completes computations for its part of the model and transfers intermediate results to other units. The requirement for communication between units can make it more complicated than data parallelism. In addition, task parallel is another approach, in its process, different tasks or operations are being carried out at the same time. Parallelism can be achieved by running different components of a machine learning pipeline, such as data preprocessing, model training, and hyperparameter tuning, to make efficient use of available resources.

4. Description of distributed machine learning

The exploration of an alternative machine learning approach, namely distributed machine learning, has been sparked by the increasing volume of data and the complexity of models that exceed the capabilities of a single machine. Distributed Machine Learning achieves more efficient computation by allocating

DOI: 10.54254/2755-2721/102/20241145

workloads to multiple machines instead of only processing and training the entire dataset and the model on a single computational. Each machine handles a subset of the data or contributes to the training process of the model. This distributed method not only has the capability to accelerate the training process, but it also enhances the resource allocation efficiency and robustness. Ultimately, distributed machine learning leverages the advantages of parallel computation, fully harnessing the capabilities of various computing resources to facilitate collaborative work [10].

In the process, multiple nodes are used to divide the data, which could include clusters of servers or distributed cloud environments. Parallel computation becomes possible due to the independent processing of a subset of data by each node. By leveraging multiple computing resources, distributed machine learning can efficiently handle large datasets and complex models, while addressing challenges in communication, synchronization, and fault tolerance.

Similar to parallel machine learning, both data parallelism and model parallelism are also crucial concepts in distributed machine learning that optimize the training process across multiple machines. Data parallelism is a suitable method for partitioning, where a model is trained on different data splits, which involves partitioning a dataset into smaller subsets and distributing these subsets across various machines. The model is trained by each machine on its local data, and when finished, the models synchronize to update the global model, frequently using techniques such as parameter averaging. The collective computational power of multiple machines is used to enhance the training speed through this approach. In addition, model parallelism entails splitting the model itself across multiple machines. This approach is especially useful for massive models that are too large to fit in one machine's memory. In order to achieve model parallelism, it is necessary to communicate between machines to exchange gradients from shared layers, ensuring that all parts of the model are updated correctly. Additionally, there is a special a distributed computing technique, tensor parallelism [10], which disseminates the computation of tensors (multidimensional arrays) across multiple processing units or devices. This method enhances the efficiency of large tensor operations by allocating the workload appropriately. It is particularly beneficial in training deep learning models with substantial parameter sets, as it leverages the collective memory and processing capabilities of multiple GPUs or machines.

5. Comparison

Parallel machine learning and distributed machine learning are both methodologies employed to address issues related to large volumes of data and enhance the efficiency of machine learning. However, they differ in their implementation and use cases. Regarding parallel machine learning, to predicting the active behaviors of extensive and intricate networks. involving extracting information from previous time series data across multiple domains, so there is a machine learning solution has been proposed, which employs a parallel architecture that simulates the topology of the target network [11]. It mentions a system schematic diagram based on the parallel machine learning, which can apply to the reservoirs. The Fig. 1 illustrates this parallel network machine learning architecture.

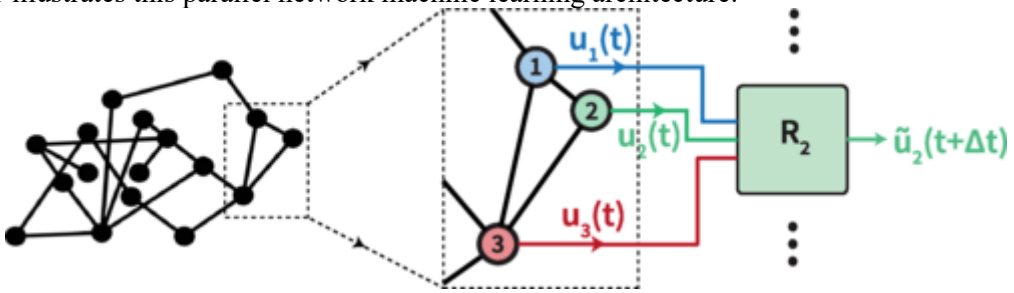

Figure 1. Parallel architecture for machine learning networks [11].

When it comes to the distributed machine learning, there are a type of application contexts where access to training data is restricted to ensure model transparency and protect user privacy. The

DOI: 10.54254/2755-2721/102/20241145

development trends in machine learning align with this demand, particularly in using the data that is highly segmented for training models.

For instance, data collaboration is a type of distributed machine learning that does not involve model sharing [12]. The use of this approach permits users to efficiently exchange and utilize data without sharing the original data directly. The privacy and security of the data are enhanced by this local processing method, which also effectively simplifies its complexity, making it easier to analyze and apply. An analyst subsequently aggregates these representations middle-ranking, after which, the feedback for each user can be created distributivity. The data's horizontal or vertical partitioning determines the variation in the incorporation process. In a vertical partitioning scenario, incorporation process is simply achieved by integrating the intermediate representations through concatenation. The Fig. 2 is the flowcharts, illustrating how data collaboration operates and processes the data in two different contexts [13].

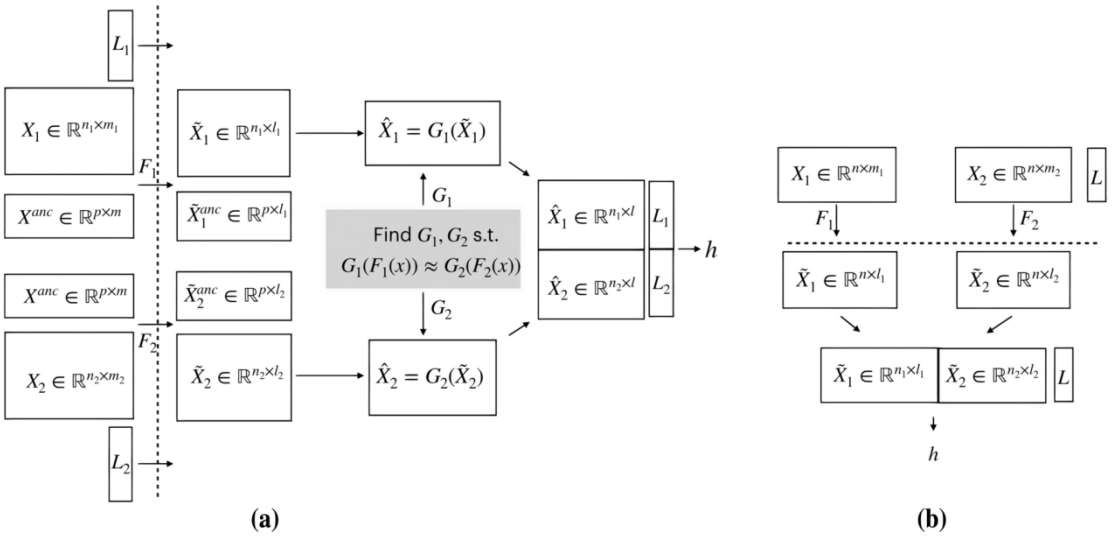

Figure 2. (a) Horizontal Data Collaboration and (b) Vertical Data Collaboration [13].

On the one hand, parallel machine learning operates on a single machine utilizing multiple processors or cores to perform calculations simultaneously. This approach is particularly effective when the dataset can fit within the memory of the individual machine. Generally, parallel machine learning involves shared datasets and is well-suited for parallelizable algorithms, which are critical in tasks like linear regression. On the other hand, distributed machine learning encompasses the collaboration of multiple machines or nodes to process large datasets when the amount of data or information that needs to be processed or stored is greater than what one computer can handle. In the process, data is partitioned and distributed across various machines, employing techniques like parameter servers and model averaging to synchronize learning across nodes. While parallel machine learning exhibits limited scalability due to its reliance on the hardware constraints of a single machine, distributed machine learning offers significant advantages in terms of both scalability and fault tolerance, making it a preferable choice for large-scale machine learning applications.

6. Limitations and prospects

Parallel and distributed machine learning make it possible to train models on large-scale datasets across multiple machines or processors. However, there are also some limitations. While there are some occasions need to exclude model sharing to protect user privacy, some algorithms that require frequent synchronization. Thus, they may perform poorly in the context of parallel and distributed machine learning, particularly when network bandwidth is limited. In such scenarios, the allocation of computational tasks among multiple nodes necessitates communication between them, which can result

DOI: 10.54254/2755-2721/102/20241145

in delays. This can lead to a situation where the time consumed by communication offsets the time savings achieved, especially when sharing models or gradients. For instance, Redis will be utilized as a distributed cluster for the storage and rapid access of data requiring real-time processing, such as session data and real-time statistics. It is suitable for scenarios that demand strong consistency. In addition, ZooKeeper, an open-source distributed coordination service, also requires a strong consistency model to ensure data consistency among multiple nodes in a distributed environment. Furthermore, parallel and distributed machine learning require more mature open-source frameworks to simplify the development and deployment processes, thereby facilitating their application.

It is well-established that in parallel and distributed machine learning, nodes need to frequently exchange gradient information. Therefore, one optimization direction is to compress communication data. This can be achieved by transmitting weight updates using fewer bits, thereby reducing bandwidth consumption. Federated Averaging serves as an example of a distributed learning method that employs techniques such as uniform quantization and binarization to substantially diminish the magnitude of data transmitted. This reduction enhances network transmission speed, alleviates network load, and improves the efficiency of model training. Additionally, the proliferation of certain algorithmic technologies and devices, such as devices of Internet of Things applied in edge computing, can significantly enhance data processing because the data transmission distance is significantly reduced by positioning it closer to the source, thereby improving system response times. Furthermore, this setting assists in reducing bandwidth utilization, which enhances overall network performance. In the future, with advancements in technology and optimization, distributed architectures will enable the analysis of real-time data streams, finding widespread applications finance, healthcare, smart manufacturing and other feasible application fields. Distributed and parallel machine learning techniques will transcend multiple domains, fostering the development of emerging applications, such as intelligent transportation and social network analysis.

7. Conclusion

Overall, both parallel and distributed machine learning are methods aimed at enhancing computational efficiency and handling large-scale data, however, there are differences in their models, algorithms, principles, and application scenarios. Parallel machine learning utilizes multi-core CPUs or GPUs on a single computer to simultaneously process multiple computational tasks, while distributed machine learning involves the collaboration of multiple computers. Both approaches must contend with challenges such as data transmission delays, frequent synchronization requirements, and the protection of user privacy. Nevertheless, parallel and distributed machine learning hold significant promise for the future. With ongoing advancements in hardware and the continuous optimization of algorithms, these methods are expected to find widespread applications across various domains. This research provides an overview and analysis of the current state of algorithms, models, and applications in parallel and distributed machine learning, and offers feasible projections for their future development, aiming to assist practitioners in this field.

References

[1] Fradkov A 2020 Early history of machine learning IFAC-PapersOnLine vol 53(2) pp 1385-1390
[2] Tyagi A, Kukreja S, Nair M M and Tyagi A K 2022 Machine Learning: Past Present and Future School of Computer Science and Engineering Vellore Institute of Technology; Terna Engineering College University of Mumbai
[3] Park D J, Park M W, Lee H, Kim Y J, Kim Y and Park Y H 2021 Development of machine learning model for diagnostic disease prediction based on laboratory tests Scientific Reports vol 11(1) p 7567
[4] Salman S A, Dheyab S A, Salih Q M and Hammood W A 2023 Parallel machine learning algorithms Mesopotamian Journal of Big Data vol 12 pp 12–15

DOI: 10.54254/2755-2721/102/20241145

[5] Wang J, Hei H, Zheng Y, et al 2024 Five-site water models for ice and liquid water generated by a series–parallel machine learning strategy Journal of Chemical Theory and Computation American Chemical Society vol 19 p 6
[6] Verbraeken J, Wolting M, Katzy J, Kloppenburg J, Verbelen T and Rellermeyer J S 2020 A survey on distributed machine learning ACM Computing Surveys (CSUR) vol 53(2) pp 1-33
[7] Muscinelli E, Shinde S S and Tarchi D 2022 Overview of distributed machine learning techniques for 6G networks Algorithms vol 15(6) p 210
[8] Hoang T M, Le Thi T L and Quy N M 2023 A novel distributed machine learning model to detect attacks on edge computing network Journal of Advanced Information Technology vol 14(1) pp 153–159
[9] Miao Q H, Lv Y S, Huang M, Wang X and Wang F Y 2023 Parallel learning: Overview and perspective for computational learning across Syn2Real and Sim2Real IEEE/CAA Journal of Automatica Sinica vol 10(3) pp 603–631
[10] Chatterjee B 2024 Distributed machine learning Proceedings of the 25th International Conference on Distributed Computing and Networking pp 4-7
[11] Srinivasan K, Coble N, Hamlin J, Antonsen T, Ott E and Girvan M 2022 Parallel machine learning for forecasting the dynamics of complex networks Physical Review Letters vol 128(16) p 164101
[12] Imakura A and Sakurai T 2020 Data collaboration analysis framework using centralization of individual intermediate representations for distributed data sets ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems Part A: Civil Engineering vol 6(2)
[13] Bogdanova A, Imakura A and Sakurai T 2023 DC-SHAP method for consistent explainability in privacy-preserving distributed machine learning Human-Centric Intelligent Systems vol 3 pp 197–210