A Systematic Literature Review of Distributed Techniques for Parallelizing Stochastic Descent Backpropagation

First Author
Department of Computer Science
University Name
firstauthor@university.edu

Second Author
Department of Computer Science
University Name
secondauthor@university.edu

Abstract

This paper presents a systematic literature review of distributed techniques for parallelizing stochastic descent backpropagation in neural networks. The review synthesizes research from the past decade, examining various approaches to distributed training, their effectiveness, and implementation challenges. Our analysis covers [X] primary studies, identifying key patterns in algorithmic design, communication strategies, and convergence properties. The findings indicate [brief summary of key findings]. This review contributes to the field by providing a comprehensive overview of current distributed training techniques, highlighting research gaps, and suggesting future research directions. The work serves as a valuable resource for researchers and practitioners working on large-scale neural network training systems.

1. Introduction

1.1. Background

Machine learning (ML) has become essential for extracting knowledge from data across diverse applications. Deep learning, a subfield of ML using artificial neural networks, is increasingly important, especially with the massive amounts of data now available [1], [6], [9]. Distributed deep learning (DDL) has become crucial due to the increasing size of datasets and model complexity [2], [6]. This literature review focuses on the core concepts, techniques, and frameworks used to implement and optimise DDL. Challenges in scaling deep learning include distributing ML programs, bridging computation with communication, and determining what to communicate between machines [11]. This systematic review will provide a comprehensive overview of the current state of DDL and address these questions, identify gaps, and give direction for future research [9], [1], [6], [10] **[TODO: Review citations]**.

1.2. Importance of the Topic

The increasing volume of data necessitates advanced analysis techniques, making distributed deep learning essential. Efficiently training deep learning models is vital for research and development [6], [5], [11]. This review is important for both practitioners and researchers, providing an overview of techniques and frameworks for implementing DDL solutions [6], [2] and also highlighting current research trends and future opportunities [2], [1]. This review addresses gaps in existing surveys by focusing on practical implementation and a detailed analysis of the performance and efficiency of different DDL techniques [2] **[TODO: review citation]**. The outcomes of this review will aid software engineers in selecting and implementing appropriate DDL techniques.

1.3. Research Questions

This review addresses the following primary research questions **[TODO: Review citations below]**:

- What are the different techniques for parallelising deep learning models and data (e.g., data and model parallelism) [2], [1], [9]?
- How do parameter update strategies impact distributed deep learning systems (e.g., Parameter Server and decentralised approaches) [1, 2, 9]?
- How is stochastic gradient descent (SGD) computed in distributed environments, and what are the associated challenges [1, 2, 9, 10]?
- What are the key frameworks currently available for implementing DDL, and how do their features and capabilities compare [2]?
- Finally, what are the main challenges and future research directions in distributed deep learning [1, 2, 6, 9]?

1.4. Scope and Methodology

This systematic review will focus on peer-reviewed articles and conference papers about DDL published from 2015 to 2023 [2]. The scope includes studies that investigate par-

allelisation of deep learning algorithms and techniques for distributed parameter updates [1, 2, 9, 10], and also the use of different frameworks and architectures for training large models [2, 9]. Studies on single-machine learning and applications not related to distributed training will be excluded, alongside papers lacking substantial technical contributions. The methodology will involve a systematic literature search across SCOPUS and Google Scholar using specific keywords related to DDL [2]. The selection process will involve screening titles, abstracts, and full texts [1], and a formal protocol will guide each stage of the analysis. Data extraction will involve recording details such as model architecture and performance metrics. Data synthesis will involve a narrative synthesis of findings.

1.5. Paper Organization

The remainder of this paper is organized as follows: Section 2 presents our systematic review protocol including the study selection framework and search process documentation, Section **??** describes our systematic review methodology, including the search strategy across digital libraries, inclusion/exclusion criteria, and quality assessment protocol. Section **??** presents our findings, Section **??** discusses the implications, and Section **??** concludes with future research directions. This structure ensures a logical flow from background to findings and recommendations.

2. Review Protocol and Process

This section outlines the methodology that will be used to conduct the systematic literature review, following the guidelines described in [3, 4, 7, 8]. A systematic literature review involves several discrete activities, and existing guidelines suggest slightly different numbers and orders of these activities. However, medical guidelines and sociological textbooks generally agree on the major stages. This document summarizes these stages into three main phases: Planning the Review, Conducting the Review, and Reporting the Review. **The primary goal is to ensure that the review process is both transparent and replicable**. This includes defining the search strategy, study selection criteria, quality assessment process, and data extraction methods. The review will investigate both distributed deep learning techniques and their parallel implementations using CUDA.

2.1. Review Workflow

Figure 1 visually outlines our systematic review process, divided into three key phases: the main workflow, studies selection, and validation. The diagram clearly presents the seven main steps of our process. This section will detail how we implement these steps.

2.2. The Review Process

2.2.1. Planning the Review

The stages associated with planning the review are:

- Identification of the need for a review (See Section **[TODO: Reference section number]**).
- Commissioning a review (See Section **[TODO: Reference section number]**).
- Specifying the research question(s) (See Section 1.3).
- Developing a review protocol (See Section 2).
- Evaluating the review protocol (See Section **[TODO: Reference section number]**).

2.2.2. Conducting the Review

The stages associated with conducting the review are:

- Identification of research (See Section **[TODO: Reference section number]**).
  - **Initial Search:** This stage involves using the defined search terms within selected databases to identify relevant studies, as further detailed in section 2.5 (Search Process Documentation). The process of how these terms are combined to create search strings is described in section 2.5, and the search results will be stored using a reference manager.
- Selection of primary studies (See Section **[TODO: Reference section number]**).
  - **Screening:** This stage involves an initial screening of titles and abstracts to remove irrelevant studies, which is part of the study selection process described in section 2.6 (Study Selection Criteria).
  - **Full-Text Review:** All potentially relevant studies will have their full texts retrieved, and the full texts will then be assessed against pre-defined inclusion and exclusion criteria (see section 2.6).
- Study quality assessment (See Section **[TODO: Reference section number]**).
- Data extraction and monitoring (See Section **[TODO: Reference section number]**).
  - **Data Extraction:** The final step is data extraction, where relevant information will be extracted from the included studies using a predefined data extraction form (detailed in section 2.8).
- Data synthesis (See Section **[TODO: Reference section number]**).

2.2.3. Reporting the Review

The stages associated with reporting the review are:

- Specifying dissemination mechanisms (See Section **[TODO: Reference section number]**).
- Formatting the main report (See Section **[TODO: Reference section number]**).
- Evaluating the report (See Section **[TODO: Reference section number]**).

We consider all the above stages to be mandatory except:

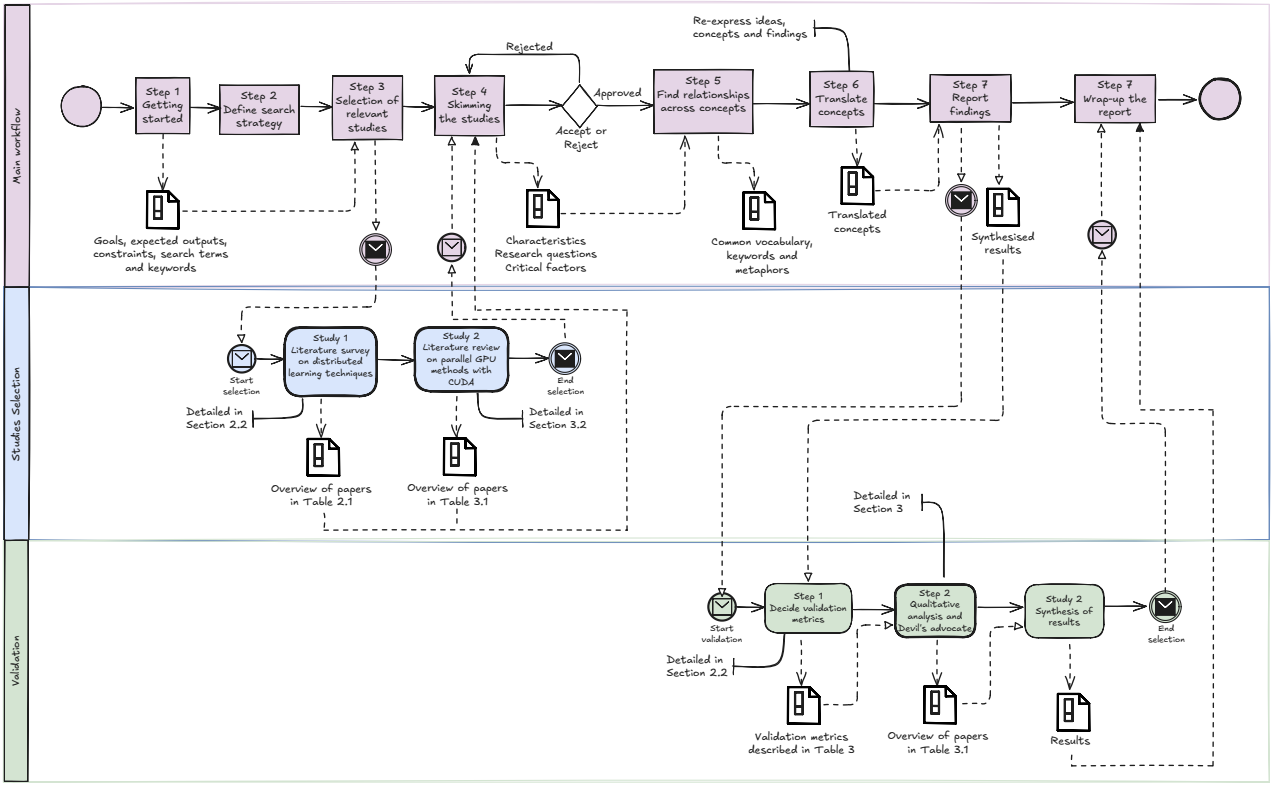

Figure 1. Systematic review workflow showing the main steps, documentation artifacts, and validation processes. The workflow is divided into three main phases: main workflow (top), studies selection (middle), and validation (bottom). Dashed lines indicate documentation and communication flows.

- Commissioning a review which depends on whether or not the systematic review is being done on a commercial basis.
- Evaluating the review protocol and Evaluating the report which are optional and depend on the quality assurance procedures decided by the systematic review team (and any other stakeholders).

The stages listed above may appear to be sequential, but it is important to recognise that many of the stages involve iteration. In particular, many activities are initiated during the protocol development stage, and refined when the review proper takes place. For example:

- The selection of primary studies is governed by inclusion and exclusion criteria. These criteria are initially specified when the protocol is drafted but may be refined after quality criteria are defined.
- Data extraction forms initially prepared during construction of the protocol will be amended when quality criteria are agreed.
- Data synthesis methods defined in the protocol may be amended once data has been collected.

2.3. Study Selection Framework

This section defines the aims and scope of the review, clarifying the types of studies to be included and the rationale for exploring distributed learning and CUDA implementations together. The specific objectives of the review are aligned with the research questions to provide a clear focus. The boundaries of the review concern the types of distributed learning techniques and CUDA implementations considered, with a time period between 2015 and 2022 to ensure currency. The review includes studies focusing on the design and analysis of distributed learning algorithms and those focusing on CUDA-based parallel implementations to understand the translation of theoretical aspects into practical implementations. This approach allows for the exploration of patterns and challenges in mapping distributed algorithms onto parallel architectures **[TODO: Different topics explored together]**.

2.3.1. Distributed Learning Techniques Review

The review will focus on distributed learning approaches, aligning with ”Study 1” in Figure 1, with the following considerations:

- Types of algorithms including **data parallelism, model parallelism, and asynchronous Stochastic Gradient Descent (SGD)** [1, 9].
- Different distributed architectures including parameter servers and peer-to-peer systems [1, 9, 10].
- Specific machine learning models such as neural networks and support vector machines.

2.3.2. CUDA-based Parallel Implementation Review

For CUDA implementations, the review will consider aspects relevant to "Study 2" in Figure 1:

- Implementation of distributed methods on NVIDIA GPUs using the CUDA framework
- Different CUDA libraries and architectures
- Specific hardware considerations including GPUs and Tensor Processing Units (TPUs)

2.3.3. Justification for Inclusion

Both distributed learning techniques and CUDA implementation studies will be included to provide a complete picture of the current state-of-the-art research in the area. By including both study types, a deeper understanding of both theoretical approaches and implementation techniques for practical applications can be reached.

2.4. Preliminary Protocol Development

This systematic review follows the guidelines proposed by Kitchenham and Charters for software engineering research. The preliminary review protocol was developed to establish the foundation for the steps visualized in Figure 1, particularly in the initial stages. An overview of the papers included after the initial selection phase (corresponding to the output of the "Studies Selection" phase in Figure 1) will be presented in Table 2.1.

2.4.1. Background and Rationale

This section provides the necessary context for the review, outlining the research gaps that will be addressed [1]. It explicitly states the need for a systematic review of the current literature to address this gap and provide a focused analysis.

2.4.2. Initial Search Strategy

The initial search strategy involves combining keywords using Boolean and proximity operators to generate search strings, based on the "Goals, expected outputs, constraints, search terms and keywords" documented as an input to Step 1 in Figure 1. Databases like Scopus, Google Scholar, and ACM Digital Library are selected for their coverage of computer science, engineering, and applied mathematics literature. Studies published between 2015-2022 will be considered to ensure recent advancements are included while maintaining a consistent period for analysis.

- **Search Terms:** Details of the search terms will be provided in Section 2.5.
- **Database Justification:** Rationale for selecting specific databases is detailed in Section 2.5.
- **Timeline:** The timeframe for including studies is 2015-2022.

2.4.3. Preliminary Selection Criteria

Preliminary criteria for inclusion will use specific examples such as "studies that evaluate the performance of synchronous distributed SGD in deep learning models" rather than general terms like "distributed computing" [1]. Preliminary quality thresholds will ensure only high-quality studies are included in the final analysis. Specific inclusion and exclusion criteria are detailed in Section 2.6.

2.4.4. Initial Data Extraction Plan

The following information will be extracted from each study:

- Details of distributed systems [1, 9]:
  1. Number of nodes
  2. Communication network
  3. Communication method
  4. Topology
- Machine learning algorithms and models used [11].
- Datasets and benchmarks [1].
- Performance metrics (training time, accuracy, speedup) [1, 9, 11].
- CUDA implementation details (libraries, optimizations) [1, 10, 11].

Further details on the data extraction strategy can be found in Section 2.8.

2.4.5. Quality Assessment Framework

Preliminary quality assessment will use specific criteria to evaluate the validity and reliability of methods, using established checklists from the literature [1]. A Likert scale will be used for a standardized approach. Detailed guidelines for reviewers will be established to ensure consistency and prevent bias, as described further in Section 2.7.

- **Criteria:** Specific criteria are detailed in Section 2.7.
- **Scoring System:** A Likert scale will be used.
- **Guidelines:** Guidelines for reviewers are detailed in Section 2.7.

2.4.6. Synthesis Approach

The synthesis approach will involve meta-analysis where appropriate, using statistical analysis to combine results from included studies with clearly defined methods [1]. Thematic synthesis will be used for narrative synthesis, allowing an in-depth understanding of themes present in selected studies.

- **Meta-analysis:** Details of the methods will be defined later.
- **Narrative synthesis:** Thematic synthesis will be employed.

2.5. Search Process Documentation

This section aims to ensure that the search process is fully transparent and replicable.

2.5.1. Digital Libraries

The digital libraries chosen for this review are Scopus, Google Scholar, and ACM Digital Library. These databases were selected based on their broad coverage and relevance to computer science, engineering, and applied mathematics literature. Plans to include grey literature, such as technical reports and conference proceedings, may be considered to ensure a wider variety of relevant studies are included [1].

2.5.2. Search Strategy

The complete search strings used for each database and how they were adapted to suit the different interfaces, including field restrictions, are presented in Table 1. This ensures the search strategy is repeatable and transparent.

Table 1. Search Strings Used for Each Database

Database	Search String
Scopus	*To be added*
Google Scholar	*To be added*
ACM Digital Library	*To be added*

2.6. Study Selection Criteria

This section specifies the detailed inclusion and exclusion criteria for the studies to be included in the review. These will be specific, measurable, and objective to ensure that all studies are assessed consistently and fairly [1].

2.6.1. Inclusion Criteria

The following criteria will be used for including studies:

- Studies published between 2013 and 2023
- Peer-reviewed articles and high-quality preprints
- Studies focusing on distributed training techniques
- Articles written in English
- Implementation details available

[1, 10].

2.6.2. Exclusion Criteria

Studies will be excluded based on the following criteria:

- Studies not focused on neural network training
- Pure theoretical papers without implementation
- Secondary studies (surveys, reviews)
- Insufficient technical details or results

2.7. Quality Assessment Process

This part of the methodology details the process used to assess the quality of the selected studies, aligning with the ”Validation” phase shown in the bottom section of Figure 1.

2.7.1. Quality Criteria

The quality of the studies will be evaluated based on the methodological rigour, clarity of reporting, limitations of the studies, and potential for bias. Established checklists, such as those provided by the CASP, will be used to address bias and validity in a rigorous and systematic way.

2.8. Data Extraction Strategy

This section details how data will be extracted from the included studies. The data extraction form will be designed to capture all necessary information, including study details, methodology, implementation specifics, dataset details, and results. The form will be piloted to ensure it captures the information effectively [1].

- **Details:** The data extraction form will capture study design, participants, interventions, and outcomes, as well as details of the implementation in distributed systems and parallel CUDA frameworks [1].
- **Piloting:** The extraction form will be piloted to ensure effectiveness [1].
- **Study Details:** The form will capture relevant details including implementation specifics.

By using this methodology, the systematic review will aim to provide a comprehensive and reliable analysis of the current state of research in distributed deep learning and CUDA implementations. The approach used here will enable the review to identify any trends and gaps in current research and make recommendations for future study.

References

[1] Tal Ben-Nun and Torsten Hoefler. Demystifying Parallel and Distributed Deep Learning: An In-depth Concurrency Analysis. *ACM Computing Surveys*, 52(4):1–43, 2020. 1, 2, 4, 5

[2] Francesco Berloco, Vitoantonio Bevilacqua, and Simona Colucci. A Systematic Review of Distributed Deep Learning Frameworks for Big Data. In *Intelligent Computing Methodologies*, pages 242–256, Cham, 2022. Springer International Publishing. 1, 2

[3] Pearl Brereton, Barbara A. Kitchenham, David Budgen, Mark Turner, and Mohamed Khalil. Lessons from applying the systematic literature review process within the software engineering domain. *Journal of Systems and Software*, 80 (4):571–583, 2007. 2

[4] David Budgen, Pearl Brereton, Sarah Drummond, and Nikki Williams. Reporting systematic reviews: Some lessons from a tertiary study. *Information and Software Technology*, 95: 62–74, 2018. 2

[5] Karanbir Chahal, Manraj Singh Grover, and Kuntal Dey. A Hitchhiker’s Guide On Distributed Training of Deep Neural Networks, 2018. arXiv:1810.11787 [cs]. 1

[6] Mohammad Dehghani and Zahra Yazdanparast. From distributed machine to distributed deep learning: a comprehensive survey. *Journal of Big Data*, 10(1):158, 2023. 1

[7] Vinicius dos Santos, Anderson Y. Iwazaki, Katia R. Felizardo, Érica F. de Souza, and Elisa Y. Nakagawa. Sustainable systematic literature reviews. *Information and Software Technology*, 176:107551, 2024. 2

[8] Barbara Kitchenham. Procedures for Performing Systematic Reviews. 2

[9] Matthias Langer, Zhen He, Wenny Rahayu, and Yanbo Xue. Distributed Training of Deep Learning Models: A Taxonomic Perspective. *IEEE Transactions on Parallel and Distributed Systems*, 31(12):2802–2818, 2020. 1, 2, 4

[10] Joost Verbraeken, Matthijs Wolting, Jonathan Katzy, Jeroen Kloppenburg, Tim Verbelen, and Jan S. Rellermeyer. A Survey on Distributed Machine Learning. *ACM Computing Surveys*, 53(2):1–33, 2021. 1, 2, 4, 5

[11] Eric P. Xing, Qirong Ho, Pengtao Xie, and Wei Dai. Strategies and Principles of Distributed Machine Learning on Big Data, 2015. arXiv:1512.09295 [stat]. 1, 4