Computational Mathematics for AI: Numerical Methods and Distributed Computing for Deep Learning on Big Data

Pouya Ataei

October 3, 2024

1 Introduction

This document outlines the protocol for a systematic literature review (SLR) on computational mathematics for AI, focusing on numerical methods and distributed computing techniques for deep learning on big data. The review will follow the PRISMA guidelines (**?**) and Kitchenham's methodology for SLRs in software engineering (**?**).

2 Background

2.1 Rationale

As deep learning models grow in complexity and data volumes continue to increase, there is a critical need to understand and optimize the computational methods underpinning these systems. This review aims to synthesize current knowledge on numerical methods and distributed computing techniques specifically applied to deep learning in big data contexts.

2.2 Objectives

The primary objectives of this SLR are:

1. To identify and categorize state-of-the-art numerical methods used in deep learning for big data.

2. To evaluate the effectiveness of various distributed computing techniques for scaling deep learning to big data problems.

3. To compare these methods and techniques in terms of computational efficiency, scalability, and accuracy.

4. To identify emerging trends and future directions in this field.

3 Research Methodology

This study employs a comprehensive approach combining two systematic literature reviews (SLRs) with subsequent meta-analysis and network analysis. The methodology is structured into seven distinct phases:

3.1 Phase 1: Planning and Protocol Development

3.1.1 Research Questions

For SLR 1 (Numerical Methods):

RQ1.1 What are the state-of-the-art numerical methods used in deep learning for big data?

RQ1.2 How do these methods perform in terms of computational efficiency and accuracy?

For SLR 2 (Distributed Computing Techniques):

RQ2.1 What distributed computing techniques are used for scaling deep learning to big data problems?

RQ2.2 How effective are these techniques in terms of scalability and performance?

3.1.2 Literature Review Classification Framework

We will use Cooper's taxonomy (**?**) to classify the literature in both SLRs:

This classification will be applied to each included study during the data extraction phase. It will help us to:

- Systematically categorize the nature and scope of each study
- Identify patterns and trends in the literature
- Ensure a balanced representation of different types of research in our review

Table 1: Adaptation of Cooper's Literature Review Taxonomy

Characteristic	Categories
(a) Focus	Research outcomes, Research methods, Theories, Practices or applications
(b) Goal	Integration, Criticism, Identification of central issues
(c) Perspective	Neutral representation, Espousal of position
(d) Coverage	Exhaustive, Exhaustive with selective citation, Representative, Central or pivotal
(e) Organization	Historical, Conceptual, Methodological
(f) Audience	Specialized scholars, General scholars, Practitioners or policymakers, General public

- Tailor our findings to different audience needs
- Guide our analysis and synthesis of the literature

The classification results will be used in Phase 6 (Study Classification and Bias Assessment) to provide additional context for interpreting our findings and identifying gaps in the current research landscape.

3.1.3 Search Strategy Development

PICO-based search strings for each SLR:

SLR 1 (TITLE AND ABSTRACT SEARCH) :

```
("deep learning"
AND
("numerical method*"
OR "computational mathematics"
OR "optimization algorithm*")
AND
("big data"))
```

IEEE Explore Search String:

((”deep learning” AND (”numerical method*” OR ”computational mathematics” OR ”optimization algorithm*”) AND (”big data”)))

Scopus Search String:

(TITLE-ABS (”deep learning”) AND (TITLE-ABS (”numerical method*” OR ”computational mathematics” OR ”optimization algorithm*”)) AND TITLE-ABS (”big data”))

Aisel Search String: ([[Title:"deep learning"] AND [Title: "numerical method*"]] OR [Title: "computational mathematics"] OR [[Title: "optimization algorithm*"] AND [Title: "big data"])

ACM Search String: ([[[Title: "deep learning"] AND [Title: "numerical method*"]] OR [Title: "computational mathematics"] OR [[Title: "optimization algorithm*"] AND [Title: "big data"]]] AND [[Abstract: "deep learning"] OR [Abstract: "numerical method*"] OR [Abstract: "computational mathematics"] OR [Abstract: "optimization algorithm*"] OR [Abstract: "big data"]])

SLR 2:

```
(("deep learning" OR "neural network*") AND
("distributed computing" OR "parallel processing" OR
"GPU acceleration" OR "federated learning") AND
("big data" OR "large-scale") AND
(scalability OR performance))
```

3.1.4 Information Sources

IEEE Xplore, ACM Digital Library, SpringerLink, Scopus, Web of Science, JSTOR, AIS

3.1.5 Eligibility Criteria

Inclusion criteria for SLR 1:

- Studies published between January 1, 2014 and September 21, 2024
- Peer-reviewed journal articles and full conference papers
- English language publications
- Studies focusing on numerical methods for deep learning in big data contexts
- Research explicitly addressing computational efficiency or accuracy of numerical methods
- Studies providing quantitative, qualitative results or comparative analyses of numerical methods

Exclusion criteria for SLR 1:

- Studies not explicitly addressing big data characteristics

- Publications without clear details on the numerical methods used
- Review papers, editorials, or opinion pieces
- Short papers (less than 10 pages), extended abstracts, or posters
- Duplicate studies or multiple publications of the same research

3.2 Phase 2: Literature Search and Study Selection

3.2.1 Search Execution

1. Execute search strategy on selected databases
2. Import results to a unified CSV file

3.2.2 Deduplication

1. Remove duplicates

3.2.3 Initial Screening

1. Initial screening of titles and abstracts
2.

3.2.4 Deeper Screening

1. Full-text assessment of potentially eligible studies

Document selection process should be done using PRISMA flow diagram.

3.3 Phase 3: Quality Assessment

The quality of individual studies will be assessed using a criteria made up of 7 elements, inspired by the CASP checklist for assessing qualitative research and Kitchenham's guidelines on empirical research in software engineering . This assessment will be applied to studies in both SLRs.

3.3.1 Quality Assessment Criteria

The criteria test literature on 4 major areas:

1. **Minimum quality threshold:**
   - Does the study report empirical research or is it merely a 'lesson learnt' report based on expert opinion?
   - Are the objectives and aims of the study clearly communicated, including the reasoning for why the study was undertaken?
   - Does the study provide adequate information regarding the context in which the research was carried out?

2. **Rigour:**
   - Is the research design appropriate to address the objectives of the research?
   - Is there a data collection method used and is it appropriate?

3. **Credibility:**
   - Does the study report findings in a clear and unbiased manner?

4. **Relevance:**
   - Does the study provide value for practice or research?

3.3.2 Assessment Process

1. The assessment will be conducted in two phases:
   - Phase 1: Assess only the minimum quality threshold criteria.
   - Phase 2: If a study passes Phase 1, assess it for rigour, credibility, and relevance.
2. reviewers will independently assess each study.
3. Each criterion will be scored as either 'yes' or 'no'.
4. A study passes the quality assessment if it receives positive responses for at least 75% of the criteria.
5. Inter-rater reliability will be assessed using Krippendorff's alpha, aiming for q ¿ 0.8.
6. Disagreements will be resolved through discussion. If consensus cannot be reached, a third reviewer will be consulted.

3.3.3 Quality Threshold

To be included in the final analysis, a study must:

- Pass all criteria in the minimum quality threshold category (Phase 1)
- Receive positive responses for at least 75% of all criteria (Phase 1 and 2 combined)
- Achieve at least 75% inter-rater reliability

This quality assessment framework will ensure that only studies meeting a minimum standard of methodological rigour and relevance are included in our analysis, thereby enhancing the reliability and validity of our findings.

Quality threshold: 75% positive responses, 75% inter-rater reliability (Krippendorff's q ¿ 0.8)

3.4 Phase 4: Data Extraction

3.4.1 Data Extraction

Using a standardized, pre-piloted form to extract:

- Bibliographic information
- Methods/techniques used
- Problem domain and dataset characteristics
- Performance metrics
- Hardware and software environment
- Key findings and limitations

3.4.2 Quality Assessment

3.5 Phase 4: Data Synthesis for Individual SLRs

For each SLR:

- Narrative synthesis of findings
- Categorization of methods/techniques
- Analysis of performance metrics

3.6 Phase 5: Combined Analysis

3.6.1 Meta-Analysis

- Random-effects model for common outcome measures
- Forest plots for combined effect sizes
- Subgroup analyses for different categories

3.6.2 Network Analysis

- Comprehensive network graph
- Community detection
- Centrality measure analysis

3.7 Phase 6: Study Classification and Bias Assessment

3.7.1 Study Classification

Classify all studies according to Cooper's taxonomy:

- Focus, Goal, Perspective, Coverage, Organization, Audience

3.7.2 Assessment of Meta-Bias

- Funnel plot examination
- Egger's test for small-study effects

3.8 Phase 7: Synthesis and Reporting

- Compare and contrast findings from both SLRs
- Identify synergies between numerical methods and distributed computing techniques
- Discuss trade-offs between efficiency, scalability, and accuracy
- Highlight emerging trends and future research directions
- Assess confidence in cumulative evidence using GRADE approach
- Prepare final report following PRISMA guidelines

This phased approach ensures a systematic and comprehensive review of computational mathematics for AI in big data contexts, combining insights from numerical methods and distributed computing techniques.

4 Discussion

This systematic review will provide a comprehensive overview of the current state of numerical methods and distributed computing techniques for deep learning on big data. The findings will be interpreted considering the strength of evidence, applicability, and generalizability. Limitations of the review and the included studies will be discussed, and implications for future research will be outlined.

5 Notes

There was a challenge among researchers to detect big data or what constitutes big data. While some studies did run their numerical method against a large set of data, it was not always clear if it was big data.

For instance one paper discussed fault prediction with a large amount of data, but it did not occur naturally to us that this data could be big data. It was only clarified during the discussion phase that the data was indeed big data.