Distributed Deep Learning in Large-Scale Datasets

Maria Koleva  
CCT CollegeDublin, Ireland  
sba23020@student.cct.ie

*Abstract*—Deep learning is a powerful subset of machine learning that can handle complex tasks such as speech recognition, computer vision, and natural language processing. However, the success of deep learning models is possible due to the availability of large-scale datasets and big computational resources. As a result, the use of distributed processing systems, such as Apache Spark, has become crucial for the efficient processing of the large amount of data required for the deep learning tasks. In this paper, the design and implementation of distributed neural network architectures will be explored, with focus on their application in handling large-scale datasets.

Keywords—distributed deep learning, neural networks, large-scale datasets, big data analytics, Apache Spark, HDFS

# Introduction

The amount of newly created data has been increasing at a remarkable rate. According to predictions cited by Forbes, the volume of data generated in 2020 was expected to grow 44 times, reaching 35 zettabytes (35 trillion gigabytes). Remarkably, just two years prior, the world had already accumulated 33 zettabytes of data, lading experts to forecast that by 2025, the global data volume could increase to 175 zettabytes (175 trillion gigabytes)(Press,2020). The term big data is usually broken down into V’s characteristics, starting from 3 up to 42 (Shafer, 2017, Elder Research). Usually, the most common ones are: Volume – refers to the magnitude of the data that is being generated and collected, Velocity – refers to the rate of generation of data, Variety – refers to different types of data that are being generated. As data is generated from various sources at unprecedented speeds in diverse formats, organizations have to adapt their infrastructure and tools in order to effectively store, process, and analyse the data. One approach for handling big data is by incorporating the use of distributed systems, which distribute data processing tasks across different nodes/machines. Additionally, advanced machine learning techniques, and more specifically neural networks, have emerged as a powerful tool for extracting insights out of big data. Neural networks make decision in a way similar to how the human brain works. It uses interconnected nodes/neurons in a layered structure that resembles the human brain, creating a system that computers use to learn from their mistakes and improve. Exploring the computational capabilities of distributed systems, such as Hadoop and Apache, this study will investigate the capacity of neural networks in processing and analysing large-scale datasets.

# Literature Review

## Introduction

Deep Learning is a subset of Machine Learning, which in turn is a subset of Artificial Intelligence. Deep learning, a type of machine learning inspired by the structure of the human brain, uses multi-layered neural network to extract features and make predictions with unprecedented accuracy. While machine learning algorithms are widely used for various tasks, such as classification, regression, and clustering, deep learning distinguishes itself by building multiple levels of representation from a series of other simple representations (Dev, 2017). In the era of big data, characterized by the unprecedented increase in data volume, velocity, and variety, the importance of scalability in machine learning algorithms has become crucial. To address this challenges, numerous authors have explored scaling up machine learning algorithms through distribution and parallelization (Dean et al., n.d.; Hegde & Usmani, n.d.). By leveraging distributed computing frameworks and parallel processing techniques, such as MapReduce and Apache Spark, machine learning tasks can be efficiently executed across multiple machines. This enables seamless handling of large-scale datasets and enhances computational efficiency.

## Neural Networks for Large-Scale Data Processing

Neural networks consist of interconnected nodes, arranged in layers. These layers include an input layer, hidden layers, and an output layer. Neural networks can be designed with various architectures, including deep neural networks (DNNs) with multiple hidden layers as seen on Figure 1. These deep architectures enable neural networks to extract complex features from raw data, making them well-suited for tasks such as image recognition, natural language processing, and speech recognition.

A diagram of a neural network

Description automatically generated

Figure 1: Neural Network architecture (Adapted from Dev, 2017)

Each neuron, or node, in a neural network performs a series of mathematical operations on the input data and produces an output signal. The mathematical representation of the processing performed by the neuron can be represented as follows:

Where:

Output is the output of the neuron

is the activation function, which introduces nonlinearity in the model

*wi*are the weights associated with each input

*xi* are the input signals

*b* is the bias term

In this formulation, the weighted sum of the inputs, along with the bias term, is passed through an activation function to produce the output of the neuron. This activation process is crucial for propagating information through the network and generating meaningful predictions or classifications (IBM, n.d.).

Biases play a crucial role in neural networks by allowing the model to capture complex relationships in the data more effectively. Each neuron in the network is associated with its bias term, which is added to the weighted sum of inputs before applying the activation function. This bias term enables neurons to exhibit varying levels of activation even when all input values are zero, enhancing the flexibility and expressiveness of the neural network.

In the training process of neural networks, optimization algorithms such as Stochastic Gradient Descent (SGD) play a crucial role. SGD is specifically designed for large-scale datasets, where computing gradients for the entire dataset is infeasible. Instead, SGD computes gradients based on small, random subsets of data, called mini-batches (Nielsen, 2015). The SGD algorithm involves several key steps: feed-forward evaluation, back-propagation, and weight updates. During feed-forward evaluation, the neural network processes input data and produces output predictions. Back-propagation computes gradients of the loss function with respect to model parameters, which are then used to adjust the model weights to minimize the loss function.

Another crucial component of the neural networks is the loss or cost function. It quantifies the disparity between the predicted outputs of the model and the truth labels in the train data. It serves as the measure of how well the model is performing on a given task. The goal of a learning algorithm is to minimize the loss function systematically. In the case of neural networks, the total loss function is a separable and differentiable function of the model parameters.

The goal of SGD is to update the model weights to minimize the loss function, which results in improving the model’s predictive accuracy and performance.

*C.* ***Distributed Deep Learning Frameworks***

The success of deep learning models relies heavily on the availability of large-scale datasets and powerful computational resources. In response to the growing demand for scalable data processing, researchers have explored the integration of deep learning with distributed computing frameworks like Hadoop and Apache Spark (Indirman, Wiriasto, & Akbar, 2023).

Hadoop is an open source software framework for storage and processing of large-scale datasets on clusters of commodity hardware. At its core, Hadoop consists of several key components:

Hadoop Distributed File System (HDFS) is the primary storage system used by Hadoop for storing large volumes of data across multiple machines in a distributed manner. It can be described as a file system that stores data not on a single Hard Disk Drive (HDD). Instead, data is broken into pieces (files are divided into blocks with a configurable size of 64 MB) and stored across a cluster consisting of several computers (Apache Hadoop, 2008). These blocks are then replicated across different DataNodes within the Hadoop cluster to ensure data reliability and availability. By spreading data blocks across numerous machines, HDFS enables parallel processing and fault tolerance, as the loss of any individual machine does not result in data loss. The HDFS architecture consists of two primary components: the NameNode and DataNodes. The NameNode is responsible for managing the file system namespace and metadata, including information about the location and replication factor of each data block. DataNodes, on the other hand, are responsible for storing and serving data blocks upon client requests.

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

# Prepare Your Paper Before Styling

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections A-D below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

## Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

## Units

* Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive”.
* Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
* Do not mix complete spellings and abbreviations of units: “Wb/m2” or “webers per square meter”, not “webers/m2”. Spell out units when they appear in text: “. . . a few henries”, not “. . . a few H”.

Identify applicable funding agency here. If none, delete this text box.

* Use a zero before decimal points: “0.25”, not “.25”. Use “cm3”, not “cc”. (*bullet list*)

## Equations

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

Number equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop. To make your equations more compact, you may use the solidus ( / ), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

*a**b* 

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1)”, not “Eq. (1)” or “equation (1)”, except at the beginning of a sentence: “Equation (1) is . . .”

## Some Common Mistakes

* The word “data” is plural, not singular.
* The subscript for the permeability of vacuum **0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter “o”.
* In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
* A graph within a graph is an “inset”, not an “insert”. The word alternatively is preferred to the word “alternately” (unless you really mean something that alternates).
* Do not use the word “essentially” to mean “approximately” or “effectively”.
* In your paper title, if the words “that uses” can accurately replace the word “using”, capitalize the “u”; if not, keep using lower-cased.
* Be aware of the different meanings of the homophones “affect” and “effect”, “complement” and “compliment”, “discreet” and “discrete”, “principal” and “principle”.
* Do not confuse “imply” and “infer”.
* The prefix “non” is not a word; it should be joined to the word it modifies, usually without a hyphen.
* There is no period after the “et” in the Latin abbreviation “et al.”.
* The abbreviation “i.e.” means “that is”, and the abbreviation “e.g.” means “for example”.

An excellent style manual for science writers is [7].

# Using the Template

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

## Authors and Affiliations

**The template is designed for, but not limited to, six authors.** A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

### For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

### For papers with less than six authors: To change the default, adjust the template as follows.

#### Selection: Highlight all author and affiliation lines.

#### Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select the correct number of columns from the selection palette.

#### Deletion: Delete the author and affiliation lines for the extra authors.

## Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named “Heading 1”, “Heading 2”, “Heading 3”, and “Heading 4” are prescribed.

## Figures and Tables

#### Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence.

1. Table Type Styles

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

1. Sample of a Table footnote. (*Table footnote*)
2. Example of a figure caption. (*figure caption*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

##### Acknowledgment *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

##### References

Press, G. (2020). 6 Predictions About Data In 2020 And The Coming Decade. Forbes. Retrieved from <https://www.forbes.com/sites/gilpress/2020/01/06/6-predictions-about-data-in-2020-and-the-coming-decade/?sh=297c7024fc36>

Shafer, C. (2017). The 42 V’s of Big Data and Data Science. Elder Research. Retrieved from <https://www.elderresearch.com/blog/the-42-vs-of-big-data-and-data-science/>

Dean, J., Corrado, G. S., Monga, R., Chen, K., Devin, M., Le, Q. V., ... Ng, A. Y. (n.d.). Large Scale Distributed Deep Networks. Google Inc., Mountain View, CA. Retrieved from <https://proceedings.neurips.cc/paper_files/paper/2012/file/6aca97005c68f1206823815f66102863-Paper.pdf>

Hegde, V., & Usmani, S. (n.d.). Parallel and Distributed Deep Learning. Stanford University. Retrieved from <https://web.stanford.edu/~rezab/classes/cme323/S16/projects_reports/hedge_usmani.pdf>

Dev, D. (2017). Deep Learning with Hadoop. Packt Publishing Ltd.

IBM. (n.d.). Neural networks. Retrieved from <https://www.ibm.com/topics/neural-networks>

Nielsen, M. A. (2015). *Neural Networks and Deep Learning*. Determination Press. Retrieved from <https://jingyuexing.github.io/Ebook/Machine_Learning/Neural%20Networks%20and%20Deep%20Learning-eng.pdf>

Indirman, M. D. C., Wiriasto, G. W., & Akbar, L. A. S. I. (2023). Distributed Machine Learning using HDFS and Apache Spark for Big Data Challenges. Retrieved from https://www.e3s-conferences.org/articles/e3sconf/pdf/2023/102/e3sconf\_icimece2023\_02058.pdf

Apache Hadoop. (2008). HDFS design. Retrieved from <https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html>

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

1. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*
2. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
3. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
4. K. Elissa, “Title of paper if known,” unpublished.
5. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
6. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
7. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord “Format” pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.