

EFFICIENTLY RUN A DEEP LEARNING MODEL ON A DISTRIBUTED CPU SYSTEM

D.D.K.K. Gnanasena

218524K

Degree of Master of Science in Artificial Intelligence

Department of Computational Mathematics
Faculty of Information Technology

University of Moratuwa
Sri Lanka

April 2024

EFFICIENTLY RUN A DEEP LEARNING MODEL ON A DISTRIBUTED CPU SYSTEM

D.D.K.K. Gnanasena

218524K

Thesis/Dissertation submitted in partial fulfillment of the requirements for the degree
Degree of Master of Science in Artificial Intelligence

Department of Computational Mathematics
Faculty of Information Technology

University of Moratuwa
Sri Lanka

April 2024

DECLARATION

I declare that this is my own work and this thesis/dissertation does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other University or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature:

Date:2024-04-26

The above candidate has carried out research for the PhD/MPhil/Masters thesis/dissertation under my supervision. I confirm that the declaration made above by the student is true and correct.

Name of Supervisor: Dr. Subha Fernando

Signature of the Supervisor:

Date:

ACKNOWLEDGEMENT

At the very first I take this opportunity to thank my research project supervisor Dr. Subha Fernando for guiding me throughout my research project by giving advices, opinions & suggestions as well as by encouraging me. Her expertise, knowledge, and unwavering commitment have been instrumental in helping me shape my ideas, refine my research questions, and develop my analytical skills.

In addition, I'm grateful to my friends, family, and coworkers for providing a supportive and cooperative work environment for me. Their enthusiasm, encouragement, and comments have been invaluable in helping me with this research.

I want to thank everyone who has contributed to this work.

ABSTRACT

In Distributed Deep Learning, Data transfer between Computer equipment might act as a bottleneck that prevents the system from growing to its full capacity. To overcome this, people have come with methods like Quantization and Finetuning which work well only on GPUs and TPUs which are costly infrastructure. However, if we could suffice from running deep learning models on CPU based Distributed System, it is a great edge in cost reduction. So, in this research such a system is created and evaluated to find out whether they can compete with GPUs using MultiWorkerMirroredStrategy. MNIST dataset on a linear model, CIFAR10 dataset on a Convolution Neural Network (CNN) model and finally a quantized Llama 2 model was run to summarize and translate an essay. They were evaluated by their accuracy and time taken to train against running them in a Google Colab T4 GPU(For Llama model, BertScore and Rouge Score for summarization and BertScore and BleuScore for Translation was used). Performance of MNIST data concluded, this system is inefficient for small models as the time taken to train on two machines was 200 times of a single machine. When running CIFAR10 dataset CNN accuracy decreases over the number of machines on multi-node and the increasing batch sizes and time taken decreases and increased after a certain batch size showing the bottleneck of communication. In Llama model it was observed that with increasing dataset size the time taken also reduces when compared with GPU time taken. None of the models were fast enough to pass the time taken for the GPU in all cases but some accuracies in Llama model were able to pass GPU accuracies. For future works Model parallelism should also be tested for further development of this CPU Distributed system.

Keywords: MultiWorkerMirroredStrategy, multi-node, CPU, Quantization, Finetuning

TABLE OF CONTENTS

Declaration	i
Acknowledgement.....	ii
Abstract	iii
Table of Contents	iv
List of Figures	vii
List of Tables.....	viii
List of Abbreviations.....	ix
Chapter 1	1
Introduction	1
1.1 Prolegomena.....	1
1.2 Objectives	1
1.3 Background and Motivation	2
1.4 Problem Definition	3
1.5 Approach of the CPU Distributed System.....	3
1.6.1 Resource Requirement	3
1.6.1 Hardware	3
1.6.2 Good Internet Access	3
1.7 Organizational Structure of the Thesis	4
1.8 Summary	4
Chapter 2	5
A LITERATURE REVIEW ON DISTRIBUTED SYSTEMS.....	5
2.1 Introduction	5
2.2 Overview of Distributed Systems.....	5
2.3 AI Frameworks used for Distributed Systems.....	6
2.4 Current applications of Distributed Systems.....	7
2.5 Major Challenge in Distributed Systems.....	11
2.6 Summary of the Literature Analysis.....	13
2.7 Research Problem	16
2.8 Summary	17

Chapter 3	18
TECHNOLOGIES ADOPTED FOR DISTRIBUTED SYSTEMS.....	18
3.1 Introduction	18
3.2 MultiWorkerMirroredStrategy for CPUs	18
3.3 Summary	18
Chapter 4	19
An approach to implement a CPU distributed system	19
4.1 Introduction	19
4.2 Hypothesis	19
4.3 Input.....	19
4.4 Output.....	19
4.5 Process.....	19
4.6 Features	20
4.7 Users	20
4.8 Summary	20
Chapter 5	21
Design of the CPU Distributed System.....	21
5.1 Introduction	21
5.2 Distributed CPU system’s Architecture	21
5.3 Create a CPU Distributed System	22
5.4 LLM improving.....	22
5.5 Comparative analysis	22
5.8 Summary	22
Chapter 6	23
The Implementation of the CPU Distributed System	23
6.1 Introduction	23
6.2 Creation of CPU Distributed System	23
6.2.1 MNIST dataset	23
6.2.2 CIFAR10 dataset.....	23
6.3 LLM improving.....	24
6.3.1 Llama model Quantization.....	24
6.3.2 Llama model evaluation.....	24

6.4	Comparative analysis	24
6.5	Summary	25
Chapter 7		26
The Evaluation of the Implemented CPU Distributed System		26
7.1	Introduction	26
7.2	MNIST Dataset Results	26
7.3	CIFAR10 Results	27
7.3.1	Accuracies	27
7.3.2	Time Taken	29
7.4	Llama Model Results for Summarization.....	31
7.4.1	Time taken for the three machines for summarization.....	31
7.4.2	Summarization results selected from the best time taken out of the three machines.....	34
7.4.3	Time taken on the Summarization results	37
7.4.4	BERTscores of the Summarization results	39
7.4.5	Rouge Scores of the Summarization results	43
7.5	Llama Model Results for Translations from English to German	52
7.5.1	Time taken for the three machines for Translations.....	52
7.5.2	Translation results selected from the best time taken out of the three machines.....	55
7.5.3	Time taken on the Summarization results	57
7.5.4	BERTscores of the Translation results.....	59
7.5.5	BleuScores of the Translation results.....	63
7.6	Summary	64
Chapter 8		65
Conclusion and FuTuRE Work.....		65
8.1	Introduction	65
8.2	Conclusion.....	65
8.3	Further work	66
8.4	Summary	67
References		68

List of Figures

Figure	Description	Page
Figure 5.1	Architecture of the CPU Distributed System	21
Figure 7.1	The training time taken for 01 Machine	26
Figure 7.2	The training time taken for 02 Machine	26
Figure 7.3	The training stopped in the 03rd epoch due to a communication error.	27
Figure 7.4	Accuracies over Batch Size	28
Figure 7.5	Time Taken over Batch Size	29
Figure 7.6	Time taken on the selected Summarization results	38
Figure 7.7	BERTscores of the selected Summarization results for the entire essay	40
Figure 7.8	BERTscores of the selected Summarization results for text under maximum tokens of Llama 2	41
Figure 7.9	BERTscores of the selected Summarization results for part of the Essay	42
Figure 7.10	Recalls of Rouge Scores of the selected Summarization results for the entire essay	43
Figure 7.11	Precisions of Rouge Scores of the selected Summarization results for the entire essay	44
Figure 7.12	F1-Scores of Rouge Scores of the selected Summarization results for the entire essay	45
Figure 7.13	Recalls of Rouge Scores of the selected Summarization results for text under maximum tokens of Llama 2	46
Figure 7.14	Precisions of Rouge Scores of the selected Summarization results for text under maximum tokens of Llama 2	47
Figure 7.15	F1-Scores of Rouge Scores of the selected Summarization results for text under maximum tokens of Llama 2	48
Figure 7.16	Recalls of Rouge Scores of the selected Summarization results for part of the Essay	49
Figure 7.17	Precisions of Rouge Scores of the selected Summarization results for part of the Essay	50
Figure 7.18	F1 Scores of Rouge Scores of the selected Summarization results for part of the Essay	51
Figure 7.19	Time taken on the selected Translation results	57
Figure 7.20	BERTscores of the selected Translation results for the entire essay	60
Figure 7.21	BERTscores of the selected Translation results for text under maximum tokens of Llama 2	61
Figure 7.22	BERTscores of the selected Translation results for part of essay	62
Figure 7.23	BleuScores on the selected Translation results	63

LIST OF TABLES

Table	Description	Page
Table 2.1	Summary of literature review	13
Table 7.1	Accuracies results	27
Table 7.2	Time taken results	29
Table 7.3	Summarization Results for the entire essay	31
Table 7.4	Summarization Results for text under maximum tokens of Llama 2	32
Table 7.5	Summarization Results for part of the Essay	33
Table 7.6	Selected Summarization Results for the entire essay	34
Table 7.7	Selected Summarization Results for text under maximum tokens of Llama 2	35
Table 7.8	Selected Summarization Results for part of the Essay	36
Table 7.9	Time taken on the selected Summarization results	37
Table 7.10	BERTscores of the selected Summarization results	39
Table 7.11	Rouge Scores of the selected Summarization results for the entire essay	43
Table 7.12	Rouge Scores of the selected Summarization results for text under maximum tokens of Llama 2	46
Table 7.13	Rouge Scores of the selected Summarization results for part of the Essay	49
Table 7.14	Translation results for the three machines for Translating the entire essay	52
Table 7.15	Translation results for the three machines for Translating text under maximum tokens of Llama 2	53
Table 7.16	Translation results for the three machines for Translating part of the essay	54
Table 7.17	Selected Translation Results for the entire essay	55
Table 7.18	Selected Translation Results for Translating text under maximum tokens of Llama 2	56
Table 7.19	Selected Translation Results for Translating part of the essay	56
Table 7.20	Time taken on the selected Translation results	57
Table 7.21	BERTscores on the selected Translation results	59
Table 7.22	BleuScores on the selected Translation results	63

LIST OF ABBREVIATIONS

Abbreviation	Description
DDL	Distributed Deep Learning
CNN	Convolutional Neural Networks