Master Thesis

Distributed Deep Learning

Joeri R. Hermans

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

at

Maastricht University
Faculty of Humanities and Sciences
Department of Data Science & Knowledge Engineering
Maastricht, The Netherlands

Preface

This thesis is submitted as a final requirement for the Master of Science degree at the Department of Data Science & Knowledge Engineering of Maastricht University, The Netherlands. The subject of study originally started as a pilot project with Jean-Roch Vlimant, Maurizio Pierini, and Federico Presutti of the EP-UCM group (CMS experiment) at CERN. In order to handle the increased data rates of LHC Run 3 and High Luminosity LHC, the CMS experiment is considering to construct a new architecture for the High Level Trigger based on Deep Neural Networks. However, they would like to significantly decrease the training time of the models as well. This would allow them to tune the neural networks more frequently. As a result, we started to experiment with various state of the art distributed optimization algorithms. Which resulted in the achievements and insights presented in this thesis.

I would like to express my gratitude to several people. First and foremost, I would like to thank my promotor [redacted] for his expertise and suggestions during my research, which drastically improved the quality of this thesis. Furthermore, I would also like to thank my friends, colleagues and scientists at CERN for their support, feedback, and exchange of ideas during my stay there. It was a very motivating and inspiring time in my life. Especially the support and experience of my CERN supervisors, Zbigniew Baranowski, and Luca Canali, was proven to be invaluable on multiple occasions. I would also like to thank them for giving me the personal freedom to conduct my own research. Finally, I would like to thank my parents and grandparents who always supported me, and who gave me the chance to explore the world in this unique way.

Joeri R. Hermans
Geneva, Switzerland 2016 - 2017

Abstract

Abstract here.

Summary

Summary here.

Contents

Abbreviations and Notation

ASGD	Asynchronous Stochastic Gradient Descent
CERN	European Organization for Nuclear Research
CMS	Compact Muon Solenoid
EASGD	Elastic Averaging Stochastic Gradient Descent
HL-LHC	High Luminosity Large Hadron Collider
LHC	Large Hadron Collider
SGD	Stochastic Gradient Descent

Chapter 1

Introduction

In this chapter we will give an introduction to Distributed Deep Learning and the problems surrounding it. A more detailed description of the subject of study is given in Chapter 2. Furthermore, we make the reader more comfortable with the notation and abbreviations used throughout this thesis. Finally, we formally define the problem statement in Section 1.2, and give an outline of this thesis in Section 1.3.

1.1 Distributed Deep Learning, an introduction

Unsupervised feature learning and deep learning has shown that being able to train large models can drastically improve model performance. However, consider the problem of training a deep network with millions or even billions of parameters. How do we achieve this without waiting for days, or even weeks? Dean et al. propose a different training paradigm which allows us to train and serve a model on multiple physical machines [1]. The authors propose two novel methodologies to distribute stochastic gradient descent.

1.1.1 Model Parallelism

1.1.2 Data Parallelism

1.2 Problem Statement

1.3 Thesis Outline

Chapter 2

Distributed Deep Learning

Bibliography

[1] Jeffrey Dean et al. “Large scale distributed deep networks”. In: *Advances in neural information processing systems*. 2012, pp. 1223–1231.

Appendices