University of Warsaw
Faculty of Mathematics, Informatics and Mechanics

Tomasz Grześkiewicz
Student no. 394317

Mateusz Kobak
Student no. 385760

Iwona Kotlarska
Student no. 394380

Krzysztof Piesiewicz
Student no. 385996

Dataloading optimisation for deep learning on NVIDIA GPUs

Bachelor's thesis
in COMPUTER SCIENCE

Supervisor:
dr Janusz Jabłonowski
University of Warsaw
Faculty of Mathematics, Informatics, and Mechanics
Institute of Informatics

May 2020

Supervisor's statement

Hereby I confirm that the presented thesis was prepared under my supervision and that it fulfils the requirements for the degree of Bachelor of Computer Science.

Date Supervisor's signature

Authors' statements

Hereby I declare that the presented thesis was prepared by me and none of its contents was obtained by means that are against the law.

The thesis has never before been a subject of any procedure of obtaining an academic degree.

Moreover, I declare that the present version of the thesis is identical to the attached electronic version.

Date Authors' signatures

Abstract

In this thesis dataloading process has been described for needs of deep learning. The concept of the process has been deliberated in terms of time consumed by computing machines. There is an overview of a current state of the domain including programming techniques used by Python deep learning frameworks: PyTorch and TensorFlow. The aim of thesis has been finding bottle necks of dataloading for common deep learning tasks and optimise some of them at the internal code level of the frameworks or the usage of the frameworks. The Analysis is limited to PyTorch and TensorFlow cooperating with NVIDIA GPUs. As a result of the analysis, several bottle necks has been identified. Some of them has been optimised at the internal framework level, a few others has been explained how to be optimised adjusting the use of the frameworks.

Keywords

deep learning, GPU, dataloader

Thesis domain (Socrates-Erasmus subject area codes)

11.3 Informatics, Computer Science

Subject classification

D. Software

Tytuł pracy w języku polskim

Optymalizacja wprowadzania danych na karty graficzne NVIDIA

Contents

Introduction

For several years the terms *Deep learning* and *Neural networks* have been getting more and more recognisable thanks to the great impact of this technology on a huge variety of business aspects. *Speech recognition*, *interpretation of natural language*, *image classification*, *computer vision* and even *self-driving vehicles* are examples of the use of the deep learning technology. The increasing demand for deep learning solutions has caused the race to maximised the performance of available computing machines. The research and practise have shown that GPUs are best suited for deep learning computation purpose because they are unrivaled in linear algebra computations. There are several manners of increasing the performance and all of them are the interests of research and development teams of information technology industries. Some of the ways are constructing more powerful computing clusters or faster computation units. Others are targeted at software issues (operational systems, programming and algorithmic techniques). *Dataloading* mostly relates to the category which has been mentioned as the last one. Let assume the following definition.

Dataloading is the process of copying data from non-volatile storage to GPU memory including processing the data in CPU.