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 performance of data loading process for deep learning on Nvidia GPUs has been analyzed. The overview of the currently used techniques of optimization are included. Over the course of the work, data loading process has been profiled to identify bottlenecks in the most common use cases in PyTorch and TensorFlow frameworks and chosen ones were mitigated, on either internal level of those frameworks or by creating examples of usage that improve the performance.

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

Topic

For several years the terms *deep learning* and *neural networks* have been getting more and more attention, thanks to the great impact of this technology on a wide variety of applications. *Speech recognition*, *natural language processing*, and *computer vision* are only a few examples of the use cases of the deep learning technology. The increasing demand for deep learning solutions has caused the race for maximizing the performance of computations. From the very beginning of this field GPUs have reached state-of-the-art performance for almost every model and they still remain unbeaten as the best general purpose deep learning computing device.

The performance of deep learning process depends on many different factors. In this thesis we focus on the data loading process, which is the first step of the model pipeline. By data loading we mean all the operations that are done on the training data before it is processed by the actual neural network. It implies data loading includes not only data preprocessing routines s.a. shuffling, augmenting and different kinds of transformations, but also data transfers (CPU to GPU, GPU to GPU, etc.) and memory alignment. On one hand, in most cases data loading is not a bottleneck of the whole deep learning model, but as the GPUs are becoming faster it will likely be the case in the near future.

The process of data loading can be optimized in many ways. Big effort is being put into writing efficient data preprocessing code which runs on GPU - one can mention Nvidia DALI as an example here. Furthermore, deep learning frameworks, s.a. PyTorch and Tensorflow, provide solutions for parallelizing the data loading flow with the training of the neural network. The right use of appropriate features of CUDA libraries is also crucial in terms of the framework's internal code. However, it appears that there are some good practices of using the deep learning framework that might drastically increase the data loading performance.

Over the course of the work we profiled various use cases for PyTorch and TensorFlow. We identified and mitigated some of the most common bottlenecks. We measured achieved performance improvements.

Functionality