Parallel Computer Systems, Fall 2022
Instructions for Databar Exercise 9: Parallelism in Machine learning

November 8, 2022

1 Introduction

You will this week explore machine learning and how machine learning algorithms could map to hardware.

You report on this exercise with up to a page in the third report. This exercise is the final exercise in the third report.

Before attempting to solve the exercise, study a tutorial on machine learning hardware, `https://arxiv.org/abs/1703.09039`.

2 Reports and rules

DTU has a zero tolerance policy on cheating and plagiarism. This also extends to the reports and indeed all your work. For example, to copy text passages from someone else without clearly and properly citing your source is considered plagiarism. You are assumed to stay informed of and follow DTU's rules.

3 Working on the exercises

We will explore a simple machine learning algorithm for image processing. When working with this, very open, exercise, it is important that you state all assumptions you make.

Input images are 300x300 pixels with three color channels, the typical red, green and blue channels.

We first explore the computational density. Assuming we use a 3x3 filter kernel and a single output channel, how many multiplications and additions are needed to generate a fmap for an input image. If we use 64 output channels, how many operations would we need?

Lets now add a fully connected layer to the outputs of all the filters so that for a single image we get 20 outputs. In total for both filters and the fully connected layer, how many multiplications and additions would we need?

What is the maximal speedup we can reach with parallelism for this machine learning model?

We now explore the hardware design space. Assume that a small memory, less than 512 values cost 6. A large memory cost 200 and an operation, multiplication or addition, cost 1. Assuming we need to store the input image, all weights and temporaries in memories, what memories would you propose to reduce cost while at the same time allow for significant speedup over sequential execution of the model? What is the resulting cost? Sketch your proposed architecture. What is the speedup you achieve?

4 Reporting

You report on this exercise with up to a page in the third report. There, you provide answers to the questions in the "Working on the exercises" section.

5 License