Теория Параллелизма

Отчет

Простая нейронная сеть

Выполнил 21932, Бабенко Егор Степанович

17.05.2022

Цель работы

Реализовать используя cuDNN или cuBLAS простую нейронную сеть и сравнить результаты с версией написаной на python с использованием pytorch

Используемый компилятор: *nvcc*

Для компиляции использовался скрипт, предоставленный в файле compile.sh:

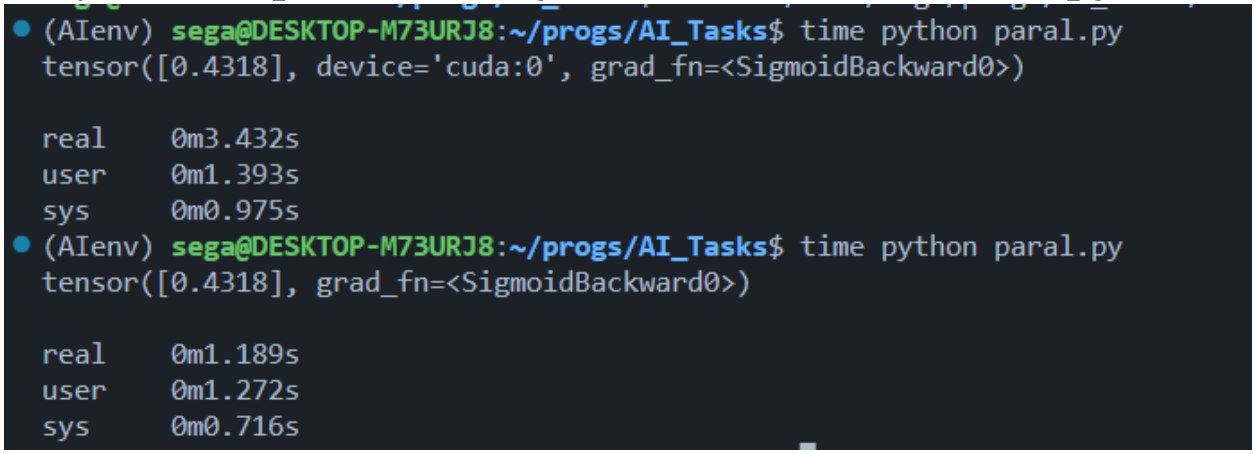
Для дополнительной профилировки: -D NVPROF_

Используемый профилировщики: *nvprof, nsys*

Nsys использовался с CUDA trace, а также NVTX trace.

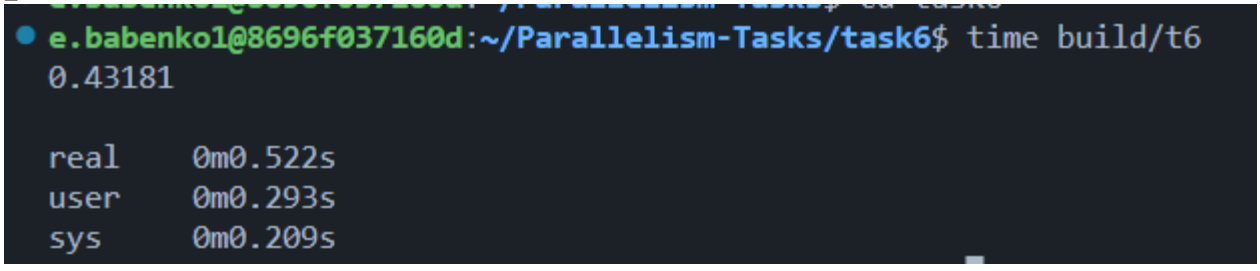
Используя библиотеку cuBLAS, а также cuda runtime, я реализовал свои классы для линейного слоя, а также слоя активации сигмоиды. Дополнительно реализовал свой контейнер для данных, чтобы можно было протягивать данные сквозь слои. Линейный слой реализован с помощью cuBLAS, а симгмоида через вызов __global__ функции. Также был реализован общий класс для создания архитектуры нейросети из слоев.

Результаты работы на Python с использованием pytorch



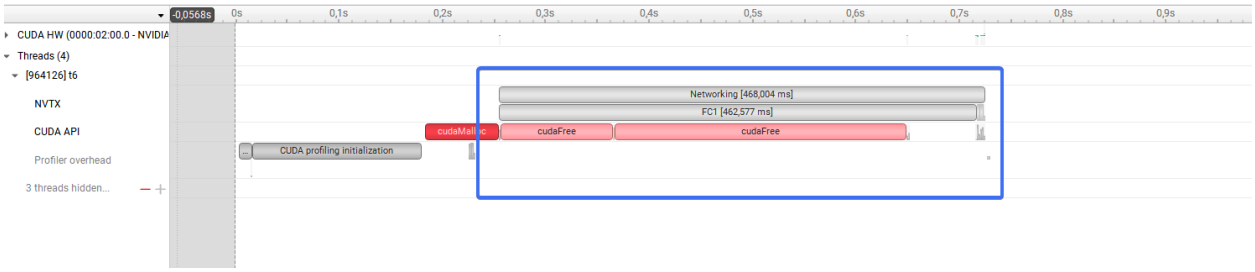
Первые запуск проводился с расчетами на видеокарте, второй с расчетами на CPU. Как можно увидеть что накладные расходы на передачу на видеокарту для такой маленькой нейронной сети.

Результаты работы на cuda C++ с собственной реализацией



Как можем видеть результат расчетов не отличается от версии на pytorch, но при этом затрачиваемое время в 7 раз быстрее.

Ниже на скриншотах можно видеть результаты из профилировщика nsys.



Вывод:

Используя cublas, cuda runtime можно реализовать свою простенькую сеть с только прямым проходом. Также за счет использования ООП и планирования архитектуры программы можно реализовать и backward. Но это занимает порядочное время для разработки и куда целесообразнее использовать framework torch

Github: https://github.com/JooudDoo/Parallelism-Tasks