МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ

(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)

Институт №8 «Информационные технологии и прикладная математика»

Кафедра 806 «Вычислительная математика и программирование»

**Лабораторная работа №3**

**по курсу «Программирование графических процессоров»**

**Классификация и кластеризация изображений на GPU.**

Выполнил: М.А. Бронников

Группа: 8О-407Б

Преподаватели: К.Г. Крашенинников,

А.Ю. Морозов

Москва, 2020

**Условие**

**Цель работы:** Научиться использовать GPU для классификации и кластеризации изображений. Использование константной памяти.

**Вариант 4.** Метод максимального правдоподобия.

**Программное и аппаратное обеспечение**

Device: GeForce GT 545

Размер глобальной памяти: 3150381056

Размер константной памяти : 65536

Размер разделяемой памяти: 49152

Регистров на блок: 32768

Максимум потоков на блок: 1024

Количество мультипроцессоров : 3

OS: Linux Mint 20 Cinnamon

Редактор: VSCode

**Метод решения**

Для начала необходимо посчитать значения средних по каждому из каналов и значения обратной матрицы ковариации для каждого из заданных признаков. Далее для каждого из пикселей изображения определяется его класс по формуле:



**Описание программы**

Для выполнения программы я реализовал собственный класс изображения в методе которого и вызывался kernel. Этот класс не потерпел значительных изменений со времени выполнения второй лабораторной работы.

Для выполнения операции я инициализировал на этапе компиляции массив константной памяти необходимого размера, а именно максимального количества возможных классов, умноженного на размер, необходимый для хранения вычислительной информации для каждого из классов.

struct class\_data{

float avg\_red;

float avg\_green;

float avg\_blue;

float cov11;

float cov12;

float cov13;

float cov21;

float cov22;

float cov23;

float cov31;

float cov32;

float cov33;

float log\_det;

};

// constant memory

\_\_constant\_\_ class\_data computation\_data[MAX\_CLASS\_NUMBERS];

Для копирования данных с host в этот участок памяти у меня есть следующий код:

throw\_on\_cuda\_error(

cudaMemcpyToSymbol(computation\_data, cov\_avg,

MAX\_CLASS\_NUMBERS\*sizeof(class\_data), 0, cudaMemcpyHostToDevice)

);

После чего Сами же данные для вычисления я расчитываю на CPU, поскольку расчет на GPU не даст значимого прироста к производительности.

После чего я вызываю kernel с заданным количеством блоков и потоков, где я преобразую изображение в кластеризованное по описанному алгоритму:

dim3 threads = dim3(MAX\_X, MAX\_Y);

dim3 blocks = dim3(BLOCKS\_X, BLOCKS\_Y);

classification<<<blocks, threads>>>(d\_data, \_height, \_widht, indexes.size());

throw\_on\_cuda\_error(cudaGetLastError());

В самом kernel мы вычисляем правдоподобие для пикселя по каждому из классов и записываем наиболее вероятный номер класса.

\_\_global\_\_ void sobel(uint32\_t\* d\_data, uint32\_t h, uint32\_t w){

int32\_t idx = blockIdx.x \* blockDim.x + threadIdx.x;

int32\_t idy = blockIdx.y \* blockDim.y + threadIdx.y;

uint32\_t step\_x = blockDim.x \* gridDim.x;

uint32\_t step\_y = blockDim.y \* gridDim.y;

for(int32\_t i = idx; i < w; i += step\_x){

for(int32\_t j = idy; j < h; j += step\_y){

// ans pixel

uint32\_t ans = 0;

// locate area in mem(32 bite)

// compute grey scale for all pixels in area

float w11 = GREY(tex2D(g\_text, i - 1, j - 1));

float w12 = GREY(tex2D(g\_text, i, j - 1));

float w13 = GREY(tex2D(g\_text, i + 1, j - 1));

float w21 = GREY(tex2D(g\_text, i - 1, j));

float w23 = GREY(tex2D(g\_text, i + 1, j));

float w31 = GREY(tex2D(g\_text, i - 1, j + 1));

float w32 = GREY(tex2D(g\_text, i, j + 1));

float w33 = GREY(tex2D(g\_text, i + 1, j + 1));

// compute Gx Gy

float Gx = w13 + w23 + w23 + w33 - w11 - w21 - w21 - w31;

float Gy = w31 + w32 + w32 + w33 - w11 - w12 - w12 - w13;

// full gradient

int32\_t gradf = (int32\_t)sqrt(Gx\*Gx + Gy\*Gy);

// max(grad, 255)

gradf = gradf > 255 ? 255 : gradf;

// store values in variable for minimize work with global mem

ans ^= (gradf << 16);

ans ^= (gradf << 8);

ans ^= (gradf);

// locate in global mem

d\_data[j\*w + i] = ans;

}

}

}

После вызова kernel я копирую данные в массив и освобождаю выделенную память.

**Результаты**

Для иллюстрации результатов работы алгоритма я выбрал 3 изображения:





****

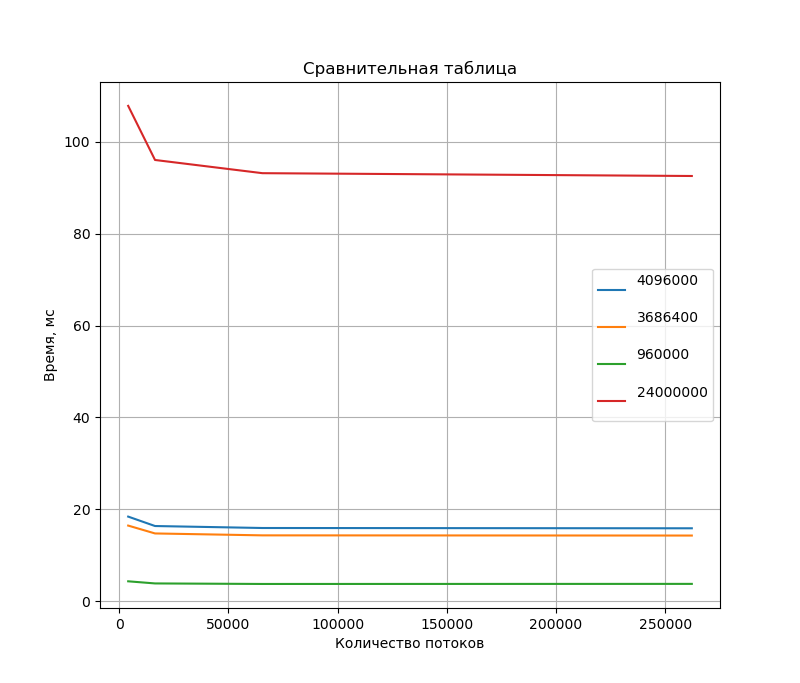
После чего я применил конвертер в заданный в задании формат, который я взял из материалов от преподавателя. После чего я применил к ним свою программу, и конвертировал обратно в jpg. Полученные результаты:



****

****

Я посмотрел как зависит время работы на этих изображениях от количества запущенных потоков:



Отсюда видно, что начиная с некоторого количества потоков, производительность работы на GPU не возрастает. При этом этот порог для разных размеров изображений разный, что довольно объяснимо, поскольку при большем размере данных требуется большее количество нитей для оптимального распараллеливания алгоритма.

Однако куда интереснее посмотреть на разницу, полученную при запуске алгоритма на этих изображениях на GPU и на CPU:

Img №1:

GPU threads 65536:

size: 4096000

time: 15.9292

CPU:

size: 4096000

time: 1643.48

Img №2:

GPU threads 65536:

size: 3686400

time: 14.3269

CPU

size: 3686400

time: 1483.8

Img №3:

GPU threads 65536:

size: 960000

time: 3.74608

CPU

size: 960000

time: 440.154

Img №4:

GPU threads: 65536

size: 24000000

time: 93.1731

CPU

size: 24000000

tiime: 9604.2

**Выводы**

Реализованный мной алгоритм широко применяется при обработке изображений, поскольку позволяет четко выделить контуры, что бывает полезно в задачах машинного обучения. При этом этот алгоритм является достаточно шумным(на изображении можно заметить светлые размывы), поэтому перед его применением рекомендуется применять сглаживающие фильтры.

Алгоритмы свертки хорошо распараллеливаются, что делает эффективным их использование на графических процессорах. Недаром, что современные свёрточные нейронные сети обучаются гораздо быстрее на GPU.

В ходе выполнения работы возникла трудность в том, как расположить данные в текстурной памяти так, чтобы влезть в ограничения, а также было совсем не очевидно, что следует решать задач в оттенках серого, а не в 3-ех каналах RGB изображения.