Zachodniopomorski Uniwersytet Technologiczny

w Szczecinie

**WYDZIAŁ INFORMATYKI**

WI_1

Juliusz Romanowski

Kierunek Informatyka

**Zastosowanie technologii CUDA w sztucznej inteligencji**

Praca dyplomowa magisterska

napisana pod kierunkiem

**dr inż. Wiesław Pietruszkiewicz**

w Katedrze Organizacji i Zarządzania

Szczecin, czerwiec 2010

**Oświadczenie**

*Oświadczam niniejszym, że przedkładaną pracę dyplomową kończącą studia podyplomowe napisałem samodzielnie. Oznacza to, że przy pisaniu pracy poza niezbędnymi konsultacjami, nie korzystałem z pomocy innych osób, a w szczególności nie zlecałem opracowania rozprawy innym osobom. Wszystkie dane, istotne myśli pochodzą z literatury i opatrzone są odpowiednim przypisem. Jednocześnie przyjmuję do wiadomości, że gdyby powyższe oświadczenie okazało się nieprawdziwe, decyzja o wydaniu mi świadectwa zostanie cofnięta.*

……………………

własnoręczny podpis

Spis treści

[1. Wstęp 7](#_Toc261192402)

[2. Równoległe przetwarzanie z zastosowaniem GPU 10](#_Toc261192403)

[2.1. Jednostki obliczeniowe GPU 10](#_Toc261192404)

[2.1.1. Różnice między CPU a GPU 10](#_Toc261192405)

[2.1.2. Struktura pamięci GPU 12](#_Toc261192406)

[2.2. Technologie GPGPU 14](#_Toc261192407)

[2.2.1. Historia GPGPU 14](#_Toc261192408)

[2.2.2. Użycie GPGPU w różnych zastosowaniach 16](#_Toc261192409)

[2.3. CUDA 17](#_Toc261192410)

[2.3.1. Architektura i podstawy CUDA 17](#_Toc261192411)

[2.3.2. Wprowadzenie do technologii 18](#_Toc261192412)

[2.3.3. Wersje CUDA 19](#_Toc261192413)

[2.3.3.1. Ograniczenia 20](#_Toc261192414)

[2.3.4. Proces kompilacji plików CUDA 21](#_Toc261192415)

[2.3.5. Tryb emulacji 22](#_Toc261192416)

[2.3.6. Model architektury CUDA 23](#_Toc261192417)

[2.3.7. Komunikacja między wątkami 23](#_Toc261192418)

[2.3.8. Proces wykonania programu CUDA 24](#_Toc261192419)

[2.3.9. Przykład kodu 24](#_Toc261192420)

[3. Sztuczne sieci neuronowe 25](#_Toc261192421)

[3.1. Model neuronu 25](#_Toc261192422)

[3.1.1. Neuron biologiczny 25](#_Toc261192423)

[3.1.2. Sztuczny neuron 26](#_Toc261192424)

[3.1.3. Funkcja aktywacji, neuron nieliniowy 27](#_Toc261192425)

[3.2. Architektury sieci neuronowych 28](#_Toc261192426)

[3.2.1. Jednokierunkowe 28](#_Toc261192427)

[3.2.2. Rekurencyjne 30](#_Toc261192428)

[3.2.3. Sieci komórkowe 30](#_Toc261192429)

[3.3. Metody uczenia 30](#_Toc261192430)

[3.3.1. Uczenie nadzorowane 30](#_Toc261192431)

[3.3.1.1. Z nauczycielem 30](#_Toc261192432)

[3.3.1.2. Z krytykiem 30](#_Toc261192433)

[3.3.2. Uczenie nienadzorowane 30](#_Toc261192434)

[3.3.2.1. Samoorganizujące się mapy 30](#_Toc261192435)

[3.3.2.2. Konkurencyjne 30](#_Toc261192436)

[3.4. Zastosowanie sztucznych sieci neuronowych 31](#_Toc261192437)

[4. Implementacja sztucznych sieci neuronowych na GPU 32](#_Toc261192438)

[4.1. Sieć neuronowa z punktu widzenia programisty 32](#_Toc261192439)

[4.1.1. Struktury danych w sieciach neuronowych 32](#_Toc261192440)

[4.1.2. Użyte algorytmy 32](#_Toc261192441)

[4.2. Zastosowanie GPU w sztucznych sieciach neuronowych 32](#_Toc261192442)

[4.2.1. Zrównoleglenie operacji na sieciach neuronowych 32](#_Toc261192443)

[4.2.2. Wewnętrzne struktury danych GPU 32](#_Toc261192444)

[4.2.3. Istniejące rozwiązania software’owe 32](#_Toc261192445)

[5. Biblioteka CNL 33](#_Toc261192446)

[5.1. Ogólny projekt aplikacji 33](#_Toc261192447)

[5.1.1. Zastosowania programu 35](#_Toc261192448)

[5.1.2. Diagramy: 35](#_Toc261192449)

[5.1.2.1. Diagram klas 36](#_Toc261192450)

[5.1.2.2. Diagram komponentów 37](#_Toc261192451)

[5.1.2.3. Diagram sekwencji uczenia sieci 38](#_Toc261192452)

[5.1.2.4. Activity 39](#_Toc261192453)

[5.1.3. Struktura plików danych 39](#_Toc261192454)

[5.1.3.1. Plik zestawu testów CSV 41](#_Toc261192455)

[5.1.3.2. Plik zestawu testów XML 41](#_Toc261192456)

[5.1.3.3. Format Sieci MLP w XML 45](#_Toc261192457)

[5.1.4. Katalogi i pliki w projekcie 46](#_Toc261192458)

[5.1.5. Logowanie 47](#_Toc261192459)

[5.2. Część CPU 48](#_Toc261192460)

[5.2.1. Kolejność operacji przy operacjach na MLP 49](#_Toc261192461)

[5.2.1.1. Diagram sekwencji uruchamiania sieci 49](#_Toc261192462)

[5.2.1.2. Diagram sekwencji trenowania sieci 49](#_Toc261192463)

[5.3. Część GPU 50](#_Toc261192464)

[5.3.1. Implementacja MLP na GPU 51](#_Toc261192465)

[5.3.1.1. Ogólny algorytm 51](#_Toc261192466)

[5.3.1.2. Diagram sekwencji uruchamiania sieci 51](#_Toc261192467)

[5.3.1.3. Diagram sekwencji trenowania sieci 52](#_Toc261192468)

[5.3.1.4. Struktury danych w pamięci GPU 54](#_Toc261192469)

[5.3.1.5. Opis działania kerneli 55](#_Toc261192470)

[5.3.1.6. Analiza zależności danych 56](#_Toc261192471)

[5.3.2. Ograniczenia wykonanego algorytmu. 56](#_Toc261192472)

[5.3.3. Użyte optymalizacje kerneli 57](#_Toc261192473)

[5.3.3.1. Dostępy do danych globalnych o dużej przepustowości 58](#_Toc261192474)

[5.3.3.2. Obsługa 2 testów na raz 58](#_Toc261192475)

[5.3.3.3. Zapisywanie danych do pamięci dzielonej 58](#_Toc261192476)

[5.3.3.4. Przewidywanie ilości uruchomionych bloków 59](#_Toc261192477)

[5.3.3.5. Użycie pamięci constant 60](#_Toc261192478)

[5.4. Testy implementacji sieci MLP 60](#_Toc261192479)

[5.4.1. Opis danych testowych 61](#_Toc261192480)

[5.4.2. Środowisko testowe 61](#_Toc261192481)

[5.4.3. Wpływ parametrów sieci na jakość uczenia 61](#_Toc261192482)

[5.4.4. Wpływ optymalizacji na wydajność kerneli 63](#_Toc261192483)

[5.4.5. Porównanie wydajności uczenia wersji CPU i GPU 64](#_Toc261192484)

[5.4.6. Informacje końcowe 66](#_Toc261192485)

[5.5. Zewnętrzne biblioteki 66](#_Toc261192486)

[5.5.1. Mersenne Twister 66](#_Toc261192487)

[5.5.2. TinyXML 66](#_Toc261192488)

[5.6. Narzędzia pomocnicze 66](#_Toc261192489)

[5.7. Możliwości rozwoju programu 67](#_Toc261192490)

[6. Wnioski 68](#_Toc261192491)

[7. Zakończenie 69](#_Toc261192492)

[8. Bibliografia 70](#_Toc261192493)

[9. Załączniki 73](#_Toc261192494)

[9.1. Kod CPU i GPU? 73](#_Toc261192495)

Spis ilustracji

[Rysunek 1. Model fizyczny CPU i GPU 10](#_Toc261274306)

[Rysunek 2. Hierarchia procesorów oraz typów pamięci w GPU 11](#_Toc261274307)

[Rysunek 3. Złączone dostępy do pamięci na urządzeniach CC 1.0 i 1.1 13](#_Toc261274308)

[Rysunek 4. Architektura CUDA 17](#_Toc261274309)

[Rysunek 5. Podział pracy na bloki i watki 18](#_Toc261274310)

[Rysunek 6. Proces kompilacji plików źródłowych CUDA 21](#_Toc261274311)

[Rysunek 7. Schemat neuronu biologicznego 25](#_Toc261274312)

[Rysunek 8. Schemat sztucznego neuronu 25](#_Toc261274313)

[Rysunek 9. Wykres funkcji y=tanh(ßx) 27](#_Toc261274314)

[Rysunek 10. Diagram klas biblioteki CNL 35](#_Toc261274315)

[Rysunek 11. Diagram komponentów biblioteki CNL 36](#_Toc261274316)

[Rysunek 12. Diagram sekwencji uczenia sieci w bibliotece CNL 37](#_Toc261274317)

[Rysunek 13. Diagram sekwencji uruchamiania sieci na CPU 48](#_Toc261274318)

[Rysunek 14. Diagram sekwencji trenowania sieci na CPU 49](#_Toc261274319)

[Rysunek 15. Diagram sekwencji uruchamiania sieci na GPU 51](#_Toc261274320)

[Rysunek 16. Diagram sekwencji trenowania sieci na GPU 53](#_Toc261274321)

[Rysunek 17. Program CUDA Visual Profiler 66](#_Toc261274322)

Spis tabel

[Porównanie technologii GPGPU 14](#_Toc261275745)

[Formaty plików danych obsługiwane przez bibliotekę CNL 38](#_Toc261275746)

[Opis użytych kerneli 54](#_Toc261275747)

[Porównanie jakości uczenia sieci w zależności od różnych parametrów 61](#_Toc261275748)

[Porównanie czasu uczenia sieci na CPU i GPU 63](#_Toc261275749)

[Porównanie czasu uczenia sieci na platformach 32 i 64-bitowych 64](#_Toc261275750)

# Wstęp

Przez ostatnie kilkadziesiąt lat, wydajność komputerów wzrastała bardzo szybko w sposób wykładniczy[[1]](#footnote-1). Sam rozwój sprzętu jest powiązany z rozwojem oprogramowania, które też wymaga coraz więcej zasobów komputera. Aby sprostać tym wymaganiom, firmy produkujące procesory zaczęły coraz bardziej zwiększać ilość rdzeni w CPU, co postawiło wyzwania przed twórcami oprogramowania – programy zaczynają używać więcej niż jednego procesora.

Jednak zaledwie 4 lata temu[[2]](#footnote-2) pojawiła się możliwość wykonywania obliczeń w zupełnie inny sposób. Zamiast używania (tak jak do tej pory) CPU oraz pamięci RAM, pojawiły się biblioteki umożliwiające przetwarzanie danych na kartach graficznych (GPU), opisane w rozdziale X.X. Najnowsze karty graficzne, stworzone do przetwarzania danych graficznych z bardzo dużą prędkością, mogą być aktualnie użyte do wykonywania obliczeń każdego rodzaju. Budowa GPU jest zdecydowanie inna od budowy CPU, przez co sposoby wykonywania na nich obliczeń bardzo się różnią. na współczesnych GPU jest bardzo dużo jednostek obliczających, a minimalna ilość wątków potrzebna do osiągnięcia wysokiej wydajności jest często liczona w tysiącach. w związku z tym, pełne wykorzystanie mocy współczesnych GPU[[3]](#footnote-3) jest możliwe tylko w przypadku zadań, które można w dużym stopniu zrównoleglić.

Do tej pory znaleziono już wiele praktycznych zastosowań GPU w programach ogólnego użytku, jest też wiele prac badawczych porównujących prędkość programów wykonywanych na CPU[[4]](#footnote-4) z programami napisanymi na GPU. Jednym z najczęściej wykonywanych porównań są operacje na sieciach neuronowych. Uruchamianie i trenowanie sieci neuronowych są operacjami, gdzie wykonywanych jest wiele podobnych działań, przez co są idealne jako możliwy program do wykonania na GPU.

W tej pracy, chciałem opisać bibliotekę CNL, której zadaniem jest wykonywanie operacji na sieciach neuronowych (na CPU oraz GPU), przy użyciu biblioteki NVIDIA CUDA. w bibliotece CNL została zaimplementowana obsługa sieci MLP[[5]](#footnote-5); sama biblioteka została zaprojektowana tak, aby była łatwo rozszerzalna o obsługę innych typów sieci. w czasie jej tworzenia, duży nacisk został też położony na uzyskanie dużej wydajności operacji na GPU.

Powyższy [rozdział 1](#_Wstęp) zawiera bardzo ogólne wprowadzenie do problematyki GPGPU[[6]](#footnote-6), sieci neuronowych oraz opisu stworzonej biblioteki. Opisuje również cele dotyczące niniejszej pracy.

[Rozdział 2](#_Równoległe_przetwarzanie_z) przedstawia informacje o kartach graficznych, ich budowę i technologie służące do użycia ich w celu wykonywania obliczeń ogólnego przeznaczenia. Są tam też wymienione programy wykorzystujące karty graficzne do przyspieszania różnych obliczeń.

Informacje o sieciach neuronowych, ich budowie i rodzajach, są opisane w [rozdziale 3](#_Sztuczne_sieci_neuronowe).

[Rozdział 4](#_Implementacja_sztucznych_sieci) jest swoistym powiązaniem między rozdziałami 2 i 3. Pokazuje, że z matematycznego punktu widzenia wiele operacji na sieciach neuronowych może być przedstawionych jako działania macierzowe. Jest tam opisane, w jaki sposób byłoby możliwe wykorzystanie możliwości wielopotokowego procesora GPU przy operacjach na sieciach neuronowych.

Najwięcej miejsca zostało poświęcone [rozdziałowi 5](#_Biblioteka_CNL), gdzie została dokładnie opisana biblioteka CNL. Zawarte są tam informacje o sposobie implementacji całej biblioteki – części uruchamianej na CPU, jak i na GPU. Opisane zostały sposoby optymalizacji operacji GPU i testy badające prędkość oraz jakość działania biblioteki.

W [rozdziale 6](#_Wnioski/spostrzeżenia) zawarte są wnioski i spostrzeżenia wyciągnięte po napisaniu i przetestowaniu całej biblioteki oraz możliwe perspektywy rozwoju i rozbudowy biblioteki, jak i całego rozwoju technologii GPGPU.

[Rozdział 7](#_Zakończenie) zawiera końcowe myśli na temat niniejszej pracy.

# Równoległe przetwarzanie z zastosowaniem GPU

W rozdziale tym chciałem przedstawić kolejne udoskonalenia w dziedzinie kart graficznych, szczególnie pod kątem możliwości uruchamiania na nich programów ogólnego zastosowania. Opisałem również budowę nowoczesnych kart graficznych oraz technologię CUDA.

## Jednostki obliczeniowe GPU

Pierwsze karty graficzne do komputerów PC powstały ok. 30 lat temu i pozwalały tylko na wyświetlanie jednokolorowego tekstu. Niedługo po nich pojawiły się udoskonalone wersje, które udostępniały możliwość wyświetlania dwuwymiarowego obrazu graficznego, a kolejne wersje pozwalały na coraz wyższą rozdzielczość i ilość kolorów. w 1995 roku pojawiły się pierwsze karty graficzne przetwarzające również tryb 3D[[7]](#footnote-7). w kolejnych latach, napędzane wymaganiami rynku, firmy produkujące chipsety graficzne tworzyły coraz szybsze karty GPU, a wzrost wydajności kart graficznych był zdecydowanie szybszy niż wzrost wydajności CPU. Obecnie najszybsze procesory CPU mają wydajność ok. 100 GigaFlops[[8]](#footnote-8), a najszybsze GPU osiągają ok. 1000 GigaFlops, mają dziesiątki rdzeni i zapewniają dużą możliwość zrównoleglania operacji.

### Różnice między CPU a GPU

Poniżej zamieszczona jest ilustracja przedstawiająca porównanie budowy obu rodzajów procesorów.

|  |
| --- |
|  |
| Rysunek 1. Model fizyczny CPU i GPU  Źródło: (NVIDIA, 2010, str. 3) |

Budowa procesorów graficznych różni się znacząco od budowy CPU. Jest to związane z zadaniami, do których są wykorzystywane. GPU zawiera wiele ALU[[9]](#footnote-9) oraz FPU[[10]](#footnote-10), ponieważ jest dostosowany do rozwiązywania zadań, które mogą być łatwo zrównoleglone – podobne operacje są wykonywane dużej ilości danych. Właśnie tego typu operacje są wykonywane przy renderowaniu obrazu – te same instrukcje są używane do przekształceń na wielu tysiącach lub milionach pikseli, tekseli lub wierzchołków. z drugiej strony, przy przekształceniach graficznych nie jest wymagana zaawansowana kontrola przepływu sterowania[[11]](#footnote-11), a ilość pamięci cache jest ograniczona do minimum. Ilość wątków uruchomionych w danym momencie na GPU jest bardzo duża, jednak narzut przy ich tworzeniu i kasowaniu jest bardzo mały, w związku z czym programy wykonywane przez nie często są krótkie.

### Struktura pamięci GPU

|  |
| --- |
| http://www.ixbt.com/video3/images/cuda/cuda5.png |
| Rysunek 2. Hierarchia procesorów oraz typów pamięci w GPU  Źródło: (Berillo, 2008, str. 5) |

Na powyższej ilustracji pokazany jest model fizycznej struktury oraz pamięci nowoczesnej karty graficznej firmy NVIDIA. Zawiera ona od kilku do kilkudziesięciu multiprocesorów[[12]](#footnote-12), pamięć globalną oraz niewielką pamięć constant. w każdym multiprocesorze znajduje się 8 procesorów skalarnych[[13]](#footnote-13), pamięć dzieloną, cache tekstur oraz cache pamięci constant. Pojedynczy procesor skalarny zawiera pewną ilość rejestrów, a jego zadaniem jest wykonywanie operacji arytmetycznych. Każdy z rodzajów pamięci ma specyficzne przeznaczenie:

* Pamięć globalna – ma ona wielkość od kilkuset MB do 4 GB, jest dostępna dla wszystkich wątków. Jej przepustowość wynosi do 100GB/s, ale opóźnienie przy dostępie wynosi kilkaset cykli zegara, więc zaleca się minimalizowanie używania tej pamięci w programach.
* Pamięć constant – niewielka (64 KB) pamięć zapisywana przed uruchomieniem wszystkich operacji na GPU, zapis do niej nie jest później możliwy. Każdy multiprocesor zawiera 8KB cache pamięci constant. Ten rodzaj pamięci jest używany, jeśli są pewne często używane dane, które nie są zmieniane w trakcie działania operacji na GPU.
* Pamięć lokalna – w tej pamięci znajdują się parametry wywołania funkcji GPU, lub zmienne lokalne (jeśli brakuje rejestrów).
* Pamięć dzielona – znajduje się na każdym multiprocesorze, ma wielkość 16-48 KB. Może być używana jako cache pamięci globalnej, lub jako sposób komunikacji między wątkami w danym multiprocesorze.
* Pamięć teksturowa – jest to pamięć cache (w każdym multiprocesorze), specjalny rodzaj pamięci mapowany na pamięć globalną.
* Rejestry – bardzo szybki rodzaj pamięci, są przypisane do każdego uruchomionego wątku. w jednym multiprocesorze znajduje się 8-32 tys. 4-bajtowych rejestrów i są one rozdzielane na uruchomione wątki.

Jak widać, na kartach graficznych jest możliwość wyboru między kilkoma różnymi rodzajami pamięci. Jedną z najważniejszych zasad przy pisaniu programu na GPU jest ograniczenie używania wolnych typów pamięci do minimum, a zamiast nich użycie szybszych (pamięć dzielona, cache pamięci teksturowej, cache pamięci constant).

Oprócz tego, dla uzyskania wysokiej wydajności przy dostępie do pamięci globalnej i dzielonej, odczyty lub zapisy muszą być wykonywane w odpowiedni sposób. na przykład, w technologii CUDA, aby uzyskać dużą przepustowość pamięci globalnej, dostęp do niej musi być przeprowadzony w odpowiedni sposób. Wymagane jest, aby każdy wątek w half-warpie (1 warp zawiera 32 wątki, half-warp zawiera 16 wątków) dostępował odpowiedniej pozycji tabeli. na GPU z compute capability[[14]](#footnote-14) 1.1, n-ty wątek powinien dostępować pozycji tabeli o indeksie n mod 16. Pozwala to na zapisanie lub odczytanie wszystkich 16 zmiennych za jednym razem (w przeciwnym wypadku te operacje mogą być wykonane sekwencyjnie). Odczyt lub zapis, w którym za jednym razem przetwarzane są zmienne z ponad jednego wątku, nazywany jest przez w dokumentach jako coalesced (złączony). Poniżej jest przykład ilustrujący „złączony” dostęp do pamięci.

|  |
| --- |
|  |
| Rysunek 3. Złączone dostępy do pamięci na urządzeniach CC 1.0 i 1.1  Źródło: (NVIDIA, 2009, str. 84) |

## Technologie GPGPU

**GPGPU (**General-Purpose computation on Graphics Processing Units) oznacza użycie GPU do obliczeń do tej pory uruchamianych na CPU(GPGPU.org). Technika ta jest możliwa do zastosowania dzięki kolejnym osiągnięciom w dziedzinie kart graficznych, opisanym poniżej.

### Historia GPGPU

Wzrost wydajności kart graficznych w ostatnich latach był powiązany z obsługą coraz większej ilości etapów renderowania przez kartę graficzną. do czasu stworzenia karty graficznej NVIDIA GeForce 256, etapy transformacji geometrii i obliczania oświetlenia były wykonywane na CPU. Karta ta jako pierwsza umożliwiała obsługę tych operacji[[15]](#footnote-15).

Kolejne karty umożliwiały już programować etapy T&L – specjalne funkcje (tzw. shadery) pozwalały na dodanie np. odbić lustrzanych. Te programowalne etapy mogły być już używane do przetwarzania danych nie-graficznych, jednak nie było to łatwe z kilku względów:

* Języki programowania shaderów (Microsoft HLSL, OpenGL Shading Language) różnią się od popularnych języków programowania
* Wiele kart nie wspierało obsługi liczb zmiennoprzecinkowych nawet pojedynczej precyzji (32 bity)
* Dopiero późniejsze karty graficzne pozwalały na użycie instrukcji warunkowych i pętli

Wkrótce potem NVIDIA zaprezentowała framework CUDA pozwalający na wykonywanie dowolnych obliczeń na najnowszych kartach graficznych tej firmy. Nie był on obarczony ograniczeniami shaderów, które opisałem powyżej. Firma ATI stworzyła podobny produkt Stream SDK[[16]](#footnote-16) działający na kartach ATI, powstała też biblioteka OpenCL, która wspiera karty graficzne obu tych producentów oraz procesory innych typów. Poniżej umieściłem porównanie tych technologii.

|  |  |  |  |
| --- | --- | --- | --- |
| Tabela . Porównanie technologii GPGPU | | | |
|  | NVIDIA CUDA | OpenCL | ATI Stream |
| Obsługiwane procesory | Karty graficzne NVIDIA, NVIDIA Tesla | Procesory CPU, karty graficzne NVIDIA, ATI, NVIDIA Tesla, , procesory DSP i inne (Wilk, 2009, str. 4) | Karty graficzne ATI |
| Popularność, wsparcie produktu | Duża. Popularne forum NVIDIA (NVIDIA) | Średnia | Niewielka |
| Częstość uaktualniania biblioteki | Dosyć duża | Średnia | Średnia |
| Składnia | Podobna do C[[17]](#footnote-17) | Podobna do C | Podobna do C |
| Styl programowania, składnia programów | Wysokopoziomowy lub niskopoziomowy[[18]](#footnote-18) | Niskopoziomowy[[19]](#footnote-19) | Niskopoziomowy |
| Źródła: (NVIDIA), , (AMD) | | | |

### Użycie GPGPU w różnych zastosowaniach

Od czasu powstania frameworków GPGPU, zauważa się coraz większą ilość komercyjnych programów pozwalających na użycie mocy obliczeniowej GPU do wykonywanie niektórych długotrwałych zadań:

* Niektóre projekty w ramach BOINC (np. SETI@home) pozwalają na użycie technologii CUDA lub Stream i dzięki temu kilkukrotnie zwiększają prędkość swojego działania.
* Jest również możliwość tworzenia filtrów w programie Adobe Photoshop przy użyciu CUDA.
* Powstał plugin do MATLABa, który wykonuje transformację Fouriera na GPU – osiągając kilkunastokrotne przyspieszenie
* PhysX – zestaw narzędzi umożliwiający wykonywanie obliczeć fizycznych na GPU, zaimplementowany w technologii CUDA

Na stronie internetowej firmy NVIDIA znajduje się pełniejsza lista aplikacji wykorzystujących CUDA. Powstało też wiele prac naukowych, w których opisano nowo stworzone programy wykonywane na GPU, np.:

* Rozpoznawanie i klasyfikacja obiektów graficznych
* Zintegrowanie NVIDIA CUDA do programów Einstein@Home i OpenSteer
* Implementacja algorytmu segmentacji livewire

W wielu z tych prac osiągnięte jest wielokrotne przyspieszenie działania w porównaniu z wersjami działającymi tylko na CPU.

## CUDA

Do stworzenia biblioteki obsługującej sieci neuronowe na GPU mogłem użyć jednego z trzech podanych wyżej frameworków (NVIDIA CUDA, OpenCL, ATI Stream). Najważniejszymi aspektami przy wyborze jednego z nich była łatwość tworzenia kodu, przenośność międzyplatformowa oraz wsparcie ze strony twórców i osób korzystających z danej biblioteki. w związku z tym wybrałem technologię NVIDIA, która najlepiej spełnia moje kryteria.

### Architektura i podstawy CUDA

Aby wykorzystać CUDA, należy posiadać kartę graficzną zgodną z tą technologią, Potrzebny jest również specjalny sterownik graficzny oraz pakiet CUDA Runtime. Istnieją też oficjalne biblioteki na tę platformę – CUBLAS (do wykonywania operacji macierzowych) i CUFFT (do wykonywania transformacji Fouriera). Poniżej jest podany schemat elementów składowych:

|  |
| --- |
| http://www.behardware.com/medias/photos_news/00/19/IMG0019322.gif |
| Rysunek 4. Architektura CUDA  Źródło: (Triolet, 2007, str. 4) |

### Wprowadzenie do technologii

Technologia CUDA ma skalowalny model programowania równoległego – program można uruchomić na każdej ilości mikroprocesorów. Jedna funkcja wykonywana na GPU – kernel – opisuje działania wykonywane w jednym wątku. Przy wywołaniu kernela wymagane jest podanie ilości tzw. bloków w gridzie, oraz ilości wątków w bloku. Liczby te określają, ile wątków zostanie uruchomionych równolegle. Zostało to zilustrowane poniżej:

|  |
| --- |
| http://www.behardware.com/medias/photos_news/00/19/IMG0019327.gif |
| Rysunek 5. Podział pracy na bloki i watki  Źródło: (Triolet, 2007, str. 4) |

Każdy blok w gridzie oraz każdy wątek w bloku ma indeks określony maksymalnie trzema wymiarami (na powyższej ilustracji dwoma). do każdego uruchomionego kernela można użyć innej ilości bloków lub wątków. Po rozpoczęciu kernela, GPU ma za zadanie przypisać każdy blok do wykonania na którymś z dostępnych mikroprocesorów strumieniowych. z tego względu, aby wykorzystać pełną moc GPU, ilość bloków powinna być przynajmniej równa ilości multiprocesorów w GPU.

### Wersje CUDA

CUDA może być wykorzystana na kartach GeForce, Quadro lub urządzeniach Tesla (specjalnych kartach GPU używanych jako procesory w superkomputerach). Każda z nich posiada określoną architekturę rdzenia (oznaczoną przez tzw. compute capability - CC). Numer wersji compute capability pozwala określić możliwości danej karty graficznej (takich jak ilość pamięci dzielonej lub to, czy dana karta wspiera liczby podwójnej precyzji). Obecne produkty mają CC 1.0, 1.1, 1.2, 1.3 oraz 2.0. Pierwsze karty wspierające compute capability 2.0 (GeForce GTX 470, GTX 480) zostały wypuszczone na rynek pod koniec marca 2010 roku mimo tego że ich premiera była planowana rok wcześniej. Były one bardzo oczekiwane przez środowisko developerów GPGPU, ponieważ produkty CC 2.0 (tzw. platforma Fermi) miały dawać dużo większe możliwości niż poprzednie produkty:

* Wykonywanie kilku różnych kerneli na raz
* Obsługa pamięci ECC (Error Correction Code) (NVIDIA, 2010)
* Możliwość debugowania wykonania kerneli
* Jednolita przestrzeń adresowa wszystkich rodzajów pamięci
* Wiele innych drobnych ulepszeń architektury

Architektura Fermi usuwa wiele ograniczeń w porównaniu ze starszymi wersjami CUDA i dzięki temu programowanie zrównoleglonych programów wykonywanych na procesorach graficznych może stać się niedużo trudniejsze niż zrównoleglenie programów wykonywanych na CPU.

#### Ograniczenia

Programy CUDA – jako że wykonywane są na GPU, układach które zostały stworzone głównie do wykonywania obliczeń graficznych – mają pewne ograniczenia oraz utrudnienia w porównaniu ze zwykłymi programami zrównoleglonymi na CPU:

* Mnożenie i dzielenie liczb całkowitych kilka razy wolniejsze od podobnych operacji na liczbach zmiennoprzecinkowych[[20]](#footnote-20)
* W GPU z compute capability 1.2 lub mniejszym, nie ma obsługi liczb zmiennoprzecinkowych podwójnej precyzji. w procesorach z CC 1.3 są one obsługiwane, ale operacje na nich są kilka razy wolniejsze niż na liczbach pojedynczej precyzji.
* Możliwość używania klas C++ w kernelach pojawiła się dopiero w urządzeniach CC 2.0.
* Uruchomienie kernela ma pewien narzut czasowy – przygotowanie kernela może zająć dłużej niż samo jego działanie.
* Jestem wiele zasad (zależnych od wersji compute capability GPU) dostępu do pamięci globalnej i dzielonej, które muszą być spełnione, by prędkość działania była optymalna. Zostały one opisane w (NVIDIA, 2010, strony 142-153).

### Proces kompilacji plików CUDA

Programy CUDA wykonywane na GPU są zapisywane w plikach o rozszerzeniu „cu”. Składają się z kerneli, funkcji wywołujących kernele i ewentualnie innych funkcji w języku C/C++. Poniżej znajduje się diagram blokowy kolejnych przekształceń plików „cu”:

|  |
| --- |
|  |
| Rysunek 6. Proces kompilacji plików źródłowych CUDA  Źródło: (Spek, 2008, str. 20) |

Pierwszym etapem przetwarzania jest rozdzielenie przez program cudafe pliku na część wykonywaną na hoście i na część GPU. Następnie część wykonywana na GPU jest kompilowana do postaci binarnej i/lub ptx (odpowiednik assemblera). Przy każdym wywołaniu kernela, odpowiednia postać binarna kernela jest ładowana do pamięci graficznej i uruchamiana na GPU.

### Tryb emulacji

Inną pomocą w czasie tworzenia programu była możliwość **emulacji** wykonania części GPU na CPU (jest włączana przez specjalny przełącznik kompilacji). Wykonanie całego programu w trym trybie jest możliwe nawet na komputerach niewyposażonych w karty graficzne wspierające CUDA. Tryb emulacji umożliwia pozwala na debugowanie kerneli (nie jest możliwe w standardowym trybie bez emulacji). Trzeba jednak pamiętać o tym, że jest to emulator, a nie symulator karty graficznej CUDA – występują pewne różnice – niektóre kernele działają poprawnie w trybie emulacji, a w trybie standardowym mogą nawet zawiesić komputer. Kod wykonany w emulatorze jest wielokrotnie wolniejszy niż na GPU.

### Model architektury CUDA

Kiedy mikroprocesor dostaje jeden lub więcej bloków do uruchomienia, dzieli on je na specjalne grupy, tzw. warpy (każdy warp zawiera 32 wątki). Wszystkie wątki w warpie wykonują w danym momencie mogą wykonywać najwyżej jedną instrukcję. w sytuacji, gdy niektóre wątki warpa wykonują pewną instrukcję warunkową a pozostałe nie, te pozostałe wątki mają zamrożone wykonanie do czasu zakończenia się pętli warunkowej.

Opisana architektura nazywana jest SIMT (Single Instruction, Multiple Threads). Jest ona podobna do architektury SIMD (Single Instruction, Multiple Data) w tym, że dane instrukcje mogą być wykonywane jednocześnie na wielu wątkach. Jednak różnią się one tym, że SIMT pozwala na użycie instrukcji warunkowych, dzięki czemu jest możliwe pisanie kodu dla pojedynczych wątków. Należy jednak pamiętać, że aby osiągnąć maksymalną wydajność programu, nie można używać zbyt wielu instrukcji warunkowych, które rozdzielają ścieżkę wykonania przez wątki w danym warpie.

### Komunikacja między wątkami

Technologia CUDA nie pozwala na bezpośrednią komunikację między wątkami, ale może to być osiągnięte w inny sposób. Często szybka pamięć dzielona jest używana do komunikacji między wątkami w danym bloku (jeden wątek ustawia wartość w tablicy, następnie inny odczytuje ją i interpretuje).

Oprócz tego CUDA udostępnia mechanizm synchronizacji wątków w bloku (jest ona wywoływana przez funkcję *\_\_syncthreads() ).*Gwarantuje ona, że dalsze instrukcje w programie będą wykonywane dopiero po dojściu do instrukcji synchronizacji (jest to często konieczne np. w przypadku ładowania danych globalnych do pamięci dzielonej). Nie ma możliwości komunikacji między wątkami w różnych blokach – dlatego trzeba tak projektować kernele, by taka komunikacja nie była konieczna.

### Proces wykonania programu CUDA

Większość operacji wykonywanych na GPU ma następującą sekwencję:

* Alokowanie wejściowej i wyjściowej pamięci na karcie graficznej
* Kopiowanie danych wejściowych do pamięci graficznej
* Wykonywanie właściwych operacji (kerneli) na GPU
* Kopiowanie danych wyjściowych z pamięci graficznej do RAM
* Dealokacja wejściowej i wyjściowej pamięci na karcie graficznej.

Można zauważyć, że jest więcej operacji niż w przypadku wykonywania obliczeń na CPU – jest to związane z kopiowaniem danych między RAM a pamięcią graficzną.

# Sztuczne sieci neuronowe

Mózg człowieka składa się z ok. 100 miliardów komórek nazywanych neuronami. Każdy neuron ma wiele wypustek łączących go z innymi neuronami, a jego działanie polega na generowaniu lub propagowaniu impulsów elektrycznych do innych neuronów lub mięśni. Sztuczne sieci neuronowe zostały stworzone jako matematyczny model próbujący symulować biologiczną sieć neuronową. Składają się ze sztucznych neuronów, które na wyjściu mają wartość zależną od wejść w tym neuronie. Sztuczna sieć neuronowa jest systemem adaptującym się, zmieniającym swoją strukturę w czasie fazy uczenia i dopasowującym do danych. Są zwykle używane do zadań, w których standardowy algorytm byłby nieefektywny lub zbyt skomplikowany, takich jak modelowanie skomplikowanych relacji między wejściem a wyjściem, lub wyszukiwanie schematów w danych.

## Model neuronu

### Neuron biologiczny

Biologiczne oraz sztuczne sieci neuronowe składają się z wielu neuronów, są one podstawą ich budowy. na poniższym rysunku znajduje się schemat pojedynczego biologicznego neuronu. w neuronie wyróżnia się somę – ciało komórki, oraz dendryty i akson. Dendryty wprowadzają informacje do neuronu, natomiast akson wyprowadza sygnały wyjściowe neuronu. Dendryty kończą się tzw. synapsami, w których sygnał wejściowy może być wzmacniany lub osłabiany. Neurony łączą się przez akson oraz dendryty z wieloma innymi neuronami, liczba połączeń może dochodzić do tysiąca.

|  |
| --- |
|  |
| Rysunek 7. Schemat neuronu biologicznego  Źródło: (Wikipedia, 2010) |

### Sztuczny neuron

Model sztucznego neuronu jest uproszczeniem neuronu biologicznego. Poniżej jest umieszczona ilustracja symbolizująca działanie sztucznego neuronu.

|  |
| --- |
|  |
| Rysunek 8. Schemat sztucznego neuronu  Źródło: (Klaus), zmodyfikowane |

Wejścia odpowiadają sygnałom nadchodzącym przez dendryty. Wagi to odpowiedniki modyfikacji dokonywanych na informacjach przez synapsy. Blok sumujący i aktywacji jest odpowiednikiem jądra, a wyjście opowiada aksonowi. Działanie sztucznego neuronu można opisać następująco: wejścia zawierają wartości, które następnie są mnożone przez odpowiadające współczynniki wag, blok sumujący następnie sumuje te pomnożone sygnały wejść oraz dodaje specjalną wagę – tzw. bias. Zsumowana wartość jest poddana działaniu funkcji aktywacyjnej w bloku aktywacji i podana na wyjście. Wyjście neuronu można więc przedstawić za pomocą następującego wzoru:

|  |  |
| --- | --- |
|  | (1) |

Zmienna oznacza k-tą wagę, oznacza k-tą wartość wejścia, a  to bias.

### Funkcja aktywacji

Funkcja aktywacji może przybierać różne postacie w zależności od modelu neuronu. Opiszę trzy z nich – neuron liniowy, perceptron Rosenblatta oraz neuron sigmoidalny.

Neuron liniowy nie modyfikuje wyjścia bloku sumującego, jest zwykle stosowany w neuronach warstwy wyjściowej sieci MLP. Wzór jego funkcji aktywacji przedstawia się następująco:

|  |  |
| --- | --- |
|  | (1) |

Perceptron Rosenblatta pozwala na klasyfikowanie danych liniowo separowanych. Jego binarna funkcja aktywacji może mieć jedną z dwóch postaci:

|  |  |
| --- | --- |
| lub | (1) |

Neuron sigmoidalny ma funkcję aktywacji przybierającą postać sigmoidy. Jest często używany w neuronach warstw ukrytych w sieciach MLP. Wzór funkcji aktywacji tego neuronu może przyjmować dwie postacie:

|  |  |
| --- | --- |
| (funkcja unipolarna)  lub (funkcja bipolarna) | (1) |

Parametr w powyższych wzorach określa skos funkcji (najczęściej używana jest wartość 1). Wykres funkcji bipolarnej (w zależności od wartości ) wygląda następująco:

|  |
| --- |
|  |
| Rysunek 9. Wykres funkcji y=tanh(ßx)  Źródło: Wykres wygenerowany przeze mnie w programie MATLAB |

## Architektury sieci neuronowych

Pojedynczy neuron mający n wejść, dzieli n-wymiarową przestrzeń na dwie półprzestrzenie oraz może być używany do klasyfikacji obiektów w przestrzeni n-wymiarowej. Może on jednak klasyfikować wektory tylko do jednej z dwóch klas, oraz radzi sobie tylko z problemami liniowo separowalnymi. Aby usunąć te ograniczenia i poszerzyć zakres odwzorowań, należy użyć bardziej skomplikowanych struktur.

### Jednokierunkowe

Sieci jednokierunkowe wielowarstwowe składają się z odpowiednio połączonych ze sobą neuronów rozmieszczonych w dwóch lub więcej warstwach. W tego typu sieciach, wyjście każdego neuronu w danej warstwie jest wejściem każdego z neuronów w warstwie kolejnej. Zwykle oprócz warstwy wejściowej lub wyjściowej występują tzw. warstwy ukryte, które pozwalają na to, by sieć miała możliwość rozwiązywania bardziej skomplikowanych problemów. Poniżej znajduje się schemat sieci jednokierunkowej z warstwą wejściową (neurony oznaczone czarnym kolorem), jedną warstwą ukrytą oraz warstwą wyjściową.



Sieci jednokierunkowe mogą mieć nieogranicznoną ilość warstw, jednak zwykle korzysta się z sieci z jedną lub dwoma warstwami ukrytymi. Poniżej znajduje się ogólny schemat sieci jednokierunkowej wielowarstwowej.



### Rekurencyjne

### Sieci komórkowe

## Metody uczenia

### Uczenie nadzorowane

#### Z nauczycielem

// (opis reguły delta i metody momentum)

#### Z krytykiem

### Uczenie nienadzorowane

#### Samoorganizujące się mapy

#### Konkurencyjne

## Zastosowanie sztucznych sieci neuronowych

// (Różne zadania - aproksymacja i klasyfikacja)

// w jakich konkretnie zadaniach są wykorzystywane

# Implementacja sztucznych sieci neuronowych na GPU

Jeżeli punkt 4 nie będzie zbyt duży, to może być dołączony jako ostatni podpunkt do punktu 3

## Sieć neuronowa z punktu widzenia programisty

### Struktury danych w sieciach neuronowych

// (Opis tego, jak dane są przedstawione w pamięci - wejścia, wagi, błąd, pochodna wyjścia, wyjście sieci).

### Użyte algorytmy

// (Algorytmy execute i train. Napisanie, że są to operacje macierzowe).

## Zastosowanie GPU w sztucznych sieciach neuronowych

### Zrównoleglenie operacji na sieciach neuronowych

// (Opisanie, że wiele operacji można wykonywać równolegle od siebie. do tego można zastosować GPU).

### Wewnętrzne struktury danych GPU

//(przedstawienie, jak dane sieci neuronowych i testów mogą być przedstawione wewnątrz pamięci GPU).

### Istniejące rozwiązania software’owe

# Biblioteka CNL

Głównym celem tej pracy magisterskiej jest stworzenie biblioteki pozwalającej na wykonywanie operacji na sieci neuronowe z wykorzystaniem CPU oraz GPU. Stworzona biblioteka CNL (CUDA Network Library) obsługuje sieci typu MLP, jednak została zaprojektowana tak, żeby dodawanie innych typów sieci nie było trudne.

Biblioteka została napisana w języku C++, może być kompilowana w środowisku Visual Studio 2008 (wersja 32bit i 64bit) i uruchamiana w systemach operacyjnych Windows XP i wyższych. w związku z brakiem czasu, nie została dodana możliwość kompilacji w kompilatorze G++.

## Ogólny projekt aplikacji

Implementacja biblioteki CNL znajduje się w plikach C++ (.cpp, .h) oraz z pliku TrainNetwork.cu (zawiera on implementacje funkcji wykonywanych na GPU). Plik CUDA.cpp nie należy do biblioteki, ale jest wykorzystywany do uruchamiania na niej różnego rodzaju testów. Aktualna implementacja biblioteki pozwala na:

* Stworzenie obiektu sieci MLP
* Załadowanie sieci MLP z pliku XML
* Zapisanie sieci MLP do pliku XML
* Skonfigurowanie ilości warstw ukrytych, ilości neuronów w tych warstwach i funkcji aktywacji neuronów w każdej warstwie
* Stworzenie obiektu zestawu testów
* Załadowanie zestawu testów z pliku CSV
* Załadowanie zestawu testów z pliku XML
* Zapisanie zestawu testów do pliku XML
* Wygenerowanie zestawu testów na podstawie określonej funkcji wyjściowej
* Uruchomienie sieci neuronowej na zestawie testów (na CPU oraz GPU)
* Uczenie sieci neuronowej na zestawie testów (na CPU oraz GPU).

W założeniu program miał umożliwić obliczanie wyników sieci tylko za pomocą GPU, ale dodać też możliwość wykonywania tych samych operacji na CPU. Zdecydowałem się na to, ponieważ wg. mnie to jest najlepszy sposób na określenie, czy część GPU działa poprawnie (przez porównywanie działania sieci uruchomionych na GPU oraz hoście). Innym możliwym sposobem sprawdzania poprawności obliczeń na GPU mogłoby być np. stworzenie sieci o tej samej strukturze w jakimś znanym programie do obsługi sieci neuronowych. Wolałem jednak mieć możliwość porównywania działania algorytmu na GPU i CPU na wszystkich etapach działania algorytmu (bardzo ułatwia to wyszukiwanie błędów).

Trenowanie sieci na hoście oraz GPU daje możliwość użycia więcej niż jednego testu w każdej iteracji. We większości bibliotek jest możliwość użycia tylko jednego testu w jednej iteracji treningu, ale w bibliotece CNL zostało to zrobione inaczej ze względu na charakter obliczeń na GPU. Przy jednym teście użytym w każdej iteracji, ilość danych przetwarzanych przed GPU byłaby zawsze niewielka, nawet w przypadku większych sieci neuronowych. Przy większej ilości testów użytych równocześnie, łatwo jest zrównoleglić algorytm na większą ilość użytych bloków.

Jednym z założeń było to, żeby biblioteka była łatwo rozszerzalna o obsługę nowych rodzajów sieci neuronowych (innych niż MLP, którą już zaimplementowałem). Aspekty biblioteki CNL, które ułatwiają dodawanie innych typów sieci:

* Rozdzieliłem implementacje sieci neuronowej oraz zestawu testów wykonywanego na tej sieci. Dzięki temu, dodanie nowego rodzaju sieci nie będzie ingerować w część aplikacji obsługującą same testy.
* Istnieją osobne klasy reprezentujące sieć neuronową, warstwę neuronów, oraz pojedynczy neuron, dzięki czemu przy dodawaniu nowego typu sieci będzie możliwe użycie istniejących klas warstwy lub neuronu. Obie te klasy posiadają też zaimplementowaną możliwość zapisywania lub wczytywania właściwości do/z pliku XML.
* Możliwe jest załadowanie danych sieci neuronowej z pliku XML bez znajomości typu sieci.

W bibliotece, warstwa wejściowa sieci nie ma odpowiadającego jej obiektu Layer. Jej rolę pełni wejście danego testu. Zrobiłem tak, ponieważ neurony wejściowe zawsze mają ten sam typ funkcji aktywacji (funkcję linearną). Dzięki temu nie jest też konieczne kopiowanie wartości wejściowych testu do danych warstwy wejściowej.

Po stronie hosta, wszystkie zmienne są przechowywane jako liczby zmiennoprzecinkowe podwójnej precyzji (double), jednak w związku z ograniczeniami technologii CUDA[[21]](#footnote-21), na komputerze który posiadam operacje po stronie GPU są przetwarzane z wykorzystaniem liczb pojedynczej precyzji (float). Gdyby biblioteka była użyta na karcie graficznej wspierającej compute capability 1.3 lub wyższą, możliwe by było użycie liczb podwójnej precyzji[[22]](#footnote-22).

### Zastosowania programu

Biblioteka służy do rozwiązywania zadań aproksymacji lub klasyfikacji. Każdy rekord w zestawie testów może mieć wiele wejść oraz wyjść. Każdy atrybut może być liczbowym atrybutem z pewnego zakresu, lub atrybutem symbolicznym (klasyfikacyjnym). Każdy atrybut symboliczny zostaje sprowadzony do jednego lub więcej atrybutów liczbowych (w szczególności – zadanie klasyfikacji jest sprowadzane w programie do zadania aproksymacji).

W bibliotece jest możliwość stworzenia sieci MLP, o dowolnej strukturze. Program obsługuje prostą funkcję aktywacji oraz sigmoidalną funkcję aktywacji w postaci unipolarnej oraz bipolarnej. Wspierane są tylko sieci gęste – w których każdy neuron w warstwie N jest połączony ze wszystkimi neuronami w warstwie N+1.

### Diagramy:

Poniżej zamieszczone są diagramy UML opisujące bibliotekę CNL.

#### Diagram klas

|  |
| --- |
|  |
| Rysunek 10. Diagram klas biblioteki CNL  Źródło: Diagram stworzony przeze mnie |

Pełna lista plików w projekcie razem z ich opisem znajduje się w rozdziale X.X.

#### Diagram komponentów

|  |
| --- |
|  |
| Rysunek 11. Diagram komponentów biblioteki CNL  Źródło: Diagram stworzony przeze mnie |

Na powyższym diagramie widać, że program używający biblioteki, zarządza osobno obiektami zestawu testów i sieci neuronowej. Obiekt sieci neuronowej (np. obiekt klasy MLP) wywołuje metody statyczne klasy CUDATools, które z kolei wywołują kernele. Biblioteki zawarte w CUDA Toolkit oraz sterowniku graficznym kopiują wersję binarną kerneli do pamięci graficznej i uruchamiają je.

#### Diagram sekwencji uczenia sieci

|  |
| --- |
|  |
| Rysunek 12. Diagram sekwencji uczenia sieci w bibliotece CNL  Źródło: Diagram stworzony przeze mnie |

Powyższy diagram ilustruje przypadek użycia biblioteki – uczenie nowo stworzonej sieci. Najpierw należy stworzyć obiekt zestawu testów (InputTestSet) oraz załadować testy z pliku. Następnie trzeba stworzyć obiekt sieci (klasy MLP), skonfigurować jego strukturę, a następnie wykonać procedurę uczenia (wykonaną na hoście lub na GPU). Po zakończeniu uczenia, można uruchomić sieć na podanym zestawie testów i zapisać zaktualizowany plik sieci neuronowej i plik zestawu testów.

#### Activity

Nie wiem, jak miałby ten diagram wyglądać…

### Struktura plików danych

Program, żeby mieć pełnię funkcji, musi mieć możliwość zapisywania i odczytywania stanu pomiędzy uruchomieniami. Poniżej jest zawarta tabela z dostępnymi formatami plików i możliwością ich obsługi:

|  |  |  |  |
| --- | --- | --- | --- |
| Tabela . Formaty plików danych obsługiwane przez bibliotekę CNL | | | |
|  | *MLP* | *InputTestSet* | *InputTestSet* |
| Format | XML | XML | CSV |
| Użycie | Read/Write | Read/Write | Read |
| Zapisane dane | Cała struktura sieci | Wszystkie dane zestawu testów | Wszystkie dane zestawu testów oprócz wyjść sieci dla podanych testów |

Ogólnodostępne zestawy testów są zwykle zapisane w formacie CSV. Chciałem też dodać przydatną możliwość wczytywania i zapisywania stanu sieci neuronowej lub zestawu testów. Zależało mi na tym, żeby format plików był czytelny również dla człowieka i zawierał wszystkie informacje o stanie danego obiektu. w związku z tym wybrałem format plików XML, w którym niektóre dane są nadmiarowe: w elemencie AttributeMapping jest informacja o minimalnej i maksymalnej wartości, a można to też wyczytać z konkretnych wartości testów. Jako potomkowie AttributeMapping jest podana lista elementów klasyfikacyjnych, która znajduje się też w potomku Tests. Dla większej czytelności, wartości liczbowe są skalowane do skali -1;1 dopiero po załadowaniu.

Poniżej opiszę strukturę podanych typów plików. w poniższej tabeli jest opisana sekwancja do wygenerowania plików XML zestawu testów i sieci MLP.

|  |
| --- |
| InputTestSet testSetCSV; // Nowy zestaw testów  vector<int> vecOutputColumns; // Tworzenie listy numerów kolumn wyjściowych (wynikowych)  vecOutputColumns.push\_back(12); // Jedyna wyjściowa kolumna - indeks 12  vector<int> vecUnusedColumns; // Lista numerów kolumn nieużywanych - pusta  testSetCSV.loadFromCSVFile // Ładowanie listy testów  ("forestfires2.csv" // Plik wejściowy z testami w formacie CSV  ,true // Pierwszy wiersz zawiera nazwy kolumn  ,',' // Określenie znaku oddzielającego elementy - przecinek  ,vecOutputColumns // Podanie listy kolumn wyjściowych  ,vecUnusedColumns); // Podanie listy kolumn nieużywanych  MLP dummyNet; // Nowa sieć MLP  dummyNet.setInputNeuronCount // Ustawienie ilości neuronów wejściowych  (testSetCSV.getInputCount());  dummyNet.addNewLayer // Dodawanie warstwie ukrytej  (6 // Ilość neuronów  ,Neuron::NT\_SIGMOID); // Funkcja aktywancji w warstwie ukrytej  dummyNet.addNewLayer // Ustawienie ilości neuronów wyjściowych ...  (testSetCSV.getOutputCount() // ... - tyle ile wyjść w zestawie testów  ,Neuron::NT\_LINEAR); // Wyjście sieci - linearne  dummyNet.randomizeWeights(0.01,NULL); // dobranie losowych wartości wag  dummyNet.trainNetwork // Uczenie sieci przez CPU  (testSetCSV // Uczenie wcześniej załadowanym zestawem testów  ,6000 // Ilość sekwencji uczenia sieci  ,0.01 // eta - czynnik uczenia  ,1 // Ilość testów uczona na raz  ,NULL); // Generator liczb pseudolosowych - niepotrzebny  dummyNet.executeNetwork(testSetCSV); // Uruchomienie sieci na wszystkich testach przez CPU  dummyNet.executeNetworkGPU(testSetCSV); // Uruchomienie sieci na wszystkich testach przez GPU  testSetCSV.saveToFile("TestSetFromCSV.xml");// Zapisywanie zestawu testów jako XML  dummyNet.saveToFile("NetworkStruct.xml"); // Zapisywanie sieci MLP jako XML |

#### Plik zestawu testów CSV

Plik forestfires2.csv użyty w powyższym kodzie ma następującą zawartość:

|  |
| --- |
| X,Y,month,day,FFMC,DMC,DC,ISI,temp,RH,wind,rain,area  7,5,mar,fri,86.2,26.2,94.3,5.1,8.2,51,6.7,1,0  7,4,oct,tue,90.6,35.4,669.1,6.7,18,33,0.9,0,0  8,4,oct,sat,90.6,43.7,686.9,6.7,14.6,33,1.3,0,1 |

Ten zestaw testów to 3 pierwsze testy z zestawu testów forestfires.csv, dostępnego w Internecie, po drobnej przeróbce. Czasem w plikach CSV pierwsza linia zawiera nazwy kolumn. Jeśli tak jest, należy podać wartość true jako drugi parametr metody InputTestSet::loadFromCSVFile. Format pliku CSV jest ogólnie znany. Jedyne różnice to znak oddzielający elementy – albo przecinek albo średnik.

Pliki CSV użyte w programie nie były stworzone przez mnie, więc metoda InputTestSet::loadFromCSVFile zawiera pewne zabezpieczenia przed niepoprawnymi danymi w pliku:

* Minimalnie 3 testy
* Ilość wartości w każdej linii jest taka sama
* Każda z wartości w każdym teście jest niepusta
* Każda kolumna ma przynajmniej 2 możliwe wartości

#### Plik zestawu testów XML

Poniżej jest zamieszczony plik TestSetFromCSV.xml wygenerowany na końcu sekwencji:

|  |
| --- |
| <?xml version="1.0" ?>  <TestSet SourceFileName="forestfires2.csv">  <AttributeMappings>  <AttributeMapping ColumnName="X" IsOutputAttribute="False" ColumnIndexInInputFile="0" ColumnIndexInStructure="0" IsLiteralAttribute="False" MinValue="7" MaxValue="8" />  <AttributeMapping ColumnName="Y" IsOutputAttribute="False" ColumnIndexInInputFile="1" ColumnIndexInStructure="1" IsLiteralAttribute="False" MinValue="4" MaxValue="5" />  <AttributeMapping ColumnName="month" IsOutputAttribute="False" ColumnIndexInInputFile="2" ColumnIndexInStructure="2" IsLiteralAttribute="True">  <ColumnElementName>mar</ColumnElementName>  <ColumnElementName>oct</ColumnElementName>  </AttributeMapping>  <AttributeMapping ColumnName="day" IsOutputAttribute="False" ColumnIndexInInputFile="3" ColumnIndexInStructure="3" IsLiteralAttribute="True">  <ColumnElementName>fri</ColumnElementName>  <ColumnElementName>tue</ColumnElementName>  <ColumnElementName>sat</ColumnElementName>  </AttributeMapping>  <AttributeMapping ColumnName="FFMC" IsOutputAttribute="False" ColumnIndexInInputFile="4" ColumnIndexInStructure="6" IsLiteralAttribute="False" MinValue="86.2" MaxValue="90.6" />  <AttributeMapping ColumnName="DMC" IsOutputAttribute="False" ColumnIndexInInputFile="5" ColumnIndexInStructure="7" IsLiteralAttribute="False" MinValue="26.2" MaxValue="43.7" />  <AttributeMapping ColumnName="DC" IsOutputAttribute="False" ColumnIndexInInputFile="6" ColumnIndexInStructure="8" IsLiteralAttribute="False" MinValue="94.3" MaxValue="686.9" />  <AttributeMapping ColumnName="ISI" IsOutputAttribute="False" ColumnIndexInInputFile="7" ColumnIndexInStructure="9" IsLiteralAttribute="False" MinValue="5.1" MaxValue="6.7" />  <AttributeMapping ColumnName="temp" IsOutputAttribute="False" ColumnIndexInInputFile="8" ColumnIndexInStructure="10" IsLiteralAttribute="False" MinValue="8.2" MaxValue="18" />  <AttributeMapping ColumnName="RH" IsOutputAttribute="False" ColumnIndexInInputFile="9" ColumnIndexInStructure="11" IsLiteralAttribute="False" MinValue="33" MaxValue="51" />  <AttributeMapping ColumnName="wind" IsOutputAttribute="False" ColumnIndexInInputFile="10" ColumnIndexInStructure="12" IsLiteralAttribute="False" MinValue="0.9" MaxValue="6.7" />  <AttributeMapping ColumnName="rain" IsOutputAttribute="False" ColumnIndexInInputFile="11" ColumnIndexInStructure="13" IsLiteralAttribute="False" MinValue="0" MaxValue="1" />  <AttributeMapping ColumnName="area" IsOutputAttribute="True" ColumnIndexInInputFile="12" ColumnIndexInStructure="0" IsLiteralAttribute="False" MinValue="0" MaxValue="1" />  </AttributeMappings>  <Tests>  <Test>  <Inputs>7;5;[-1][mar];[1;-1;-1][fri];86.2;26.2;94.3;5.1;8.2;51;6.7;1</Inputs>  <CorrectOutputs>0</CorrectOutputs>  <NetworkOutputs>-3.52944e-006</NetworkOutputs>  <NetworkOutputsGPU>-3.45707e-006</NetworkOutputsGPU>  </Test>  <Test>  <Inputs>7;4;[1][oct];[-1;1;-1][tue];90.6;35.4;669.1;6.7;18;33;0.9;0</Inputs>  <CorrectOutputs>0</CorrectOutputs>  <NetworkOutputs>1.0557e-005</NetworkOutputs>  <NetworkOutputsGPU>1.06096e-005</NetworkOutputsGPU>  </Test>  <Test>  <Inputs>8;4;[1][oct];[-1;-1;1][sat];90.6;43.7;686.9;6.7;14.6;33;1.3;0</Inputs>  <CorrectOutputs>1</CorrectOutputs>  <NetworkOutputs>0.999976</NetworkOutputs>  <NetworkOutputsGPU>0.999976</NetworkOutputsGPU>  </Test>  </Tests>  </TestSet> |

Jak widać, plik w formacie XML jest zdecydowanie większy od pliku CSV. Głównym powodem większego rozmiaru jest to, że każda wartość jest opisana nazwą. Oprócz tego, zestaw testów w formacie XML zawiera dodatkową informację niedostępną w pliku CSV – zawarte są w nim również wyjścia sieci obliczone przez CPU i GPU.

Struktura pliku jest taka, że główny element to TestSet. Zawiera on atrybut SourceFileName, który jest wypełniony, jeśli obiekt InputTestSet, który zapisywał dane, był wczytany z pliku CSV. Pierwszym potomkiem wlementu TestSet jest AttributeMappings – element zawierający informacje o wszystkich atrybutach (kolumnach) w danym zestawie testów. Jego potomkami są elementy AttributeMapping (po jednym na każdą kolumną testu). Atrybuty elementu AttributeMapping i ich opisy:

* ColumnName – nazwa danej kolumny w zestawie testów. Mogła być ustawiona w pliku CSV, który został wczytany, a następnie zapisany do pliku XML.
* IsOutputAttribute – wartość true/false określająca, czy atrybut jest wyjściowy, czyli czy ma odpowiednik w neuronach wyjściowych.
* ColumnIndexInInputFile – jeśli zestaw testów pochodzi z pliku CSV, to ten atrybut zawiera indeks kolumny w pliku CSV zawierającej ten atrybut. Jeśli któryś atrybut jest atrybutem klasyfikacyjnym, to ten atrybut oraz ColumnIndexInStructure mogą się różnić.
* ColumnIndexInStructure – atrybut zawiera indeks danej kolumny w strukturze wewnętrznej programu. Dana kolumna może odpowiadać więcej niż jednej kolumnie w strukturze wewnętrznej programu. Oprócz tego, ta wartość określa indeks danej kolumny w pliku XML w elementach Inputs, CorrectOutputs, NetworkOutputs, NetworkOutputsGPU
* IsLiteralAttribute – atrybut false/true określający, czy atrybut jest klasyfikacyjny.
* MinValue – atrybut występuje tylko przy atrybutach liczbowych (IsLiteralAttribute = false). Określa minimalną wartość danej kolumny.
* MaxValue - atrybut występuje tylko przy atrybutach liczbowych (IsLiteralAttribute = false). Określa mmaksymalną wartość danej kolumny.

Atrybuty klasyfikacyjny (IsLiteralAttribute = false) mają potomków o nazwie ColumnElementName. Każdy z tych potomków zawiera literalną wartość klasyfikacyjną. Jeśli są możliwe tylko dwie wartości, wtedy danemu atrybutowi testu odpowiada jeden neuron wejściowy/wyjściowy (wartość pierwsza odpowiada wartości -1, a druga wartości 1). Jeśli jest więcej wartości, wtedy każda możliwa wartość odpowiada jednemu neuronowi wejściowemu/wyjściowemu. w tym przypadku test, którego ta kolumna ma n-ta wartość, powoduje, że wszystkie neurony wejściowe/wyjściowe przypisane do tego atrybutu mają wartość -1 oprócz n-tego, który ma wartość 1.

Tests - element ten zawiera informacje o wszystkich testach w danym zestawie testów. Jego potomkami są elementy Test (po jednym na każdy test w zestawie). Elementy potomne elementu AttributeMapping i ich opisy:

* Inputs – wejścia testu
* CorrectOutputs – wyjścia wzorcowe
* NetworkOutputs – wyjścia wygenerowane przez CPU. Nadpisywane przez metodę NeuralNetwork::executeNetwork() lub NeuralNetwork::trainNetwork().
* NetworkOutputsGPU - wyjścia wygenerowane przez GPU. Nadpisywane przez metodę NeuralNetwork::executeNetworkGPU() lub NeuralNetwork::trainNetworkGPU().

Każdy element w wejściach lub wyjściach jest albo wartością liczbową albo klasyfikacyjną. Są one poprzedzielane średnikami. Wartości liczbowe są przedstawione jako liczby. Wartości klasyfikacyjne są przedstawione liczba/liczby oraz nazwa klasyfikacyjna. Nazwa klasyfikacyjna jest określana po tym, która wartość jest największa. Jeśli k-ta wartość, wtedy „wybrana” została k-ta wartość klasyfikacyjna. Np. jeśli dana kolumna wyjściowa przyjmuje wartości ‘a’, ‘b’, ‘c’, a wartości neuronów wyjściowych odpowiadających tej kolumnie wynoszą -0.8, 0.3, 0.1, to oznacza, że w tym teście została wybrana druga wartość tej kolumny (czyli ‘b’).

#### Format Sieci MLP w XML

Poniżej jest plik sieci MLP stworzony po poprzedniej sekwencji operacji w pliku NetworkStruct.xml:

|  |
| --- |
| <?xml version="1.0" ?>  <NeuralNetwork Type="MLP">  <Layer>  <Neuron NeuronType="NT\_SIGMOID">  <Weights>-0.644995;0.0517328;-0.044401;0.0508847;0.553819;-0.640026;-0.0499197;-0.340208;-0.0756364;-0.0542719;0.154846;0.049415;0.0122903;0.0387111;0.0363326;</Weights>  </Neuron>  <Neuron NeuronType="NT\_SIGMOID">  <Weights>-0.492991;0.0634698;-0.0655245;0.0716414;0.426053;-0.491632;-0.0600052;-0.262257;-0.0701903;-0.0706763;0.110501;0.0650343;0.0256953;0.0695995;0.014993;</Weights>  </Neuron>  <Neuron NeuronType="NT\_SIGMOID">  <Weights>-0.597295;0.0589075;-0.0505662;0.0570438;0.509406;-0.600241;-0.0560901;-0.306225;-0.0646669;-0.050627;0.138164;0.0468326;0.0200996;0.0558004;0.0230561;</Weights>  </Neuron>  </Layer>  <Layer>  <Neuron NeuronType="NT\_LINEAR">  <Weights>-1.10815;-0.736504;-0.978474;1.238;</Weights>  </Neuron>  </Layer>  </NeuralNetwork> |

Główny element pliku XML ma nazwę NeuralNetwork. Ma on jeden atrybut Type. Określa on typ sieci neuronowej. Sieć MLP ma w sobie jednego lub więcej potomków Layer reprezentujących warstwę ukrytą lub wyjściową. Warstwy są uporządkowane w kolejności od pierwszej warstwy ukrytej do warstwy wyjściowej. Ilość neuronów wejściowych w testach przyjmowana przez tę sieć jest równa ilości wag w pierwszej warstwie ukrytej (lub wyjściowej, jeśli nie ma warstw ukrytych) minus jeden. Każdy element Layer ma jeden lub więcej potomków Neuron. Element Neuron ma jeden atrybut NeuronType o wartości NT\_LINEAR lub NT\_SIGMOID określający funkcję aktywacji. Element Neuron posiada jednego potomka o nazwie Weights, który zawiera wagi neuronu oddzielone średnikiem.

### Katalogi i pliki w projekcie

Poniżej znajduje się lista katalogów i plików użytych w projekcie Visual Studio 2008:

|  |  |
| --- | --- |
| **Wstaw FileIndex.txt.**  **Pogrub (albo pochyl) nazwy plików i katalogów** |  |

### Logowanie

W programie dodałem obsługę logowania wiadomości do pliku lub na konsolę, które służy do chronologicznej rejestracji różnych zdarzeń. Każda wiadomość ma określony „poziom” logowania (typ wiadomości), np. LT\_INFORMATION (informacje o przebiegu programu), LT\_WARNING (ostrzeżenia o potencjalnych błędach), lub LT\_ERROR (błędy lub niespełnione warunki asercji). Aplikacja umożliwia dynamiczne określenie, które poziomy logowania są wyświetlane w konsoli, a które w pliku logowania. Logowanie umożliwia:

* określanie sekwencji działań w bibliotece
* mierzenie czasu wykonywania pewnych operacji
* kilka razy w czasie pisania programu, logowanie pewnych danych bardzo ułatwiło mi rozwiązanie problemów z niestabilnym działaniem biblioteki
* w czasie wczytywania zestawu testów z pliku CSV, mogą być logowane informacje o wszystkich kolumnach testów
* można logować informacje o tym, jak duże są różnice między poprawnymi wyjściami testów a wyjściami sieci neuronowej (metoda InputTestSet::printVectorDifferenceInfo)

Prawie wszystkie informacje są logowane przy pomocy metody Logging:: logTextFileLine, jednak ograniczenia kompilatora CUDA uniemożliwiają użycie go wewnątrz pliku TrainNetwork.cu – z tego powodu logowanie w tym pliku jest wykonywane przez funkcję printf (tylko do konsoli). Logowanie wewnątrz kerneli jest też ograniczone tylko do trybu emulacji.

## Część CPU

W bibliotece została zaimplementowana możliwość uruchamiania i uczenia sieci MLP wykonywana bez użycia GPU. Każdy obiekt sieci MLP (sieć neuronowa, warstwa neuronów, pojedynczy neuron), posiada reprezentującą go klasę (odpowiednio klasy MLP, Layer, Neuron). Każdy obiekt jednej z tych klas zawiera następujące informacje:

* Obiekt MLP zawiera informacje o warstwach (obiekty Layer) w nim zawartych
* Obiekt Layer zawiera informacje o neuronach w nim zawartych
* Obiekt Neuron zawiera właściwości danego neuronu - typ neuronu, wartości wag, wartości ostatnich wyjść i błędów.

### Kolejność operacji przy operacjach na MLP

Poniżej zamieściłem dwa diagramy przedstawiające uruchamianie i trenowanie sieci MLP razem z opisami.

#### Diagram sekwencji uruchamiania sieci

|  |
| --- |
|  |
| Rysunek 13. Diagram sekwencji uruchamiania sieci na CPU  Źródło: Diagram stworzony przeze mnie |

Na powyższym diagramie opisano sekwencję operacji przy uruchomieniu sieci przy użyciu CPU. Najpierw są pobierane wejścia testu, później obiekt MLP uruchamia po kolei wszystkie warstwy sieci, a obiekt każdej warstwy oblicza wyjście każdego neuronu w tej warstwie. Później, zapisywane są obliczone wyjścia testu.

#### Diagram sekwencji trenowania sieci

|  |
| --- |
|  |
| Rysunek 14. Diagram sekwencji trenowania sieci na CPU  Źródło: Diagram stworzony przeze mnie |

W powyższym diagramie opisano sekwencję operacji dla jednej iteracji trenowania sieci MLP przy użyciu CPU. Pierwszym etapem jest wybranie losowych testów z zestawu testów oraz uruchomienie ich. Następnie określany jest „błąd” wyjściowy każdego neuronu. Końcowym etapem jest zaktualizowanie wszystkich wag.

## Część GPU

Wszystkie kernele (funkcje wykonywane na GPU) oraz funkcje języka C służące do wywoływania kerneli, zawarte są w pliku TrainNetwork.cu. Kernele są oznaczone przez słowo kluczowe \_\_global\_\_, a ich nazwy kończą się słowem „Kernel”, natomiast funkcje wywołujące kernele mają nazwy z zakończeniem „CUDA”. Pozostałe funkcje używające wywołań CUDA znajdują się w pliku CUDATools.cpp. Zawarte są tam metody statyczne odpowiedzialne za pozostałą komunikację z GPU. Są to operacje alokowania, dealokowania, ustawiania, pobierania pamięci GPU, oraz uruchamiania funkcji z zawartych w TrainNetwork.cu.

### Implementacja MLP na GPU

#### Ogólny algorytm

Algorytmy uruchamiania i trenowania sieci przez wsteczną propagację błędu są identyczne na CPU i GPU (jedyne różnice to kolejność dodawania kolejnych czynników). Mimo to, wykonanie tych algorytmów na procesorze graficznym jest bardziej skomplikowane ze względu na konieczność kopiowania danych do tablic w pamięci graficznej oraz równoległy charakter operacji na GPU. W kolejnych podrozdziałach są zamieszczone sekwencje działań (z punktu widzenia obiektów i klas) wykonywane przy uruchamianiu i uczeniu sieci MLP.

#### Diagram sekwencji uruchamiania sieci

|  |
| --- |
|  |
| Rysunek 15. Diagram sekwencji uruchamiania sieci na GPU  Źródło: Diagram stworzony przeze mnie |

Na powyższym diagramie opisano sekwencję operacji przy uruchomieniu sieci przy użyciu GPU. Dla każdej warstwy jest wykonywanych kilka operacji. Najpierw alokowana jest pamięć potrzebna dla aktualnej warstwy (wagi, wyjścia) i zwalniana jest niepotrzebna pamięć używana wcześniej. Następnie jest wykonywana metoda statyczna executeLayerGPU, która uruchamia kernel wykonujący równolegle operacje uruchamiania neuronów na danej warstwie. Sam kernel opisany jest w rozdziale X.X.

#### Diagram sekwencji trenowania sieci

|  |
| --- |
|  |
| Rysunek . Diagram sekwencji trenowania sieci na GPU  Źródło: Diagram stworzony przeze mnie |

Na powyższym diagramie opisano sekwencję operacji dla jednej iteracji uczenia sieci MLP przy użyciu GPU. Pierwszym etapem jest wybranie losowych testów z zestawu testów oraz uruchomienie ich. Następnie określany jest „błąd” wyjściowy każdego neuronu. na końcu każdej iteracji wszystkie wagi sieci są aktualizowane.

#### Struktury danych w pamięci GPU

Po stronie CPU jest określona hierarchia klas, a wszystkie dane sieci neuronowej lub zestawu testów znajdują się w obiektach tych klas. Technologia CUDA też umożliwia używanie złożonych typów danych (struct). Jest to ułatwienie, jednak nie jest to zalecane – w większości przypadków spowodowałoby to kilkukrotne spowolnienie pracy kerneli. Żeby uzyskać wysoką wydajność, konieczne jest czytanie oraz zapisywanie danych z/do globalnej pamięci przy pomocy tablic.

Z tego względu, po stronie GPU używane są tylko zmienne całkowite, zmienne zmiennoprzecinkowe oraz tablice jednowymiarowe zmiennych tych typów. Alokowanie i dealokowanie pamięci na GPU zajmuje dużo więcej czasu niż w pamięci RAM, więc nie ma osobnych tablic dla pojedynczego neuronu lub pojedynczego testu – tablice zawierają informacje o całej warstwie neuronów lub zestawie testów. Poniżej znajduje się lista tablic używanych po stronie GPU oraz ich opis:

* TestsInput – tablica zawierająca wartości wejść testów
* TestsOutput – tablica zawierająca wartości wyjść testów
* LayerWeights – wagi w danej warstwie neuronów
* DerivativeOfLastOutput – pochodna wyjść neuronów w danej warstwie
* LastOutputWithOutputFunction – wyjścia neuronów po wykonaniu funkcji aktywacji
* LastError – różnica między oczekiwanymi a rzeczywistymi wyjściami neuronów
* TestIndices – tablica określa indeksy testów używanych przy trenowaniu sieci, znajduje się w pamięci stałej (constant).

Niektóre z powyższych tablic to są tzw. padded tables (zobacz rozdział x.x), gdzie niektóre komórki nie są używane, żeby odczyt z tych tablic lub zapis do nich był wykonywany szybciej. Oprócz tych tablic, kernele używają tablic znajdujących się w pamięci dzielonej, która jest używana jako szybka pamięć podręczna.

#### Opis działania kerneli

Wszystkie operacje na GPU są wykonywane w kernelach (odpowiednikach funkcji wykonywanych na CPU). Poniżej są wymienione i opisane kernele użyte w bibliotece.

|  |  |  |
| --- | --- | --- |
| Tabela . Opis użytych kerneli | | |
|  | executeLayerKernel | calculateErrorInLastLayerKernel |
| Opis | Oblicza wyjście neuronów w danej warstwie na podstawie wejść i wag neuronów. | Oblicza błąd w ostatniej warstwie sieci |
| Użycie | Uruchamianie i trenowanie sieci. | Trenowanie sieci |
| Ilość bloków | W przypadku uruchamiania sieci - liczba testów w zestawie testów.  W przypadku trenowania sieci – ilość uczonych elementów | Ilość uczonych elementów |
| Ilość wątków | Ilość neuronów w danej warstwie | Ilość wyjściowych neuronów |
| Użycie shared memory | Wejścia warstwy i wagi neuronów. | Nie |
|  | calculateErrorInNotLastLayerKernel | updateWeightsInTrainingKernel |
| Opis | Oblicza i zapisuje błąd w warstwie sieci innej niż ostatnia. | Uaktualnia wartości wag w danej warstwie |
| Użycie | Trenowanie | Trenowanie |
| Ilość bloków | Ilość uczonych elementów | Ilość neuronów w danej warstwie |
| Ilość wątków | Ilość neuronów w danej warstwie | Ilość neuronów w poprzedniej warstwie |
| Użycie shared memory | Błąd wyjścia oraz wagi neuronów kolejnej warstwy | Błąd wyjścia oraz pochodne wyjścia neuronów danej warstwy |

Powyżej zostało podane, w jakich sytuacjach jest używany dany kernel, od czego zależy ilość bloków oraz wątków, oraz opisane, czy jest użyta pamięć dzielona (a jeśli tak, to na co). Jak widać, w kernelu updateWeightsInTrainingKernel ilość bloków i wątków jest określana w inny sposób niż w innych kernelach. Zostało to tak zrobione, żeby przyspieszyć dostęp do pamięci globalnej.

#### Analiza zależności danych

Operacje są wykonywane na GPU przez wiele wątków na raz, a prawie zawsze do tych operacji jest wykorzystywana pamięć współdzielona przez te wątki. Jeśli dana komórka pamięci może być w jednym momencie zapisywana przez jeden wątek, i odczytywana przez inny, program staje się niedeterministyczny – uruchomienie programu wiele razy może dawać różne wyniki (zależność „read after write”, szerzej opisana w (Pfeiffer, 2005)). w kernelach użytych w CNT jest wiele tego typu zależności przy czytaniu/zapisywaniu do pamięci dzielonej. Aby mieć pewność, że zależne od siebie operacje zawsze będą wykonane w odpowiedniej kolejności, użyta jest komenda synchronizująca \_\_syncthreads*[[23]](#footnote-23)*.

### Ograniczenia wykonanego algorytmu.

Chociaż GPU pozwalają na wykonywanie programów niegraficznych, to cały czas ich głównym zastosowaniem jest rendering grafiki. Architektura GPU i różnego typu pamięci graficznych jest głównie przygotowana do tych zastosowań – przez co programy ogólnego zastosowania wykonywane na GPU muszą brać pod uwagę pewne ograniczenia. Ograniczenia mojej biblioteki są właśnie związane z ograniczeniami GPU firmy NVIDIA. Te ograniczenia to:

* Każda z warstw może zawierać nie więcej niż 511 neuronów. Jest to związane z maksymalną ilością wątków w bloku.
* Jeden zestaw testów może zawierać maksymalnie 65535 testów. Jest to związane z maksymalną ilością bloków w jednym wymiarze gridu.

Możliwe byłoby ominięcie tych ograniczeń, ale zmiana kodu byłaby żmudna, a modyfikacje mogłoby zwolnić działanie programu.

Innego rodzaju ograniczeniem (ograniczeniem samej biblioteki CUDA) jest to, że nawet w trybie emulacji nie można wywoływać funkcji innych niż standardowych, przez co logowanie wewnątrz kerneli, w trybie emulacyjnym, jest inne niż w reszcie programu.

Pamięć GPU jest zwykle mniejsza niż RAM, a wszystkie dane muszą się na niej zmieścić, jednak nie jest to problemem w przypadku biblioteki CNL. Maksymalna możliwa wielkość jednocześnie zadeklarowanej pamięci graficznej wynosi około 140 MB[[24]](#footnote-24), czyli mniej od minimalnej wymaganej pamięci graficznej potrzebnej do używania CUDA (256 MB).

### Użyte optymalizacje kerneli

Pierwsza wersja pliku TrainNetwork.cu zawierała niezoptymalizowane kernele służące do wykonywania i uczenia sieci. Używając programu CudaProf zauważyłem, że przy krótko działających kernelach, czas uruchomienia kernela jest często większy niż jego wykonania. Jest to problemem i powoduje, że wywołania niewielkich kerneli na GPU mogą działać dużo dłużej niż na CPU. w przypadku biblioteki CNL moc obliczeniowa potrzebna do wykonania kernela nie jest znana przed jego uruchomieniem. w przypadku programów wykonywanych na GPU przy dokonywaniu modyfikacji (optymalizacji), dobrze jest ją przetestować dla kilku różnych zestawów danych wejściowych – często się zdarza, że modyfikacja w jednym przypadku przyspiesza działanie kernela, a w przypadku innych danych wejściowych spowalnia.

W przypadku niniejszej biblioteki, w czasie optymalizacji, nacisk był kładziony na jak najlepsze zoptymalizowanie działania kerneli w przypadku dużej ilości danych wejściowych[[25]](#footnote-25).Poniżej są opisane optymalizacje, których użyłem, żeby przyśpieszyć działanie kerneli przy dużej ilości danych

#### Dostępy do danych globalnych o dużej przepustowości

W rozdziale Struktura pamięci GPU jest opisany, czym są tzw. złączone (coalesced) zapisy/odczyty globalnej pamięci globalnej. Aby zwiększyć przepustowość komunikacji między pamięcią graficzną a GPU, dostępy złączone zostały użyte w następujących przypadkach:

Wejścia testów, wyjścia testów oraz wyjścia wszystkich warstw są wyrównane (aligned) do 16 elementów. w prawie wszystkich kernelach (oprócz updateWeightsInTrainingKernel),kolejne wątki w bloku reprezentują kolejne neurony. w związku z czym n-ty wątek czyta/zapisuje n-ty element pamięci, dostęp do pamięci jest złączony.

Wagi sieci nie są wyrównane do 16 elementów, ale w kernelach calculateErrorInNotLastLayerKernel i executeLayerKernel odczyty tych wag są złączone (wagi te najpierw są przepisywane do pamięci dzielonej, a później odczytywane).

#### Obsługa 2 testów na raz

W kernelach executeLayerKernel i calculateErrorInNotLastLayerKernel jeden wątek reprezentuje dwa testy, a w kernelu updateWeightsInTrainingKernel jeden wątek reprezentuje dwie wagi. Ta optymalizacja powoduje, że dane skopiowane w każdym bloku mogą być użyte do wykonania dwóch razy więcej obliczeń. Jednak zmiana prędkości działania kernela executeLayerKernel po tej optymalizacji jest bardzo zależna od wielkości sieci. w przypadku sieci, gdzie 2 kolejne warstwy mają po 500 neuronów, przyspieszenie jest prawie dwukrotne. w przypadku sieci, gdzie 2 kolejne warstwy mają po 20 neuronów, kernel działa ok. 10% wolniej.

#### Zapisywanie danych do pamięci dzielonej

Pamięć dzielona (shared memory) jest o wiele szybsza w dostępie od pamięci globalnej (nawet przy złączonym dostępie) i ma dużo krótszy czas dostępu (2 cykle zegara w porównaniu z kilkuset cyklami). Oprócz tego umożliwia łatwe dzielenie się danymi i komunikację między różnymi wątkami wewnątrz bloku. Zmniejszenie ilości operacji w pamięci globalnej a operowanie na lokalnej jest opisywane w poradnikach o technologii CUDA jako pierwszy krok optymalizacji kerneli. Kernele executeLayerKernel, calculateErrorInNotLastLayerKernel, updateWeightsInTrainingKernel używają pamięci dzielonej do przechowywania wag, wejść neuronów, błędów na wyjściu neuronu oraz pochodnej wyjścia neuronu. Kernel executeLayerKernel jako jedną z pierwszych operacji, wykonuje kopiowanie globalnych danych do tablicy w pamięci dzielonej:

|  |
| --- |
| // first, we copy d\_LayerInputThisTest to s\_InputNeurons  for(int iInputIndex = threadIdx.x;iInputIndex < p\_iNumInputNeurons; iInputIndex+=blockDim.x)  {  s\_InputNeurons[iInputIndex] = d\_LayerInputThisTest[iInputIndex];  s\_InputNeurons2[iInputIndex] = d\_LayerInputThisTest2[iInputIndex];  PRINT\_MEMORY\_INFO(dp\_pLayerInput,&d\_LayerInputThisTest[iInputIndex]);  }  // we have to make sure that all data was written to shared memory  \_\_syncthreads(); |

Po kopiowaniu danych globalnych do pamięci dzielonej konieczne jest zsynchronizowanie między wszystkimi wątkami jednego bloku (opisane jest to w podrozdziale Analiza zależności danych). Bardzo podobne odbywa się to w dwóch pozostałych kernelach.

#### Przewidywanie ilości uruchomionych bloków

Kernel executeLayerKernel używa pamięci dzielonej – jest ona przeznaczana do przechowywania wejść oraz części wag neuronów. Ilość pamięci przeznaczona na przechowywanie wejść jest z góry określona, ale ta na przechowywanie wag – nie. Nie ma gwarancji, że pamięć potrzebna na przechowywanie wag nie przekroczy maksymalnej pamięci dzielonej jednego multiprocesora (16kb). z tego względu w każdej iteracji pętli, część tablicy wag jest wczytywana do pamięci dzielonej, później przetwarzana, a w kolejnej iteracji pamięć dzielona jest nadpisywana kolejnym fragmentem tablicy wag i przetwarzana. w tym przypadku można zdecydować o wielkości pamięci dzielonej poświęconej na wagi w każdym bloku. Im większa wielkość pamięci dzielonej dla jednego bloku, tym mniej będzie potrzebnych do wykonania iteracji pętli, mniej synchronizacji, ale jednocześnie może być mniejsza ilość bloków równocześnie wykonywanych przez multiprocesor GPU (ponieważ ilość pamięci dzielonej na multiprocesorze wynosi tylko 16kb). w funkcji executeLayerCUDA są wykonywane obliczenia optymalnej ilości pamięci dzielonej dla wag. Obliczenia te bazują testach empirycznych, tak żeby kernel działał szybko w przypadku różnej wielkości danych.

#### Użycie pamięci constant

W przypadku uczenia sieci, każde wywołanie executeLayerKernel dla pierwszej warstwy sieci wymaga przekazania tablicy indeksów testów, które mają być użyte w danym kroku uczenia. Zwykle wielkość tej tablicy nie jest duża, ale nie może ona być przekazana jako parametr wywołania kernela – biblioteka CUDA nie pozwala na to. Jednym z rozwiązań byłoby to, że indeksy te byłyby wczytywane za każdym razem z pamięci globalnej. w tym wypadku jednak wymagałoby to użycia pamięci dzielonej, a poza tym każdy blok musiałby kopiować tę samą tablicę globalną. z tego względu użyłem specjalnego rodzaju pamięci – stałej (constant). Dzięki temu, że ilość pamięci użytej na przechowanie tabeli jest mniejsza niż ilość cache wewnątrz każdego multiprocestora (8kb), dane te będą kopiowane tylko raz do multiprocesora, a później kopiowane z szybkiej pamięci cache.

## Testy implementacji sieci MLP

Jednym z zadań w ramach pisania tej pracy było wykorzystanie zestawów testów do uczeniu sieci. Dzięki temu można:

* określić, czy uczenie przez CPU i GPU daje takie same wyniki
* określić najlepsze parametry uczenia sieci
* porównać prędkość działania operacji na CPU i GPU oraz prędkość działania na różnych platformach

Dla zwiększenia wiarygodności wyników, każdy scenariusz zostaje przeprowadzony 3 razy, a statystyki błędów oraz czasy działania operacji są uśredniane.

### Opis danych testowych

Wybrałem następujące zestawy testów:

* „Concrete” – baza z 7 wejściami opisującymi skład oraz wiek betonu, oraz parametrem wyjściowym określającym jego twardość.
* “Iris Plants Database” – w tej bazie są 4 wejścia określające atrybuty irysa. Wyjściem jest jeden atrybut – określający jeden z 3 typów irysów.
* “Wisconsin Diagnostic Breast Cancer (WDBC)” – Baza określająca na podstawie 30 atrybutów wejściowych, czy guz piersi jest łagodny czy złośliwy.

Wszystkie te trzy zestawy posiadają tylko jedną zmienną wyjściową i nie mają brakujących wartości. Zostały one ściągnięte z ogólnodostępnego repozytorium (Asuncion & Newman, 2010).

### Środowisko testowe

Wszystkie testy zostały przeprowadzone na laptopie z następującą konfiguracją:

* Procesor Intel Core 2 Duo P8400, 2.26Ghz
* 4GB RAM DDR2, 800Mhz
* Karta graficzna z chipsetem NVIDIA GeForce 9600M GT, 512 MB RAM DDR2
* Windows Vista 32bit SP2 oraz Windows 7 64bit[[26]](#footnote-26)
* CUDA Toolkit 2.3, CUDA SDK 2.3[[27]](#footnote-27)

### Wpływ parametrów sieci na jakość uczenia

W poniższej tabeli zostały zamieszczone wyniki jakości uczenia sieci neuronowych dla trzech powyższych zestawów testów. Wyniki te zostały wygenerowane przez funkcję makeTrainingWithManyPossibilities, która trenuje sieci neuronowe przy użyciu różnych kombinacji parametrów uczenia. Sieć MLP używana w tej funkcji zawiera jedną warstwę ukrytą z sigmoidalną funkcją aktywacji.

Wyjścia sieci dla wszystkich tych konfiguracji przy przeprowadzaniu uczenia sieci na GPU były praktycznie identyczne do tych przeprowadzanych na hoście (różnice wynosiły najwyżej 0.000001).

W pierwszym wierszu zostały podane informacje o statystykach błędów przy uruchomieniu sieci przed uczeniem (dane oznaczone kursywą). Są tu też wszystkie 24 kombinacje różnych wartości parametrów:

* Ilości iteracji w czasie uczenia sieci (40000, 80000 lub 160000)
* Współczynnik uczenia Eta (0.01 lub 0.03)
* Ilość testów użytych w czasie jednej iteracji (1 lub 4)
* Ilość neuronów w warstwie ukrytej sieci (32 lub 64)

W kolumnach został umieszony maksymalny i średni bezwzględny błąd danych z pliku Concrete\_Data.csv, oraz część niepoprawnie sklasyfikowanych testów dla danych z plików iris.data i wdbc.data. Kolorem żółtym oznaczono najlepsze kombinacje parametrów (o najmniejszych błędach) dla każdej z trzech ilości iteracji, w których był użyty jeden test w iteracji (koloru zielonego użyto odpowiednio przy kombinacjach z 4 testami w iteracji). Zestawy danych iris.data i wdbc.data są zadaniami klasyfikacji, w związku z czym ilość błędów (niepoprawnie zaklasyfikowanych elementów) dla różnych konfiguracji sieci często się powtarza.

|  |
| --- |
| Tabela . Porównanie jakości uczenia sieci w zależności od różnych parametrów |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Ilość iteracji | Eta | Testów w iteracji | Neurony w warstwie ukrytej | Concrete\_Data.csv | | iris.data | wdbc.data |
| Maks. | Średni |
| *1* | *0.01* | *1* | *32* | *1.055382* | *0.355913* | *0.666667* | *0.542472* |
| 40000 | 0.01 | 1 | 32 | 0.881323 | 0.210257 | 0.022222 | 0.025190 |
| 40000 | 0.01 | 1 | 64 | 0.948090 | 0.226863 | 0.022222 | 0.025776 |
| 40000 | 0.01 | 4 | 32 | 0.862221 | 0.208165 | 0.022222 | 0.016989 |
| 40000 | 0.01 | 4 | 64 | 0.865360 | 0.211914 | 0.020000 | 0.017575 |
| 40000 | 0.03 | 1 | 32 | 0.885246 | 0.213265 | 0.017778 | 0.019332 |
| 40000 | 0.03 | 1 | 64 | 0.971700 | 0.235784 | 0.022222 | 0.019332 |
| 40000 | 0.03 | 4 | 32 | 0.749446 | 0.144566 | 0.020000 | 0.016403 |
| 40000 | 0.03 | 4 | 64 | 0.844297 | 0.147203 | 0.020000 | 0.016403 |
| 80000 | 0.01 | 1 | 32 | 0.874994 | 0.209155 | 0.026667 | 0.019918 |
| 80000 | 0.01 | 1 | 64 | 0.906669 | 0.214904 | 0.026667 | 0.024605 |
| 80000 | 0.01 | 4 | 32 | 0.817314 | 0.186735 | 0.020000 | 0.013474 |
| 80000 | 0.01 | 4 | 64 | 0.913231 | 0.216725 | 0.020000 | 0.014060 |
| 80000 | 0.03 | 1 | 32 | 0.883803 | 0.201784 | 0.020000 | 0.018746 |
| 80000 | 0.03 | 1 | 64 | 0.924605 | 0.211393 | 0.024444 | 0.017575 |
| 80000 | 0.03 | 4 | 32 | 0.718555 | 0.117593 | 0.017778 | 0.011716 |
| 80000 | 0.03 | 4 | 64 | 0.688783 | 0.117480 | 0.020000 | 0.011131 |
| 160000 | 0.01 | 1 | 32 | 0.867153 | 0.208844 | 0.024444 | 0.016989 |
| 160000 | 0.01 | 1 | 64 | 0.876605 | 0.212444 | 0.020000 | 0.017575 |
| 160000 | 0.01 | 4 | 32 | 0.771856 | 0.140492 | 0.020000 | 0.012888 |
| 160000 | 0.01 | 4 | 64 | 0.785761 | 0.146911 | 0.020000 | 0.012302 |
| 160000 | 0.03 | 1 | 32 | 0.761427 | 0.144541 | 0.022222 | 0.015817 |
| 160000 | 0.03 | 1 | 64 | 0.859998 | 0.156688 | 0.020000 | 0.016403 |
| 160000 | 0.03 | 4 | 32 | 0.677923 | 0.123546 | 0.017778 | 0.009373 |
| 160000 | 0.03 | 4 | 64 | 0.659576 | 0.113889 | 0.020000 | 0.008787 |

Wpływ czynników na jakość sieci:

* w wynikach testów można zauważyć prawidłowość, że najlepsze kombinacje parametrów z 4 testami w iteracji dają lepsze wyniki (mniejszy błąd) niż te z 1 testem w iteracji
* wszystkie sieci neuronowe po nauczeniu wykazywały mniejszy błąd niż sieć nienauczona, a większa ilość iteracji uczenia zmniejsza błąd wyjścia sieci
* w przypadku podanych wyżej testów, współczynnik uczenia eta o wartości 0.03 dawał lepsze wyniki niż współczynnik 0.01
* trudno określić wpływ neuronów w warstwie ukrytej. Czasem większa ilość zwiększa, czasem zmniejsza błąd

### Wpływ optymalizacji na wydajność kerneli

Głównym celem pisania programów na karty graficzne jest uzyskanie większej prędkości niż przy uruchamianiu ich na CPU. Aby rezultaty były zadowalające, często trzeba dokonać odpowiednich optymalizacji. Poniżej zamieściłem porównanie prędkości działania uruchamiania sieci neuronowej na zestawach testów dla zoptymalizowanej oraz niezoptymalizowanej wersji kernela GPU, oraz dla wersji CPU. Testy zostały przeprowadzone dla sieci z dwoma warstwami ukrytymi o tej samej ilości neuronów (liczba w pierwszej kolumnie poniższej tabeli określa ilość tych neuronów).

// 2 -> X (NT\_SIGMOID)-> X (NT\_SIGMOID)-> 2(linear) neurony (wstaw ilustracje)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Tabela . Porównanie czasu uczenia sieci na CPU i GPU | | | | | |
| Ilość neuronów | Czas uruchamiania sieci na CPU | Całkowity czas uruchamiania sieci na GPU (executeNetworkGPU) | | Czas wywołania samego kernela | |
| Niezoptymalizowane | Zoptymalizowane | Niezoptymalizowany (executeLayerKernel\_OLD) | Zoptymalizowany (executeLayerKernel) |
| 32 | 143 | 239 | 271 | 19 | 5 |
| 64 | 241 | 312 | 241 | 57 | 11 |
| 128 | 494 | 486 | 302 | 237 | 78 |
| 256 | 1274 | 1320 | 722 | 1074 | 449 |
| 510 | 4462 | 3701 | 2000 | 3457 | 1771 |
| 511 | 4492 | 6396 | 2090 | 6169 | 1865 |

Można zauważyć, że czas alokacji i dealokacji pamięci (różnica między czasem wykonania metody executeNetworkGPU a sumą czasów wykonania samych kerneli) nie jest zależna od ilości zadeklarowanych danych i wynosi ok. 200ms. Czas działania zoptymalizowanego algorytmu na GPU przy małej ilości neuronów jest podobny do tego wykonywanego na hoście, ale im więcej neuronów, tym różnica się zwiększa na korzyść algorytmu GPU. Przy maksymalnej wielkości sieci, wersja CUDA jest 2.15 razy szybsza od wersji CPU (oraz ok. 1.8 razy szybsza od wersji niezoptymalizowanej). Co ciekawe, wersja niezoptymalizowana jest bardzo wrażliwa na ilość danych wejściowych – czas działania dla 511 neuronów jest prawie 2 razy większy niż dla 510[[28]](#footnote-28).

### Porównanie wydajności uczenia wersji CPU i GPU

Na zestawie danych Concrete\_Data.csv zostały porównane prędkości uczenia sieci. Tak jak w rozdziale Wpływ optymalizacji na wydajność kerneli, sieć ta zawiera 2 warstwy ukryte o tej samej ilości neuronów. w następującej tabeli zostały umieszczone obliczone czasy w milisekundach 100 iteracji uczenia tej sieci, z podziałem na czas uruchomienia na procesorze CPU oraz GPU, oraz na platformę (Win32 oraz Win64):

|  |
| --- |
| Tabela . Porównanie czasu uczenia sieci na platformach 32 i 64-bitowych |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Ilość neuronów* | *Testów w iteracji* | *Czas uczenia sieci na CPU w ms.* | | *Czas uczenia sieci na GPU w ms.* | |
| *x86* | *x64* | *x86* | *x64* |
| 32 | 1 | 7 | 10 | 142 | 233 |
| 64 | 1 | 19 | 10 | 123 | 114 |
| 128 | 1 | 63 | 46 | 187 | 234 |
| 256 | 1 | 250 | 182 | 389 | 571 |
| 510 | 1 | 1086 | 769 | 1158 | 452 |
| 32 | 16 | 98 | 88 | 116 | 99 |
| 64 | 16 | 289 | 223 | 157 | 145 |
| 128 | 16 | 997 | 774 | 238 | 275 |
| 256 | 16 | 4099 | 2984 | 650 | 1035 |
| 510 | 16 | 19835 | 12470 | 2498 | 764[[29]](#footnote-29) |

Łatwo jest zauważyć zależność, że czas uczenia na CPU jest proporcjonalny do kwadratu ilości neuronów i do ilości testów w iteracji, natomiast przy uczeniu na GPU dużo trudniej znaleźć taką zależność. Zwiększenie ilości testów w iteracji nie zwiększa w dużym stopniu czasu działania algorytmu[[30]](#footnote-30), dzięki czemu w przypadku dużej ilości neuronów w obu warstwach ukrytych i dużej ilości testów w jednej iteracji, uczenie na GPU jest kilka razy szybsze niż wykonane na CPU.

Na platformie 64-bitowej można zauważyć przyspieszenie 20-30% w wykonywaniu operacji na CPU oraz spowolnienie na GPU. Przyspieszenie na CPU może być tłumaczone tym, że program jest wykonywany na procesorze z architekturą 64-bitową, dzięki czemu operacje na 64-bitowych liczbach zmiennoprzecinkowych są wykonywane szybciej. Spowolnienie w wykonaniu programu na GPU może być spowodowane nieoptymalnie napisanym sterownikiem graficznym na tę platformę.

### Informacje końcowe

Napisz o najnowszych procesorach i kartach graf. – ile maja razy wiecej mocy niż moje.

// (Wytłumaczenie, czemu są inne wyniki CPU i GPU - są różnice, bo float zamiast double i inna kolejność obliczeń). (może we wcześniejszym rozdziale)

// Kod na CPU nie był zoptymalizowany (vectory są wolne)

## Zewnętrzne biblioteki

W implementacji biblioteki CNL użyłem dwóch różnych bibliotek zewnętrznych – Mersenne Twister i TinyXML.

### Mersenne Twister

Generator liczb pseudolosowych wbudowany w język C++, rand(), jest bardzo niskiej jakości – działa wolno, a wylosowane liczby nie spełniają wielu testów na losowość. Przez to badania wykonane z tym generatorem mogą nie być wiarygodne. Dla zadań wykonywanych w tej bibliotece, bardzo dobrym rozwiązaniem jest generator Mersenne Twister, który jest polecany do programów ogólnego użytku. Wybrałem bibliotekę obiektową stworzoną przez Richarda J. Wagnera, opisaną w (Wagner, 2009).

### TinyXML

Jako format przechowywania danych sieci i zestawów testów, wybrałem XML. Zależało mi na znalezieniu biblioteki do obsługi XML, która przedstawia dokument w postaci struktury obiektów (Document Object Model, DOM, opisany w (W3C, 2009)), oraz jest prosta w obsłudze. TinyXML to jedna z najpopularniejszych darmowych bibliotek dla C++, opisana w (Thomason).

## Narzędzia pomocnicze

NVIDIA dostarcza razem z biblioteką CUDA aplikację **CUDA Visual Profiler**. Jest to tzw. profiler dynamiczny kerneli CUDA. Program ten uruchamia aplikację CUDA i określa różne parametry kerneli i operacji na pamięci GPU, m.in.:

* Czas wykonania przez CPU, GPU
* Ilość danych zapisanych lub odczytanych z pamięci globalnej, z podziałem na operacje złączone (coalesced) i nie złączone
* Ilość danych zapisanych lub odczytanych z pamięci dzielonej
* Ilość wykonanych instrukcji

Poniżej jest zrzut ekranu programu:

|  |
| --- |
|  |
| Rysunek 17. Program CUDA Visual Profiler  Źródło: Zrzut ekranu |

Pole 1 przedstawia listę poprzednich symulacji programu. Pole 2 zawiera listę wywołań kerneli oraz operacji na pamięci GPU, razem ze wszystkimi ich właściwościami. Pole 3 jest to okno logowania zawierające pomocnicze dane.

Program udostępnił mi wiele informacji o programie CNL, które były mi bardzo pomocne w procesie optymalizacji kerneli oraz pozwoliły na przyspieszenie działania części GPU programu.

## Możliwości rozwoju programu

**// popraw to…**

Biblioteka CNL mogłaby być poprawiona na wiele sposobów. do lepszej oceny wiarygodności uczenia, możnaby dzielić zestaw testów na zestaw uczący oraz testujący (weryfikujący). Inną zmianą byłaby możliwość zmiany współczynnika uczenia w trakcie procesu uczenia (często daje to lepsze wyniki uczenia niż jedna wartość). Jak napisano wcześniej, biblioteka została zaprojektowana tak, żeby umożliwić łatwe dodanie nowych typów sieci neuronowych – byłoby oczywiście możliwe dodanie sieci RBF lub Hopfielda do istniejącego kodu.

Aktualnie istnieją ograniczenia związane z ograniczeniami GPU (np. takie, że maksymalna ilość neuronów w warstwie to 511). Możnaby je usunąć, ale spowodowałoby to spadek wydajności (a poza tym, chyba te ograniczenia są rzadko przekraczane).

Zmiana w kernelach z operatora mnożenia liczb całkowitych na użycie funkcji \_\_[u]mul24 zwiększyłaby trochę prędkość ich działania[[31]](#footnote-31). Aktualnie brakuje obsługi kilku równolegle działających GPU.

// Teraz widze, ze jak w kernelu jest wiele galezi, to jest prawie tak zle jak czytanie z globalnej pamieci

# Wnioski

**// popraw**

Jak wspomniano w niniejszej pracy, GPU bardzo dobrze dają się zastosować w przypadkach algorytmów dających się łatwo zrównoleglić. Uruchamianie oraz uczenie sieci neuronowych jest takim właśnie zadaniem, a wewnątrz tej pracy zostało udowodnione, że prędkość działania takiej aplikacji uruchomionej na GPU może być większa niż na CPU. To szybkie działanie zostało osiągnięte dzięki kolejnym optymalizacjom, coraz bardziej przyspieszającym algorytm. Problemem jest to, ze nie ma znanej ilości danych – w wielu przykładach rozmiary są znane.

Problemem jest to, że przy deklaracjach pamięci i kopiowaniu danych RAM <-> pamięć graficzna, trzeba dużo pisać...

Gdybym teraz miał za zadanie zrobić ten program jeszcze raz, zrobiłbym trochę prostszy, bez tak wielu instrukcji skoku – bo one też zwalniają program, nie tylko dostępy do globalnej pamięci. Widzę, że żeby napisać program w CUDA, nie jest to trudne. Dużo większy kłopot to napisanie tak, żeby działało szybko – jest kilka typów pamięci do wykorzystaniua, a optymalizację często trzeba robić w kilku krokach. Widać, że w przypadku sieci neuronowych użycie float nie pogarsza jakości wynikow programu.

Co jest ważne, przyspieszenie na GPU jest duże tylko przy dużych danych wejściowych, gdzie można wykorzystać moc GPU.

Wg. mnie wybrałem dobrą bibliotekę, zawsze mogłem łatwo znaleźć rozwiązanie moich problemów. Można pisać w Wysokopoziomowy języku, łatwo też się profiluje (łatwe do zoptymalizowania). Pomagała dokumentacja i forum NVIDIA.

Wg. mnie GPGPU ma przyszłość – Fermi i inne technologie pokazują, że NVIDIA stawia na to dużą wagę (opisz dokładniej).

# Zakończenie

**Twarde spacje „z tym” albo „i coś tam” – Shift + spacja.**

# Bibliografia

AMD. (brak daty). *ATI Stream Technology.* Pobrano z lokalizacji AMD: http://www.amd.com/US/PRODUCTS/TECHNOLOGIES/STREAM-TECHNOLOGY/Pages/stream-technology.aspx

Asuncion, A., & Newman, D. (2010). *UCI Machine Learning Repository.* Pobrano z lokalizacji Irvine, CA: University of California, School of Information and Computer Science: http://mlr.cs.umass.edu/ml

Baggio, D. L. (2007). *GPGPU Based Image Segmentation Livewire Algorithm Implementation.* Pobrano z lokalizacji GPGPU.org: http://gpgpu.org/2008/04/01/gpgpu-based-image-segmentation-livewire-algorithm-implementation

Berillo, A. (2008, Październik 21). *NVIDIA CUDA - Non-graphic computing with graphics processors.* Pobrano z lokalizacji iXBT Labs: http://ixbtlabs.com/articles3/video/cuda-1-p1.html

Breitbart, J. (2008, Sierpień 7). *Case studies on GPU usage and data structure design.* Pobrano z lokalizacji Geeks3D: http://www.geeks3d.com/downloads/200808/Jens\_Breitbart\_thesis.pdf

Fung, J., & Murray, T. (2008, Grudzień). *Building CUDA Photoshop Filters for the GPU.* Pobrano z lokalizacji NVIDIA: http://developer.download.nvidia.com/compute/cuda/Photoshop/CUDAFilters4.pdf

GPGPU.org. (brak daty). *About GPGPU.org.* Pobrano z lokalizacji GPGPU.org: http://gpgpu.org/about

Harris, M., & Buck, I. (2005, Kwiecień). *GPU Gems 2, Chapter 34. GPU Flow-Control Idioms.* Pobrano z lokalizacji NVIDIA Developer Zone: http://http.developer.nvidia.com/GPUGems2/gpugems2\_chapter34.html

Harvey, J. P. (2009, Wrzesień). *GPU acceleration of object classification algorithms using NVIDIA CUDA.* Pobrano z lokalizacji Digital media library - RIT Scholars: https://ritdml.rit.edu/handle/1850/10894

Jorden. (2010, Kwiecień 17). *FAQ: Nvidia CUDA & ATI Stream (CAL) FAQ.* Pobrano z lokalizacji BOINC: http://boincfaq.mundayweb.com/index.php?language=1&view=471

Lavignotte, F. (2010, Marzec 23). *CUDA Toolkit 3.0 released.* Pobrano z lokalizacji NVIDIA Forums: http://forums.nvidia.com/index.php?showtopic=163511&st=20&p=1025290&#entry1025290

Mattb3. (2010, Kwiecień 14). *Poor cgemm performance with cuda 3.0.* Pobrano z lokalizacji NVIDIA Forums: http://forums.nvidia.com/index.php?showtopic=166184&st=0&p=1039424&#entry1039424

NVIDIA. (2010, Kwiecień 21). *CUDA 3.0 Downloads.* Pobrano z lokalizacji NVIDIA: http://developer.nvidia.com/object/cuda\_3\_0\_downloads.html

NVIDIA. (brak daty). *CUDA GPU Computing.* Pobrano z lokalizacji NVIDIA Forums: http://forums.nvidia.com/index.php?showforum=62

NVIDIA. (brak daty). *CUDA-Accelerated Applications.* Pobrano z lokalizacji NVIDIA: http://www.nvidia.com/object/cuda\_app\_tesla.html

NVIDIA. (brak daty). *GeForce 256.* Pobrano z lokalizacji NVIDIA: http://www.nvidia.com/page/geforce256.html

NVIDIA. (2010, Luty 8). *MATLAB plug-in for CUDA.* Pobrano z lokalizacji NVIDIA Developer Web Site: http://developer.nvidia.com/object/matlab\_cuda.html

NVIDIA. (2009, Sierpień 26). *NVIDIA CUDA Programming Guide 2.3.* Pobrano z lokalizacji NVIDIA: http://developer.download.nvidia.com/compute/cuda/2\_3/toolkit/docs/NVIDIA\_CUDA\_Programming\_Guide\_2.3.pdf

NVIDIA. (2010, Luty 20). *NVIDIA CUDA Programming Guide 3.0.* Pobrano z lokalizacji NVIDIA: http://developer.download.nvidia.com/compute/cuda/3\_0/toolkit/docs/NVIDIA\_CUDA\_ProgrammingGuide.pdf

NVIDIA. (brak daty). *NVIDIA PhysX.* Pobrano z lokalizacji NVIDIA Polska: http://www.nvidia.pl/object/nvidia\_physx\_pl.html

NVIDIA. (2009). *NVIDIA’s Next Generation CUDA Compute Architecture: Fermi, v1.1.* Pobrano z lokalizacji NVIDIA: http://www.nvidia.com/content/PDF/fermi\_white\_papers/NVIDIA\_Fermi\_Compute\_Architecture\_Whitepaper.pdf

NVIDIA. (brak daty). *Supercomputing at 1/10th the Cost.* Pobrano z lokalizacji NVIDIA: http://www.nvidia.com/object/tesla\_computing\_solutions.html

NVIDIA. (brak daty). *What is CUDA?* Pobrano z lokalizacji NVIDIA: http://www.nvidia.com/object/what\_is\_cuda\_new.html

Pfeiffer, J. (2005, Luty 23). *CS473 - Pipeline Hazards.* Pobrano z lokalizacji Joe Pfeiffer - Home Page: http://www.cs.nmsu.edu/~pfeiffer/classes/473/notes/hazards.html

Sandhu, T. (2010, Marzec 26). *NVIDIA's GeForce GTX 480 finally unleashed. Reviewed and rated.* Pobrano z lokalizacji HEXUS.net: http://www.hexus.net/content/item.php?item=24000&page=3

Spek, J. v. (2008, Kwiecień 1). *The CUDA Compiler Driver - NVCC.* Pobrano z lokalizacji Frank Mueller Homepage: http://moss.csc.ncsu.edu/~mueller/cluster/nvidia/2.0/nvcc\_2.0.pdf

The Khronos Group. (brak daty). *OpenCL - The open standard for parallel programming of heterogeneous systems.* Pobrano z lokalizacji The Knronos Group: http://www.khronos.org/opencl/

Thomason, L. (brak daty). *Lee Thomason's Homepage .* Pobrano z lokalizacji TinyXML Main Page: http://sourceforge.net/projects/tinyxml

Triolet, D. (2007, Marzec 21). *Nvidia CUDA: preview.* Pobrano z lokalizacji BeHardware: http://www.behardware.com/articles/659-1/nvidia-cuda-preview.html

W3C. (2009, Styczeń 6). *W3C Document Object Model.* Pobrano z lokalizacji World Wide Web Consortium (W3C) Website: http://www.w3.org/DOM

Wagner, R. J. (2009, Wrzesień 28). *Mersenne Twister Random Number Generator.* Pobrano z lokalizacji Creations of Rick Wagner: http://www-personal.umich.edu/~wagnerr/MersenneTwister.html

*Wikipedia.* (brak daty). Pobrano z lokalizacji Video card: http://en.wikipedia.org/wiki/Graphic\_card

Wikipedia. (brak daty). *GPGPU.* Pobrano z lokalizacji Wikipedia: http://en.wikipedia.org/wiki/GPGPU

Wilk, A. (2009, Październik 5). *Mity i fakty: komputery bez procesora.* Pobrano z lokalizacji Trochetechniki.pl: http://www.trochetechniki.pl/Mity-i-fakty%3A-komputery-bez-procesora,t,929,page,1.html

# Załączniki

## Kod CPU i GPU?

(poniżej to był podpunkt 2.3.7)

### Przykład kodu

Kod wykonywany równolegle na GPU znajduje się zawsze w funkcjach zwanych kernelami. Są natomiast dwa sposoby Przykładowy kernel i wywolanie go.

Cuda ma tryb driver i ten inny, który używam

// przeklej kody vectorAdd.cu i vectorAddDrv.cu

// Kernel definition

\_\_global\_\_ void VecAdd(float\* A, float\* B, float\* C)

{

int i = threadIdx.x;

C[i] = A[i] + B[i];

}

int main()

{

// Kernel invocation with N threads

VecAdd<<<1, N>>>(A, B, C);

}

Zrodlo: (NVIDIA, 2010, str. 7)

1. Wg. tzw. Prawa Moore’a ilość tranzystorów w procesorach podwaja się co 2 lata. Podobnie jest z mocą obliczeniową komputerów. [↑](#footnote-ref-1)
2. Technologia CUDA powstała w 2006 roku – zobacz [↑](#footnote-ref-2)
3. Aktualnie (kwiecień 2010), najbardziej wydajne karty graficzne mają kilkukrotnie wyższą moc obliczeniową od najbardziej wydajnych procesorów (daj linka) [↑](#footnote-ref-3)
4. Operacje wykonywane na procesorze CPU (w odróżnieniu od tych wykonywanych na GPU) są też nazywane operacjami wykonywanymi na hoście. [↑](#footnote-ref-4)
5. Multi Layer Perceptron – jeden z najpopularniejszych typów sieci neuronowych. Został opisany w rozdziale X.X [↑](#footnote-ref-5)
6. General Purpose GPU – użycie GPU do ogólnych zastosowań [↑](#footnote-ref-6)
7. do tej pory operacje 3D musiały być obliczane przez główny procesor komputera. [↑](#footnote-ref-7)
8. Floating Point Operations Per Second *– liczba operacji zmiennoprzecinkowych na sekundę.* [↑](#footnote-ref-8)
9. *Arithmetic Logic Unit,* **Jednostka arytmetyczno-logiczna, układ wykonujący operacje na liczbach całkowitych** [↑](#footnote-ref-9)
10. *Floating Point Unit,* **jednostka zmiennoprzecinkowa, układ wykonujący operacje na liczbach zmiennoprzecinkowych** [↑](#footnote-ref-10)
11. Nie ma konieczności używania wielu instrukcji sterujących for, do, while [↑](#footnote-ref-11)
12. ang. Streaming Multiprocessor (SM) [↑](#footnote-ref-12)
13. Ang. Scalar Processor (SP) [↑](#footnote-ref-13)
14. Opisane w rozdziale Wersje CUDA [↑](#footnote-ref-14)
15. Technologia Hardware T&L - Tranform and Lighting [↑](#footnote-ref-15)
16. Software Development Kit [↑](#footnote-ref-16)
17. Przykładowy kernel oraz jego wywołanie zostało zamieszczone w późniejszym rozdziale [↑](#footnote-ref-17)
18. Więcej na temat składni programów CUDA jest w rozdziale X.X [↑](#footnote-ref-18)
19. Do uruchomienia kernela lub skopiowania pamięci między hostem a pamięcią graficzną jest wymagane więcej instrukcji niż w przypadku CUDA [↑](#footnote-ref-19)
20. w nowej architekturze Fermi nie ma już takich różnic [↑](#footnote-ref-20)
21. Opisane są one dokładniej w rozdziale „Ograniczenia” [↑](#footnote-ref-21)
22. Wymagałoby to zakomentowania linii „#define REAL\_GPU\_IS\_FLOAT 1” wewnątrz pliku stdafx.h oraz zmienienia docelowej architektury pliku TrainNetwork.cu na sm\_13 [↑](#footnote-ref-22)
23. Szerzej jest ona opisana w rozdziale [Komunikacja między wątkami](#_Komunikacja_między_wątkami) [↑](#footnote-ref-23)
24. w przypadku 65535 testów, przy dwóch kolejnych warstwach zawierających 511 neuronów, ilość zadeklarowanej pamięci graficznej wynosi nieco powyżej 65535 \* 511 \* 4 B = 128 MB [↑](#footnote-ref-24)
25. Duże sieci neuronowe, zestawy testów z dużą ilością wejść i wyjść, duża ilość testów użytych w trenowaniu w każdej iteracji [↑](#footnote-ref-25)
26. Wszystkie testy przy których nie zaznaczono inaczej, były uruchamiane w systemie Vista 32bit. [↑](#footnote-ref-26)
27. w marcu 2010 udostępniono framework CUDA w wersji 3.0, jednak nie jest ona dopracowana i w wielu sytuacjach programy w nim uruchomione działają wolniej niż w wersji 2.3 – więcej informacji w  i  [↑](#footnote-ref-27)
28. Prawdopodobnie jest to spowodowane tym, że przy 510 neuronach, dostęp do pamięci globalnej był częściej złączaony (coalesced). [↑](#footnote-ref-28)
29. Wyniki w tym przypadku są błędne. Może to być spowodowane błędnie działającą biblioteką CUDA Toolkit lub sterownikiem graficznym. [↑](#footnote-ref-29)
30. Jest to związane z tym, że różne testy w iteracji mogą być jednocześnie przetwarzane na różnych mikroprocesorach GPU. [↑](#footnote-ref-30)
31. Oprócz platform Fermi – tam byłoby to wolniejsze – zobacz [↑](#footnote-ref-31)