Perceptron na GPU

B4M39GPU 2019/2020 zimní semestr

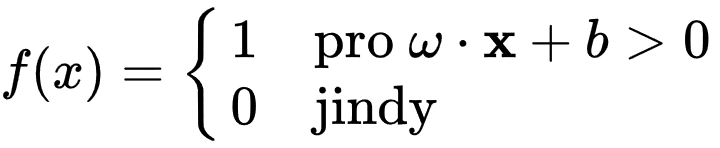
Jan Skala

# Abstrakt

Semestrální práce se zaměřuje na paralelní implementaci Perceptronu. Zahrnuje implementaci trénování a klasifikace dat. Obsahuje naivní CPU implementaci a paralelní GPU implementaci. Cílem je zjistit, zda se dosáhne, a pokud ano tak jakého, zrychlení výpočtu při použití grafické karty a paralelní implementace.

# Popis algoritmu

Perceptron je nejjednodušší model dopředné neuronové sítě. Sestává se pouze z jednoho neuronu a byl vynalezen Frankem Rosenblattem v roce 1957. Jedná se o binární klasifikátor, který mapuje vektor sloupců x na výstupní hodnoty f(x).



Rovnice 1 Klasifikace vektoru vstupů x

Kde **ω** je vektor vah a **b** je konstanta(někdy se jí také říká bias nebo práh) a **x** je vektor vstupních dat. Pro jednodušší implementaci se místo 0 vrací -1.

Hlavní nedostatek perceptronu spočívá v tom, že aby správně fungoval musí být data lineární separovatelná. Někdy lze tento nedostatek řešit částečnou úpravou vstupních dat. Máme-li dvourozměrná data a je patrná že existuje kružnice, která by data oddělila, můžeme data převést do polárních souřadnic. Dalším rozšířením je vícevrstevný perceptron, který má již mnohem širší možnosti použití. (Mařík, 1993)

## Trénování

V této práci bylo využito učení s učitelem. Síti byla nejdříve prezentována tréninková množina, která odpovídá správnému chování. Trénování probíhá v iteracích. Jejich počet můžeme stanovit. V každé iteraci je aktualizován vektor vah a bias. Zároveň pracujeme s konstantou **r**,která značí learning rate (rychlost učení, ale typicky se tento pojem nepřekládá).Dosazením menších hodnot docílíme pomalejšího učení, tzv fine tuning. Dosazením větších hodnot zrychlíme proces učení, ale riskujeme, že optimum nespočítáme správně. Hodnoty jsou v rozmezí od 0 do 1.

Nejprve je zvolen náhodný vektor vah (může se jednat o nulový vektor). Každý vektor vstupních dat klasifikujeme a porovnáme s očekávaným výsledkem. Pokud se výstupy neshodují provedeme upravíme váhy a bias. Učící algoritmus u perceptronu je konečný, pouze pokud řešení existuje, a navíc pokud je učící množina lineárně separovatelná.

Váhy perceptronu a bias jsou adaptovány dle následujícího pravidla:

**p** značí learning rate  
**a** značí očekávanou klasifikaci  
**b** značí **bias  
Dj**značí řádek trénovacích dat  
**wi** značí vektor vah  
**f(x)** značí skokovou přenosovou funkci  
Kde Wi+1 značí nový vektor vah

Skoková přenosová funkce

Skoková přenosová funkce vrací pro vstup menší než daná mez -1, pro větší vrací 1.

Klasifikace

Klasifikace spočívá ve vynásobení vektoru s daty (D) vektorem vah(W) skalárně. Následně se přičte bias (b). Výsledek se použije jako argument skokové funkce (f).

# Struktura projektu

* Data – složka obsahuje trénovací a testovací Datasety
* CL11.hpp – Header only knihovna pro zpracovávání vstupních argumentů
* CSV.cpp/h – pomocná třída, slouží k načítání CSV dat do struktury std::vector
* Debug.h – obsahuje skupinu funkcí pro debugování – vypisování dat z GPU do konzole
* default.h – obsahuje definici assert makra pro spouštění CUDA funkcí
* kernels.cu – implementace všech kernelů
* kernels.cuh – hlavičky a implementace spouštěcí funkcí pro kernely
* main.cpp – hlavní vstup programu – načítání argumentů a spuštění perceptronu
* Perceptron.cpp/h – Hlavní třída implementující funkcionalitu perceptronu

## Datasety

Všechna data byla převedena do CSV pro snazší načítání. K trénování a testování byly použity následující Datasety:

* Iris
* Minst číslice 1
* Mnist číslice 2
* Mnist číslice 3

# CPU Implementace

CPU implementace je realizovaná ve třídě Perceptron. Konstruktor přijímá na vstupu 3 parametry:

* Počet iterací trénování
* Learning Rate
* Slovní výstup – true pokud požadujeme logování zpět do konzole

## Trénování

Metoda **fit** ve třídě Perceptron. Jako parametry přijímá dva vektory jeden s daty a druhý s očekávanými klasifikacemi. Začíná s nulovými váhami. Celá metoda je implementována sekvenčně není použita žádná paralelizace, ať už v podobě OpenMP nebo SIMD. Vypočítané váhy a bias si uloží do privátních polí.

## Klasifikace

Metoda **predict** ve třídě Perceptron. Skalárně vynásobí vektor vstupních dat s vektorem vah, následné přičte bias a na výsledek aplikuje skokovou přenosovou funkci.

# GPU Implementace

GPU implementace je realizovaná ve třídě Perceptron. Konstruktor přijímá na vstupu 3 parametry:

* Počet iterací trénování
* Learning Rate
* Slovní výstup – true pokud požadujeme logování zpět do konzole

Zároveň využívá CUDA kernely a jejich pomocné a spouštěcí funkce.

# Použité Kernely

Ke každému kernelu je definovaná pomocná funkce, která provádí alokaci a volá kernel. Třída Perceptron tak nemusí volat kernel přímo, nemusí provádět alokaci na GPU a vývojové prostředí Visual Studio nehází chybová hlášení o nevalidní syntaxi při používání streamové notace <<<,>>>.

## Skalární součin

Skalární součin je jedna z esenciálních operací, která je použita jak při trénování, tak při klasifikaci. Implementoval jsem jej jako device funkci, kterou mohu volat z více kernelů. Nejprve vynásobím příslušné složky obou vektorů a uložím data do sdílené paměti. Následně data se sdílené paměti sečtu pomocí paralelní redukční sumy. Tuto funkci volají dva kernely. K\_dot, který dělá obyčejný skalární součin. A k\_update, který po vektorovém součinu, ještě počítá chybu klasifikace a úpravu hodnotu proměnné bias.

## Update

Nejprve skalárně vynásobí vstupní data a váhy (device funkce d\_dot). K výsledku přičte bias a aplikuje skokovou přenosovou funkci. Výstupem je proměnná **u** (update), která je použita v dalším kernelu. Bohužel nelze paralelizovat na úrovni dat, protože pro trénování na datech di se využívají výsledky z trénování dat di-1.

Kde u značí proměnnou **u** (update)  
**p** značí learning rate  
**a** značí očekávanou klasifikaci  
**b** značí **bias\_gpu  
Dj**značí řádek trénovacích dat  
**w** značí vektor vah  
**f(x)** značí skokovou přenosovou funkci

## Násobení skalárem

Nyní zbývá korekce vah. Vektor s daty vynásobím skalárem **u**, který jsem spočítal v předchozím kernelu. Výsledek uložím jako nový vektor vah. Následně překopíruji skalár **u** do proměnné **bias\_gpu**.

Kde Wi+1 značí nový vektor vah  
Dj značí řádek s daty, pomocí kterých se provádí trénink

## Klasifikace

Vzhledem k tomu, že dat ke klasifikaci je typicky více než rozměr vah. Například dataset mnist potřebuje 784 vah, zatímco dat je více než 6 tisíc. V tomto kernelu dává smysl paralelizovat na úrovni dat. Každé vlákno tedy provede skalární součin řádku dat a vektoru vah. Následně přičte bias a spočítá hodnotu skokové přenosové funkce.

# Naměřené výsledky

Veškeré časy jsou uvedeny v sekundách.

## Test 1

|  |  |  |  |
| --- | --- | --- | --- |
| CPU - Intel Core i5 6300HQ | | | |
| Trénovací dataset | Testovací dataset | Doba trénování | Doba klasifikace |
| iris\_train.csv | iris\_test.csv | 0.0004642 | 1.23e-05 |
| iris\_test.csv | iris\_train.csv | 0.0001206 | 4.72e-05 |
| mnist\_train1.csv | mnist\_test1 | 3.96954 | 0.0560375 |
| mnist\_test1 | mnist\_train1.csv | 0.658736 | 0.33868 |
| mnist\_train2.csv | mnist\_test2.csv | 3.49075 | 0.055833 |
| mnist\_test2.csv | mnist\_train2.csv | 0.599298 | 0.292574 |
| mnist\_train3.csv | mnist\_test3.csv | 3.55048 | 0.0580637 |
| mnist\_test3.csv | mnist\_train3.csv | 0.591221 | 0.30822 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| GPU 1 NVIDIA GeForce GTX 950M | | | | | |
| Trénovací dataset | Testovací dataset | Doba trénování | Doba trénování a kopírování | Doba klasifikace | Doba klasifikace a kopírování |
| iris\_train.csv | iris\_test.csv | 0.0741891 | 0.0827714 | 8.58e-05 | 0.0015304 |
| iris\_test.csv | iris\_train.csv | 0.020647 | 0.0232225 | 8.01e-05 | 0.0055496 |
| mnist\_train1.csv | mnist\_test1 | 4.18036 | 4.49103 | 0.0011362 | 0.048448 |
| mnist\_test1 | mnist\_train1.csv | 0.714684 | 0.769929 | 0.0070984 | 0.292959 |
| mnist\_train2.csv | mnist\_test2.csv | 3.73057 | 4.0088 | 0.0011439 | 0.0452453 |
| mnist\_test2.csv | mnist\_train2.csv | 0.661603 | 0.717859 | 0.0043072 | 0.306409 |
| mnist\_train3.csv | mnist\_test3.csv | 3.81819 | 4.12242 | 0.001128 | 0.0513694 |
| mnist\_test3.csv | mnist\_train3.csv | 0.641257 | 0.690492 | 0.0046894 | 0.307401 |

## Test 2

|  |  |  |  |
| --- | --- | --- | --- |
| CPU - Intel(R) Core(TM) i7-8700 CPU | | | |
| Trénovací dataset | Testovací dataset | Doba trénování | Doba klasifikace |
| iris\_train.csv | iris\_test.csv | 7.035e-06 | 3.02e-07 |
| iris\_test.csv | iris\_train.csv | 2.459e-06 | 6.88e-07 |
| mnist\_train1.csv | mnist\_test1 | 0.0291863 | 0.000887468 |
| mnist\_test1 | mnist\_train1.csv | 0.00464948 | 0.00501863 |
| mnist\_train2.csv | mnist\_test2.csv | 0.025638 | 0.000776762 |
| mnist\_test2.csv | mnist\_train2.csv | 0.00427623 | 0.00440546 |
| mnist\_train3.csv | mnist\_test3.csv | 0.0263798 | 0.000749905 |
| mnist\_test3.csv | mnist\_train3.csv | 0.00415912 | 0.00453386 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| GPU 2 NVIDIA GeForce RTX 2060 | | | | | |
| Trénovací dataset | Testovací dataset | Doba trénování | Doba trénování a kopírování | Doba klasifikace | Doba klasifikace a kopírování |
| iris\_train.csv | iris\_test.csv | 0.00908131 | 0.00950981 | 1.0818e-05 | 0.00012986 |
| iris\_test.csv | iris\_train.csv | 0.00236556 | 0.00248393 | 1.1334e-05 | 0.000449044 |
| mnist\_train1.csv | mnist\_test1 | 0.519207 | 0.545449 | 0.00014074 | 0.0045435 |
| mnist\_test1 | mnist\_train1.csv | 0.0836026 | 0.088188 | 0.000228141 | 0.0263691 |
| mnist\_train2.csv | mnist\_test2.csv | 0.46322 | 0.486196 | 0.000135957 | 0.00419546 |
| mnist\_test2.csv | mnist\_train2.csv | 0.075468 | 0.0796296 | 0.000241631 | 0.0270068 |
| mnist\_train3.csv | mnist\_test3.csv | 0.476223 | 0.499672 | 0.000136311 | 0.00397208 |
| mnist\_test3.csv | mnist\_train3.csv | 0.0741683 | 0.0781258 | 0.000240048 | 0.0233982 |

# Závěr

Z naměřených dat vyplývá, že trénování se paralelizací vůbec nezrychlilo. Započítáme-li i čas kopírování na paměť GPU výsledky nejsou vůbec příznivé. Při testování výkonosti klasifikace dat pozorujeme mírné zlepšení, nicméně pokud přičteme i čas kopírování opět je CPU výkonnější.

Nízká efektivita GPU implementace při fázi trénování je primárně způsobena vektory malé dimenze. Zároveň při trénování se musí data načítat postupně. Dataset mnist obsahuje vektory dimenze 784 oproti datasetu iris. Lze pozorovat, že efektivita GPU implementace se zvyšuje se zvyšujícími se rozměry vstupních vektorů.

Při klasifikaci, zejména velkých datasetů mnist\_train1/2/3.csv lze pozorovat nárust efektivity GPU implementace. Nicméně pokud uvažujeme i čas kopírování, CPU implementace je opět efektivnější. Efektivnějšího kopírování na paměť GPU lze docílit implementací návrhového paradigmatu Producer Consumer. Na paměť GPU by se nakopírovala část dat, GPU by je začalo zpracovávat a současní s tím by se na jiný úsek paměti nakopírovala další část dat. Jakmile by GPU první část zpracovalo, pokračovalo by ve zpracování dalších částí. Nealokovala by se nová paměť, ale pouze by došlo k prohození ukazatelů.

# Reference

Kelleher, C. (nedatováno). *curran.* Načteno z Github: https://gist.github.com/curran/a08a1080b88344b0c8a7

Mařík, V. (1993). *Umělá inteligence I.* Praha: Academia.

Redmon, J. (nedatováno). *projects/mnist-in-csv*. Načteno z Mnist dataset CSV: https://pjreddie.com/projects/mnist-in-csv/