VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV TELEKOMUNIKACÍ
DEPARTMENT OF TELECOMMUNICATIONS

HLUBOKÉ NEURONOVÉ SÍTĚ PRO PROSTŘEDÍ SUPERPOČÍTAČE
DEEP NEURAL NETWORK FOR SUPERCOMPUTER ENVIRONMENTS

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE
AUTHOR
Bc. Samuel Bronda

VEDOUCÍ PRÁCE
SUPERVISOR
doc. Ing. Radim Burget, Ph.D.

BRNO 2019

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Diplomová práce

magisterský navazující studijní obor **Telekomunikační a informační technika**
Ústav telekomunikací

Student: Bc. Samuel Bronda ***ID:*** 173620
Ročník: 2 ***Akademický rok:*** 2018/19

NÁZEV TÉMATU:

Hluboké neuronové sítě pro prostředí superpočítače

POKYNY PRO VYPRACOVÁNÍ:

Seznamte se s problematikou hlubokých konvolučních neuronových sítí a hardwarovými výpočetními kartami dostupnými na trhu. Navrhněte sadu výkonnostních zátěžových testů, které spustíte a změřte výkonnost, elektrickou spotřebu a paměťové požadavky. Identifikujte nejvýhodnější varianty z pohledu výkonu a poměru výkon/cena. Výsledky vhodně komentujte a vyneste do výsledného grafu.

DOPORUČENÁ LITERATURA:

[1] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

[2] Karpathy, Andrej, et al. "Large-scale video classification with convolutional neural networks." Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2014.

Termín zadání: 1.2.2019 ***Termín odevzdání:*** 16.5.2019

Vedoucí práce: doc. Ing. Radim Burget, Ph.D.
Konzultant:

prof. Ing. Jiří Mišurec, CSc.
předseda oborové rady

ABSTRAKT

Hlavným prínosom práce je optimalizácia hardvérovej konfigurácie pre výpočet neurónových sieti. Teoretická časť popisuje neurónové siete, frameworky hlbokého učenia a hardvérové možnosti. Ďalšia časť práce sa venuje implementácií výkonnostných testov, ktoré zahŕňajú aplikovanie modelov Inception V3 a ResNet. Modely siete sú aplikované na rôzne grafické karty a výpočetný hardvér. Výstupom diplomovej práce je implementovaný model siete Inception V3, ktorý skúma grafické karty a ich výkon, časovú náročnosť výpočtov a ich efektivitu. Model siete ResNet je aplikovaný do časti, ktorý skúma ostatné vplyvy na výpočet neurónových sieti ako použitý disk, operačná pamäť a pod. Každá praktická časť obsahuje diskusiu, kde sú vysvetlené poznatky k danej časti. V prípade merania spotreby bol identifikovaný nesúlad medzi deklaráciou výrobcom a nameranými hodnotami.

KĽÚČOVÉ SLOVÁ

grafické karty, hardvér, Inception V3, neurónové siete, NVIDIA, ResNet

ABSTRACT

The main benefit of the work is the optimization of the hardware configuration for the calculation of neural networks. The theoretical part describes neural networks, deep learning frameworks and hardware options. The next part of the thesis deals with implementation of performance tests, which include application of Inception V3 and ResNet models. Network models are applied to various graphics cards and computing hardware. The output of the thesis is the implemented model of the network Inception V3, which examines the graphics cards and their performance, time-consuming calculations and their efficiency. The ResNet model is applied to a section that examines other impacts on neural network computing such as used disk, operating memory, and so on. Each practical part contains a discussion where the knowledge of the given part is explained. In the case of consumption measurement, a mismatch between the declaration by the manufacturer and the measured values was identified.

KEYWORDS

graphics cards, hardware, Inception V3, neural networks, NVIDIA, ResNet

BRONDA, Samuel. *Hlboké neuronové siete pre prostredie superpočítača*. Brno, 2019, 70 s. Diplomová práca. Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav telekomunikací. Vedúci práce: doc. Ing. Radim Burget, Ph.D.

VYHLÁSENIE

Vyhlasujem, že som svoju diplomovú prácu na tému „Hlboké neuronové siete pre prostredie superpočítača" vypracoval samostatne pod vedením vedúceho diplomovej práce, využitím odbornej literatúry a ďalších informačných zdrojov, ktoré sú všetky citované v práci a uvedené v zozname literatúry na konci práce.

Ako autor uvedenej diplomovej práce ďalej vyhlasujem, že v súvislosti s vytvorením tejto diplomovej práce som neporušil autorské práva tretích osôb, najmä som nezasiahol nedovoleným spôsobom do cudzích autorských práv osobnostných a/alebo majetkových a som si plne vedomý následkov porušenia ustanovenia § 11 a nasledujúcich autorského zákona Českej republiky č. 121/2000 Sb., o práve autorskom, o právach súvisiacich s právom autorským a o zmene niektorých zákonov (autorský zákon), v znení neskorších predpisov, vrátane možných trestnoprávnych dôsledkov vyplývajúcich z ustanovenia časti druhej, hlavy VI. diel 4 Trestného zákoníka Českej republiky č. 40/2009 Sb.

Brno ....

podpis autora

POĎAKOVANIE

Rád by som poďakoval vedúcemu diplomovej práce pánovi doc. Ing. Radimovi Burgetovi, Ph.D. za odborné vedenie, konzultácie, trpezlivosť, cenný čas a podnetné návrhy k práci.

Brno ....

podpis autora

Tato práce vznikla jako součást klíčové aktivity KA6 - Individuální výuka a zapojení studentů bakalářských a magisterských studijních programů do výzkumu v rámci projektu OP VVV Vytvoření double-degree doktorského studijního programu Elektronika a informační technologie a vytvoření doktorského studijního programu Informační bezpečnost, reg. č. CZ.02.2.69/0.0/0.0/16_018/0002575.

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

Projekt je spolufinancován Evropskou unií.

Obsah

Zoznam obrázkov

Zoznam tabuliek

Úvod

Konvolučné neuronové siete sú v dnešnej dobe najviac využívané hlavne v odbore umelej inteligencie. Mnoho veľkých korporácií sa snažia byť prvý napríklad v autonómnych autách. Mnohé mestá sa snažia pomocou umelej inteligencie dosiahnuť väčšiu bezpečnosť vo verejných priestranstvách, letiskách a pod.

Nevýhodou hlbokých neurónových sieti je potrebný výkon na výpočty, a preto sa tento nedostatok rieši použitím vysoko výkonného hardvéru. Niektoré firmy investujú milióny do postavenia superpočítača v objeme niekoľko metrov štvorcových. Iné firmy sa snažia pomocou produktov, ktoré ponúkajú, dosiahnuť čo najlepší výkon v najmenšom balení.

Hlavným prínosom práce je optimalizácia hardvérovej konfigurácie pre výpočet neurónových sieti. Teoretická časť popisuje neurónové siete, frameworky hlbokého učenia a hardvérové možnosti. Ďalšia časť práce sa venuje implementácií výkonnostných testov, ktoré zahŕňajú aplikovanie modelov Inception V3 a ResNet. Modely siete sú aplikované na rôzne grafické karty a výpočetný hardvér. Výstupom diplomovej práce je implementovaný model siete Inception V3, ktorý skúma grafické karty a ich výkon, časovú náročnosť výpočtov a ich efektivitu. Model siete ResNet je aplikovaný do časti, ktorý skúma ostatné vplyvy na výpočet neurónových sieti ako použitý disk, operačná pamäť a pod. Každá praktická časť obsahuje diskusiu, kde sú vysvetlené poznatky k danej časti. V prípade merania spotreby bol identifikovaný nesúlad medzi deklaráciou výrobcom a nameranými hodnotami.

V dnešnej dobe poznáme niekoľko druhov konvolučných neurónových sieti. Aby sa jednotlivé siete nemuseli programovať, využívajú sa predprogramované frameworky, ktoré uľahčujú programátorom prácu a môžu sa tak zamerať na dôležitejšie záležitosti. Samozrejme nie je možné vyskúšať všetky možnosti a preto kvôli svojej obľúbenosti bol použitý framework Keras s backendom TensorFlow. Vybrané modely siete Inception V3 a ResNet sú v dnešnej dobe taktiež veľmi obľúbené modely a dosahujú veľké úspechy na uskutočnených súťažiach.

Hlboké konvolučné siete využívajú na výpočty hlavne grafický procesor. Väčšina programátorov siahne po najlepšom produkte,ktorý je určený pre herný priemysel. Ak majú však programátori väčší finančný balík je možné taktiež využiť grafické karty pre vedecké účely, ktoré sú však niekoľko násobne drahšie. V práci bude taktiež porovnaný výkon k cene grafickej karty. Samozrejme nie je možne v práci otestovať všetky možnosti hardvéru a preto je vybraných len niekoľko dôležitých z nich.

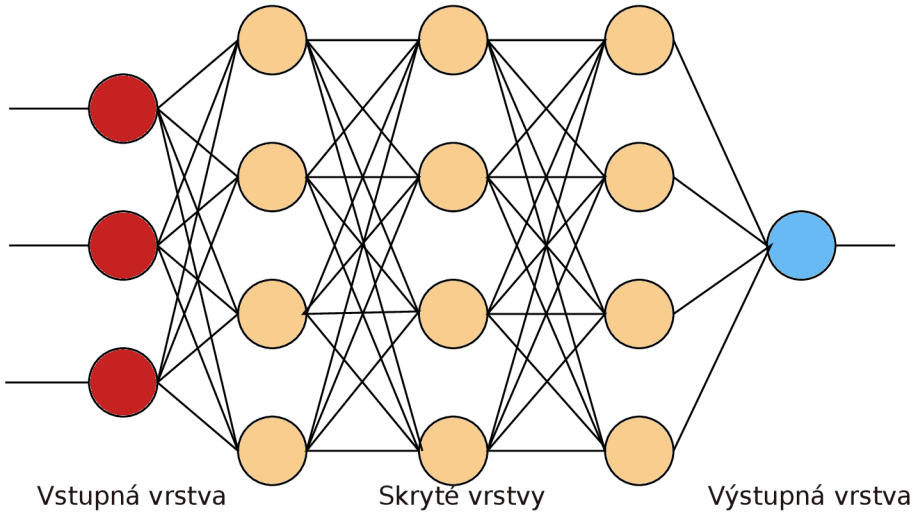
Na internete sa nájde dostatok porovnaní výkonu, žiaľ veľakrát ide o neoverené výsledky a preto záverom mojej práce budú jednoznačné a relevantné výsledky výkonu jednotlivých kariet a zariadení určených na výpočty hlbokých neurónových sietí.

1 Neurónové siete

Pojem neurónové siete pochádza z biológie človeka. Ľudský mozog obsahuje približne 80 miliard neurónov. Umelá neurónová sieť je paradigma na spracovanie informácií, ktorá je inšpirovaná spôsobom, akým biologický nervový systém, ako je mozog, spracováva informácie. Kľúčovým prvkom tejto paradigmy je nová štruktúra systému spracovania informácií. Skladá sa z veľkého množstva prepojených procesných prvkov (neurónov), ktoré pracujú spoločne na riešení špecifických problémov. V dnešnej dobe neexistuje metóda, ktorá by dokázala sledovať prirodzenú neurónovú sieť, no existujú metódy, ktoré sledujú len jednu alebo niekoľko nervových buniek [1].

1.1 Úvod do umelých neurónových sieti

Neurónová sieť je skupina neurónov, ktoré sú medzi sebou prepojené. Schopnosť spracovania informácie je uložená v internej jednotke. Neurónová sieť je typicky organizovaná do vrstiev. Jednotlivé vrstvy pozostávajú z množstva prepojených uzlov, ktoré obsahujú aktivačnú funkciu [1]. Vzory, učiace sa prvky, vstupujú do vstupnej vrstvy, ktorá komunikuje s jednou alebo viacerými skrytými vrstvami, kde sa spracovanie uskutočňuje prostredníctvom systému vážených spojení. Skryté vrstvy sa potom odkazujú na výstupnú vrstvu [2].



Obr. 1.1: Ukážka neuronovéj siete

Umelé neurónové siete (ANN) poväčšine obsahujú nejakú formu učiaceho sa pravidla, ktorá modifikuje váhy spojení podľa vstupných vzorov, s ktorými je prezentovaná. Jednoduchším spôsobom povedané, že neuronové siete sa učia rozpoznávať

určité objekty z konkrétnych príkladov objektov [1].

Neurónové siete (NN) sú univerzálnym nelineárnym aproximátorom funkcií. Ak máme dáta, ktoré vstupujú do systému a k nim odpovedajúce výstupy, NN sa môže naučiť chovať ako sledovaný systém pomocou trénovacích údajov. Toto je kľúčový moment pre aplikovanie NN do praxe [2].

ANN sa využívajú na určenie približnej hodnoty a najlepšie fungujú v systémoch, kde je vysoká tolerancia na chybu. Používateľ by preto nemal používať neuronové siete v citlivých a dôležitých systémoch.

Neuronové siete sa však odporúča používať v prípade, kedy objem dát je príliš veľký na spracovanie, kde vzťahy medzi premennými sú nejasné alebo vzťahy je príliš ťažké opísať adekvátne bežnými spôsobmi [1].

Využitie neurónových sieti

Neuronové siete sa dajú aplikovať do širokého spektra odvetí. V reálnom živote sa môžu využiť pri riešení problémov, akými sú validácie údajov, prognózy predaja, riadenia rizík a výskumy zákazníkov [2].

V maloobchodnom priemysle je dôležitý každý zárobok. Neurónové siete sú schopné zvážiť naraz viaceré parametre obchodu, ako napr. dopyt po produkte, príjem zákazníka, cena produktu a počet obyvateľov. Pre tých, ktorí využívajú neurónové siete to môže byť veľkou výhodou. Ako príklad môžeme uviesť vzťah medzi dvoma predmetmi v priebehu času pri nákupe tlačiarne s tonerom. Pri využití neurónových sieti sa dá predpokladať, že v priebehu 3 až 4 mesiacoch pravidelného používania sa toner v tlačiarni minie. Ako maloobchodník môže tento poznatok využiť na kontaktovanie zákazníka [1].

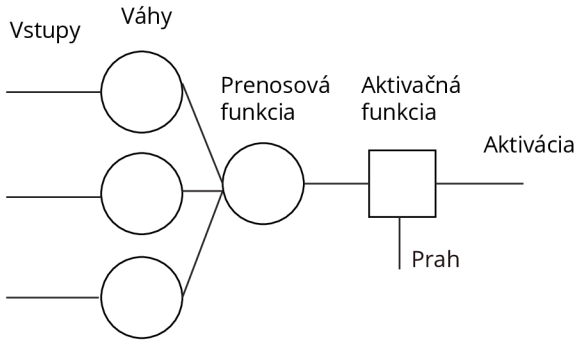
Správne nastavené neurónové siete v bankovníctve môžu priniesť neuveriteľné zisky. Využívajú sa na derivatové oceňovanie cenných papierov, prognózy cien, prognózy výmenných lístkov, prehľad akcií, sledovanie finančných búrz, metóda riadenia rizík [1].

V dnešnej dobe sa neurónové siete v medicíne aplikujú hlavne na rozpoznávanie lekárskych obrazov. Ide o modelovanie častí ľudského tela, sledovanie diagnóz a pod. Dnešné moderné systémy však predpokladajú rozsiahlu aplikáciu na väčšinu biomedicinských systémov [1].

Neuronové siete nám pomáhajú sa pozrieť na „ľahké problémy“ z pohľadu človeka. Počítač sa ale zatiaľ v žiadnom prípade nedokáže pozrieť na vnútorný život človeka a jeho sebapoznanie. Neuronové siete spôsobili, že počítačové systémy sa stávajú bližšie ľudskými [1].

1.2 Neurón

Neurón je základnou stavebnou a procesnou jednotkou. Vzhľadom na poprednú technológiu, dnešné systémy sú oveľa rýchlejšie v simulácií neurónov ako je samotný neurón v ľudskom tele. Základné časti neurónu sú vstup do neurónu, prah neurónu, aktivačná funkcia neurónu, vstupná funkcia neurónu a synaptické váhy [2].



Obr. 1.2: Zobrazenie jedného neurónu

Neuróny môžme rozdeliť podľa toku signálu po synapsii na predsynaptické (zdrojové) a postsynaptické (cieľové) [2].

Vstup do neurónu

Súčet vstupov od predsynaptických neurónov. Väčšinou je uvažovaný s určitými váhami. Kde X je matica vstupných premenných a W je vektor váh – majú rôznu orientáciu [2][3].

$$XW \tag{1.1}$$

Prah neurónu

Neuróny, ktoré nemajú vstup z iného neurónu, ale majú vstup z vonkajšieho sveta. Tieto neuróny môžu byť tiež nazývané sigma neuróny. Kde b je vstup z vonkajšieho sveta [2][3].

$$XW + b \tag{1.2}$$

Aktivačná funkcia neurónu

Neurónové siete sú dynamický systém, ktorý je závislý na čase. V okamihu vstupu do neurónu hovoríme o aktivačnej funkcii neurónu. Aktivačné funkcie delíme podľa

tvarov na napr. lineárnu funkciu, funkciu signum, po častiach lineárnu funkciu, sigmoidálnu funkciu a atď [2][3].

$$f(XW + b) \tag{1.3}$$

Výstupná funkcia neurónu

Výstupná funkcia neurónu je dôležitá súčasť procesnej jednotky. Pri neurónových sietiach musíme počítať s neidentickou vstupnou funkciou [2][3].

$$Y = (XW + b) \tag{1.4}$$

Synaptické spojenie a váhy

Pri neurónových sieťach je jedným z najdôležitejších aspektov prepojenie medzi neurónmi. Na orientovaných prepojeniach sú dôležité tzv. synaptické váhy. Ich hlavnou funkciou je ovplyvňovanie celej neurónovej sieti. Synaptické váhy dokážu ovplyvňovať vstupy do neurónov a stavy neurónov [2][3].

1.3 Hlboké učenie

Hlboké učenie je technika strojového učenia, ktorá učí zariadenia robiť tak, ako je prirodzené pre človeka, učí sa príkladom. Hlboké učenie prináša kľúčovú technológiu do autonómnych vozidiel, kde vozidlá samé dokážu rozpoznávať značky a chodcov. Taktiež je to kľúč ku ovládaniu hlasom v spotrebiteľských zariadeniach, ako sú telefóny, tablety a pod. Hlboké učenie dnes dosahuje výsledky, ktoré v minulosti nebolo možné dosiahnuť [4].

V hlbokom učení sa počítačový model učí vykonávať klasifikačné úlohy priamo z obrázkov, textu alebo zvuku. Modely hlbokého učenia môžu dosiahnuť stav presnosti, ktorý niekedy presahuje ľudský výkon. Modely sú vyškolené pomocou veľkej sady dát a pomocou viacvrstvových neurónových sieti [4].

V súčasnosti hlboké učenie dosahuje presnosť rozpoznávania na vyšších úrovniach než kedykoľvek predtým. To pomáha spotrebnej elektronike splniť očakávania používateľov a je rozhodujúca pre aplikácie, ktoré sú kritické z hľadiska bezpečnosti, ako sú vozidlá bez vodiča. Hlboké učenie si vyžaduje množstvo definovaných údajov. Pri vývoji autonómnych vozidiel je potrebné mať databázu miliónov obrázkov a tisíce hodín videa [4].

Nevýhodou hlboké učenia je značný výpočetný výkon. Vysoko výkonné grafické karty majú paralelnú architektúru, ktorá je efektívna pre hlboké učenie. V kombinácii s výpočetnými klastrami alebo cloudovým počítaním to umožňuje vývojovým tímom skrátiť čas potrebný na učenie z týždňov na hodiny [4].

Hlboké učenie je špeciálna forma strojového učenia. Pracovný tok strojového učenia sa začína manuálnym extrahovaním relevantných funkcií z obrázkov. Funkcie sa potom použijú na vytvorenie modelu, ktorý kategorizuje objekty v obraze. S hlbokým pracovným tokom učenia sa príslušne funkcie automaticky extrahujú z obrázkov. Okrem toho, hlboké učenie vykonáva end-to-end learning - kde sú k sieti pridelené nespracované dáta a úlohy, ktoré sa majú vykonať, napríklad klasifikácia obrázkov [4].

Väčšina metód hlbokého učenia používa neurónové siete, preto sú modely hlbokého učenia často označované ako hlboké neuronové siete. Hlboké neuronové siete môžu mat až 150 vrstiev [5]. Jedny z najpoužívanejších typov hlbokých neurónových sieti sú konvolučné neurónové siete. Táto architektúra siete používa vstupné údaje a 2D konvolučné vrstvy, preto je táto architektúra vhodná na spracovanie obrázkov [4].

1.4 Hlboké konvolučné neurónové siete

Hlboká neurónová sieť je neurónová sieť s určitou úrovňou zložitosti. ktorá používa sofistikované matematické modelovanie na spracovanie údajov zložitými spôsobmi. DNN sa skladá zo vstupnej vrstvy, niekoľko skrytých vrstiev a z výstupnej vrstvy [6].

Hlboké konvolučné neurónové siete sú tiež známe len ako konvolučné neurónové siete alebo ako konvolučné siete. Konvolučné siete sú špecializovaným druhom hlbokých neurónových sieti na spracovanie dát použitých na známu topológiu, ktorá je podobná mriežke [5][6].

Dáta prichádzajú v pravidelných časových intervaloch. Mriežky môžeme rozdeliť do dvoch skupín. 1-D mriežka, ktorá v pravidelných intervaloch odoberá vzorky a 2-D mriežku pixelov, ktorá zbiera obrazové dáta [5][6].

Názov konvolučné siete vznikol z matematického prostredia – konvolučné riešenie. Konvolučné siete namiesto klasického násobenia matíc využívajú konvolúciu [5][6].

Riedka interakcia, zdieľanie parametrov, Equivariant representations sú tri dôležité myšlienky, prečo sú populárne konvolučné siete [5][6].

Nevýhodou násobenia matíc v tradičných neurónových sietiach je, že každá vstupná jednotka komunikuje s každou výstupnou jednotkou. Konvolučné siete však využívajú riedku interakciu. Pri spracovaní obrazu, treba spracovať milióny pixelov. Pri konvolúcii však môžeme odhaliť malé, zmysluplné prvky ako hrany, ktoré zaberajú oveľa menej pixelov [5][6].

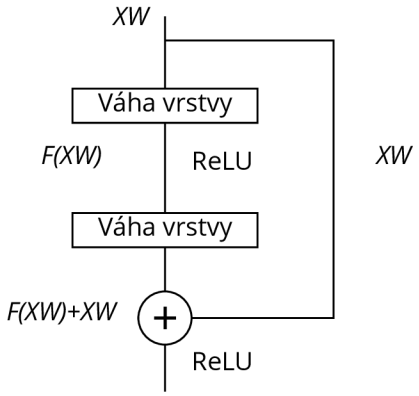
Zdieľanie parametrov je myšlienka, ktorá zdieľa ten istý parameter na ďalšiu funkciu v modeli. V neurónových sieťach sa každý parameter alebo hodnota matice použije presne jedenkrát pri výpočte výstupu vrstvy. Vynásobí sa jedným vstupom

a potom sa už nepoužíva. Avšak pri konvolučných sietiach sa každý člen používa v každej polohe vstupu [5][6].

1.4.1 Typy neurónových sieti

ResNet

V tradičných neurónových sieťach sa každa vrstva dodáva do nasledujúcej vrstvy. V sieti typu Residual Network (zvyškové siete) alebo skr. ResNet sa každá vrstva napája do nasledujúcej vrstvy a priamo do vrstiev o 2 až 3 skoky [7].



Obr. 1.3: Zobrazenie zbytkového bloku

Pri zvyšovaní počtu vrstiev je pozorovateľné, že presnosť sa začne saturovať na jednom mieste a nakoniec degraduje,čo je spôsobené pretrénovaním. Trénovanie niekoľkých vrstiev je možné preskočiť pomocou vynechania alebo zvyšných spojení. Z obr.1.3 vyplýva možnosť priameho učenia funkcie zhodnosti spoliehaním sa len na preskočenie spojenia. ReLU označuje usmernenú lineárnu jednotku,ktorá je jedna z najvyužívanejších funkcií v modeloch hlbokého učenia. Funkcia vráti 0 ak je hodnota záporná, inak vráti maximálnu hodnotu $f(x) = max(0, x)$ [7].

Blok neurónovej siete, ktorého vstup je x a je potrebné naučiť distribúciu $H(x)$. Rozdiel alebo zvyšok teda označíme ako,

$$R(x) = X_{\text{OUT}} - X_{\text{IN}} = H(x) - x \tag{1.5}$$

po úprave,

$$H(x) = R(x) + x \tag{1.6}$$

Zvyškový blok sa pokúša naučiť skutočný výstup $H(x)$. Na obr.1.3 vidieť spojenie so zhodnosťou, ktoré sa deje v dôsledku x, vrstvy sa snažia trénovať zostávajúce

$R(x)$. Tradičné siete sa trénujú reálnemu výstupu $H(x)$, zatiaľ čo vrstvy v ResNet sa učia zvyšku $R(x)$ [7].

Preskakovanie trénovania v niektorých zostávajúcich blokoch vrstvách je možné vnímať aj z optimistického hľadiska. Vo všeobecnosti nepoznáme optimálny počet vrstiev potrebných pre neurónovú sieť, ktorá by mohla závisieť od zložitosti súboru údajov. Niektoré vrstvy nie sú užitočné pre kompletnú neurónovú sieť a nepridávajú hodnotu, preto preskočeným niektorých vrstiev robí neurónové siete dynamickejšími, takže sa môže optimálne naladiť počet vrstiev počas tréningu [7].

Zvyškové bloky v podstate umožňujú tok pamäte presmerovať z počiatočných vrstiev do posledných vrstiev. ResNet urýchľuje učenie sa neurónových sieti, rastúca hĺbka obsahuje menej extra parametrov, znižuje efektivitu miznúceho gradientu a zvyšuje presnosť vo výkone siete [7].

Inception

Sieť s názvom Inception bola významným míľnikom vo vývoji klasifikátorov CNN. Pred príchodom Inception, väčšina populárnych CNN boli len konvolúcie uložené hlbšie a hlbšie aby získali, čo najväčší výkon. Inception je sieť zložitého typu, ktorá využíva veľa trikov na to, aby posunula svoj výkon ďalej. Neustály vývoj siete viedlo k vytvoreniu niekoľkých typov tejto siete: Inception v1, v2, v3 a Inception–ResNet. Každá verzia je iteratívne zlepšenie oproti predchádzajúcej. Chápanie aktualizácií môže pomôcť vytvoriť vlastné klasifikátory, ktoré sú optimalizované v rýchlosti a presnosti [8].

Prenesené učenie je metóda strojového učenia, ktorá využíva vopred natrénovanú neurónovú sieť. Model rozpoznávania obrazu Inception v3 sa skladá z dvoch častí. Prvá časť je extrakcia prvkov s konvolučnou neurónovou sieťou a druhá časť je klasifikácia s plne pripojenými a softmax vrstvami. Pretrénovaný model Inception v3 dosahuje najmodernejšiu presnosť na rozpoznávanie všeobecných objektov. Model odoberá všeobecné prvky v prvej časti a klasifikuje ich na základe znakov v druhej časti [9].

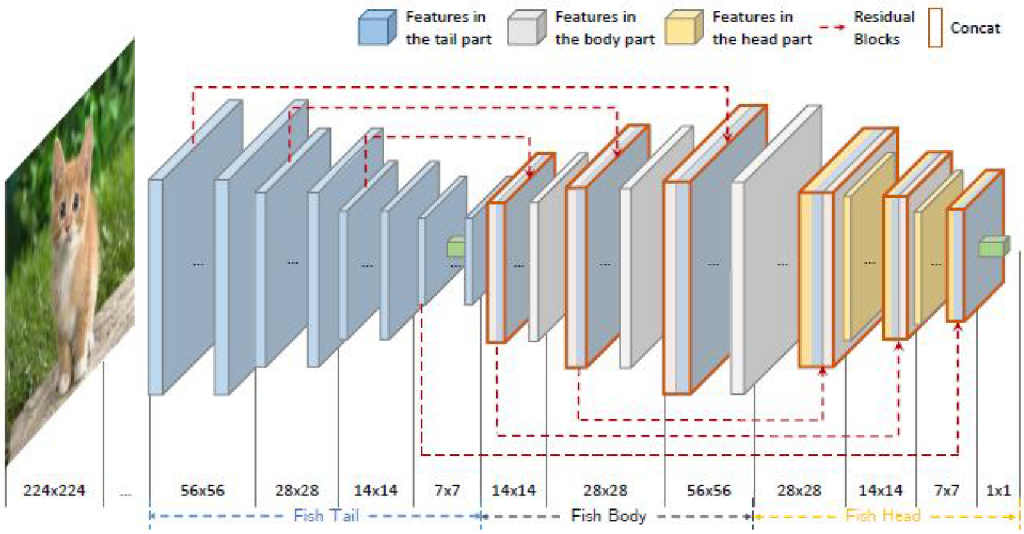
SeNet

Squeeze and Excitation Networks (SeNet) zavadzajú stavebný blok pre CNN, ktorý zlepšuje vzájomnú závislosť kanálov pri takmer žiadnych výpočtových nákladoch. SeNet boli použité v súťaži ImageNet 2017 a pomohli zlepšiť výsledok z predošlého ročníka o 25 %. Výhodou tohoto bloku je pridanie do existujúcich architektúr. Základom je pridanie parametru ku každému kanálu konvolučného bloku tak, aby sieť mohla adaptívne upraviť váhu každej mapy prvkov [10].

CNN používajú svoje konvolučné filtre na extrahovanie hierarchických informácií z obrázkov. Nižšie vrstvy nájdu triviálne kusy kontextu, ako sú hrany alebo vysoké frekvencie, zatial čo horné vrstvy môžu detekovať tváre, text alebo iné zložité geometrické tvary. Toto všetko funguje spojením priestorových a kanálových informácií obrazu. Rôzne filtre najprv nájdu priestorové funkcie v každom vstupnom kanále pred pridaním informácií do všetkých dostupných výstupných kanálov [10].

Neurónová sieť si váži každý zo svojich kanálov rovnako pri vytváraní výstupných mapových prvkov. SeNet je všetko o zmene uvedomelého mechanizmu, ktorý každý kanál váži adaptívne. Autori získajú globálne pochopenie každého kanála stlačením mapy funkcií na jednu číselnú hodnotu. Výsledkom je vektor veľkosti n, kde n sa rovná počtu konvolučných kanálov. Potom sa privádza cez dvojvrstvovú neurónovú sieť, ktorá vydáva vektor rovnakej veľkosti. Tieto hodnoty n sa použijú ako váhy na mapách pôvodných prvkov, pričom každý kanál sa môže meniť podľa jeho dôležitosti [10].

FishNet



Obr. 1.4: Zobrazenie siete typu FishNet[11]

Na obr.1.4 je možné vidieť prehľad siete FishNet. FishNet je rozdelený do troch častí, ktoré sú nazývané podľa časti tela ryby. Skladá sa z hlavy, tela a chvostu. Rybí chvost je existujúca CNN, napr. ResNet s rozlíšením funkcií menšie ako CNN, ktoré ide hlbšie. Rybie telo má niekoľko blokov na odber a zjemnenie zjemňovacích tvarov z chvosta a tela. Hlava ryby má niekoľko odberov a zjemňovacích blokov na

konzervovanie a rafináciu prvkov z chvosta, tela a hlavy. Rafinované funkcie na poslednej konvolučnej vrstvy hlavy sa používajú pre konečnú úlohu. Každá časť v sieti FishNet by mohla byť rozdelená do niekoľkých etáp podľa uznesenia výstupných funkcií. Keď sa rozlíšenie zmenši, ID štádia sa zväčšia. V sieti FishNet existujú dva druhy blokov pre odoberanie vzoriek a to nahor a nadol. Up-sampling & Refinement block (UR-blok) a Down-sampling & Refinement blok (DR-blok) [11].

Fishnet nie je sieť na klasifikáciu obrazu. Sieť dokáže obraz rozdeliť na niekoľko častí a pomenovať, čo sa v ktorej časti nachádza, tzv. segmeovať obraz. Návrh siete FishNet je stavaný pre riešenie problému gradientného šírenia. Všetky súčasti do všetkých stupňov v sieti FishNet sú spojene v hlave. Vrstvy sú starostlivo navrhnuté tak, aby sa v nich nenachádzali žiadne I-konv. Vrstvy v hlave sú zložené z reťazenia, konvolúcie s mapovaním identity a max-poolingu. Preto problém gradientového šírenia z predchádzajúcej chrbticovej siete v chvoste je riešený pomocou FishNet (výber funkcie vzorkovania hore/dole). Veľkosť jadra je nastavená na 2x2 pred podvzorkovanie s krokom 2, aby sa zabránilo prekrývaniu medzi pixelmi [11].

1.4.2 Framework pre hlboké učenie

Framework, v preklade tiež rámec, alebo softvérový rámec je platforma používaná pre vývoj aplikácií. Programátori nemusia poznať všetko od základov. Framework poskytuje základ, na ktorom môžu vývojári vytvárať programy pre konkrétnu platformu. Framework môže obsahovať triedy alebo funkcie, ktoré sa dajú použiť na spracovanie vstupov, správu hardvérových zariadení a pod. Výhodou použitia frameworku je čas ušetrení pri vývoji novej aplikácii [12].

Existuje veľa rôznych typov frameworkov pre hlboké učenie. Za väčšinu najznámejších frameworkov sú zodpovedné veľké korporácie ako sú Google, Microsoft, Facebook a pod.

TensorFlow

TensorFlow[1] je open source softvérová knižnica pre vysoko výkonné numerické výpočty. Flexibilná architektúra umožňuje jednoduché nasadenie výpočtovej techniky na rôznych platformách ako CPU, GPU, TPU. TensorFlow vyrástol z patentovanej knižnice hlbokých neurnových sieti DistBelief V2, ktorá bola vyvinutá v rámci projektu Google Brain.

V okamihu, keď Google otvoril zdrojové kódy frameworku TensorFlow si získal veľkú vývojársku pozornosť. Jeho využitie je veľmi široké, napríklad ako porovnávanie obrazu, rozpoznávanie rukopisu, rozpoznávanie reči, spracovanie a prognóza prirodzeného jazyka. TensorFlow sa vydáva pod licenciou Apache 2.0 open source .

[1] https://www.tensorflow.org/ – Otvorená zdrojová učebná knižnica pre výskum a výrobu.

TensorBoard slúži na vizuálizáciu sieťového modelovania a výkonnosti. TensorFlow Serving slúži na uľahčenie implementácie nových algoritmov a experimentov pri rovnakej serverovej architektúry a API.

Programovacie rozhrania frameworku TensorFLow sú Python a C++. S verziou 1.0 sú taktiež podporované Java, GO, R a Haskell API.

Framework podporuje Windows 7, 10 a Server 2016. Knižnice je možné kompilovať a optimalizovať na architektúre ARM pretože používa knižnicu C++ Eigen. To znamená, že sa môžu implementovať naučené modely na serveroch ale mobilných zariadeniach bez implementácie Python interpreta.

Každý výpočtový tok musí byť konštruovaný ako statický graf a nemá symbolické slučky. To sťažuje niektoré výpočty. TensorFlow neobsahuje žiadny 3-D konvolučný kanál a preto nie je vhodný na rozpoznávanie videí, aj napriek tomu, že je niekoľkokrát rýchlejší ako jeho pôvodná verzia [13].

Caffe

FrameWork Caffe[2] založil Yangqing Jia, ktorý je momentálne inžinierom platformy Facebook AI. Caffe patrí medzi hlavné nástroje pre hlboké učenie v priemysle od roku 2013. Vďaka svojmu konvolučnému modelu je obľúbeným nástrojom v komunite počítačových vízií. Caffe dosahuje veľmi dobré výsledky pre výskumné experimenty a komerčné nasadenie. Dokáže spracovať s jedným GPU Nvidia K40 každú 1ms jeden obrázok, to znamená približne 60 miliónov obrázkov denne.

Caffe je založený na programovacom jazyku C++, ktorý je možné spustiť na rôznych zariadeniach. Caffe navyše podporuje programovacie jazyky Matlab a Python. Má vlastnú komunitu, ktorá prispieva k úložisku pod menom Model Zoo.

Ide o framework, ktorý sa používa v hlbokých sieťach na rozpoznávanie vízií. Caffe však nepodporuje niektoré vrstvy, ktoré podporuje TensorFLow, CNTK alebo Theano. Vytváranie komplexných typov vrstiev musí byť preto vykonané v programovacom jazyku na nízkej úrovni. Jeho podpora pre rekurzívne siete a jazykové modelovanie je v základe slabá, súvisí to s jeho architektúrou.

Theano

Aj keď podpora Theana[3] je momentálne pozastavená stojí za zmienku. Tvorcom frameworku je Yoshua Bengio a celý framework je veľkým prispievateľom v oblasti hlbokého učenia.

[2]https://caffe.berkeleyvision.org/ – framework hlbokého učenia

[3]http://deeplearning.net/software/theano/ – pozastavený dňa 28.9.2017

Framework nie je upravený ako najpoužívanejší TensorFlow, no poskytuje možnosti API, tzv. Scan, čo umožňuje efektívnu a jednoduchú implementáciu rekurzívnej neurónovej siete.

Výhodou Theana je, že podporuje 3-D konvolúciu, ktorá sa využíva pri klasifikácií videa. Tak isto sa využíva aj pri klasifikácií obrázkov, vrátane tých v medicíne a ručného písania. Theano podporuje rozšírenie pre paralizáciu multi–GPU a ma distribuovaný framework pre trénovanie modelov postavených v Theano. Theano je obľúbený framework pre akademické účely.

Caffe 2

Caffe 2[4] poskytuje jednoduchý a priamočiary spôsob, ako experimentovať s hlbokým učením a využiť komunitné príspevky na vytvorenie nových modelov a algoritmov. Framework umožňuje flexibilné hlboké vzdelávanie. Caffe2 vychádza z Caffe. Za vznikom Caffe 2 je podobne ako pri Caffe Yangqing spolu s tímom z Facebooku. Od apríla roku 2017 otvorili zdroje Caffe 2 pod licenciou BSD. Caffe 2 je viac modulárny a vyniká v mobilných a veľkých nasadeniach. Rovnako ako aj TensorFlow bude Caffe 2 podporovať architektúru ARM pomocou knižnice C++ Eigen. Modely frameworku Caffe sa môžu ľahko konvergovať na modely Caffe 2 pomocou skriptov. Framework Caffe bol určený na problémy týkajúce sa vízií. Caffe 2 pokračuje v silnej podpore týchto problémov, navyše ale pridáva podporu rekurzívnych sieti a dlhej krátkodobej pamäte (LSTM) na spracovanie prirodzeného jazyka, predpovedanie časových radov a rozpoznávanie rukopisu.

Aj keď zatiaľ Caffe 2 nepredbehlo Caffe, v blízkej budúcnosti sa tak určite stane, lebo jeho komunita ho zdokonaľuje v komunite hlbokého učenia.

Keras

Keras[5] je framework na vysokej úrovni neurónových sieti. Je písaný v jazyku Python a je schopný pracovať nad TensorFlow, Theano alebo CNTK. Bol vyvinutý s cieľom rýchleho experimentovania. Keras umožňuje jednoduché a rýchle vytváranie prototypov. Podporuje konvolučné aj rekurzívne siete. Taktiež je možné ho využívať ako na CPU, tak aj na GPU.

Keras oproti TensorFlow, má mierne náročnejšie rozhranie kvôli nízko-úrovňovej knižnici, ktorá môže byť zložitá pre nových používateľov. Knižnica je postavená tak, aby poskytovala jednoduché rozhranie na účely rýchleho prototypu vytvorením efektívnej neurónovej siete, ktorá môže pracovať s TensorFlow.

[4] https://caffe2.ai/ – Nový ľahký, modulárny a škálovateľný framework hlbokého učenia

[5] https://keras.io/ – Keras je vysokoúrovňové API na vytváranie a školenie modelov hlbokého učenia

Hlavné využitie Kerasu spočíva v klasifikácií, generovaní a sumarizácie textu. Taktiež v značkovaní, prekladaní spolu s rozpoznávaním reči a iné [14].

PyTorch

PyTorch[6] je open source platforma hlbokého učenia vytvorená výskumnou skupinou AI spoločnosti Facebook. Pytorch je knižnica, ktorá realizuje tenzorové operácie, ale pridáva podporu pre GPU a ďalšie hardvérové urýchľovanie a efektívne nástroje pre výskumníkov AI, aby preskúmali rôzne domény. Zatiaľ, čo PyTorch začal ako nástupca systému Python, ktorý je základom frameworku Lua Torch, rozšíril sa tak, aby nebol len výskumnou platformou, ale aj reálnou platformou nasadenia.

Tým frameworku PyTorch tvorí 100 hlavných členov, komunita obsahuje vyšše 900 prispievateľov k open source a disponuje šiestimi administrátormi. Knižnica umožňuje stovky nadväzujúcich akademických a komerčných projektov. Aktuálne tým PyTorch pracuje na vývoji nástroju, ktorý prispieva k zlepšeniu životného prostredia [15].

1.4.3 Databáza ImageNet

ImageNet[7] je projekt, ktorý poskytuje výskumníkom na celom svete ľahko prístupnú databázu obrázkov. ImageNet je súbor údajov usporiadaný podľa hierarchie WordNet. Vo WordNete je viac ako 100 000 synsetov, väčšina z nich sú podstatné mena. V systéme ImageNet sa poskytuje približne 1000 obrázkov na ilustráciu každej skupiny. V dnešnej dobe je to približne 3,2 milióna obrázkov vo viac ako 5200 kategóriach. Obrázky každého typu sú kontrolované kvalitou a sú anotované človekom.

Databáza ImageNet slúži ako užitočný zdroj pre vedcov, prípadne pre edukačné skupiny. ImageNet avšak nevlastní práva k obrázkom, len poskytujú adresy k obrázkom.

[6] https://pytorch.org/docs/stable/index.html – Tenzorové a dynamické neurónové siete v Pythone so silným zrýchlením GPU

[7] http://www.image-net.org/ – Obrazová databáza

2 Hardvérové možnosti riešenia hlbokých neurónových sieti

Existujú dve možnosti riešenia, a to buď učenie novej neurónovej siete alebo pomocou naučenej siete vyvodiť niektoré atribúty o novej dátovej vzorke. Vzhľadom na to, že zadováženie si superpočítačov a grafických kariet je finančne náročné, veľké spoločnosti ako Google, Amazon, Microsoft a pod. ponúkajú cloudové možnosti zdieľania kvalitného hardvéru. Na začiatku sa používali procesory, no zistilo sa, že pomocou grafických kariet sa v dnešnej dobe dostávame až na 50–násobok rýchlosti riešenia neurónových sieti [16].

2.1 Centrálna riadiaca jednotka

CPU je centrálna riadiaca jednotka, ktorá sa na dnešnom trhu vyskytuje väčšinou vo viac-jadrovom prevedení. CPU vykonáva efektívne zložité operácie, strojové učenie však predstavuje opačnú výzvu. V trénovanom procese prebieha násobenie matíc, čo je pomerne jednoduchá úloha avšak rozsiahla. Výpočty sú veľmi malé a ľahké, ich množstvo je priveľké, teda CPU je preťažený, ale nedostatočne zamestnaný. Samotné CPU sa dajú rozdeliť do dvoch skupín – CPU pre použitie v domácnosti a serverové CPU pre vedecké účely. [16].

2.1.1 Serverové CPU

Pri samotných CPU stačí spomenúť jedného výrobcu a jednu rodinu CPU. Ako bolo vyššie spomenuté CPU nie sú vhodné na hlboké učenie. Procesory z rodiny Intel Scalable od spoločnosti Intel Xeon optimalizujú vzájomné prepojenie so zameraním na rýchlosť bez toho, aby sa ohrozila bezpečnosť dát [16].

Intel AVX-512 je súbor nových pokynov, ktoré môžu urýchliť výkon pri pracovných zaťaženiach a použití, ako sú vedecké simulácie, finančná analýza, umelá inteligencia, hlboké učenie a pod. Tento súbor umožňuje paralelné vykonávanie veľkého počtu operácií a väčšieho počtu jadier, čim sa v podstate stáva minisuperpočítač [16][17].

Procesory z rodiny Intel Scalable sa na trhu nachádzajú v štyroch verziách a to Platinum, Gold, Silver a Bronze. Dokonca aj jednotlivé verzie sa ďalej rozlišujú. Dokopy ponúkajú rozdielny počet jadier, rozdielnu pracovnú frekvenciu a počet vlákien [16]. V dnešnej dobe prijali korporácie práve procesory pre hlboké učenie. Datatonic uverejnil, že pri 11 násobných nákladoch na použitie sa zlepší výkon platformy

o 57 %, ktorú poháňa Intel Xeon Scalable. CPU tiež vyhovujú veľkým pamäťovým modelom, ktoré sú potrebné v mnohých doménach. Farmaceutická spoločnosť Novartis použila procesory Intel Xeon Scalable, aby urýchlila školenie pre viacúrovňovú konvolučnú neurónovú sieť (M-CNN) pre 10 000 obrázkov s vysokým obsahom bunkových mikroskopických snímok, ktoré sú omnoho väčšie ako typické obrazy ImageNet [16][17].

Zákazníci HPC používajú procesory Intel Xeon na distribuované školenia, ktoré boli prezentované na Supercomputing 2018. Vedecké centrum CERN predstavil distribuovaný tréning s použitím 128 uzlov klastra TACC Stampede 2, kde boli použité procesory Intel Xeon Platinum 8160 s 3D Generative Adversarial Network, ktoré dosiahli 94 % efektivitu škálovania [16][17].

Výkon hardvéru a softvéru CPU na hlboké učenie sa v posledných rokoch zvýšil o niekoľko rádov. Tréning, ktorý sa používal niekoľko dní alebo dokonca týždňov, sa teraz môže uskutočniť v hodinách alebo dokonca v minútach. Táto úroveň zlepšenia výkonnosti sa dosiahla kombináciou hardvéru a softvéru. Napr. súčasná generácia procesorov Intel Xeon Scalable s technológiou Intel Deep Learning Boost ponúka vyššiu priepustnosť, nižšie numerické precízne inštrukcie na zvýšenie výkonu hlbokého učenia. Na strane softvéru môže byť rozdiel vo výkone medzi základným softvérom hlbokého učenia s otvoreným zdrojovým kódom a softvérom optimalizovaným pre Intel až 275x na rovnakom procesore Intel Xeon Scalable [16][17].

Počas posledných rokov sa rozhodol výrobca CPU Intel spolupracovať s vývojármi frameworkov hlbokého učenia na optimalizácií mnohých populárnych open source frameworkov ako sú TensorFlow, Caffe, a ďalšie spomenuté v kapitole 1.4.2. Keďže základným výpočtom je lineárna algebra Intel vytvoril vlastnú knižnicu Intel Math Kernel Library pre hlboké neurónové siete, špeciálne pre hlboké učenie, založené na dlhoročných skúsenostiach s jadrom Intel Math Kernel. Integrácia Intel MKL-DNN do rámcov a ďalšie optimalizácie prispeli, aby plne využili základné hardvérové možnosti, ktoré sú kľúčovým dôvodom softvérového zvýšenia výkonu [17].

Aj keď vo svete hlbokého učenia sú stále obľúbené urýchľovače, nie je treba zabúdať ani na klasické CPU. Práve pre rôznorodosť hlbokého učenia môžu byť niekedy CPU také rýchle, a ak nie aj rýchlejšie, ako urýchľovače, pričom si zachováva flexibilitu, ktorá je základom pre návrh hodnôt CPU. Spoločnosť Intel hovorí, že ich procesory spolu s ideálnym softvérom sa dokážu vyrovnať niekoľko násobne drahším urýchľovačom [17].

2.2 Grafická riadiaca jednotka

Hlboké učenie je oblasť s veľkými výpočtovými požiadavkami a správny výber GPU zásadne určuje hlbokú skúsenosť s učením. Bez GPU by sa pravdepodobne na výsledky čakalo niekoľko mesiacov, prípadne by došlo k skorému ukončeniu experimentu kvôli chybám. So správne vybranými GPU je možné rýchlo opakovať vzory a parametre neurónových sietí a výsledky sa môžu dostaviť v niekoľkých dňoch. Následne sa grafické karty dajú rozdeliť do dvoch skupín – grafické karty pre herný priemysel a grafické karty pre vedecké účely [16].

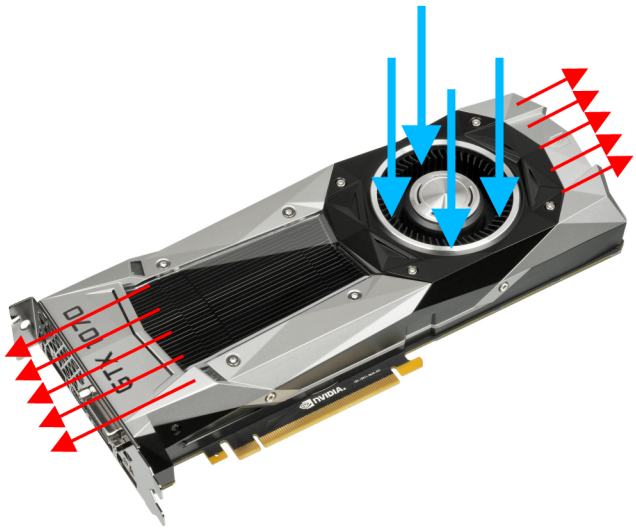
Násobenie matíc A x B = C je viazané na šírku pásma pamäte. Kopírovanie pamäte A a B na čip je náročnejšie než ich samý výpočet A x B. To znamená, že šírka pásma pamäte je najdôležitejším aspektom GPU pri používaní LSTM a iných opakujúcich sa sietí, ktoré robia množstvo malých maticových násobení. Čím menšie sú maticove násobenia, tým dôležitejšia je šírka pásma pamäte [16].

Opačným prípadom je konvolúcia, ktorá je viazaná výpočtovou rýchlosťou. FLOP na GPU je najlepším indikátorom pre výkon napr. ResNet alebo iných konvolučných architektúrach. Tenzorové jadrá dramaticky zvyšujú FLOPy [16].

2.2.1 Grafické karty vyrobené pre herný priemysel

Jedná sa o najčastejšie využívané karty vo výpočetných klastroch. Ich veľkou výhodou je pomer cena/výkon. Tieto karty sú označované od výrobcu NVIDIA ako GTX (pre staršie modely) a RTX (pre najnovšie modely). Rovnako sa dajú využiť grafické karty od výrobcu AMD označované ako RX. Cena týchto kariet sa pohybuje do približne 25 tisíc korún [16].

Populárny DNN model VGGNet bol pôvodne vyškolený na štyroch NVIDIA GeForce GTX Titan Black. Taktiež Caffe a TensorFLow sú dva široko používané frameworky a boli hodnotené na NVIDIA GeForce GTX 1080 a NVIDIA GeForce GTX Titan X v čase ich uverejnenia. Herné GPU vykonávajú porovnateľne, a niekedy aj lepšie, operácie ako GPU určené na vedecké účely, avšak majú určité obmedzenia pre použitie v HPC a hlbokom učení. Herné GPU a ich chladiče nie sú určené pre systémy s vysokou hustotou. Vyššie rady herných grafických kariet disponujú vždy aspoň jedným aktívnym chladičom. Na obrázku 2.1 je zobrazený prietok vzduchu, kde karta nasáva vzduch z priestoru skrini počítača a vyfukuje mimo skriňu. Keď je v jednom systéme nainštalovaných viac herných GPU, je to náročné na odstránenie množstva vytvoreného tepla. Neodvodené teplo môže spôsobiť chybné výpočty, preto herné GPU neobsahujú opraviteľné chyby pamäti ECC. Zlá prevádzková teplota tiež zmenšuje životnosť karty [16][18].



Obr. 2.1: Vizualizácia toku vzduchu hernej grafickej karte

Herné GPU väčšinou obsahujú menej operačnej pamäte v porovnaní s kartami určených pre vedecké účely, ako je uvedené v tab.2.1. DNN sa stále stávajú čoraz hlbšie a širšie, preto sa vyžaduje väčšia kapacita pamäte GPU. VGG-16 vyžaduje približne 10 GB RAM. Môže byť natrénovaní na jednej GPU, napr. na NVIDIA Tesla P100, ale už nedokáže byť natrénovaní na jednej NVIDIA RTX 2070. Vo väčších výpočetných klastroch býva niekedy vhodná kombinácia herných a vedeckých grafických kariet [18].

Obr. 2.2: Ukážka grafickej karty NVIDIA RTX 2080Ti[19]

Tab. 2.1: Tabuľka porovnávajúca veľkosť pamäte kariet a ich vlastnosti

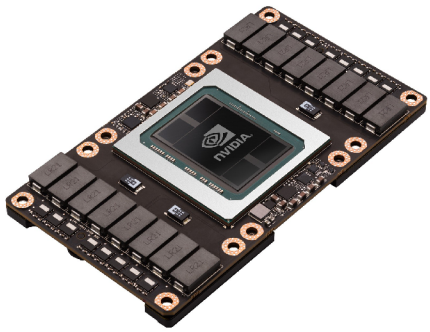
Grafická karta	Veľkosť operačnej pamäte [GB]	Šírka zbernice [Bit]	Frekvencia pamäte [GHz]
Herné grafické karty			
NVIDIA GTX 1080 Ti	11	352	11
NVIDIA RTX 2060	6	192	14
NVIDIA RTX 2070	8	256	14
NVIDIA RTX 2080 Ti	11	352	14
Vedecké grafické karty			
NVIDIA Tesla P40	24	384	10
NVIDIA Tesla P100	12/16	4096	1,430
NVIDIA Tesla T4	16	256	10
NVIDIA Tesla V100	16/32	4096	1,752

Na spomenutý problém pamäte hernej grafickej karty bola navrhnutá softvérová virtualizácia. VMDNN je skratka pre virtuálnu pamäť GPU pre DNN. Tento softvér implementuje pamäť GPU mechanizmus výmeny podobný pagingovému mechanizmu konvenčných operačných systémov. Keď je na GPU spustený framework pre hlboké učenia, ako napríklad TensorFLow alebo Caffe, systém prideľuje veľa pamäte objektom na uloženie všetkých parametrov modelu DNN, ako napr. mapy funkcií, vstupné údaje, prechody a váhy. Potom sa spúšťajú CUDA jadrá na GPU jeden po druhom. Avšak, každé jadro nemá prístup ku všetkým pamäťovým objektom GPU, pristupuje zvyčajne iba k pamäťovým objektom súvisiacim s jednou vrstvou neurónovej siete. VMDNN automaticky prepína niektoré zo zbytočných objektov GPU pamäte na hlavnú pamäť pre aktuálne jadro. Ak je celková veľkosť objektov GPU pamäte s prístupom k jednému jadru väčšia ako pamäť GPU, VMDNN rozdelí objekty pamäte na menšie kúsky a niekoľkokrát sprístupní jadro každým kúskom v tom istom čase [18]. VMDNN je implementovaný ako zdieľaná knižnica a transparentný pre cieľový framework hlbokého učenia. Nie je potrebná modifikácia alebo kompilácia hlbokého učenia. Zdieľaná knižnica sa vykonáva s frameworkom nastavením premennej prostredia LD PRELOAD, to zachytáva všetky volania jadra CUDA z frameworku a uskutočňuje správu pamäte. V dôsledku toho VMDNN je kompatibilný s rôznymi verziami frameworkov hlbokého učenia a knižníc backend ako cuDNN. Navyše VMDNN je k dispozícií pre používateľa, ktorý si preberá framework hlbokého učenia ako binárny správca balíkov alebo Docker obraz [18].

2.2.2 Grafické karty vyrobené na vedecké účely

Tieto karty spoločnosti sú určené práve na vedecké účely. Narozdiel od herných kariet výrobcovia tvrdia, že sú schopné fungovať 24 hodín denne, 7 dní v týždni. Prúdenie vzduchu je vhodné pre uloženie do serverov. Cena týchto kariet je pri AMD Radeon Instinct MI25 približne 110 tisíc korún a cena NVIDIA Tesla karty je približne 250 tisíc korún [16][19].

NVIDIA Tesla V100. Grafická karta je označovaná ako najdokonalejší akcelerátor pre dátové centrum. Jej výkon sa využíva pri AI a HPC. Postavená je na najnovšej architektúre Volta. Disponuje 640 Tenzorovými jadrami, ktoré sú oveľa výkonnejšie na počítanie ako CUDA jadrá, ktorých je 5120. Grafická karta má možnosť dvoch vyrovnávacích pamätí a to 16 alebo 32 GB. Na obr. 2.3 je zobrazené karta pre DGX systémy, táto karta je osadená pasívnym chladičom bez plastového krytu a prietok vzduchu sa stará o aktívne chladenie servera. Na obr. 2.4 je zobrazená PCIe verzia karty, táto karta taktiež neobsahuje aktívne chladenie a k správnemu odvádzaniu tepla je potrebné aktívne chladenie serverovej jednotky [16][19].



Obr. 2.3: Ukážka grafickej karty NVIDIA Tesla v100 edícia pre DGX-1 a DGX 2[19]

NVIDIA nedávno uverejnila na trh svoju ďalšiu grafickú kartu, pre vedecke účely. NVIDIA Tesla T4 zrýchľuje rôznorodé cloudové pracovné zaťaženia, vrátane vysoko výkonných výpočtových systémov hlbokého učenia, strojového učenia, dedukcie a analýzy dát. Model T4 je založený na novej architektúre NVIDIA Turing. Pripojenie urýchľovača je možné cez zbernicu PCIe. Výhodou tejto karty je úspora energie, jej výkon je iba 70W. Karta je optimalizovaná pre mainstreamové výpočtové prostredia a je vybavená presnejšími Turingovými tenzorovými jadrami a novými RT jadrami. Karta má 16 GB operačnej pamäte a disponuje 320 Turingovými tenzorovými jadrami a 2560 cuda jadrami. Karta má približnú cenovú hodnotu 70 tisíc korún [19].

Obr. 2.4: Ukážka grafickej karty NVIDIA Tesla v100 edícia PCIe[19]

Grafická karta na vedecké účely vyrobená firmou AMD je AMD Radeon Instinct MI25. Táto karta je založená na architektúre VEGA Graphics Architecture postavenej na spracovanie veľkých množín údajov a rôznorodých výpočtových zaťažení. Disponuje 64 nCU počítačových jednotiek na urýchlenie náročných pracovných zaťažení. Dosahuje až 12,3 TFLOPS pre FP32 a až 24,6 TFLOPS pre výkon FP16. Najmodernejšia pamäťová technológia obsahuje 16 GB pamäte HBM2 s ECC2. Táto karta obsahuje pasívne chladenie, preto je vhodná do serverov. Výkon AMD Radeon Instinct MI25 karty je 300W a obsahuje MxGPU pre virtualizovanie výpočtovej záťaže [20].

2.3 Tenzorová procesorová jednotka

TPU je vlastný druh čipu navrhnutý od základov spoločnosťou Google. Slúži pre pracovné zaťaženie hlbokého alebo strojového učenia. Na tomto čipe spoločnosti Google funguje niekoľko služieb ako fotografie, gmail, asistent, preklad a pod. Existuje možnosť Cloud TPU, ktorá poskytuje výhody TPU pre všetkých vývojárov a dátových vedcov, ktorí používajú modely učenia v službe Google Cloud [21].

Architektúra TPU vznikla ako špecifická doména. TPU teda nevznikol ako univerzálny procesor alebo je navrhnutý ako maticový procesor špecializovaný na pracovné zaťaženie neurónovej siete. TPU nemôžu spúšťať textové procesory, ovládať raketové motory ani vykonávať bankové transakcie, ale zvládajú masívne multiaplikácie a dodatky pre neurónové siete a to s veľkými rýchlosťami, pričom spotrebujú oveľa menej energie [21].

Primárnou úlohou tohto procesora je spracovanie matíc. Hardvérový dizajnér

TPU poznal každý krok výpočtu na vykonanie tejto operácie, tým pádom bol schopný umiestniť tisíce multiaplikátorov a pripojiť ich priamo k sebe [21].

Prvým krokom TPU je načítanie parametrov z pamäte do matíc multiplikátorov a addérov. Potom načítajú dáta z pamäte. Keď sa vykoná každé násobenie, výsledok sa odovzdá ďalším multiplikátorom, pričom sa súčasne sčíta sumácia. Výstupom bude teda súčet všetkých multiplikačných výsledkov medzi údajmi a parametrami. Počas celého procesu masívnych výpočtov a odovzdávania údajov sa vôbec nevyžaduje prístup do pamäte [21].

2.4 Operačná pamäť RAM

Výrobcovia operačnej pamäte RAM odporúčajú kúpiť pamäte s veľkou pracovnou frekvenciou. Tieto pamäte avšak nie sú najvhodnejšie pre hlboké učenie. Dôležitá vec, ktorú je potrebné si uvedomiť, že táto rýchlosť nie je dôležitá pre prenosy typu CPU RAM do GPU RAM. Dôvodom je fakt, že ak sa použila pripnutá pamäť, dáta budú prenesené do GPU bez zapojenia procesoru a ak sa nepoužila pripnutá pamäť, výkonové prírastky sa zvýšia na rýchlych RAM o 0 – 3 % oproti pomalým RAM [22].

Veľkosť pamäte neovplyvňuje výkon hlbokého učenia. Mohlo by to mať za následok plynulosť prechodu kódu do GPU. Veľkosť RAM by mohla dosahovať veľkosti RAM na grafickej karte. Ak zariadenie využíva viacej grafických kariet, nie je potrebné väčšie množstvo operačnej pamäte [22].

2.5 HDD vs. SSD

Rýchlosť disku nebýva prekážkou v hlbokom učení. Ak sú údaje z disku čítané vtedy keď sú potrebné (blokovací prístup), trvá približne 185 milisekúnd načítanie dávky z databázy ImageNet. Ak sú dáta čítané asynchrónne pred ich použitím, potom stačí túto dávku načítať v trvaní 185 milisekúnd, lebo výpočtový čas pre väčšinu neurónových sieti databázy ImageNet je 200 milisekúnd [22].

SSD disk sa však používa pre pohodlie produktivity. Programy reagujú rýchlejšie a predbežné spracovanie veľkých súborov je taktiež o niečo rýchlejšie [22].

Preto sa vo väčšine prípadoch odporúča veľký a pomalý disk na uloženie dát a SSD na produktivitu a pohodlnosť [22].

2.6 Napájací zdroj

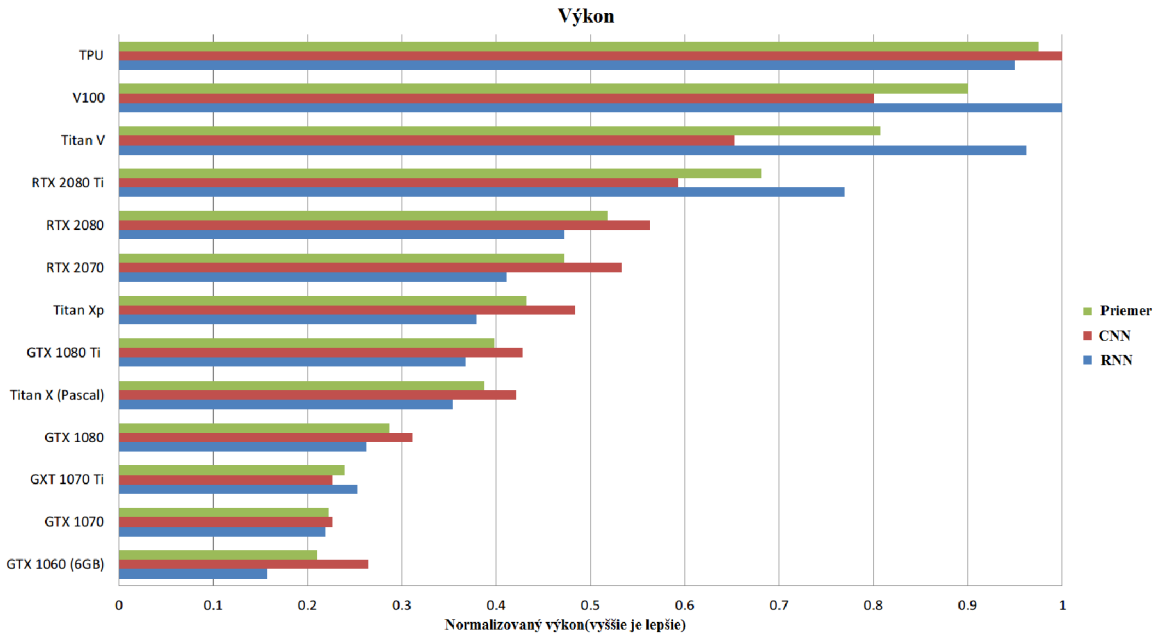
Základným parametrom je pri výbere zdroja je dostatok výkonu pre všetky grafické karty. Zdroje sa typicky menia až po dlhšej dobe, preto investícia do zdroja sa oplatí na dlhší čas. Pri výbere správneho zdroja okrem požadovaného výkonu je potrebné myslieť aj na počet potrebných pinov, ktoré sa budú pripájať na grafické karty [22].

Jeden z najdôležitejších parametrov pri výbere správneho zdroja je hodnotenie energetickej účinnosti. Predpokladá sa, že zariadenie, ktoré rieši hlboké učenie beží istú dobu. Zariadenie, ktoré v sebe obsahuje štyri grafické karty so spotrebou 1000 až 1500 wattov, čo na trénovanie konvolučnej neurónovej sieti na dva týždne vychádza asi na 300 až 500 kWh, cena v niektorých krajinách môže byť 60 až 100 € so 100 % účinnosťou. Ak by táto účinnosť bola 80 %, cena by mohla byť vyššia o 18 až 26 € [22].

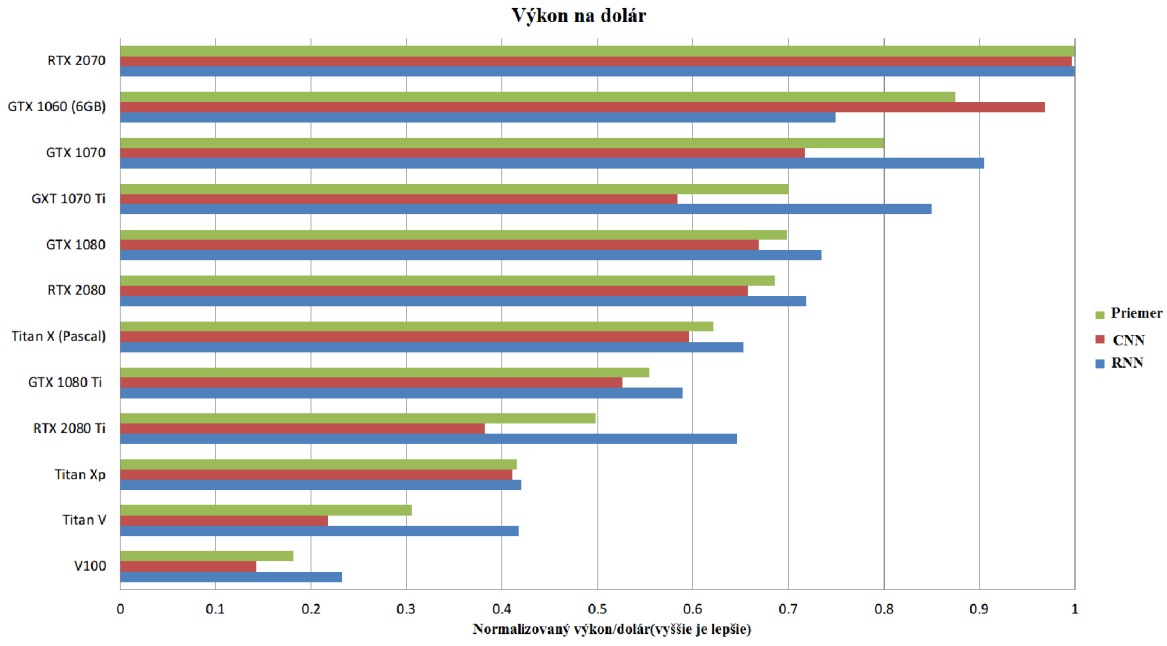
2.7 Výsledky hardvérových možností

Jednotlivý výkon sa dá len ťažko určiť, pretože vo veľa prípadoch záleží na použitej neurónovej sieti. Na obr. 2.5 sú zobrazené výkony grafických kariet na rôzne druhy NN. Tesla V100 disponuje tenzorovými jadrami. Ak chceme používať tieto tenzorové jadrá mali by sa používať 16-bitové dáta a váhy. Pri výbere kariet je dôležité dopredu vedieť aký typ sieti sa budú karty využívať, aby sa zbytočne neplatilo za veľký počet tenzorových jadier ak sieť bude potrebovať CUDA jadrá. Na obr. 2.5 je zobrazený normalizovaný výkon kariet pre CNN a RNN. V tejto časti jednoznačne vedie výkon TPU [23].

Cenová efektívnosť je z jedným najdôležitejších faktorov pre výber GPU. Na obr. 2.6 je zobrazený výkon grafickej karty k jednému doláru. Za cenu jednej karty NVIDIA Titan sa dajú zaobstarať dve karty NVIDIA GTX 1080 Ti. NVIDIA GTX 1080 Ti je jedna z najpoužívanejších kariet pre hlboké učenie. Jej výhodou je nízka cena a vysoký výkon ako aj množstvo pamäte RAM a vysoká priepustnosť. Obmedzený rozpočet môže hrať veľkú rolu vo výbere správnych kariet. Výkon grafickej karty NVIDIA Tesla V100 je jeden z najlepších, avšak jeho výkon v pomere na dolár je na poslednom mieste [23].



Obr. 2.5: Zobrazenie výkonu grafických kariet[23]



Obr. 2.6: Zobrazenie výkonu stiahnutého k jednému doláru[23]

2.8 Superpočítač

Mnohé organizácie považujú svoj vlastný HPC za nevyhnutnú súčasť obchodného úspechu. Z dôvodu, že tak vysoký výpočetný výkon sa využíva na návrh nových produktov, ťažobné údaje a prioritne simuláciu obchodného procesu, preto vlastnenie HPC môže byť tajnou zbraňou [24].

Cena týchto systémov je príliš veľká a preto si veľa menších spoločností takýto superpočítač nemôže dovoliť [24].

Poznáme niekoľko typov superpočítača. Jedným typom superpočítača sa nazýva Komoditný HPC klaster. Takýto klaster, zamestnáva stovky, tisíce a dokonca aj desiatky tisíc štandartných rackových serverov, ktoré sú medzi sebou vysoko rýchlostne prepojené. Tie poskytujú typický špičkový, efektívny výkon HPC. Tieto servere pracujú na vyriešení jedného problému. Tento typ HPC je veľmi populárny vďaka vysokému výkonu a pomerne nízkym nákladom [24].

Ďalším typom je tzv. dedikovaný superpočítač. V minulosti, bol toto jediný spôsob ako pomocou veľmi veľa cyklov vyriešiť jediný problém. Tieto počítače sú stále vyrábané a používajú ne-komoditné komponenty. Záleží od potreby používateľa a napriek cenovej nevýhode, to môže byť stále najlepšie riešenie [24].

HPC cloud computing, v preklade HPC internetové počítanie, je pomerne nová a využívaná metóda. Používa internet ako základ pre službu "cyklus ako služba"pre model na počítanie. Príslušné výpočtové cykly sa nachádzajú v internetovom úložisku, kde má používateľ vzdialený prístup. Cloud HPC poskytuje dynamické a škálovateľné zdroje koncovým používateľom ako službu [24].

Grid computing, je podobná služba ako Cloud computing, s rozdielom, že tento typ HPC sa využíva na akademické projekty. Ide o pripojenie vlastných klastrov, ktoré sú zdielané na národnej alebo medzinárodnej úrovni [24].

2.8.1 Hardvér

Základným prvkom celého HPC je výber správneho procesora. Inštalácia celého klastra spolieha na výkon procesora. Pokrok v architektúre x86, kde patria 32 a 64 bitové procesory, ktoré poháňajú komoditné klastre vpred. Pri výrobcoch CPU treba spomenúť Intel a AMD. AMD je výrobca, ktorý svoje procesory AMD Opterom vložil do HPC Sierra. Sierra je umiestnený v Lawrence Livermore laboratory a ide o druhý najlepší superpočítač na celom svete. Intel svoje procesory Xeon E5 vložil do Tianhe-2A, ktorý je štvrtý najlepší superpočítač na svete [24][25].

Ďalším dôležitým parametrom pri stavbe HPC je dostatok operačnej pamäte. Aby každé jadro mohlo spustiť samostatný program, potrebuje preto dostatok pamäte. Najvýkonnejší superpočítač na svete je Summit, umiestnený v Oak Ridge

National Laboratory a jeho operačná pamäť je približne 2,802 TB [26].

Dôležitou vecou pri riadení celého HPC je Intelligent Platform Management Interface. Správcovia môžu pomocou IPMI monitorovať stav systému a spravovať ho. Funkcie IPMI sa často spravujú cez sieťové rozhranie, čo je veľmi dôležité pri veľkých výpočtových uzloch [24].

HPC sa skladá z viacerých serveroch, tie sa označujú ako uzly. Aby boli všetky tieto uzly zaneprázdnené je potrebné ich správne prepojenie. Vysoko výkonné vzájomne prepojenia sú hodnotené latenciou a najrýchlejším časom, ktorý odošle jeden bajt. Rýchle pripojenie je dôležité, aby každý uzol mohol byť prepojený s iným uzlom a vymeniť si tak informáciu. Pri HPC môžeme hovoriť o dvoch hlavných technológiách na prepojenie a to InfiniBand a 10 Gigabit Ethernet [24][25].

V neposlednom rade k hardvérovej súčasti HPC patrí ukladanie dát. Často sa na tento aspekt zabúda. Aby bol superpočítač plne funkčný potrebuje vysoký výkon na ukladanie dát. Jednou z možností je NFS Server. Ďalšou možnosťou je paralelný súborový systém Luster. Luster je testovaný ako open-source riešenie, ktoré poskytuje škálovateľné I/O od klustrov. HPC vytvára veľké množstvo dát a zabezpečenie dostupnosti archivačného systému, čo je pre mnohých veľmi dôležité. Dobrý archivačný systém automaticky presunie dáta z jedného úložného zariadenia na iné, podľa nastavených pravidiel používateľom [24][25].

2.8.2 Softvér

Pre správne fungovanie HPC je potrebné, aby na ňom bežal operačný systém. Linux je jeden z najpoužívanejších OS, pre ktorý sa používatelia HPC rozhodnú a akýkoľvek iný softvér musí vedieť dobre spolupracovať s Linuxom. Medzi možnosti patrí Solaris, ktorý je voľne dostupný a má plnú podporu pre binárnu kompatibilitu systému Linux a Microsoft HPC Server [24].

Základný softvér GNU/Linux je open-source a môže byť voľne kopírovaný a používaný každým. Kódy sú zdieľané a preto je ideálnym OS pre HPC. Je možné voľne pozerať a meniť zdrojový kód, ale ak je potrebná podpora, je nevyhnutné za ňu zaplatiť. Medzi komerčné GNU/Linux distribúcie patri Red Hat, SUSE a pod [24].

2.9 NVIDIA DGX systémy

NVIDIA sa stala jedinečným výrobcom portfólia systémov, ktoré sú určené pre hlboké učenie. Systém DGX sú postavené na novej revolučnej platforme GPU NVIDIA Volta. Tieto grafické karty spolu s kombináciou inovatívným a optimalizovaným softvérom prinášajú plne integrované riešenia, prvotriedny výkon a výsledky. Systémy

DGX boli vyvinuté pre vedeckých pracovníkov, aby im boli poskytnuté najsilnejšie nástroje pre AI [19].

Dátová veda dáva podnikom po celom svete právomoc analyzovať a optimalizovať obchodné procesy, dodávateľské reťazce, vedecký výskum, produkty a digitálne skúsenosti. Počítanie pomocou GPU prináša revolúciu v oblasti dátovej vedy s open-source analýzou údajov a platformou zrýchlenia strojového učenia [19].

2.9.1 Typy DGX

Existujú tri možnosti DGX systémov. Jednotlivé systémy sú rozdielne samozrejme vo výkone, ktorý sa líši nie len v použitých CPU ale aj v rozdielnom počte použitých GPU. Vo Všetkých DGX systémov sa nachádzajú NVIDIA Tesla V100 s 32GB operačnou pamäťou. Ďalším rozdielom je počet RAM, prepojení GPU kariet, vo veľkosti úložiska, v prepojení so sieťou, v maximálnom príkone, v prevedení a samozrejme v cene [19].

NVIDIA DGX Station

Prvý a najmenší DGX systém sa volá DGX Station. Ide o osobný superpočítač pre vývoj AI. Vedecký tím je závislý od výpočtového výkonu, aby získal informácie a inovoval rýchlejšie prostredníctvom schopnosti hlbokého učenia a analýzy údajov. Doposiaľ neexistoval stroj, ktorý by bol výkonom superpočítač a bolo by možné ho umiestniť v kancelárií. V minulosti bol každý superpočítač umiestnený v dátovom centre. Hlučnosť neprekračuje hluk bežného počítača. Vedci pomocou tohoto systému dokážu zvýšiť svoju produktivitu s pracovnou stanicou. Nemusia čakať, kým sa im uvoľní miesto vo veľkom dátovom centre. Potrebný výkon dosiahnu pomocou hardvéru umiestneného pod stolom a optimalizované softvéru pre hlboké učenie [19].

Stanica DGX prelomí obmedzenia v budovaní vlastnej platformy pre hlboké učenie. Niektorí trávia mesiace obstarávaním, integrovaním a testovaním vlastného hardvéru a softvéru pre hlboké učenie. Následne sú potrebné ďalšie odborné znalosti a úsilie na optimalizáciu rámcov, knižníc a ovládačov. To je cenný čas a peniaze vynaložené na systémovú integráciu a softvérové inžinierstvo, ktoré by mohli byť použité na školenie a experimentovanie [19].

Stanica NVIDIA DGX Station je navrhnutá tak, aby spustila iniciatívu s umelou inteligenciou. Hlboké učiace platformy vyžadujú odborné znalosti v oblasti softvérového inžinierstva [19].

NVIDIA DGX Station obsahuje rovnaký softvérový balík ako zvyšné DGX. Softvérový balík je popísaný v ďalšej kapitole [19].

Stanica prináša neuveriteľný výkon superpočítača v pracovnom priestore, ktorý využíva inovatívne inžinierstvo a vodou chladený systém, ktorý je tichý. Stanica je

postavená na štyroch urýchľovačoch NVIDIA Tesla V100 prepojených NVLinkom. Karty spolu obsahujú 2560 tenzorových jadier, úložisko zabezpečujú 4 1,92TB SSD disky a tento hardvér dodáva výkon takmer o,5 petaFLOPSu pri maximálnej spotrebne len 1,5kW [19].

NVIDIA DGX-1

Systém hlboké učenia NVIDIA DGX-1 obsahuje kombináciu hardvéru a softvéru, ktorá poskytuje rýchlejší a presnejší tréning neurónových sieti. DGX-1 je určená pre hlboké učenie a analytické analýzy AI a poskytuje výkon ekvivalentný 250 bežným serverom s CPU [19].

O výkon DGX-1 sa starajú najnovšie karty NVIDIA Tesla V100 v počte 8 kusov. Jednotlivé karty sú prepojené pomocou NVLinku, ten zabezpečuje rýchlu komunikáciu kariet. DGX-1 prináša štvornásobne vyššiu rýchlosť tréningu ako iné systémy na báze GPU pomocou NVIDIA GPU Cloud. Všetky karty spolu obsahujú 5120 cuda jadier a 640 tenzorových jadier [19].

NVLink otvára plný výkon ôsmich kariet Tesla V100, takže DGX-1 poskytuje takmer jeden petaFLOPS s polovičnou presnosťou, čo je najbežnejší formát používaný pri výpočtoch hlbokého učenia [19].

Ak je potrebné ešte viac výpočtového výkonu, je možné spojenie viacerých systémov DGX-1 do jedného výpočetného klastra. Hardvér DGX-1 obsahuje štyri 25Gb Infiniband EDER porty a dva 10Gb Ethernet porty. Pridávanie viacerých serverov DGX-1 do klastru si vyžaduje vysokovýkonné ukladanie a sieťovanie, aby sa mohli vyrovnať s vysokými požiadavkami na I/O operácie [19].

NVIDIA DGX-2

Najnovším a zároveň najvýkonnejším systémom do rodiny DGX je DGX-2. Hlboké neuronové siete rýchlo rastú vo veľkosti a zložitosti. NVIDIA DGX-2 je prvým dvoj petaFLOPSovým systémom na svete. DGX-2 obsahuje 16 najpokrokovejších a najdrahších GPU kariet na svete. Výkon DGX-2 sa porovnáva s výkonom 300 serverov pre trénovanie ResNet-50. Tento unikátny systém obsahuje duálne procesory Intel Xeon Gold v hodnote viac ako 2,7 milióna dolárov [19].

Medzi kartami sa musí uskutočňovať veľmi rýchla a bezchybná komunikácia, aby sa dosiahlo bezproblémového modelu paralelizmu. NVIDIA, kvôli tomuto vyvinula NVSwitch. Rovnako ako vývoj od dial-up až po ultra-vysokú rýchlosť. NVSwitch dodá sieťový materiál pre budúcnosť. DGX-2 poskytuje 2,4TB šírku pásma. Spotreba tohoto zariadenia pri plnom nasadení je 10kW [19].

O úložisko tohoto systému sa starám osem 3,84 TB rýchlych NVMe diskov. Tento priestor slúži len na ukladanie lokálnych dát, s dodatočným ukladaním na hlavnú

dosku pre OS a aplikačný kód. DGX-2 obsahuje high-bandwith sieťové rozhranie pre pripojenie viacerých zariadení a tým pádom je schopný vybudovať ešte väčší výpočetný klaster [19].

Softvér DGX systémov

Spolu s DGX systémom si používateľ kupuje aj softvérovú výbavu. Softvérový balík obsahuje operačný systém Ubuntu Serer Linux s nainštalovanými ovládačmi GPU kariet. Taktiež obsahuje nástroj na ovládanie kontejnerov NVDocker, nástroj pre hlboké učenie SDK, službu NVIDIA Cloud Management Service a službu NVIDIA DIGITS, ktorá sa používa na spustenie frameworkov ako sú Caffe, Torch, TensorFlow a iné [19].

Operačný systém je optimalizovaný tak, aby využíval hardvérové a softvérové funkcie systému a CUDA 8, najmä čo sa týka správy pamäte a hardvérovej komunikácie [19].

2.9.2 NVIDIA GPU cloud

NGC je úložisková platforma zameraná na zrýchlenie GPU, ktorá je optimalizovaná pre vedecké účely a hlboké učenie. Po prihlásení sú dostupné NGC kontajnery, register kontajnerov NGC a platforma na implementáciu a spustenie jednotlivých kontajnerov na hlboké učenie [19].

NGC kontajnery sú navrhnuté tak, aby umožnili softvérovú platformu zameranú na minimálne požiadavky na operačný systém, ďalej na inštaláciu Docker a ovládačov na serveri alebo pracovnú stanicu a poskytovanie všetkých aplikácií a softvéru SDK v kontajneroch NGC [19].

NGC spravuje katalóg plne integrovaných a optimalizovaných učiacich frameworkových kontajnerov, ktoré plne využívajú NVIDIA GPU v konfiguráciách s jedným alebo viacerými GPU. Zahŕňa sa tu CUDA Toolkit, DIGITS workflow a frameworky ako NVCaffe, Caffe2, CNTK, MXNet, Pytorch, TensorFlow a pod. Tieto kontajnery sú dodávané v stave pripravenosti, vrátane všetkých potrebných závislostí, ako CUDA runtime, knižnice NVIDIA a operačný systém [19].

Každý kontajner obsahuje zdrojový kód rámca, ktorý umožňuje vlastné modifikácie a vylepšenia spolu s kompletným balíkom vývoja softvéru. NVIDIA tieto kontajnery aktualizuje raz mesačne, aby zabezpečila, že budú ďalej poskytovať špičkový výkon [19].

3 Implementácia projektov hlbokých neurónových sieti do rôznych systémov

Praktická časť práce je zameraná na porovnanie grafických kariet pre výpočet neurónových sieti. Praktická časť taktiež porovnáva ostatné vlastnosti systému na výpočet hlbokých neurónových sieti. Rôzne druhy grafických kariet boli testované na sieti Inception V3. Vlastnosti celého systému boli testované sieťou ResNet. Použitý hardvér pri postavení testovacieho stroja je CPU: Intel(R) Xeon(R) CPU E5-2667 v4 @ 3.20GHz. Matičná doska je Supermicro X11-SRA-O a zariadenie disponuje 64 GB operačnej pamäti. Použité grafické karty sú spomenuté neskôr. Nainštalovaný operačný systém je CentOS 7.6, ide o Linuxovú verziu, tá ma v sebe implementovaný programovací jazyk Python, čo je vhodné pre túto prácu. Pri výpočtoch je využitá open-source knižnica Keras, ktorá ako svoj backend využíva TensorFlow. Backend sa využíva, aby sa nemuseli znovu programovať všetky potrebné časti, týmto pádom sa jednoducho implementujú. Okrem knižnice Keras a TensorFlow je za potrebu do operačného systému doinštalovať dôležité súčasti, akými sú napríklad oficiálne ovládače od výrobcov grafických kariet. Pre správne fungovanie celého systému je vždy potrebné hľadať najnovšie ovládače. Ďalšími doinštalovanými súčasťami systému sú doplnkové knižnice pre Python, doplnková knižnica CUDA, ktorá je určená pre paralyzáciu výpočtov a na koniec doplnkovú cuDNN. V praktickej časti sa meraním zisťuje rýchlosť výpočtov grafických kariet (čas), spotreba celého systému pri rôznych kartách a hodnotené je rovnako aj pomer cena k výkonu.

Pri inštalácií systému je potrebné dávať pozor na verzie inštalovaných doplnkových programov a knižníc. V prípade inkompatibility testovací program nebude funkčný. V prípade Kerasu, verzia nehrá veľký rolu. Verzia TensorFlow je tensorflow-gpu-1.11.0, TensorFlow sa ponúka aj vo verzii bez gpu. Tento TensorFlow je však stavaný na používanie pri systémoch bez GPU. Tensorflow-gpu-1.11.0 je kompatibilný s už nainštalovanou verziou Pythonu 2.7. Od verzie TensorFlow sa odvíja verzia CUDA, ktorá v našom prípade je verzia 9. Doplnok cuDNN vo verzii 7.

Testovací program do záverečnej práce poskytol Bc. Michal Kuvík, ktorému veľmi pekne ďakujem. Program slúži na automatické rozpoznávanie druhov jedál. Program je naprogramovaný s knižnicou Keras a TensorFlow ako backend. Dáta, na ktorých sa program učí sú druhy jedál, kde trénovacie aj validačné súbory obsahujú 211 tried s viac ako desiatkami tisíc obrázkov. Model konvolučnej neurónovej sieti je Inception V3. Táto sieť bola použitá na testovanie rôznych grafických kariet [27].

Sieť ResNet je mojím vlastným programom, ktorá využíva databázu použitú v testovaní Inception V3. V tejto časti sa práca zaoberá ostatnými vplyvmi na rýchlosť výpočtov ako napr. pamäť RAM, použitý úložný priestor a rozdiel medzi použi-

Obr. 3.1: Zapojenie pracovnej stanici do tzv. pavúka

tou jednou kartou a viacerými. Jednotlivé zmeny v systéme sú popísane v kapitole 3.2. Hlavnou testovacou kartou pre sieť ResNet bola herná grafická karta NVIDIA RTX 2080 Ti, ktorá disponuje 11 GB operačnej pamäte a 4352 CUDA jadrami.

3.1 Implementácia siete Inception V3

Táto čas implementácie siete typu Inception V3 má za úlohu zistiť jednotlivý výkon kariet. Meranie prebieha v rýchlosti výpočtov. Jednotlivé výpočty budú viditeľné v tabuľkách a grafoch. Použité grafické karty sú zvolené náhodne. Podľa dostupnosti od firmy MComputers s.r.o., ktorá tieto karty zapožičala. Herné grafické karty sú označované písmenami GTX a RTX. Tieto grafické karty sú od rôznych výrobcov, avšak rozdiely sú minimálne, preto je uvádzaný len hlavný názov. Prídavok Ti za názvom grafickej znamená Titanium. Ide o vylepšenú verziu pôvodnej grafickej karty bez označenia Ti. Herné grafické karty nedisponujú tenzorovými jadrami, disponujú iba základnými Cuda jadrami. Inno3D P102-100 je špeciálna grafická karta typu mining, tzn. firma Inno3d si základnú grafickú kartu NVIDIA GTX 1080 prispôsobila na ťažbu kryptomien. Táto karta disponuje rovnakým čipom ako NVIDIA

GTX1080, taktiež disponuje desiatimi 1 GB RAM modulmi ale firma Inno3D ich firmwerom zmenšila na celkový počet 5 GB. Pre porovnanie výkonnosti a podobnosti s grafickou kartou NVIDIA 1080Ti bol tento firmware prepísaný, a karta tak dostala 10GB celkovej operačnej pamäti. Do teraz spomenuté grafické karty sú pod konštrukčnou architektúrou Pascal. Najdrahšia testovacia grafická karta je NVIDIA Tesla V100. Táto grafická karta je pod architektúrou Volta. Práve do tejto architektúry výrobcovia vložili Tenzorové jadrá, ktorými táto karta disponuje v počte 640.

Počet Cuda jadier ovplyvňuje celkový výkon karty. Veľkosť operačnej pamäte a typ operačnej pamäte je dôležitým faktorom pri načítaní dát. Niektoré zložité neuronové siete sa kvôli malému množstvu pamäti v karte nemusia spustiť. Priepustnosť pamäte je dôležitým faktorom pri zápise do nej. Priemerná cena sa určila z internetových stránok, kde boli jednotlivé karty dostupné.

Základné vlastnosti grafických kariet, ktoré boli použité v testovaní siete Inception V3 sú popísané v tab. 3.1 a 3.2 .

NVIDIA GTX 1060 je najmenšia grafická karta, ktorá sa použila. Veľkosť jej operačnej pamäte sa dodáva v dvoch verziách, v testovaní sme použili menšiu operačnú pamäť, a to 3 GB.

Grafická karta NVIDIA Tesla V100 je najdrahšou profesionálnou kartou na trhu. Jej výkon momentálne nemá konkurenciu. Aby táto karta správne fungovala a využívala svoj potenciál je potrebné doprogramovať používanie Tenzorových jadier. V mojej záverečnej práci som však túto zmenu nestihol vykonať z dôvodu oneskorenia karty na moje testovanie a preto je karta otestovaná základným programom. Tento program nevyužíval grafickú kartu na jej plný výkon ale používal len Cuda jadrá. Rozdiely vo výkone sú aj tak patričné. Herná grafická karta NVIDIA RTX 2080 Ti, ktorá bola postavená na architektúre Touring dostala taktiež tenzorové jadrá a to v počte 544.

Tab. 3.1: Základne informácie počtu jadier grafických kartách

Grafická karta	Počet Cuda jadier	Počet Tenzor Jadier
NVIDIA GTX 1060	1280	-
NVIDIA GTX 1070 Ti	2432	-
NVIDIA GTX 1080 Ti	3584	-
Inno3D P102-100	3200	-
NVIDIA RTX 2080 Ti	4352	544
NVIDIA Tesla V100	5120	640

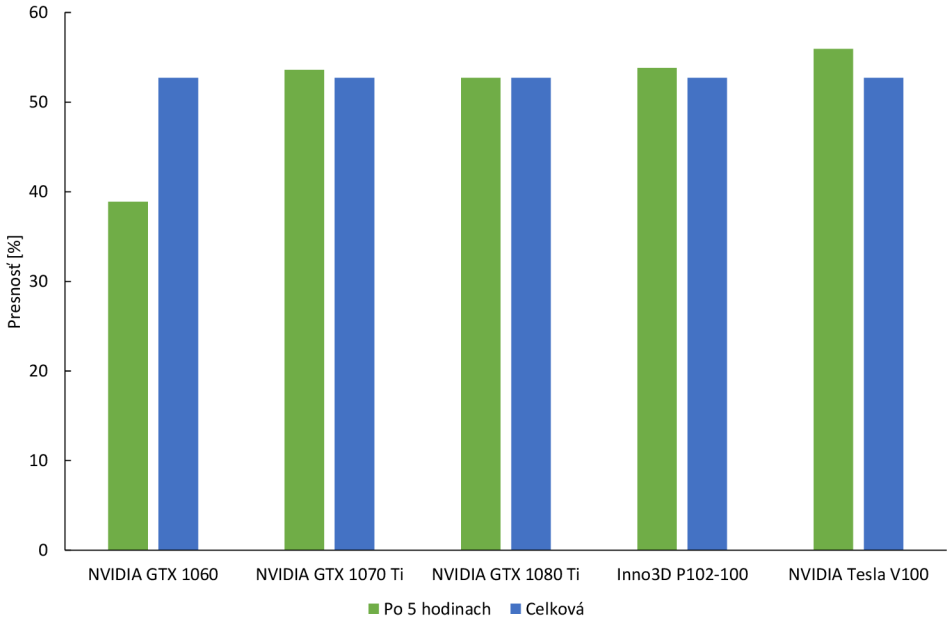
Tab. 3.2: Základne informácie o pamäti a cene grafických kartách

Grafická karta	Operačná pamäť [GB]	Priepustnosť pamäte [GB/s]	Priemerná cena [Kč]
NVIDIA GTX 1060	3 GB GDDR5	192	6000
NVIDIA GTX 1070 Ti	8 GB GDDR5	256	12 500
NVIDIA GTX 1080 Ti	11 GB DDR5X	484	25 000
Inno3D P102-100	5 GB GDDR5X	440	19 000
NVIDIA RTX 2080 Ti	11 GB GDDR6X	616	30 000
NVIDIA Tesla V100	32 GB	920	283 000

Trénovacie dáta sú neodmysliteľnou súčasťou pre proces učenia sa. Pri trénovacích dátach sú dôležité váhy. Dôležitým faktorom je, aby sa neurónová sieť nepretrénovala. Výsledky presnosti trénovacích dát sú umiestnené v tab. 3.3 Celková presnosť je spriemerovaná a je stanovená odchýlka. Rôzne celkové presnosti majú za následok náhodne generované čísla v testovacom programe. Ak by sa podarilo vždy vygenerovať rovnaké čísla, presnosť by bola rovnaká. Najmenšiu hodnotu dosiahla karta NVIDIA GTX 1060, kde odchýlka je pomerne veľká a jej výkon nestačil, aby dosiahol takú presnosť ako ostatné karty.

Tab. 3.3: Presnosť trénovacích dát

Grafická karta	Po 5 hodinách [%]	Celková [%]	Odchýlka
NVIDIA GTX 1060	44,76	81,46	8,07
NVIDIA GTX 1070 Ti	63,78	81,46	2,21
NVIDIA GTX 1080 Ti	70,86	81,46	1,23
Inno3D P102-100	61,04	81,46	2,05
NVIDIA Tesla V100	77,89	81,46	2,56

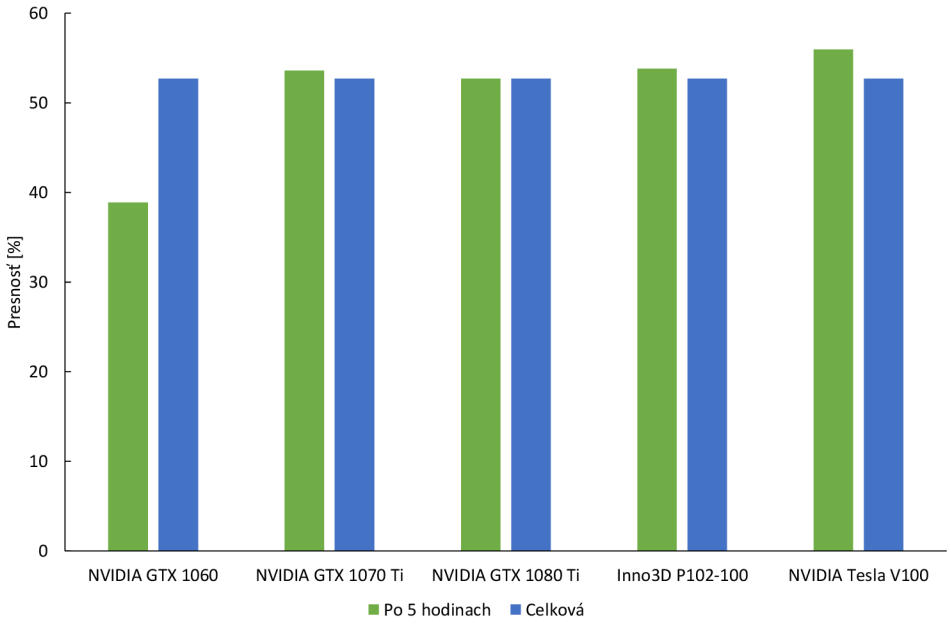


Obr. 3.2: Graf zobrazujúci presnosť trénovacích dát

Validačné dáta slúžia na ladenie neurónovej sieti. Práve tieto dáta sa používajú aby nedošlo k pretrénovaniu. Vzhľadom na to, že v testovacom programe nie sú použité labely, tieto validačné dáta nám ukazujú celkovú presnosť siete. Výsledky sú uložené v tab. 3.4. Výsledky sú opäť spriemerované a je určená odchýlka. V tejto časti si najlepšie výsledky dosiahla NVIDIA GTX 1080 Ti, kde sa odchýlka od priemernej hodnoty líšila len o 1,04. Aj keď momentálne sú na trhu už novšie herné grafické karty, práve táto karta sa používa vo veľa prípadoch vo výpočtových centrách.

Tab. 3.4: Presnosť validačných dát

Grafická karta	Po 5 hodinách [%]	Celková [%]	Odchýlka
NVIDIA GTX 1060	38,88	52,70	6,75
NVIDIA GTX 1070 Ti	53,60	52,70	2,61
NVIDIA GTX 1080 Ti	52,71	52,70	1,04
Inno3D P102-100	53,80	52,70	1,31
NVIDIA Tesla V100	55,96	52,70	1,79



Obr. 3.3: Graf zobrazujúci presnosť validačných dát

Na obr. 3.3 sú zobrazené presnosti po piatich hodinách testovania. Celková doba počítania pri NVIDIA GTX 1080 Ti bola 6,6 hodiny. V tabuľke 3.4 je vidieť ako sa minimálne zmenili presnosti za tento čas. Samozrejme by sa tieto presnosti zmenili pri väčšom počte epoch, ale toto nebola náplň práce.

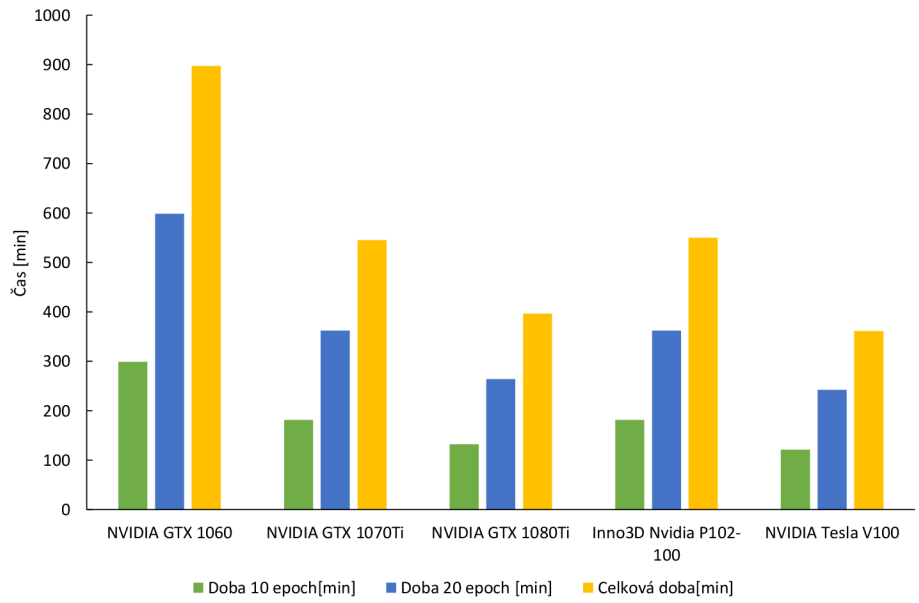
Doba počítania siete Inception V3

Záverečná práca sa zaoberala práve dobou počítania neurónových sieti na jednotlivých kartách. Výsledok jasne ukazuje, že čím karta obsahuje viac CUDA jadier alebo Tenzorových jadier, tým sú jej výpočty rýchlejšie.

Tab. 3.5: Doba počítania implementácie Inception V3

Grafická karta	Doba 10 epoch [min]	Doba 20 epoch [min]	Celková doba [min]
NVIDIA GTX 1060	299	598	897
NVIDIA GTX 1070 Ti	181	362	545
NVIDIA GTX 1080 Ti	132	264	396
Inno3D P102-100	181	362	550
NVIDIA Tesla V100	121	242	361

Na obr.3.4 je viditeľný graf, v ktorom najrýchlejšie počítanie dosiahla vedecké grafická karta NVIDIA Tesla V100. Táto karta je považovaná za jednu z najlepších kariet na svete. Na druhom mieste skončila bývala najlepšia karta NVIDIA GTX 1080 Ti. Na poslednom mieste je najmenšia karta NVIDIA GTX 1060.



Obr. 3.4: Graf znázorňujúci dobu počítania siete Inception V3

Chladenie grafických kariet

Aj keď herné grafické karty disponujú vlastným aktívnym chladičom, bolo v tomto prípade nutné doplniť o ďalšie ventilátory, ktoré boli schopné zvýšiť odvod tepla, nakoľko herné grafické karty aj pri spracovávaní náročných hier, nevykonávajú nepretržite takú námahu, ako pri počítaní neurónových sietiach.

3.2 Implementácia siete ResNet

V tejto časti testovania bola použitá neurónová sieť typu ResNet. Dôvodom výberu práve tohoto typu je všeobecná obľúbenosť a veľmi dobré výsledky na súťažiach. Implementácia tejto siete testuje jednotlivé vplyvy na výpočetný výkon siete. Zaoberá sa vplyvom operačnej pamäte, vplyvom použitého disku a v neposlednom rade viacerými grafickými kartami.

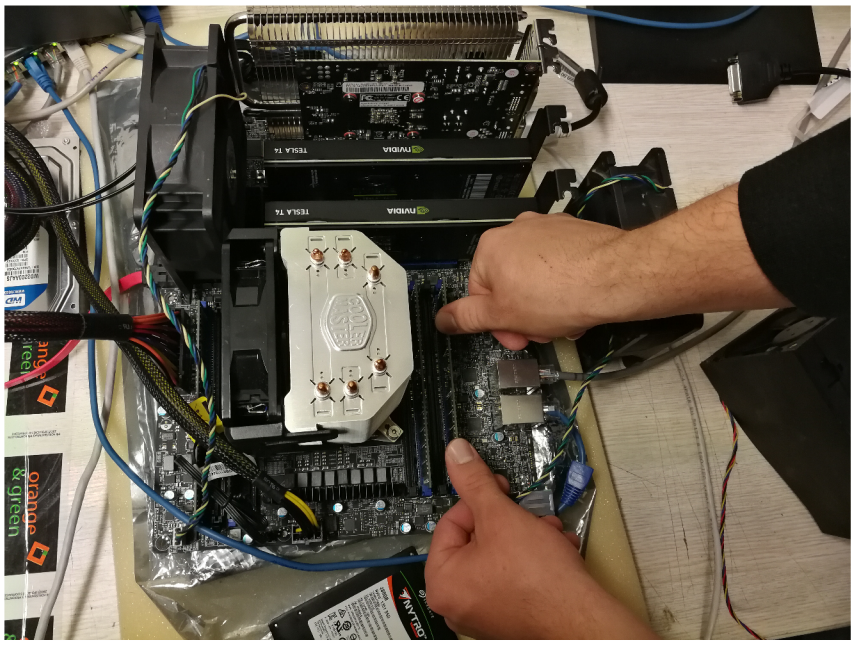
Operačná pamäť

Operačná pamäť systému je tvorená štyrmi 16 GB modulmi s frekvenciou 2400 MHz. Postupne boli tieto moduly odoberané, a tým sa znižovala veľkosť operačnej pamäte. V našom testovacom programe vplyv na výpočtový čas nebol rozdielny, rádovo sa tento rozdiel pohyboval v minútach. Výsledky aj s priebežnými hodnotami sú zapísané v tabuľke 3.6 a zobrazené v grafe na obr.3.6. Na obr. 3.5 je pozorovateľné zapojovanie pamäťového modulu.

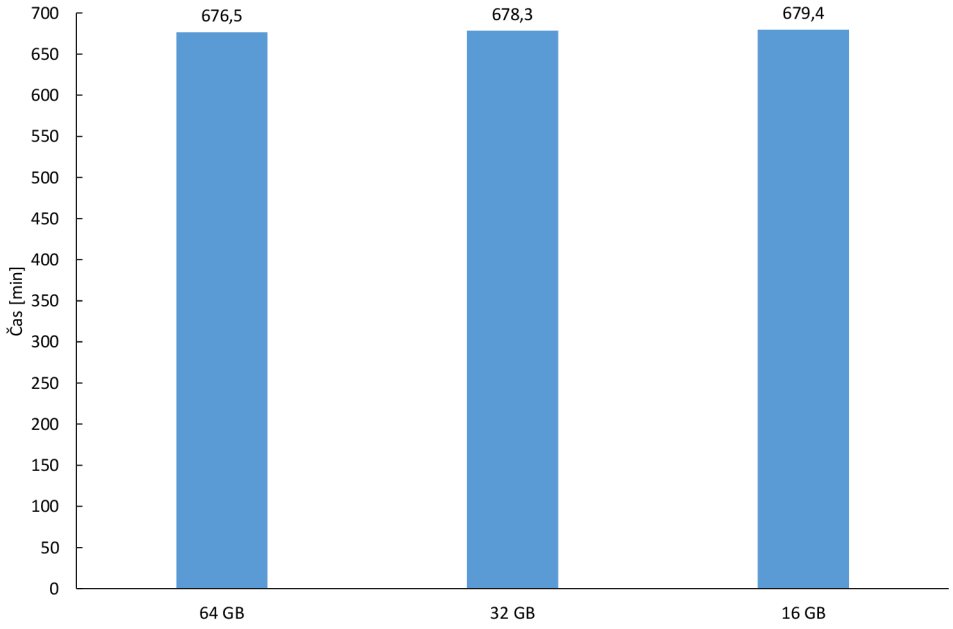
Tab. 3.6: Vplyv veľkosti operačnej pamäte na výpočet siete

Veľkosť operačnej pamäte [GB]	Stav po 1 hodine [epocha/krok]	Stav po 2 hodinách [epocha/krok]	Stav po 3 hodinách [epocha/krok]	Celkové trvanie [min]
64	5/1208	9/1680	13/1397	676,5
32	5/985	9/1733	13/1843	678,3
16	5/956	9/1749	14/484	679,4

V tejto časti sa ukázalo, že veľkosť operačnej pamäte nemala vplyv na náš testovací program pre výpočet neurónových sieti. 16 GB modul je dostačujúci na procesorovú komunikáciu.



Obr. 3.5: Zapojenie operačnej pamäte

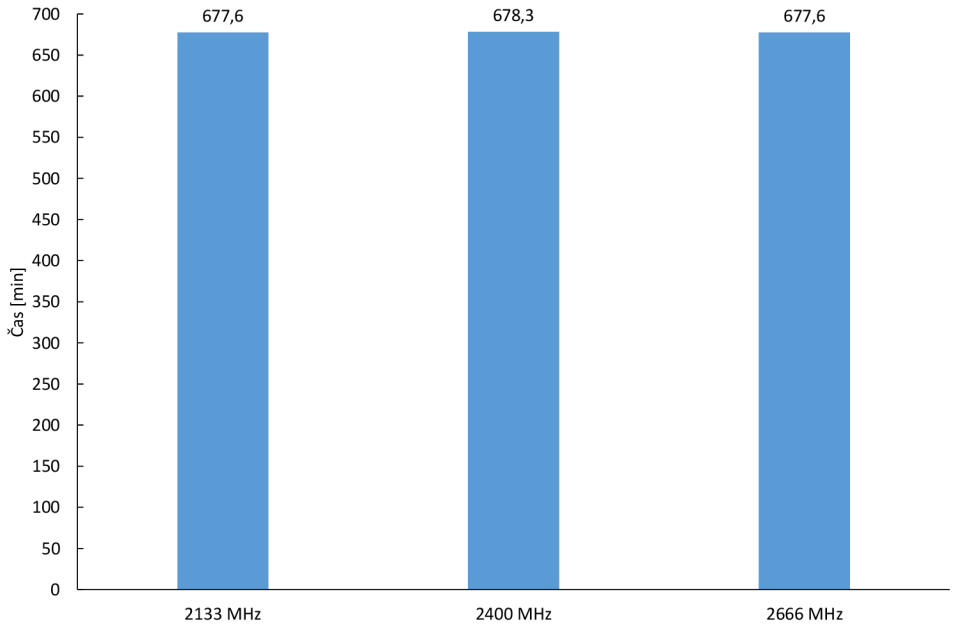


Obr. 3.6: Graf znázorňujúci vplyv veľkosti RAM na dobu počítania

V teoretickej časti záverečnej práce bolo spomenuté vplyv frekvencie na výpočet neurónovej siete. Nasledujúca časť sa zaoberá práve týmto problémom. Kde podľa dostupnosti boli testované 16 GB moduli RAM s rôznymi frekvenciami. V tabuľke 3.7 sú zobrazené konkrétne frekvencie spolu s výsledkami.

Tab. 3.7: Vplyv frekvencie operačnej pamäte na výpočet siete

Frekvencia pamäte [MHz]	Stav po 1 hodine [epocha/krok]	Stav po 2 hodinách [epocha/krok]	Stav po 3 hodinách [epocha/krok]	Celkové trvanie [min]
2133	5/972	9/1999	14/902	677,6
2400	5/956	9/1733	13/1843	678,3
2666	5/1043	9/1999	14/887	677,6



Obr. 3.7: Graf zobrazujúci vplyv frekvencie RAM na výpočet siete

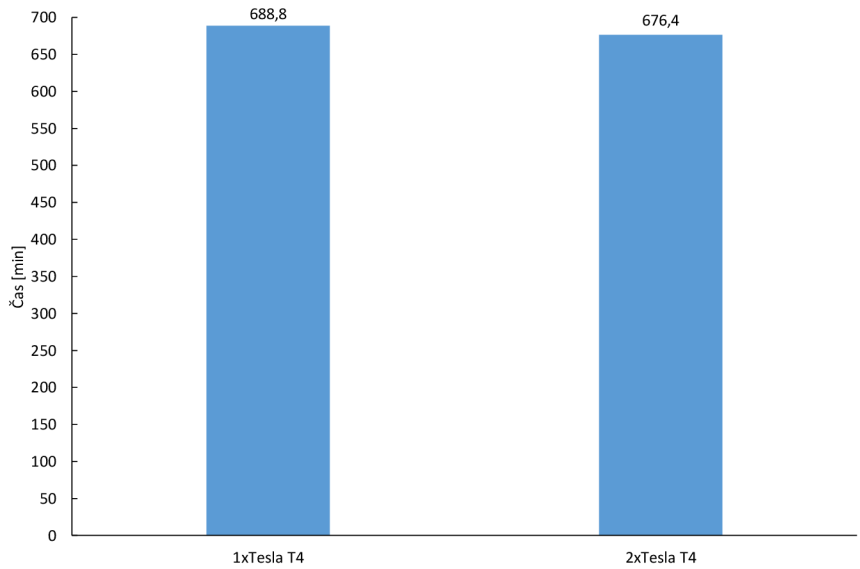
Vplyv frekvencie pamäte nebol v mojom systéme dokázaný. Pri výbere správnych modulov je potrebná komunikácia s matičnou doskou. Vplyv frekvencie by sa pravdepodobne ukázal vo väčších a komplexnejších neurónových sieťach.

Jedna GPU vs. Viac GPU

Táto časť testovania sa zaoberá rozdielom medzi použitím jednej grafickej karty a viacerých grafických kariet. Použitie viacerých grafických kariet sa využíva za účelom zvyšovania výkonu. V tejto časti testu boli k dispozícií dve grafické karty NVIDIA Tesla T4. NVIDIA Tesla T4 su jedny z najnovších kariet na trhu. Tieto karty sú postavené na architektúre Touring. Operačná pamäť týchto kariet je 16 GB. Nevýhodou týchto kariet je ich serverové riešenie a teda nedisponujú aktívnym chladením. Tento problém bolo treba vyriešiť, aby nedošlo k poškodeniu karty tzv. vzduchovým tunelom, ktorý je vidno na obr.3.9. Veľkou výhodou tejto karty je výkon 70 W, ktorý je naozaj minimálny. Ďalšou výhodou tejto karty je jej veľkosť, rozmery tejto karty sú minimálne a narozdiel od väčšiny herných grafických kariet alebo od NVDIA Tesla V100, tak NVIDIA Tesla V4 zaberá len jednu pozíciu. Karta disponuje 2560 CUDA jadrami a 320 Tenzorovými jadrami. V tomto prípade pre niektoré výpočty, ktoré využívajú CUDA jadrá, môže byť táto karta nevýhodná. V tabuľke 3.8 sú zobrazené výkony pri použití jednej a dvoch kariet.

Tab. 3.8: Tabuľka porovnávajúca hodnoty 1 vs. 2 GPU NVIDIA Tesla T4

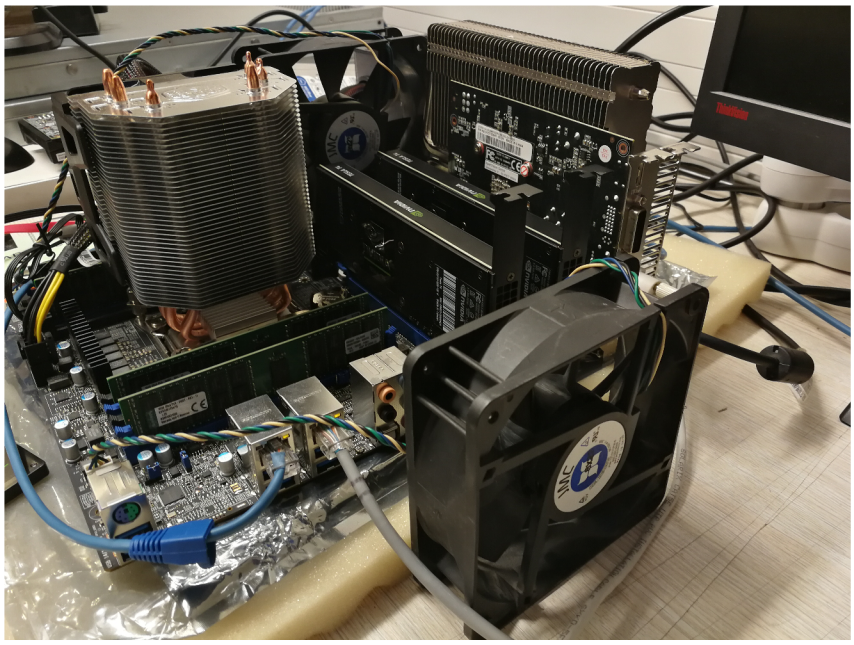
Počet grafických kariet	Stav po 1 hodine [epocha/krok]	Stav po 2 hodinách [epocha/krok]	Stav po 3 hodinách [epocha/krok]	Celkové trvanie [min]
1	5/839	9/1809	14/304	688,8
2	5/1002	9/1802	14/613	676,4



Obr. 3.8: Graf zobrazujúci porovnanie viacerých použitých kariet

Z grafu na obr. 3.8 je viditeľné, že pri využití viacerých grafických kariet je rýchlosť výpočtov menšia. V našom prípade na 50 epochách sa čas skrátil približne o 12 minút. Toto číslo môže však exponenciálne rásť s použitým grafických kariet s rastúcou náročnosťou siete.

Z tohoto pohľadu na graf vyplýva, že použitím o jednu kartu naviac získame necelé dve percenta navýšenia. V tomto prípade nesúhlasím s výsledkami a implementovanie Keras, modelu multi_GPU, bolo pre ResNet nevhodné. Nakoľko toto nebolo cieľom diplomovej práce a tieto karty boli zapožičané len na krátku dobu toto meranie nebolo možné opakovať.



Obr. 3.9: Obrázok znázorňujúci aktívne chladenie pre pasívne karty

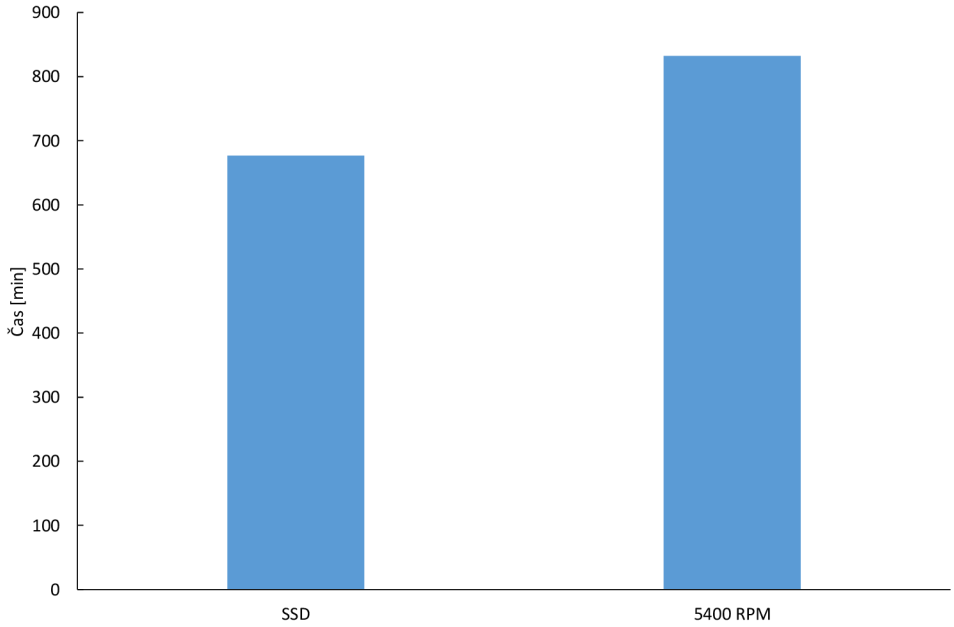
SSD vs HDD

Táto časť sa venuje porovnávaniu klasického rotačného disku a disku typu solid-state disk. Kvôli presnému meraniu som chcel pôvodný disk, na ktorom bežala väčšina testov SSD len nakopírovať na rotačný disk. Nakoľko sa mi tento krok nepodarilo spraviť, rotačný disk bol preto nainštalovaný s novou inštaláciou. Inštalácia bola rovnaká ako predchádzajúca inštalácia, preto v tab. 3.9 nie je rozdiel spôsobený inštaláciou.

Tab. 3.9: Vplyv rýchlosti disku na výpočet siete

Typ disku	Stav po 1 hodine [epocha/krok]	Stav po 2 hodinách [epocha/krok]	Stav po 3 hodinách [epocha/krok]	Celkové trvanie [min]
SSD	5/1208	9/1680	13/1397	676,5
5400 RPM	5/1335	10/1346	16/1223	832,4

Rotačný disk sa po prvej hodine javil veľmi podobne ako SSD, avšak ďalej už začal strácať výpočtovú rýchlosť a jeho doba počítania sa tým predlžovala. Doba počítania sa pravdepodobne líšila kvôli prístupu k dátam na disku. Rotačný disk nemusí mať uložené trénovacie a validačné dáta za sebou, ako to je pri SSD disku, a preto prístup k ním zaberie viac času.



Obr. 3.10: Graf porovnávajúci rotačný disk a solid-state disk

Rozdiel v diskoch po 50 epochách je viditeľný. Tento čas s pribúdajúcou náročnosťou môže veľmi rýchlo rásť, ak nebude použitý efektívny typ disku.

3.3 Spotreba systému

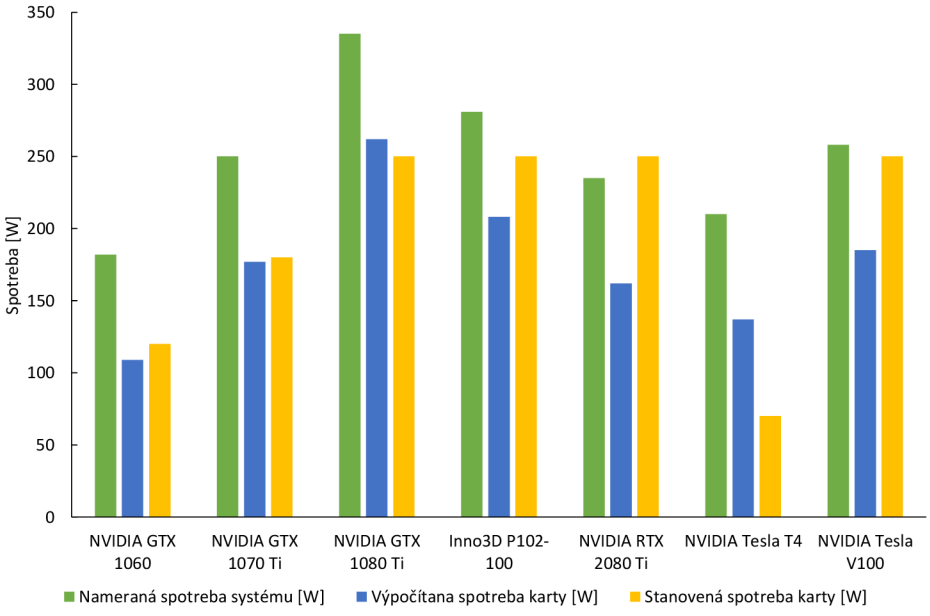
Priemerná spotreba bola meraná klasickým meračom spotreby, kde bol pripnutý celý zdroj celého systému, preto tieto výsledky sú merané na celú záťaž systému. Najmenšiu spotrebu ukázala najmenšia karta NVIDIA GTX 1060, ktorej však tieto výpočty aj trvali najdlhšie. Druhú najnižšiu spotrebu v modeli sieti Inception V3 ukázala o niečo väčšia grafická karta a to NVIDIA GTX 1070 Ti, ktorá vo výpočtoch dosiahla tretie miesto. Najvyššiu spotrebu dosiahla NVIDIA GTX 1080Ti a to 262 W. Táto výkonnostná hodnota mala za následok druhý najlepší výsledok v celkovej dobe počítania.

Kvôli špeciálnemu napájaciemu konektoru na NVIDIA Tesla V100, ktorý nebol na pôvodnom napájacom zdroji musel byť pridaný ešte jeden zdroj. Napriek dvom použitým zdrojom NVIDIA Tesla V100 mala spotrebu 258 W, čo je menej ako NVIDIA GTX 1080Ti aj Inno3D P102-100. Nevýhodou konštrukcii kariet NVIDIA Tesla

V100 a NVIDIA Tesla T4 je je pasívne chladenie. Nemá žiadne aktívne chladenie, preto bolo potrebné dodatočnými ventilátormi doplniť prietok vzduchu cez pasívny chladič, ktorý je stavaný na serverové používanie. Ostatné karty mali vlastné aktívne chladenie, no aj pri nich boli použité dodatočné ventilátory.

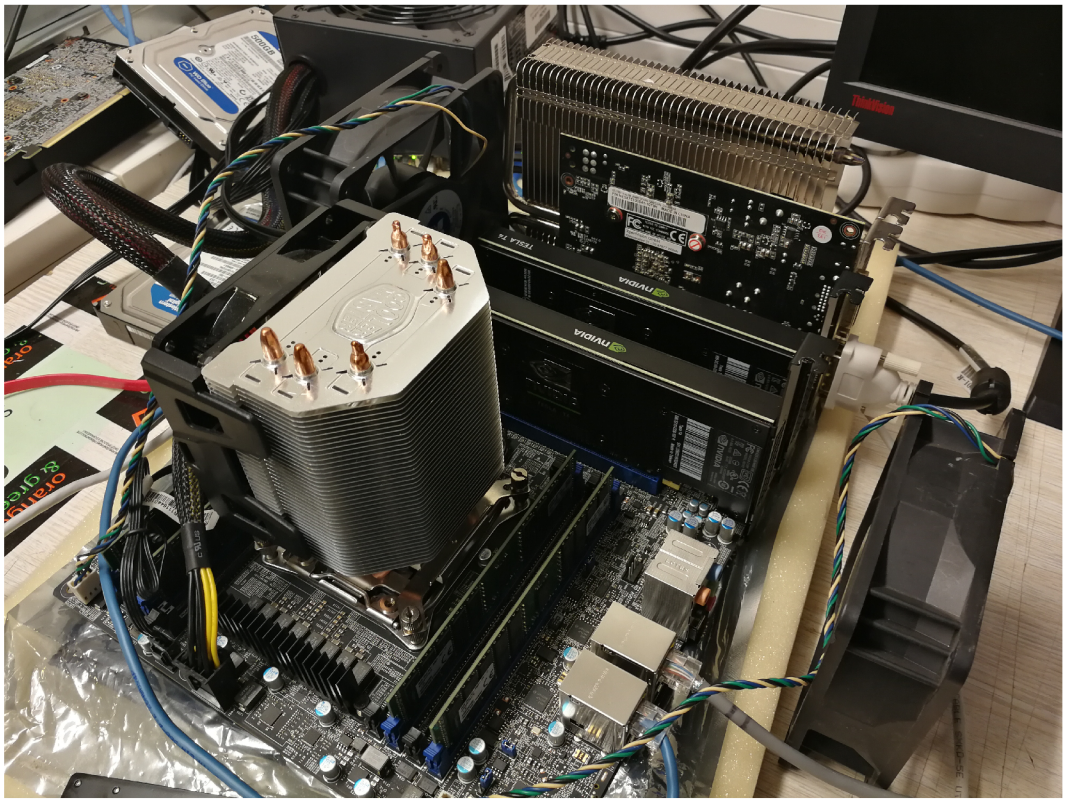
Tab. 3.10: Spotreba systému

Grafická karta	Nameraná spotreba systému [W]	Vypočítaná spotreba karty [W]	Stanovená spotreba karty [W]
NVIDIA GTX 1060	182	109	120
NVIDIA GTX 1070 Ti	250	177	180
NVIDIA GTX 1080 Ti	335	262	250
Inno3D P102-100	281	208	250
NVIDIA RTX 2080 Ti	235	162	250
NVIDIA Tesla T4	210	137	70
NVIDIA Tesla V100	258	185	250



Obr. 3.11: Graf znázorňujúci spotrebu grafických kariet

Aktuálnu spotrebu kariet sme vedeli zistiť aj pomocou príkazu v systéme `nvidia-smi`, avšak tento údaj sa veľmi rýchlo menil vzhľadom na aktuálne vyťaženie karty. Spotrebu karty bolo teda vhodnejšie dopočítať. Priemerná hodnota systému bez záťaže bola 73 W. Táto hodnota bola teda odpočítaná od celkovej hodnoty systému. Reálne hodnoty s hodnotami stanovenými výrobcami sú porovnané na obr. 3.11.



Obr. 3.12: Obrazok pracovnej stanici

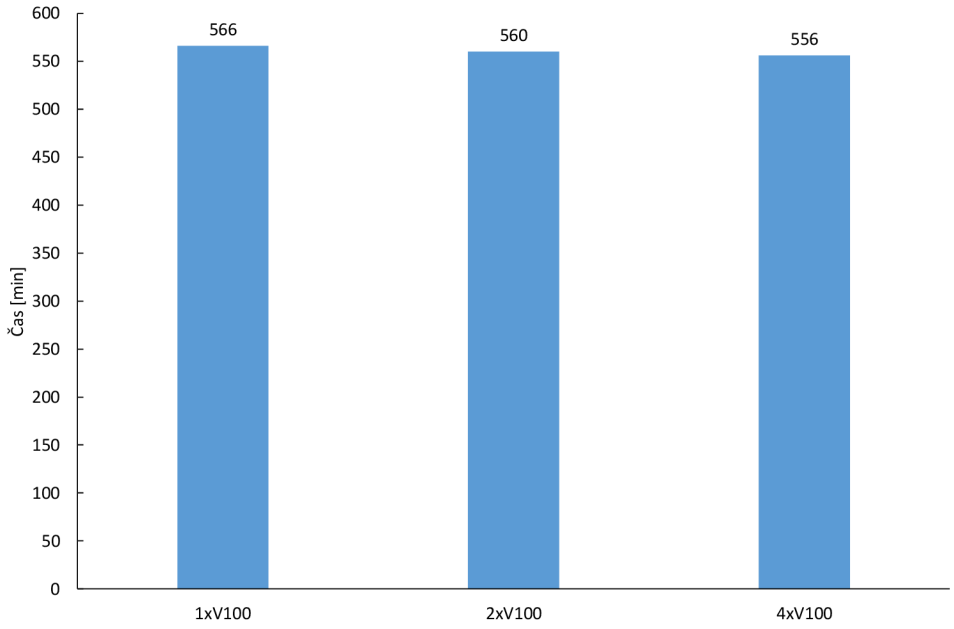
3.4 NVIDIA DGX Station

NVIDIA DGX Station, pracovná stanica od firmy NVIDIA ktorá bola detailnejšie popísaná v teoretickej časti, bola zapožičanú len na veľmi krátku dobu. V tejto časti som použil testovací kód s modelom ResNet a postupne bol aplikovaný na jednu, dve a štyri karty NVIDIA Tesla V100. Testovací kód pre multi GPU bol rovnaký ako pri NVIDIA Tesla T4. Avšak toto nebolo cieľom mojej práce a preto som tomu nevenoval veľkú pozornosť. Program bol doplnený a model, ktorý využíva viacero kariet, podľa manuálu na oficiálnych stránkach [1].

[1] www.keras.io

Tab. 3.11: Testovanie NVIDIA DGX Station

Počet GPU NVIDIA Tesla V100	Stav po 1 hodine [epocha/krok]	Stav po 2 hodinách [epocha/krok]	Stav po 3 hodinách [epocha/krok]	Celkové trvanie [min]
1xV100	4/493	7/907	10/1330	566
2xV100	4/415	7/888	10/1217	560
4xV100	4/364	7/784	10/1187	556



Obr. 3.13: Graf znázorňujúci dobu počítania na NVIDIA DGX Station

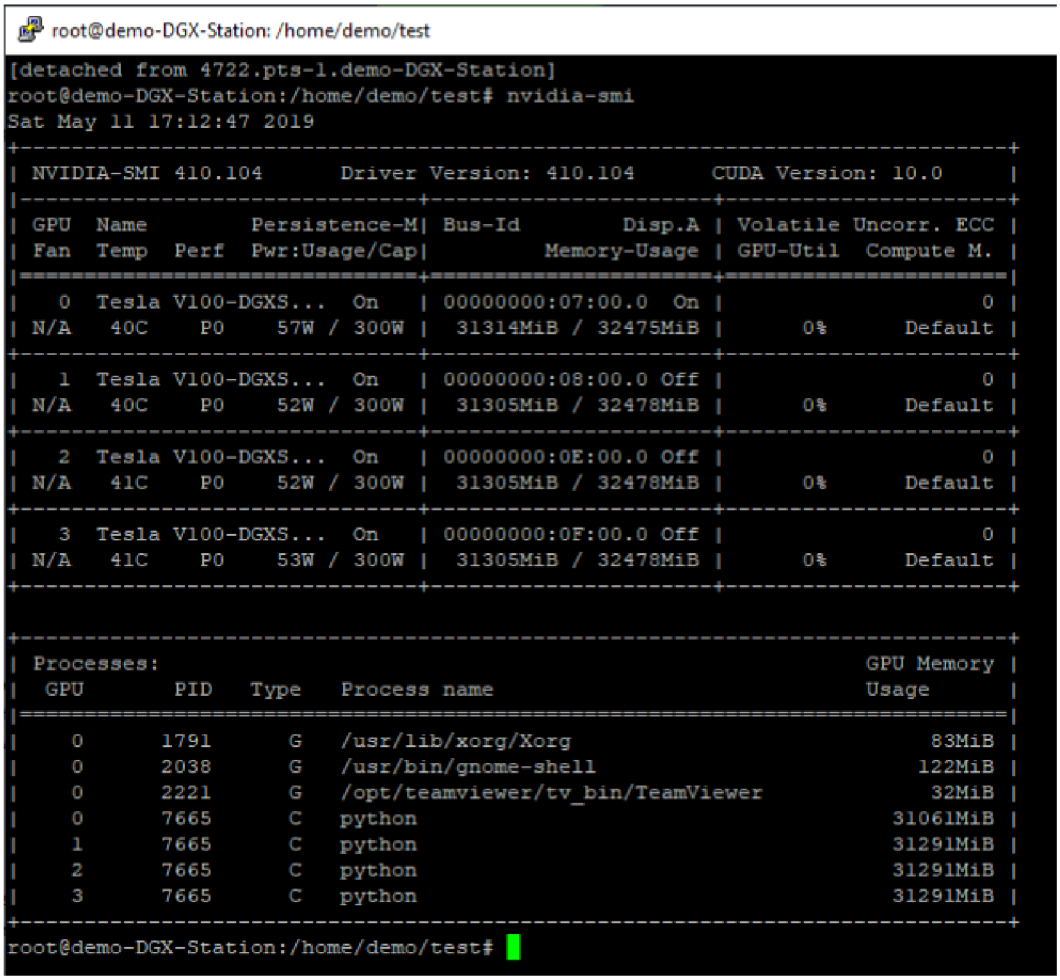
V dôsledku veľmi krátkej doby na testovanie bol použitý rovnaký kód ako pri NVIDIA Tesla V4. Ako bolo spomenuté tento kód nebol správne optimalizovaný pre použitie viacerých kariet. Z tohoto dôvodu je vidieť v tab. 3.11 a na obr. 3.13 že výsledky pri použitý jednej a viacerých kariet sú veľmi podobné. Preto výsledky tohto testovania považujem za nerelevantné.

Obr. 3.14: NVIDIA DGX Station

3.5 Pamäťové požiadavky

Práca sa taktiež zaoberala pamäťovými požiadavkami na karty. Táto časť bola skúmaná pomocou príkazu `nvidia-smi`, ktorý je k dispozícií v operačnom systéme po nainštalovaní aktuálneho ovládača pre danú kartu. Na obr.3.15 je vidieť výpis príkazu s NVIDIA DGX Station. Príkaz je z testovania štyroch GPU.

Model ResNet aj napriek nastavenej veľkosti `batch_size`= 64 si pre každú kartu alokoval celú pamäť. Tento prípad nastal aj pri modeli Inception V3 pri všetkých použitých kartách.



Obr. 3.15: Výpis obrazovky po príkaze nvidia-smi

4 Výsledky a diskusia

Výkon grafických kariet

Pri výkone kariet si najlepšie stála najdrahšia karta NVIDIA Tesla V100. Táto karta disponuje najviac CUDA jadrami. Karta disponuje aj tenzorovými jadrami, avšak aj keď vysoko úrovňový Keras by mal tieto jadra využívať automaticky. Pravdepodobne pri testovaní sa tak nestalo, pretože čas, ktorý dosiahla je veľmi podobný s časom NVIDIA GTX 1080 Ti, ktorá neobsahuje tenzorové jadrá. Na niektoré typy výpočtov sa hodí práve tento typ karty a a nie je potrebná zadovážiť si kartu s vysokou hodnotou. Táto herná grafická karta mala druhý najkratší čas počítania. Najmenšia karta bola veľmi slabá a doba jej počítania bola príliš dlhá.

Po piatich hodinách výpočtov, herná karta NVIDIA GTX 1080 Ti, dosiahla najmenšiu presnosť okrem NVIDIA GTX 1060. Túto presnosť vo zvyšných výpočtoch dobehla a dosiahla najmenšiu odchýlku od priemernej hodnoty.

Operačná pamäť

Vplyv operačnej pamäte sa v mojej práci neukázal významne rozdielny. Pri použití rozdielnych frekvencií pamäti sa zmeny ukázali minimálne. Pri výbere frekvencie pamäte je dôležitá správna komunikácia s matičnou doskou.

Veľkosť operačnej pamäte sa tak isto v moje práci ukázala ako nedôležitá. Táto operačná pamäť len komunikovala s CPU a preto pri veľkosti 16, 32 a 64 GB nebola zmena v časovej náročnosti výpočtov.

Multi GPU

Pri použití frameworku Keras a implementácií modelu pre multi GPU bol použitý návod s oficiálnych stránok. Ani tento návod však nepomohol a pri použití ďalšej karty sa rýchlosť výpočtu zvýšila o necelé 2%. Vysoko úrovňoví framework Keras pre tento typ úlohy pravdepodobne nepodporoval využitie viacerých kariet. Pri správnej implementácií by to výpočet mohlo zvýšiť niekoľko násobne.

Disk

V porovnaní klasického rotačného disku a disku typu SSD sa určite oplatí investovať do SSD disku. Tento typ disku niekoľko násobne urýchli proces učenia. SSD disk taktiež znižuje spotrebu systému.

Rotačný disk je o niekoľko percent pomalší z dôvodu pomalšieho čítania a zápisu. Tak isto uložené dáta na rotačnom disku sú uložené na rotačnej platni náhodne a trvá

istú dobu, kým tieto dáta nájde a následne načíta do pamäte karty. V opačnom prípade SSD disk tieto dáta ukladá na pamäťové čipy a prístup k ním je rýchlejší.

Spotreba

Pri spotrebe systému je pozorovateľné pri niektorých kartách pomerne veľká odchýlka od výrobcom udávanej hodnoty. Táto zmena však môže byť spôsobená aj nepresnosťou merania, keďže karta nemusela byť práve v 100% záťaži.Túto chybu som sa snažil eliminovať na minimum a vždy som hodnotu odčítal v polovici epochy. Najnižšiu spotrebu dosiahla najmenšia karta, ktorá ale nedisponovala dostatočným výkonom. Ďalšiu najlepšiu spotrebu s dostatočným výkonom dosiahla karta NVIDIA Tesla T4, aj keď sa reálna spotreba od udávanej líši, dosiahla veľmi dobrú spotrebu k jej výkonom. Najhoršie dopadla NVIDIA GTX 1080 Ti, ktorá dosiahla väčšiu spotrebu ako udávanú výrobcom.

Najlepšia varianta čo sa týka zriaďovacej ceny a spotreby je z môjho pohľadu NVIDIA GTX 1070 Ti. Táto karta disponuje nízkou spotrebou, pomerne vysokým výkonom a slušnou cenou.

Pamäťové požiadavky

Parameter `batch_size`, ovplyvňuje veľkosť potrebnej pamäte. Tento parameter, určuje koľko prvkov sa nahrá do pamäte karty natrénuje sa a znova sa zoberie taký isto počet prvkov. Napriek nastavenej hodnote 64 si program alokoval celú pamäť karty, či už išlo o 3 GB pri NVIDIA GTX 1060 alebo o 4x32 GB pri NVIDIA DGX Station.

5 Záver

Záverečná práca je zameraná na zoznámenie sa s neurónovými sieťami pre prostredie superpočítača. Práca sa zaoberala grafickými kartami, ktoré sa v dnešnej dobe využívajú na tento proces. Samotné neurónové siete je pomerne ťažko generovať a preto sú nám k dispozícií určité druhy frameworkov, ktoré sú akýmisi pred pripravenými sieťami a programátori sa tak zameriavajú na dôležitejšie detaily.

Práca sa v teoretickej časti zaoberá úvodom do neurónových sieti a používanými frameworkami. Najznámejšie frameworky boli rozobraté a popísané, v ktorých typoch neurónových sieti sa ich oplatí použiť. Teoretická časť sa taktiež zaoberá hardvérovými možnosťami riešenia NN. Jednou z možností je použitie superpočítača, ktorý je veľmi náročný na použitý hardvér, miesto v ktorom je HPC uložený a taktiež je pomerne náročný na elektrickú energiu. Ďalšou komplexnou možnosťou počítania je NVIDIA DGX systém, ktorý využíva najmodernejšie vedecké karty NVIDIA Tesla V100. Táto karta je najmodernejšia karta na dnešnom trhu a vďaka je architektúre Volta disponuje tenzorovými jadrami, ktoré proces s neurónovými sieťami niekoľko násobne urýchlia.

Hlavným prínosom práce je optimalizácia hardvérovej konfigurácie pre výpočet neurónových sieti. Teoretická časť popisuje neurónové siete, frameworky hlbokého učenia a hardvérové možnosti. Ďalšia časť práce sa venuje implementácií výkonnostných testov, ktoré zahŕňajú aplikovanie modelov Inception V3 a ResNet. Modely siete sú aplikované na rôzne grafické karty a výpočetný hardvér. Výstupom diplomovej práce je implementovaný model siete Inception V3, ktorý skúma grafické karty a ich výkon, časovú náročnosť výpočtov a ich efektivitu. Model siete ResNet je aplikovaný do časti, ktorý skúma ostatné vplyvy na výpočet neurónových sieti ako použitý disk, operačná pamäť a pod. Každá praktická časť obsahuje diskusiu, kde sú vysvetlené poznatky k danej časti. V prípade merania spotreby bol identifikovaný nesúlad medzi deklaráciou výrobcom a nameranými hodnotami.

V pomere cena a spotreba najlepšie hodnoty vykazovala grafická karta NVIDIA GTX 1070 Ti. Táto karta dosiahla podobný výkon ako niektoré drahšie a výkonnejšie karty. Ak je dôležitým parametrom čas, je potrebné siahnuť pre kartu NVIDIA GTX 1080Ti, táto karta dosiahla rýchlejšie výsledky ako NVIDIA GTX 1070 Ti.

Najdrahšia testovaná karta NVIDIA Tesla V100 a jej cena je pre mnohých neprijateľnou, no jej výrobcovia vedia, čo táto karta dokáže ponúknuť. Pri nevyužitých Tenzorových jadrách karta dosiahla podobnej presnosti za rýchlejší čas ako iné testované karty, avšak pri využití svojho potenciálu by táto karta dosiahla niekoľko násobne lepšie výsledky oproti iným testovaným kartám.

Konkrétne pamäťové požiadavky sa nepodarilo spracovať, pri spustenom procese si modely sieti alokovali celú operačnú pamäť systému.

Literatúra

[1] KVASNIČKA, Vladimír. *Úvod do teórie neurónových sietí.* , Slovenská republika: IRIS, 1997. ISBN 80-887-7830-1.

[2] SINČÁK, Peter a Gabriela Andrejková. *Neurónové siete Inžiniersky prístup* [online], [cit. 10. 10. 2018] Dostupné z URL:<http://www.student.ui.fei.tuke.sk/ZNS?action=AttachFile&do=view&target=NS1.pdf

[3] STERGIOU, Christos, a Dimitrios Siganos. *NEURAL NETWORKS* [online], [cit. 10. 10. 2018] Dostupné z URL:<https://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html#Conclusion

[4] Ian Goodfellow, Yoshua Bengio a Aaron Courville. *Deep Learning.* 2016 MIT Press

[5] GURNEY, Kevin. *An introduction to neural networks*, University of Sheffield. ISBN 0-203-45151-1

[6] LI, Chao, Yi Yang, Min Feng a Srimat Chakradhar *Optimizing Memory Efficiency for Deep Convolutional Neural Networks on GPUs* [online], [cit. 10. 10. 2018] Dostupné z URL:<https://arxiv.org/ftp/arxiv/papers/1610/1610.03618.pdf

[7] HE, Kaiming, Xiangyu ZHANG, Shaoqing REN a Jian SUN. *Deep Residual Learning for Image Recognition.* 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016, 2016, , 770-778. DOI: 10.1109/CVPR.2016.90. ISBN 978-1-4673-8851-1.Dostupné z URL:<http://ieeexplore.ieee.org/document/7780459/

[8] SZEGEDY, Christian, WEI LIU, YANGQING JIA, et al. *Going deeper with convolutions.* 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015, 2015, , 1-9. DOI: 10.1109/CVPR.2015.7298594. ISBN 978-1-4673-6964-0.Dostupné z URL:<http://ieeexplore.ieee.org/document/7298594/

[9] SZEGEDY, Christian, Vincent VANHOUCKE, Sergey IOFFE, Jon SHLENS a Zbigniew WOJNA. *Rethinking the Inception Architecture for Computer Vision.* 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016, 2016, , 2818-2826. DOI: 10.1109/CVPR.2016.308. ISBN 978-1-4673-8851-1. Dostupné z URL:<http://ieeexplore.ieee.org/document/7780677/

[10] HU, Jie, Li SHEN a Gang SUN. *Squeeze-and-Excitation Networks*. 2018 IEEE-/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018, 2018, , 7132-7141. DOI: 10.1109/CVPR.2018.00745. ISBN 978-1-5386-6420-9. Dostupné z URL:<https://ieeexplore.ieee.org/document/8578843/

[11] Sun, Shuyang and Pang, Jiangmiao a Shi, Jianping and Yi, Shuai and Ouyang, Wanli. *FishNet: A Versatile Backbone for Image, Region, and Pixel Level Prediction*. 2018 Advances in Neural Information Processing Systems 31. Pages 754-764. Publikoval: Curran Associates, Inc. Dostupné z URL:<https://bit.ly/2PTjN3C

[12] FELICE, De Mitch. *Which deep learning network is best for you?* Published: IDG Contributor Network.Dostupné z URL:<https://www.cio.com/article/3193689/artificial-intelligence/which-deep-learning-network-is-best-for-you.html

[13] ABADI Martin, Paul Barham, Jianmin Chen, Zhifeng Chen,. *TensorFlow: A System for Large-Scale Machine Learning*, 2016 ISBN 978-1-931971-33-1.

[14] GULLI, Antonio. *Depp Learning with Keras*, Packt Publishing Limited, 2017 ISBN 9781787128422.

[15] *Deep learning with Python: a hands-on introduction*. New York, NY: Springer Science Business Media, 2017. ISBN 978-1-4842-2765-7.

[16] JAWANDHIYA. Pooja. *HARDWARE DESIGN FOR MACHINE LEARNING* 63-84. DOI: 10.5121/ijaia.2018.9105. ISSN 09762191. .Dostupné z URL:<http://aircconline.com/ijaia/V9N1/9118ijaia05.pdf

[17] Intel AI *Intel AI* Dostupné z URL:<https://www.intel.ai//

[18] JO, Gangwon, Jungho PARK a Jaejin LEE. *Using Gaming GPUs for Deep Learning*. 2018 HPC Asia 2018, January 2018, Tokyo, Japan

[19] NVIDIA. *Deep Learning AI* Dostupné z URL:<https://www.nvidia.com/en-gb/deep-learning-ai/

[20] AMD *AMD* Dostupné z URL:<https://www.amd.com//

[21] Google Cloud *Google Cloud Platform* Dostupné z URL:<https://cloud.google.com

[22] *Machine Learning - Advanced Techniques and Emerging Applications*. InTech, 2018. ISBN 978-1-78923-752-8. Dostupné

z URL:<http://www.intechopen.com/books/machine-learning-advanced-techniques-and-emerging-applications/hardware-accelerator-design-for-machine-learning

[23] DETTMERS, Tim. *Which GPU(s) to Get for Deep Learning: My Experience and Advice for Using GPUs in Deep Learning* [online], [cit. 10. 10. 2018] Dostupné z URL:<http://timdettmers.com/2018/11/05/which-gpu-for-deep-learning/

[24] EADLINE, Douglas PhD. *High Performance Computing for Dummies* [online], [cit. 10. 10. 2018]. Wiley Publishing, Inc. ISBN: 978-0-470-49008-2. Dostupné z URL:<http://hpc.fs.uni-lj.si/sites/default/files/HPC_for_dummies.pdf

[25] GRAHAM, Susan L, Marc SNIR a Cynthia A PATTERSON. *Getting up to speed: the future of supercomputing.* Washington, DC: National Academies Press, c2005. ISBN 03-090-9502-6.2.

[26] TOP 500 *TOP 500 LIST* Dostupné z URL:<https://www.top500.org/lists/2018/11/

[27] Bc. KUVIK, Michal. *Rozpoznávání druhu jídla s pomocí hlubokých neuronových sítí.* 2018

Zoznam symbolov, veličín a skratiek

AI	Umelá inteligencia – Artificial Intelligence
AMT	Asynchrónne veľa úloh –Asynchronous Many-Tasks
ANN	Umelá neurónová sieť – Artificial Neural Network
API	Programovacie prostredie aplikácií –Application Programming Interface
ARM	Pokročilá znížená sada inštrukcií počítača – Advanced RISC Machines
CPU	Centrálna riadiaca jednotka –Central processing unit
DNN	Hlboká neurónová sieť – Deep Neural Network
FLOPS	Operácie s pohyblivou rádovou čiarkou za sekundu – Floating Point Operations Per Second
GAN	Generatívna nepriateľská sieť – Generative Adversarial Network
GPU	Grafická riadiaca jednotka –Graphics processing unit
HPC	Vysoko-výkonné počítanie –High-performance computing
IPMI	VInteligentná platforma pre rozranie manažmentu – Intelligent Platform Management Interface
LSTM	Dlhá krátkodobá pamäť – Long short-term memory
MIT	Massachusettský technologický inštitút – Massachusetts Institute of Technology
MKL	Matematická knižnica jadra – Math Kernel Library
NFS	Sieťový súborový systém – Network File System
NGC	NVIDIA GPU CLOUD
OCR	Open Community Runtime
OS	Operačný systém – Operating System
PCIe	Periférne prepojenie komponentu-expres – Peripheral Component Interconnect-Express
RAM	Náhodný prístup do pamäte – Random Access Memory
ReLU	Usmernená lineárna jednotka – The Rectified Linear Unit
RISC	Znížená sada inštrukcií – Reduced Instruction Set Computer
SSD	Mechanika s nepohyblivým médiom – Solid state drive
TPU	Tenzorová riadiaca jednotka –Tensor processing unit
VMDNN	Virtuálna pamäť GPU pre DNN – Virtual GPU Memory for DNNs

Zoznam príloh

A Testovací kód modelu Inception V3

```
from keras.applications.inception_v3 import InceptionV3,
preprocess_input
from keras.layers import GlobalAveragePooling2D, Dense,
BatchNormalization, Dropout
from keras.models import Model
from keras.preprocessing.image import ImageDataGenerator
from keras.metrics import top_k_categorical_accuracy
from keras.optimizers import SGD, Adam
import matplotlib.pyplot as plt
import json

def top_3_accuracy(y_true, y_pred):
    return top_k_categorical_accuracy(y_true, y_pred, k=3)



img_width, img_height = 256, 256
n_channels = 3
train_dir = '/home/test/data/train_set'
val_dir = '/home/test/data/val_set'

# 101 733
nb_train_samples = 101733
# 10 323
nb_val_samples = 10323
epoch = 30
batch_s = 16

model_base_name = 'model_base.h5'
history_base_name = 'model_history_base.json'
model_name = 'model_full.h5'
history_name = 'model_history_full.json'

train_datagen = ImageDataGenerator(
    rescale=1./255,
    rotation_range=15,
    width_shift_range=0.2,
    height_shift_range=0.2,
```

```
    zoom_range=0.2,
    horizontal_flip=True,
    vertical_flip=True)

val_datagen = ImageDataGenerator(rescale=1./255)

train_generator = train_datagen.flow_from_directory(
    train_dir,
    target_size=(img_width, img_height),
    color_mode='rgb',
    batch_size=batch_s,
    shuffle=True,
    class_mode='categorical')

val_generator = val_datagen.flow_from_directory(
    val_dir,
    target_size=(img_width, img_height),
    color_mode='rgb',
    batch_size=batch_s,
    shuffle=True,
    class_mode='categorical')

    base_model = InceptionV3(weights=None, include_top=False,
    input_shape=(img_width, img_height, n_channels))

x = base_model.output
x = BatchNormalization()(x)
x = GlobalAveragePooling2D()(x)
x = BatchNormalization()(x)
x = Dense(1024, activation='relu')(x)

predictions = Dense(211, activation='softmax')(x)

model = Model(inputs=base_model.input, outputs=predictions)

model.compile(optimizer='sgd', loss='categorical_crossentropy',
metrics=[top_3_accuracy])

history_base = model.fit_generator(
```

```
    train_generator,
    steps_per_epoch=nb_train_samples // batch_s,
    epochs=epoch,
    validation_data=val_generator,
    validation_steps=nb_val_samples //batch_s,
    max_queue_size=100,
    workers=100)

model.save(model_base_name)

with open(history_base_name, 'w') as f:
    json.dump(history_base.history, f)
```

B Testovací kód modelu ResNet

```
from keras.layers import GlobalAveragePooling2D, Dense,
BatchNormalization, Dropout
from keras.models import Model
from keras.preprocessing.image import ImageDataGenerator
from keras.metrics import top_k_categorical_accuracy
from keras.optimizers import SGD, Adam
import matplotlib.pyplot as plt
from keras import applications
import json

img_height, img_width = 128, 128
num_classes = 211
train_dir = "/home/test/data/train_set"
val_dir = "/home/test/data/val_set"

train_datagen = ImageDataGenerator(
    rescale=1. / 255,
    shear_range=0.2,
    zoom_range=0.2,
    horizontal_flip=True)

val_datagen = ImageDataGenerator(rescale=1. / 255)

train_generator = train_datagen.flow_from_directory(
    train_dir,
    target_size=(img_width, img_height),
    color_mode='rgb',
    batch_size=64,
    shuffle=True,
    class_mode='categorical')

val_generator = val_datagen.flow_from_directory(
    val_dir,
    target_size=(img_width, img_height),
    color_mode='rgb',
    batch_size=64,
    shuffle=True,
```

```
    class_mode='categorical')


# If imagenet weights are being loaded,
# input must have a static square shape (one of (128, 128),
(160, 160), (192, 192), or (224, 224))
base_model = applications.resnet50.ResNet50(weights=None,
include_top=False, input_shape=(img_height, img_width, 3))

x = base_model.output
x = GlobalAveragePooling2D()(x)
x = Dropout(0.7)(x)
predictions = Dense(num_classes, activation='softmax')(x)
model = Model(inputs=base_model.input, outputs=predictions)


# sgd = SGD(lr=lrate, momentum=0.9, decay=decay, nesterov=False)
adam = Adam(lr=0.0001)
model.compile(optimizer=adam, loss='categorical_crossentropy',
metrics=['accuracy'])


model.fit_generator(
    train_generator,
    steps_per_epoch=2000,
    epochs=50,
    validation_data=val_generator,
    validation_steps=800)
```