Hodnocení vedoucího závěrečné práce

České vysoké učení technické v Praze **Fakulta informačních technologií**

Student: Bc. Tomáš Peterka
Vedoucí práce: RNDr. Petr Škoda, CSc.
Název práce: Machine Learning in Astroinformatics Using Massively Parallel Data Processing
Obor: Znalostní inženýrství (magisterský)

Datum vytvoření: 28. 5. 2015

Hodnotící kritérium:	*Způsob hodnocení - následující škálou 1 až 5:*
1. Náročnost a další komentář k zadání	***1=mimořádně náročné zadání,*** *2=náročnější zadání, 3=průměrně náročné zadání, 4=lehčí, ale ještě dostatečně náročné zadání, 5=nedostatečně náročné zadání*

Popis kritéria:
Podrobněji charakterizujte diplomovou (bakalářskou) práci a její případné návaznosti na předchozí nebo běžící projekty. Dále posuďte, čím je zadání této ZP náročné. (U obtížnější ZP lze dále tolerovat některé nedostatky, které by u ZP standardní obtížnosti tolerovány nebyly; a naopak u jednoduché ZP mohou být zjištěné nedostatky hodnoceny přísněji.)

Komentář:
Zadání práce je velmi náročné. Oproti modelům strojového učení navrhovaným v zadání se student odvážně pustil do velmi málo probádané, avšak komerčně silně propagované metody mnohavrstvých konvolučních sítí. Cílem byla příprava technologie pro řešení našeho projektu zaměřeného na klasifikaci velkých spektroskopických archivů pomocí strojového učení. Kromě rešerše dostupné literatury (i velmi nové) a osobní instalace (s kompilací) mnoha SW balíků bylo třeba vytvořit i několik nových modulů (pro systém Caffe). Práce zahrnovala práci s paralelními implementacemi včetně praktického portování a instalaci na GPU s CUDA architekturou. Vlastní experimenty vyžadovaly mnoho času při porozumění a nastavování parametrů jednotlivých algoritmů jakož i praktickou práci s různými formáty astronomických spekter a jejich pre-processing a vizualizaci, na což si student vytvořil řadu malých procedur. Část experimentů probíhala i v gridovém prostředí systému DAME Neapolské univerzity, který byl na základě studentových požadavků výrazně vylepšen.

Hodnotící kritérium:	*Způsob hodnocení - následující škálou 1 až 4:*
2. Splnění zadání	***1=zadání splněno,*** *2=zadání splněno s menšími výhradami, 3=zadání splněno s většími výhradami, 4=zadání nesplněno*

Popis kritéria:
Posuďte, zda předložená ZP splňuje zadání. V komentáři uveďte body zadání, které nebyly zcela splněny, případně rozšíření ZP oproti původnímu zadání. Nebylo-li zadání zcela splněno, pokuste se posoudit závažnost, dopady a případně i příčiny jednotlivých nedostatků.

Komentář:
Zadání bylo plně splněno. Student vyčerpávajícím způsobem analyzoval veřejně (open source) dostupná řešení masivně paralelních technik strojového učení z hlediska zpracování opravdu velkých objemů dat na paralelních procesorech, přičemž se správně zaměřil hlavně na GPU, kam se v současnosti upírá většina nadějí. Největší potenciál pak identifikoval v metodách mnohavrstvých konvolučních sítí, které představují velmi nový a bouřlivě se rozvíjející obor strojového učení, kde však většina výzkumu je předmětem know-how významných komerčních subjektů jako Facebook či Google. O to cennější je ucelený přehled sestavený z rozptýlených většinou elektronických zdrojů, který podává první kapitola práce. V závěru pak po mnoha experimentech na dvou typech úloh z astronomie velmi přesvědčivě ukázal vysokou spolehlivost a výkon metod mnohavrstvých neuronových sítí. Podrobně diskutuje i časovou náročnost jednotlivých implementací a jejich slabiny (např. nevýhodnost poměrně komplikované vícevrstvé architektury pro řešení lineárně separovatelného problému). Student v průběhu řešení musel napsat některé části používané knihovny Caffe a značně přispěl svými pokusy k modifikaci gridového systému DAME, který byl na základě řešené práce značně vylepšen – i když bohužel nemohl být použit k přímému srovnání výkonu.

Hodnotící kritérium:	*Způsob hodnocení - následující škálou 1 až 4:*
3. Rozsah písemné zprávy	***1=splňuje požadavky,*** *2=splňuje požadavky s menšími výhradami, 3=splňuje požadavky s většími výhradami, 4=nesplňuje požadavky*

Popis kritéria:
Porovnejte rozsah předložené písemné zprávy s požadovaným rozsahem, viz Směrnice děkana č. 9/2011, článek 3. Pro hodnocení ZP je také důležité, zda všechny části písemné zprávy jsou informačně bohaté a pro práci nezbytné. Text ZP by neměl obsahovat zbytečné části.

Komentář:
Práce má požadovaný rozsah 60 stran textu, 2 strany příloh a asi 14 stran rejstříků a tiráží. Celkový počet stran je 76. Většina textu je nabitá informacemi a vyžaduje pečlivé a detailní čtení k pochopení poměrně nestandardní problematiky.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*

4. Věcná a logická úroveň práce — *100 (A)*

Popis kritéria:
Posuďte, zda předložená ZP je po věcné stránce v pořádku, případně vyskytují-li se v práci věcné chyby nebo nepřesnosti. Zhodnoťte dále logickou strukturu ZP, návaznosti jednotlivých kapitol a pochopitelnost textu pro čtenáře.

Komentář:
Práce je velmi přehledná a logická. Po krátkém úvodu do problematiky astronomických Velkých Dat, poskytuje vyčerpávající přehled moderních metod strojového učení schopných běhu i na GPU (včetně tak exotických jako jsou biologicky inspirované metody hejn či genetické algoritmy) , a řadí je podle snadnosti paralelizace, přičemž jsou uvedeny i údaje o zrychlení GPU verze oproti klasické . Těžiště práce ale spočívá v metodách typu deep learning – strojovém učení pomocí mnohavrstvých konvolučních sití, pro které jsou shrnuty dosavadní open source implementace. Dvě následující kapitoly pak podrobněji seznamují s poměrně neznámou problematikou konvolučních sítí a aspekty jejich paralelizace. Detailně jsou představeny použité technologie a terminologie architektury CUDA a knihovny Caffe, jež je předmětem dalšího zkoumání v závěrečných dvou praktických kapitolách. Zde se používají na dvou typech astronomických experimentů – klasifikaci pul miliónu objektů na základě fotometrických měření daných tabulkou a klasifikaci skoro dvou tisíc spekter o délce 2000 bodů. Tomu odpovídají i dvě odlišné architektury sítí, které student navrhl a odladil. Podrobně jsou v závěru diskutovány dosažené přesnosti klasifikace i rychlost zpracování na CPU i GPU.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
5. Formální úroveň práce	*99 (A)*

Popis kritéria:
Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku ZP, viz Směrnice děkana č. 9/2011, článek 3.

Komentář:
Práce je velmi příjemně čitelná, ilustrovaná množstvím grafů a tabulek a přehledná i přes velký podíl matematických výrazů (ve kterých jsou řádně vysvětleny a popsány proměnné). Experimenty jsou popsány formou řady tabulek. Za každou tabulkou se skrývá mnoho experimentů spojených s vyladěním parametrů k dosažení optimálních výsledků. Terminologie je korektně použitá a zkratky řádně vysvětleny. Práce je psaná pěknou srozumitelnou angličtinou, působí dojmem profesionálního dokumentu, který poslouží jako materiál pro budoucí odborný článek i jako stručný úvod do moderních metod strojového učení. Práce je psaná pečlivě a s invencí. Při čtení jsem nenašel překlepy. Nepatrnou výhradu mám vůči způsobu citace elektronických zdrojů, kde není konzistentně uváděno URL včetně datumu přístupu ve všech případech (ze zdrojového kódu práce je ale zjevně vidět nekonzistence poskytovatelů BiBTeX záznamů a nepružnost použitého bibliostylu)

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
6. Práce se zdroji	*98 (A)*

Popis kritéria:
Vyjádřete se k aktivitě studenta při získávání a využívání studijních materiálů k řešení ZP. Charakterizujte výběr studijních pramenů. Posuďte, zda student využil všechny relevantní zdroje nebo zda se pokoušel řešit již vyřešené problémy. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků a úvah, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami.

Komentář:
Myslím, že analýza dostupných zdrojů je vyčerpávající vzhledem k novosti problematiky. Reference na bohatou literaturu (36 pramenů) jsou konzistentní, u většiny je uvedeno i DOI a elektronický link. U elektronických odkazů není většinou uveden datum ověření (Accesed) a pro některé práce s Accessed zase chybí link, ale vzhledem k neexistenci jasného doporučení či předepsaného stylu v BibTeXu to nepovažuji za velký problém. Tématika je příliš nová a publikované zdroje jsou většinou v podobě on-line manuálů a webových stránek.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
7. Hodnocení výsledků, publikační výstupy a ocenění	*98 (A)*

Popis kritéria:
Vyjádřete se k úrovni dosažených hlavních výsledků ZP, např. k úrovni teoretických výsledků, nebo k úrovni a funkčnosti technického nebo programového vytvořeného řešení, apod. Případně také zhodnoťte, zda software nebo zdrojové texty, které nevytvořil sám student, byly v ZP použity v souladu s licenčními podmínkami a autorským právem. Popište případnou publikační činnost a získaná ocenění související s řešením této ZP.

Komentář:
Všechen použitý software je poskytován v rámci otevřených licencí GNU a jim podobných, stejně tak veškerá studentova práce je k dispozici pod licencí GPL. Dosažené výsledky jsou vynikající. Podařilo se prokázat vhodnost a dobrou paralelizovatelnost nejen v astronomii poměrně neznámé metody typu deep learning na mnohavrstvých neuronových sítích s konvoluční vrstvou při hledání emisních objektů ve velkých přehlídkách spekter a připravit modul spojující několik open-source balíků do ucelené aplikace nasaditelné v cloudovém prostředí. Výsledky projeku budou integrovány v brzké době do systému VO-CLOUD jako další model strojového učení . Práce bude zmíněna během prezentace příspěvku o VO-CLOUDu na mezinárodním workshopu Virtuální observatoře v Itálii v červnu 2015. Kromě toho bude část výsledků publikována v připravovaném odborném článku do některého recenzovaného astronomického časopisu.

Hodnotící kritérium:	*Způsob hodnocení - nehodnotí se*
8. Komentář o využitelnosti výsledků	

Popis kritéria:
Uveďte, zda hlavní výsledky ZP rozšiřují již publikované známé výsledky a/nebo přinášející zcela nové poznatky. Uveďte možnosti využití výsledků ZP v praxi.

Komentář:
Vytvořený SW bude v brzké době nasazen jako modul pro cloudový astronomický systém VO-CLOUD, vyvíjený na Astronomickém ústavu AVČR v Ondřejově ve spolupráci několika bývalých i současných studentů z FIT ČVUT. V brzké době na něm pak chceme provést hledání emisních objektů (Be hvězd a kvasarů) v archivu bezmála 3 milionů spekter přehlidky LAMOST. V jednaní je i nasazení vytvořeného softwaru jako modulu pro mnohem vetší gridový systém DAME vyvíjený na univerzitě v Neapoli (k jehož vylepšení už práce studenta vedla nyní). V jednání je zařazení vytvořených modulů do oficiální distribuce knihovny Caffe. Z vylepšení obou systémů bude mít prospěch celá profesionální veřejnost. Kromě toho je plánován vědecký článek.Výsledky jsou unikátní , neboť se jeví o první známý úspěšný experiment s mnohavrstvými sítěmi při klasifikaci astronomických spekter (konkurenční práci P. Hály student cituje – jakožto její oponent ji nepovažuji za srovnatelnou).

Hodnotící kritérium:	*Způsob hodnocení - následující škálou 1 až 5:*
9. Aktivita a samostatnost studenta v průběhu řešení	9a: ***1=výborná aktivita,*** *2=velmi dobrá aktivita,* *3=průměrná aktivita,* *4=slabší, ale ještě dostatečná aktivita,* *5=nedostatečná aktivita* 9b: ***1=výborná samostatnost,*** *2=velmi dobrá samostatnost,* *3=průměrná samostatnost,* *4=slabší, ale ještě dostatečná samostatnost,* *5=nedostatečná samostatnost*

Popis kritéria:
Posuďte, zda byl student během řešení aktivní, zda dodržoval dohodnuté termíny, jestli své řešení průběžně konzultoval a zda byl na konzultace dostatečně připraven (9a). Posuďte schopnost studenta samostatné tvůrčí práce (9b).

Komentář:
Student byl velmi aktivní při získávání informací, pravidelně se zúčastňoval konzultací cca každé 2-3 týdny a bleskově odpovídal na e-mailovou komunikaci, přes kterou probíhala velká část dodatečných konzultací spojených s astronomickou problematikou. Vetšinu materiálu použitých v analýze však našel sám. Několikrát se zúčastnil porad týmu grantového projektu řešeného naším ústavem ve spolupráci s MFF UK a VŠB-TU Ostrava, a také se seznámil s pořizováním a analýzou spekter na 2M dalekohledu na Ondřejově, jež byla klíčovou částí zadané práce. Spolupracoval i s dalšími členy týmu – mými PhD. studenty na MU a VUT Brno, na jejichž předzpracování spekter navazoval. Komunikoval samostatně i s autory jednotlivých balíků i experty na problematiku velkých dat. Také se podílel na neformálních schůzkách malého týmu projektu VO-CLOUD a staral se o správu projektu na git-hubu.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
10. Celkové hodnocení	*100 (A)*

Popis kritéria:
Shrňte stránky ZP studenta, které nejvíce ovlivnily Vaše celkové hodnocení. Celkové hodnocení **nemusí** být aritmetickým průměrem či jinou hodnotou vypočtenou z hodnocení v předchozích jednotlivých kritériích 1 až 9.

Text hodnocení:
Podle mého názoru je tato práce špičková a unikátní ve světovém měřítku. Podařilo se prakticky ověřit použitelnost relativně neznámé metody na specifické astronomické problémy a vytvořit ucelený programový balík pro provádění experimentů strojového učení na rozsáhlých astronomických datech.
Student prokázal, že se umí dobře zorientovat v nové problematice, poradit si s náročným problémem. Aktivně se účastnil na vývoji profesionálního programového balíku i spolupracoval při modifikacích velkého gridového systému. Velmi mě překvapil odvahou řešit neznámé problémy a profesionalitou, s jakou spolupracoval s ostatními členy našeho malého výzkumného týmu. Proto dávám plný počet bodů.

Podpis vedoucího práce: