



Posudek oponenta diplomové práce

Student: Bc. Petr Tománek

Název práce: Distributed training of neural networks

Předkládaná práce se zabývá velice důležitým tématem, což je distribuované učení neuronových sítí. Jde o stále nevyřešený problém v oblasti strojového učení, který je ale zásadní. Pokud bychom dokázali efektivně trénovat neuronové sítě například na velkých superpočítačích, byli bychom schopni pracovat s mnohem většími a komplexnějšími modely.

Autor nejprve v první kapitole nastiňuje problém enormního objemu dat, který vzniká v dnešní době. Tato data je potřeba umět zpracovat a často i podrobněji analyzovat. Zrychlené trénování neuronových sítí by tomuto mohlo velice pomoci. V druhé kapitole autor stručně popisuje právě analýzu data, strojové učení a učení neuronových sítí pomocí metody SGD. Právě paralelizací této metody se zabývá v hlavní části své práce. Před tím ale ještě ve třetí kapitole stručně uvádí paralelní zpracování dat. Ve čtvrté kapitole pak uvádí přehled existujících paralelních algoritmů. Některé z nich pak autor implementuje pomocí jazyka Julia a v páté kapitole je porovnává mezi sebou.

Text práce je psaný velice dobrou angličtinou. Zejména první dvě kapitoly na mě působily dojmem, jako bych četl knížku. Úvod je napsaný velice pěkně a díky němu je dobře vysvětleno, jakou úlohu autor řeší a jaká motivace za ní stojí.

Osobně vidím největší problém ve čtvrté kapitole, která je stěžejní částí a je bohužel napsána velice nesrozumitelně. Potíž nastává už v části 3.3.4, kde je nedostatečně popsána paralelizace SDG metody. Není vůbec jasné, co se jak distribuuje, zda se provádí distribuce parametrů θ nebo se provádí distribuce podle vstupních dat. Nakonec jsem pochopil, že se zřejmě učí několik sítí paralelně vedle sebe a jejich parametry se vzájemně synchronizují. Na straně 29 se píše "communication of the gradients $G_{t,i}$ to server either from any worker $i \in workers()$ or from ...". Bohužel není vysvětleno, co značení $G_{t,i}$ znamená. V části 4.1 je uvedeno značení G_t^i jako "worker- i gradient in t -iteration", což nepovažuji za dostatečné vysvětlení, ale domnívám se, že jde zřejmě o distribuci vstupních vektorů. Bohužel zde není ani srozumitelně popsán sekvenční algoritmus, pouze vztah (2.4), kde se ale žádný index i nevyskytuje, a pak je tu algoritmus 1, což je ale kód v jazyku Julia, který používá jiné značení. Stejně tak je na straně 29 dále uvedeno značení $\tilde{\theta}_t$ aniž by byl vysvětlen význam. Do čtvrté kapitoly jsem tedy vstupoval s nejasnostmi toho, jak je SDG metoda paralelizována.

Jednotlivé paralelní algoritmy jsou bohužel popsány velmi povrchně a nebyl jsem schopný pochopit do detailu, jak skutečně fungují. Autor používá paralelní podporu v jazyku Julia, ale bohužel ji v textu nevysvětluje. To považuji za zásadní problém, protože oproti například rozhraní MPI nejde o standardní nástroj, který by byl běžně známý. Autor pak jednotlivé algoritmy uvádí opět v jazyku Julia (osobně bych upřednostnil pseudokód opírající se o značení

používaném v textu), ale z nich opět nejsou vůbec poznat detaily paralelizace, zejména není vidět komunikace mezi procesy. Přitom právě efektivní komunikace se zdá být velmi důležitá.

V části 4.5 je metoda elastic averaging opět popsána velmi nejasně. Ve vztazích (4.6) - (4.9) se parametr τ vyskytuje vždy jen na levé straně, takže vůbec není jasné, jak metoda na tomto parametru závisí. Autor v celé čtvrté kapitole používá pěkná schémata, která ilustrují běh paralelních algoritmů, ale bez základního pochopení jednotlivých algoritmů ani toto příliš nepomůže.

V páté kapitole jsou jednotlivé paralelní algoritmy vzájemně porovnávány. Bohužel zde není ukázáno, jak jednotlivé paralelní algoritmy škálují a rostoucím počtem procesů. Ukazuje se, že s osmi procesy lze dosáhnout urychlení až 7.94, což je pěkný výsledek. Ukázat urychlení a efektivitu s větším počtem procesů by ale bylo ještě cennější. Testy navíc probíhaly na vícejadrovém počítači, tj. na architektuře se sdílenou a nikoliv distribuovanou pamětí. Právě vliv skutečně distribuované architektury by měl být otestován.

Na autora bych měl následující dotazy:

1. Jakým způsobem je paralelizována SGD metoda?
2. Jaké paralelní nástroje jazyka Julia se používají a jak fungují? Bylo by možné implementované algoritmy pouštět na architektuře s distribuovanou pamětí?

Práce se zabývá velice zajímavým tématem. Jako řešerše dané tematiky by splňovala svůj účel velmi dobře, kdyby jednotlivé algoritmy byly lépe vysvětleny. Dále bych v tomto textu velice ocenil popis podpory paralelizace v jazyku Julia, který je zde použit. Výpočetní testy jsou spíše zaměřeny na samotné učení, ale neukazují paralelní škálovatelnost. Vzhledem k výše uvedenému navrhuji diplomovou práci ohodnotit známkou **C** tedy **dobře**.

V Praze, 19. srpna 2021.

Ing. Tomáš Oberhuber, Ph.D.
katedra matematiky
Fakulta jaderná a fyzikálně inženýrská
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE
Trojanova 13
120 00 PRAHA 2