ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

POSUDEK OPONENTA ZÁVĚREČNÉ PRÁCE

Název práce: Paralelizace sítí typu Sum-Product-Transform pro architekturu GPU
Typ práce: diplomová
Jméno autora: Bc. Ondřej Poláček
Fakulta: Fakulta jaderná a fyzikálně inženýrská

Oponent práce: Ing. Milan Papež, Ph.D.
Pracoviště oponenta: Centrum umělé inteligence FEL ČVUT

Předložená práce pojednává o vývoji algoritmů pro paralelizaci výpočtu ve speciálním typu pravděpodobnostních sítích hlubokého učení nazývaných „Sum-Product-Transform networks“. Tento typ sítí byl vyvinut teprve nedávno a proto zatím nelze předpokládat existenci algoritmů zkracujících výpočetní čas využitím grafických procesorů a strukturálních vlastností těchto sítí. Téma zadání je tedy poměrně náročné.

Práce, o celkovém rozsahu padesáti stran, využívající šestnácti literárních zdrojů, je přehledně a systematicky členěna do třech kapitol, má solidní jazykovou úroveň, a vynikající grafické zpracování. První kapitola se zabývá matematickým modelem sítě. Student zde popisuje jednotlivé typy uzlů sítě, přípustné architektury sítě, a učení metodou maximální věrohodnosti za použití gradientního sestupu a algoritmu zpětné propagace. Druhá kapitola obsahuje popis navržených algoritmů. Zejména je zde uveden detailní popis implementace jednotlivých vrstev sítě pro použití v inferenci, učení, vzorkování a marginalizaci. To vše je provedeno s důrazem na seskupení opakujících se výpočtů pro paralelizaci a homogenního využití paměti. Ve třetí kapitole student porovnává výpočetní čas při použití běžného jedno-jádrového a grafického procesoru. Toto je provedeno na synteticky-generovaných a reálných datech.

Zvolený způsob řešení, zejména vymezení podmnožiny přípustných sítí a zvolené programovací prostředky, považuji za správný a dobře zdůvodněny. Diplomant představuje nápady jak výpočetní čas urychlit nejen pomocí opakovatelnosti dílčích výpočtů, ale také za pomocí vhodných matematických vlastností sítě. Oceňuji že, pro nedostupnost některých funkcí ve zvolené knihovně pro automatické derivování, student samostatně naimplementoval některé rutiny pro výpočet derivací.

Práce je celkově na vysoké odborné úrovni. Chválím, že diplomant dokázal jasně a srozumitelně objasnit řadu komplikovaných postupů a metod. Nicméně, třetí kapitola je poměrně krátká. Je třeba ale podotknout, že časová náročnost implementace všech náležitostí i pro takto krátkou kapitolu musela být značná. Hodnocení výpočetního času se zaměřuje převážně na dimenzi dat a uvažuje pouze dva typy sítě. Uvítal bych detailnější prozkoumání výpočetní náročnosti v závislosti na počtu uzlů sítě. Za nedostatek práce považuji použití pouhých šestnácti zdrojů literatury. Většina těchto zdrojů se týká programovacích prostředků. V některých částech práce zjevně schází bibliografické citace. Uvažovaný typ sítě je následníkem nedávného předchůdce „Sum-Product network“ pro který existuje řada literárních pramenů. Například popis sítě v kapitole 1, a konverzi sítě na směs hustot pravděpodobnosti v kapitole 2, je třeba vhodně ocitovat. Uvítal bych kdyby práce byla napsána v anglickém jazyce.

Student musel prokázat schopnost nastudovat a pochopit rozsáhlou problematiku nestandardního typu sítí hlubokého učení, seznámit se s metodami výpočtů na grafických procesorech a posléze přijít na

vhodný způsob implementace. Zadání práce považuji za splněné. Některé části, například výpočet marginalizace, jsou z mého pohledu zpracovány i nad rámec zadání. Za velmi pozitivní považuji i to, že se student podílel na vzniku kvalitní publikace pojednávající o sítích Sum-Product-Transform. I přes výše uvedené nedostatky doporučuji hodnotit práci známkou: **A** (výborně).

Otázky k obhajobě:

- Kde vidíte největší slabinu v paralelizaci výpočtu pro součtově-transformační vrstvu sítě?
- Můžete nějak komentovat jak velkou síť a dimenzi zpracovávaných dat musíme uvažovat aby pro nás bylo výhodné použití grafických procesorů oproti běžným procesorům s více jádry (například s přihlédnutím k času potřebnému pro přenos dat mezi těmito procesory)?

Datum: 17.5.2021 Podpis: