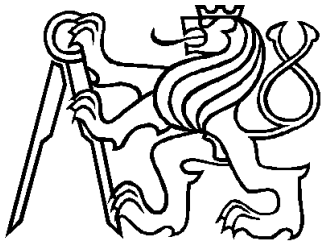

ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ

Fakulta jaderná a fyzikálně inženýrská

Posudek školitele diplomové práce

Student: Bc. Ondřeje Poláček
Název práce: Paralelizace sítí typu Sum-Product-Transform pro architekturu GPU

Předkládaná práce se zabývá sítěmi typu Sum-Product-Transform (SPTN), které tvoří strukturu pro učení se pravděpodobnostních distribucí. Jejich modelování resp. učení se z dat je jedním ze základních problémů strojového učení. Tyto sítě vznikají rozšířením sítí typu Sum-Product o transformační uzly. Z důvodů nutnosti výpočtu determinantu je nutné transformace vyjadřovat pomocí SVD rozkladu. To sice výrazně snižuje výpočetní náročnost, ale i tak jde stále o metodu vyžadující velký výpočetní výkon. Pro jejich rozumné využití v praxi je téměř nezbytné odvodit implementaci na výpočetních kartách GPU. Právě to je náplní této diplomové práce.

Autor nejprve v první kapitole podrobně popisuje konstrukci SPTN sítí. Popisuje jednotlivé typy uzlů a zejména detailněji vysvětluje transformační uzly. Ty využívají unitární transformace vyjádřené pomocí Givensových rotací. Autor ukazuje přeuspořádání těchto rotací do tzv. Butterfly vzoru. Ten tvoří základ pro paralelní algoritmus. Následně se autor zabývá učením SPTN sítí, což znamená zejména odvození algoritmu zpětné propagace. Tato část je technicky náročnější, jde o poměrně zdlouhavé odvození, ale i tak autor vše rozepsal do detailů. Samotná inference je poměrné přímočará.

Druhá kapitola se věnuje implementaci paralelního algoritmu na GPU v jazyku Julia. Nejprve jsou stručně shrnuty základy programování GPU a následně se autor hned pouští do popisu samotného algoritmu. Začíná popisem datové struktury. SPTN síť je potřeba uložit na GPU tak, aby přístupy k jejím parametrům bylo možné provádět efektivně z pohledu GPU. Dále autor popisuje učení SPTN sítí. Zde je potřebné popsat i samotnou inferenci, jejíž implementace na GPU je netriviální. Na základě vztahů odvozených v předchozí kapitole je pak popsána i implementace zpětné propagace. V závěru kapitoly autor zmiňuje i možnost práce s podstromy dále možnosti samplování a marginalizace.

Závěrečná třetí kapitola prezentuje dosažené výsledky. Jde o velmi krátkou kapitolu, ale jelikož hlavním cílem byla implementace SPTN sítí na GPU, nejpodstatnějším výsledkem je ukázat dosažené urychlení. To dosahuje hodnoty téměř 120x v porovnání se sekvenční implementací na CPU. Autor následně ještě ukazuje výsledky samplování síti naučené na datové sadě MNIST.

Na vedení této práce se významně podílel doc. Ing. Tomáš Pevný, Ph.D., který je autorem SPTN sítí, a který autorovi konzultoval pozadí týkající se strojového učení. Já jako školitel jsem byl zodpovědný hlavně za vedení v rámci paralelizace na GPU. Zde ale autor prokázal výraznou samostatnost. Vlastně jsme se jen na počátku dohodli na základní představě paralelního algoritmu a student odvodil celý zbytek zcela samostatně. Implementovaný algoritmus

je výrazně netriviální a brzy plánujeme sepsat publikaci do odborného časopisu s impakt faktorem o paralelizaci SPTN sítí. Dosažené výsledky jsou tedy mimořádně dobré. Samotný text práce je také na velmi dobré úrovni a nevím o ničem, co bych mohl vytknout, ačkoliv je jasné, že takto technicky náročnější text by bylo možné zlepšovat ještě dlouho.

Na autora mám následující dotaz:

1. Bylo by možné implementovat SPTN pro běh na více GPU současně resp. na GPU klastrech?

Vzhledem k výše uvedenému navrhuji diplomovou práci ohodnotit známkou **A** tedy **výborně**.

V Praze, 18. května 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