Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Bc. Petr Šťavík
Název práce	Analysing and Optimizing GPU Kernels with Machine Learning
Rok odevzdání	2020
Studijní program	Informatika **Studijní obor** Softwarové systémy

Autor posudku	RNDr. Martin Kruliš, Ph.D. **Role** Vedoucí
Pracoviště	Katedra softwarového inženýrství, MFF-UK

Text posudku:

Práce se zabývá možnostmi použití hlubokého učení na optimalizaci GPU kernelů na platformě CUDA. Autor provedl nejprve rozsáhlou rešerši a následně vybral dva existující přístupy jak trénovat neuronové sítě s použitím zdrojového kódu GPU programů jako vstupu. Hlavní myšlenkou je použití embedingu, který je dobře znám a prozkoumán z NLP problémů, a jeho adaptace na zdrojový kód za pomoci technik známých z překladačů.

Autor využil existující experimentální nástroje, které pracují s mezikódem v LLVM a navrhl vlastní úpravu pro PTX (mezikód používaný přímo v CUDA). Tato úprava vyžaduje preprocesing instrukcí PTX a tvorbu XFG grafu incidence instrukcí, který je následně použit při trénování rekurentních neuronových sítí. Pro testování tohoto přístupu byly zvoleny dvě klasifikační úlohy - predikce, zda daný kód poběží lépe na CPU nebo GPU a predikce tzv. occupancy (vytížení GPU jader) pro daný kernel. V první úloze byly výsledky řádově srovnatelné s již existujícím přístupem (který používa LLVM). Druhá úloha také nabídla poměrně optimistické výsledky, avšak zde jakékoli srovnání chybí.

Asi nejpodstatnějším nedostatkem práce je, malé množství provedených experimentů. Uvedené dva testovací scénáře nejsou dostatečně průkazné pro vyhodnocení daného přístupu. V celém postupu je navíc celá řada parametrů, které mohou ovlivnit např. kvalitu embedingu, takže ani není jasné, zda by nebylo možné dosáhnout lepších výsledků jen jejich úpravou.

Výše uvedený nedostatek je důsledkem omezené doby, kterou měl autor na vypracování, v kombinaci s náročností tématu, kde bylo nutné nastudovat jak principy hlubokého učení (což nebylo součástí původního studijního plánu autora), detaily implementace CUDA kernelů, ale také provést poměrně náročnou rešerši. Ve výsledku tedy autor více než dostatečně demonstroval, že je schopen samostatné analytiké, implementační i experimentální práce na úrovni očekávané u absolventů MFF-UK.

Text práce je dobře strukturovaný a snadno čitelný, velmi pravděpodobně poslouží jako základ pro navazující výzkum nebo další diplomové práce.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

Pokud práci navrhujete na zvláštní ocenění (cena děkana apod.), prosím uveďte zde stručné zdůvodnění (vzniklé publikace, významnost tématu, inovativnost práce apod.).

Datum 22.6.2020 **Podpis**