Review of Master's Thesis

Student: Matula Tomáš, Bc.
Title: Exploiting Approximate Arithmetic Circuits in Neural Networks Inference (id 22230)
Reviewer: Mrázek Vojtěch, Ing., Ph.D., UPSY FIT VUT

1. **Assignment complexity** — **more demanding assignment**
   Jedná se o obtížnější zadání, jelikož se vyžaduje se detailně seznámit s vnitřní strukturou jednoho z používaných frameworků pro neuronové sítě a rozšířit jej o nekonvenční operace.
2. **Completeness of assignment requirements** — **assignment fulfilled**
   Student zadání splnil, mírné rozšíření vidím v podrobné analýze odolnosti vůči chybám v jednotlivých vrstvách
3. **Length of technical report** — **in usual extent**
   Práce je v obvyklém rozmezí a obsahuje všechny potřebné informace.
4. **Presentation level of technical report** — **70 p. (C)**
   Práce je dobře strukturovaná, kapitoly na sebe navazují a text je pro čtenáře pochopitelný. V textu práce se však vyskytuje několik faktických chyb - například vzorce 3.1 a 3.2 nejsou kompletní (chybí podstatný operátor *"pro všechny vstupy"*), v kapitole 4.3 autor uvádí, že vyhledávací tabulku pro 9-bitovou násobičku je možné uložit do 4 MB paměti (pro uložení 2048x2048 18-bitových hodnot potřebujeme 16 MB). U vzorce 4.3 došlo k poměrně velkým aproximacím konstant (např. 83.5% => 80%), které nejsou v textu dostatečně vysvětlené.
5. **Formal aspects of technical report** — **55 p. (E)**
   Práce nevyužívá oficiální šablonu fakulty a díky tomu došlo k celé řadě problémů - např. v úvodu se objevilo *"(c) Tomáš Matula, 2019"*, které neodpovídá ani starším šablonám. V některých případech není psaní velkých a malých písmen popisu obrázků a grafů jednotné. Práce je psaná anglicky, bez výraznějších chyb, v některých částech se však objevují neformální fráze (např. kapitola 4.3).
6. **Literature usage** — **85 p. (B)**
   Autor cituje podstatnou literaturu, která se zadaným tématem zabývá. V některých případech autor cituje přímo webovou stránku, přestože v popisu nástrojů je odkaz na vědecký článek, který software popisuje ([10], [17]).
7. **Implementation results** — **95 p. (A)**
   Autor implementoval aproximační operace ve frameworku TensorFlow Lite určený pro vestavěná zařízení. Díky využití kvantizovaných modelů jsou aproximace zavedeny tak, že vyhodnocení přesnosti klasifikace odpovídá reálnému chování systému. Vzhledem k časové náročnosti simulace bylo nutné vyhodnocovat pouze podmnožinu testovacích dat, ale autor v práci jasně demonstroval, že tato optimalizace nemá zásadní vliv na vzájemné porovnávání metod.
8. **Utilizability of results**
   Výsledné rozšíření frameworku TensorFlow Lite přináší nové výsledky pro aproximací existujících modelů neuronových sítí a je využitelné jako základ pro další výzkum odolnosti neuronových sítí vůči chybám.
9. **Questions for defence**
   1. Jaké kroky je nutné provést pro aproximaci existujícího modelu neuronové sítě?
   2. Souvisí počet operací v aproximované konvoluční vrstvě s poklesem přesnosti celé sítě? (obrázky 5.7 a 5.8)
10. **Total assessment** — **80 p. very good (B)**
    Autor podstatně rozšířil existující framework o aproximační operace. Práce navazuje na výzkumná témata zkoumaná na naší fakultě a vytváří nástroj použitelný pro další zkoumání vlastností velmi hlubokých neuronových sítí. Práce však obsahuje drobné nedostatky v textu a také nesplňuje všechny formální požadavky. Proto navrhuji **hodnocení B**.

In Brno 2. June 2019

Mrázek Vojtěch, Ing., Ph.D.
reviewer