Posudek oponenta diplomové práce

Student: Slouka Lukáš, Bc.
Téma: Implementace neuronové sítě bez operace násobení (id 19117)
Oponent: Baskar Murali K., UPGM FIT VUT

1. **Náročnost zadání** — **průměrně obtížné zadání**
   The difficulty is scalable with regard to the topic. Given the interest of time, the attempted work has implemented a considerably difficult task.
2. **Splnění požadavků zadání** — **zadání splněno**
   The thesis fulfills the requirements and is based on the current works in literature.
3. **Rozsah technické zprávy** — **je v obvyklém rozmezí**
   The pages are in the desired range and the figures are of proper resolution.
4. **Prezentační úroveň předložené práce** — **90 b. (A)**
   The thesis is well written. It starts with basic introduction of neural networks and why speeding them up is mandatory. The existing techniques to accelerate neural networks is then discussed. Chapter 4 introduces the binarization of neural networks and thoroughly explains the binarization algorithms such as Binary Connect and XNOR networks. Chapter 5 is quite well explained as it consolidates the basics of Tensorflow and XGEMM with respect to binarization.
5. **Formální úprava technické zprávy** — **90 b. (A)**
   The language used is understandable and grammatically correct.
6. **Práce s literaturou** — **95 b. (A)**
   The contents of the thesis are quite informatory and shows the ground work done by the student to gather most of the relevant topics related to the field. There are adequate references and the writing enables uninterrupted flow of reading.
7. **Realizační výstup** — **95 b. (A)**
   CUDA programming (c++) is used to implement backend part of XGEMM kernel and is the code is properly structured and is understandable. The frontend code written in c++ registers the module in Tensorflow. The integration of Tensorflow along with XGEMM makes it easier to extend it to various other experimentation.
8. **Využitelnost výsledků**
   The work is satisfactory with respect to the topic as it implements the skeleton of the Binary NN idea. There is ample scope to extend it to recurrent models such as LSTM and different other tasks such as Language modeling. The thesis shows considerable speedup of Binary NN over the standard matrix multiplication in Tensorflow which is nice. The work also requires much more experimentation with different datasets and models, but the current work is sufficient considering the constraint of time. Graphical representation of distribution of weights in different layers and how it differs from full-precision models could have been more illustrative.
9. **Otázky k obhajobě**
   Any suggestions on how these binarized NN models can be scaled to large-scale datasets ?
   What were the glitches in the current implementation which lead to degradation in performance compared to full-precision models (From results in Chaper-6) ?
10. **Souhrnné hodnocení** — **92 b. výborně (A)**
    Implementation of XGEMM module in CUDA and its integration to Tensorflow requires more understanding of the architecture and the code implemented substantiates it.
    The thesis is easy to read and consolidates the important topics and GPU programming techniques in a nice way.
    The conclusion clearly states whats done and the extensions mentioned are more helpful if executed.

Prohlášení: Uděluji VUT v Brně souhlas ke zveřejnění tohoto posudku v listinné i elektronické formě.

V Brně dne: 7. června 2018

.....

podpis