ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

POSUDEK OPONENTA ZÁVĚREČNÉ PRÁCE

I. IDENTIFIKAČNÍ ÚDAJE

Název práce:	**Relaxed quantization and Binarization for Neural Networks**
Jméno autora:	**Martin Mraz**
Typ práce:	bakalářská
Fakulta/ústav:	Fakulta elektrotechnická (FEL)
Katedra/ústav:	Katedra kybernetiky
Oponent práce:	Teymur Azayev
Pracoviště oponenta práce:	Katedra kybernetiky

II. HODNOCENÍ JEDNOTLIVÝCH KRITÉRIÍ

Zadání — náročnější

The work consists of theoretical and programming work in a niche field which is an active area of research. This puts the difficulty to above average due to scarcity of materials and well established methods as well as the requirement of understanding new concepts which can only be found in scientific publications.

Splnění zadání — splněno s většími výhradar

The first (main) point is fulfilled, the student implemented a framework for the proposed methods and compared the performance and shows that a simpler method (SR) is superior to more complex SQ method. Experiments show that probabilistic pretraining does not help, but the author suggests that it might, in larger networks.
Points 2 & 3 are not implemened. They seem to be stand-alone projects and would require a significant amount of time and work. Point 2 would require the student‘s creativity and experimentation to implement so it is a shame that it wasn‘t realised, but Point 3 is just an additional experiment to show performance so I don‘t think that it‘s important. In general, the supervisor likely underestimated the amount of time which would be required to implement all the points.

Zvolený postup řešení — správný

The proposed solution is ok.

A few minor comments:
It seems that mostly only training time and accuracy has been considered as the main criterium for the various approaches. It would be nice to see memory usage and runtime comparisons on several different systems (normal pc, single board computer (SBC), etc), since this is the main selling point of these quantization methods.

Odborná úroveň — B - velmi dobře

The technical level is very good. Explanations in intro and theoretical parts are usually comprehensive. Good illustrations and mathematical description. Figures could be better described in some cases.

ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

POSUDEK OPONENTA ZÁVĚREČNÉ PRÁCE

Formální a jazyková úroveň, rozsah práce B - velmi dobře

The formality and level of language is generally very good at the beginning, so it shows that the student is capable of it, but the experiments and discussion section seems a bit rushed, with some errors in language, sentence formulation and some lacking explanations.

Výběr zdrojů, korektnost citací A - výborně

The citations for the used and relevant work is comprehensive and satisfactory.

Další komentáře a hodnocení

A lot of what is in the discussion probably belongs in the intro. The experiments could be organised a bit better. Even though the proposed probabilistic pretraining didn't help, it would be nice to see the results in a table as well.

III. CELKOVÉ HODNOCENÍ, OTÁZKY K OBHAJOBĚ, NÁVRH KLASIFIKACE

- Overall assessment: The work generally looks well done, with decent quality explanations, illustrations and mathematical derivations, however due to some un implemented points and rushed ending I would prefer to give the work a C (good) grade, but I suspect that the negative aspects of the work can be attributted to a mismanagement of the supervisor so I will give the work an official B (very good) grade.

- Question: For me the initial curiosity was not how to quantize a 1 or 2 bit network, but if such a network has the learning capacity to perform the trained task and how the degree of quantization has an effect on learning capacity in comparison to using high precision floats. Is there any way to quantify this or any work that looks at it?

Předloženou závěrečnou práci hodnotím klasifikačním stupněm B - velmi dobře.

Datum: 1.6.2022 Podpis: