Oponentský posudek diplomové práce Teodora Kováče Komprese konvolutorních vrstev neuronových sítí

Vypracoval: Ing. Jiří Vomlel, Ph.D.
Pracoviště: Ústav teorie informace a automatizace AV ČR, v.v.i.
Pod Vodárenskou věží 4, Praha 8, 182 00

Diplomová práce se zabývá kompresí vrstev konvolučních neuronových sítí pomocí metod kanonického rozkladu tenzorů. V práci bylo experimentálně ukázáno, že i při významné redukci paměťových nároků (více jak 50%) je většinou možné udržet relativně vysokou přesnost a malou chybu klasifikace (méně než 5%). Pro získání stabilních rozkladů je v práci navržena, implementována a otestována modifikace Levenberg-Marquardtova algoritmu, která pracuje s omezením sensitivity.

Práce nejprve obsahuje stručný úvod do umělých neuronových sítí. Vzhledem k tomu, že práce se zabývá výhradně konvolučními neuronovými sítěmi (nikoliv obecnými neuronovými sítěmi), bylo by vhodné úvod rozšířit o konvoluční sítě – detailněji vysvětlit a popsat jaký je základní rozdíl oproti obecným neuronovým sítím. K vysvětlení by například také bylo možné použít motivaci ze zpracování obrazové informace – např. popsat, co v takovém případě znamenají parametry M, N, S a T.

V sekci 3.2.2 je popsána implementace navrženého řešení a způsob jeho testování. V kroku 1 je proveden rozklad tenzoru vah. Překvapuje mne, že pokud prostor řešení obsahuje velké množství lokálních minim, tak aproximace byla počítána pouze třikrát. Pokud rozumím správně kroku 4 implementace, tak se při vlastním výpočtu již nevyužívá nalezený CP rozklad, tj. dojde k převodu na tenzor Y_2, který je stejných rozměrů jako původní tenzor. Pro praktické využití CP rozkladů by bylo tedy ještě vhodné implementovat a otestovat využití rozkladů při vlastním výpočtu pomocí naučené neuronové sítě. Bylo by vhodné udělat i podrobné srovnání doby výpočtu a paměťových nároků obou metod a výsledky prezentovat například formou grafů.

V kapitole 4 je detailně popsána a odvozena úprava Levenberg-Marquardtova algoritmu, která pracuje s omezením sensitivity. Tato část je nejvíce teoretická a dokazuje, že autor získal dobrý teoretický náhled do řešené problematiky.

Závěrečná kapitola je věnována experimentům na rozsáhlých konvolučních neuronových sítích Res-Net 18 a VGG 16. Při těchto experimentech se ukázalo, že modifikace Levenberg-Marquardtova algoritmu dosahuje výrazně lepších výsledků než standardní metoda z tenzorové knihovny Tesorlab 3.0 Matlabu.

Detailní komentáře:
str. 12 Česky se píše hardwarová nikoliv hardwareová.
str. 18 Píše se kdybychom nikoliv když bychom.
str. 19 $X^{(K)}$ je tenzor třetího řádu, takže popis "vstupní matice" je zavádějící.

str. 19 V Obrázku 2.3 není zřejmé, co jednotlivé bloky představují. Chybí jeho popis vzhledem k prezentaci uvedené v textu.
str. 21 V poslední větě kapitoly 3.1 je uvedeno, že výše uvedeným vzorcem dostáváme návod, jak výpočet z jedné konvoluční vrstvy rozdělit na výpočet ve čtyřech hypoteticky menších vrstvách. Myslím, že by vyvětlení rozkladu pomohlo grafické znázornění tohoto rozdělení.
str. 23 Uvádí se zde, že natrénovaná síť AlexNet je volně dostupná a je zde uvedena reference na článek. Z uvedené reference ale není zřejmé, kde lze síť AlexNet nalézt.
str. 26 V pseudokódu Levenberg-Matrquardtova algoritmu by bylo dobré označovat chybu stejně (err vs. chyba).
str. 30 Odkaz na přílohu je porušený.
str. 31 Byte se píše s velkým písmenem B, malé písmeno b značí bit.
str. 36 ... jsou k nalezení v příloze D.
str. 51 V textu je používán termín úspěšnost, v tabulkách termín Accuracy. Bylo by vhodné sjednotit.
str. 55 Artificial nikoliv Artifiacial
str. 55 Formátování i způsob reference 24 není dobrý, chybí autoři, jméno časopisu, atd. Myslím si, že správně by měla být citace článku: *Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks, Advances in Neural Information Processing Systems 25 (NIPS 2012)*

Otázky:

- Na straně 13 je uvedeno, že tato práce bude využívat rozklady v tělese reálných čísel. Jaké další těleso by mohlo být užitečné pro rozklady tensorů pro konvoluční neuronové sítě a proč?
- Vysvětlete, proč metoda kanonického rozkladu byla použita pro konvoluční neuronové sítě a nikoliv pro obecné neuronové sítě.
- Lze nějak vysvětlit pozorování, že u testované sítě VGG 16 (na rozdíl od sítě Res-Net 18) nedochází k výraznému zhoršení přesnosti při kompresi všech vrstev sítě?

Závěr:

Výsledky prezentované v diplomové práci jsou zajímavé. Autor prokázal schopnost aplikovat výsledky z oblasti kanonických rozkladů tenzorů v současné době hojně využívané oblasti umělých neuronových sítí. Oceňuji praktickou implementaci navržených algoritmů i jejich otestování na rozsáhlých reálných neuronových sítích. Diplomovou práci navrhuji hodnotit známkou **A (výborně)**.

V Praze, dne 25. ledna 2021.

Jiří Vomlel