Ing. Petr Tichavský, DSc.
ÚSTAV TEORIE INFORMACE A AUTOMATIZACE AV ČR, v.v.i.
Pod Vodárenskou věží 4, 182 00 Praha 8

Posudek školitele diplomové práce

Student: Bc. Teodor Kováč

Název práce: Komprese konvolutorních vrstev neuronových sítí

Předložená diplomová práce se týká problematiky strojového učení, a jak její název prozrazuje, komprese konvolutorních vrstev neuronových sítí. Neuronové sítě jsou známy už desítky let, ale teprve relativně nedávno zaznamenaly veliký rozvoj vzhledem k tomu, že se ukázaly být velice efektivním nástrojem například u počítačového vidění, rozpoznávání a klasifikace. Jedná se zejména o hluboké neuronové sítě s relativně velkým počtem skrytých vrstev, a konvolutorní sítě, kde dochází k opakované filtraci zpracovávaných obrazů a extrakci užitečné informace z nich.
Snahou v poslední době je přenést umělou inteligenci, kterou neuronové sítě vytvářejí, do běžné spotřební elektroniky, do kamer a fotoaparátů. Zde je pak z energetických a výpočetních důvodů snaha omezit paměťové a výpočetní nároky těchto sítí při zachování jejich efektivity, přesnosti rozpoznávání a klasifikace. Toto je motivací předložené práce.

Neuronové sítě a zvláště jejich konvolutorní vrstvy obsahují matematické struktury nazývané tenzory. Myšlenka náhrady těchto tenzorů odpovídajícími kanonickými rozklady rovněž není nová. Ovšem ukazuje se, že ne všechny algoritmy pro rozklad tenzorů jsou stejně vhodné pro tento účel. Toto dokazuje i předložená diplomová práce.

Po nezbytném úvodu do kanonických rozkladů tenzorů v kapitole 1 a úvodu do neuronových sítí v kapitole 2 se práce v kapitole 3 věnuje náhradě konvolutorních vrstev tenzorovými rozklady. Princip je ukázán na příkladě jednoduchých konvolutorních sítí pro rozpoznávání ručně psaných číslic a konvolutorní sítí AlexNet, která je v tomto oboru již klasikou.

V kapitole 4 je představen pojem sensitivity kanonického rozkladu tenzoru a metoda jak rozklady s nízkou senzitivitou hledat. Kapitola 5 popisuje experimenty s již skutečně velkými neuronovými sítěmi známými pod zkratkami ResNet 18 a VGG-16. Student zvládnul technologii a práci s neuronovými sítěmi, transfer znalostí, tj. přenos informace z částečne naučených sítí do nového prostředí, zvládl techniku doučování sítí na grafických kartách na velkém počítači, a tak systematicky dokázal hypotézy, ze kterých jsme vycházeli, totiž že běžné tenzorové rozklady se pro kompresi neuronových sítí nehodí, a je třeba používat rozklady s nízkou senzitivitou.
Ukázal, ze při nepatrném snížení přesnosti rozpoznávání, např. 3%, lze uspořit až kolem 80% paměťových nároků sítí.

Jedinou slabinou práce je, že měla být dokončena asi o tři čtvrtě roku dřív, aby bylo možno ji prezentovat na některé vědecké konferenci týkající se oboru, byť konané online, v dnešní pandemické době. Konkurence totiž nespí a této problematice se věnuje řada výzkumných skupin po celém světě. Přinejmenším kolegové v Moskvě, se kterými spolupracuji, a tamní studenti, byli rychlejší. Předložená práce tak spíše potvrzuje jejich výsledky. Je zde ovšem rozdíl, že kolegové v Moskvě používali knihovnu KERAS v programovém prostředí Python, kdežto student Kováč používal Matlab. Už proto jsou výsledy Bc. Kováče cenné.

S ohledem na výše uvedené se domnívám, že předložená práce splňuje požadavky kladené na diplomové práce a navrhuji klasifikaci A, výborně.

V Praze, 14.1. 2021

Ing. Petr Tichavský, DSc.