Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	David Vondrák
Název práce	Rozpoznávání a klasifikace učebnic pomocí hlubokého učení
Rok odevzdání	2022
Studijní program	Informatika **Studijní obor** Softwarové systémy

Autor posudku	Michal Novák **Role** oponent
Pracoviště	ÚFAL MFF UK

Text posudku:

Popis práce: Tématem předložené práce je rozpoznávaní a klasifikace učebnic pomocí strojového učení. Práce konkrétně řeší tři úlohy: (1) rozpoznávání jestli daná publikace je nebo není učebnicí, (2) zařazení učebnice do jedné z předdefinovaných kategorií tematicky na základě školního předmětu a (3) zařazení učebnice do jedné z předdefinovaných tříd na základě úrovně školského systému, pro nějž je určena. K rozhodování v těchto úlohách přitom dochází na základě názvu knihy, autora, nakladatelství a slovního popisu knihy.

Vlastní práce začíná přípravou dat. Ta zahrnuje stažení záznamů z několika internetových knihkupectví a online databází knih, jejich čištění a automatické párovaní kategoriálních položek autora a vydavatelství. V podmnožině záznamů obsahující 15 tisíc knih jsou následně dvěma anotátory ručně anotovány odpovědi na tři zkoumané úlohy a výsledná anotovaná data jsou rozdělena na trénovací a testovací sadu.

Na řešení úloh autor práce navrhuje čtyři architektury hlubokých neuronových sítí kombinující hustě propojené, konvoluční a rekurentní vrstvy. Kromě toho na textové položky (název a popis knihy) aplikuje tři druhy slovních embeddingů: předtrénované vlastní, předtrénované stažené a nepředtrénované. Přínos jednotlivých druhů embeddingů je vyhodnocen pomocí křížové validace na trénovací sadě. Jednotlivé architektury jsou následně porovnány mezi sebou a s Naive Bayes baseline modelem vyhodnocením na testovací sadě. V závěru autor provádí podrobnější chybovou analýzu i manuální inspekci vzorku chyb.

Práce je psaná česky, má 65 stran včetně seznamu použité literatury, obrázků a tabulek. Příloha práce obsahuje zdrojové kódy experimentů, data v různých fázích předzpracování a výsledky experimentů.

Hodnocení: Po formální stránce je práce ukázková. Je přehledně strukturovaná, úvodní a rešeršní sekce jsou jasně odděleny od vlastní práce. Použitý jazyk odpovída odbornému stylu, terminologie je řádně vysvětlena a používaná konzistentně. Zejména v úvodních částech jsou pasáže

precizně ozdrojovány, včetně zdrojů převzatých ilustrací. Navzdory pozornému čtení jsem v celé práci narazil jen na dva překlepy („narozdíl“ na str. 21, „Klasifikce“ na str. 58) a jednu formální chybu (chybějíci číslo tabulky v posledním odstavci na str. 39).

Po obsahové a metodologické stránce je práce rovněž na vysoké úrovni. Jednotlivé úlohy jsou dobře motivovány a definovány. Všechny použité metody jsou dostatečně vysvětleny v samostatné sekci. Tady bych jenom uvítal o něco delší popis souvisejících prací. I když autor tvrdí, že úlohu rozpoznávaní a klasifikace učebnic nikdo před ním nedělal, klasifikací knih např. podle žánru se určitě zabývalo víc než jen čtyři publikace (např. Thakur and Patel: An Improved Dictionary Based Genre Classification Based on Title and Abstract of E-book Using Machine Learning Algorithms (2021); Ozsarfati et al.: Book Genre Classification Based on Titles with Comparative Machine Learning Algorithms (2019)).

Autor si data sám připravil od stažení až po manuální anotaci, kde oceňuji detailně nastavená anotační pravidla (např. předem stanovená hierarchie úrovní u sporných případů). Celá trénovací a testovací sada byly dokonce označkovány i druhým anotátorem, což umožnilo spočítat mezi-anotátorskú shodu a porovnat rozdíly v rozpoznávaní mezi algoritmy s rozdíly mezi anotátory. Možná bych zvolil jiný poměr u dělení na trénovací a testovací sadu (např. 80:20 místo 67:33), ale to je jenom drobnost.

Architektury sítí tak, jak jsou navrženy, dávají pro dané vstupy a požadované výstupy smysl. Autor rovněž provedl dostatek experimentů v různych konfiguracích. Jelikož se v těchto úlohách zpracovává text, uvítal bych však experimenty s architekturou typu Transformer (nebo alespoň zmínku o ní), která zpracování přirozeného jazyka v posledních pěti letech dominuje. Rovněž mám pocit, že se autor dopustil jedné zásadnejší metodologické chyby. Z popisu hyperparametrů na str. 43 usuzuji, že počet trénovacích epoch se určoval na základě křížové validace na trénovacích datech. Křížová validace se ale podle str. 51 prováděla jenom u konvoluční architektury, nikoliv u zbylých tří architektur. V textu se to už víc nezmiňuje, ale vypadá to tak, že tento „ideální“ počet epoch pro konvoluční architekturu se použil i pro natrénovaní zbylých architektur. Tím se ale dle mého názoru zanáší do výsledků experimentů systematická chyba ve prospěch konvoluční architektury. Výsledky experimentů to potvrzují. Konvoluční architektura nebo vrstva se ukázala jako nejlepší. To ale může být jenom důsledek této metodologické chyby.

Další poznámky a otázky:

- **str. 20**: rekurentní sítě jsou dle mého názoru samostatným typem, a ne podtypem dopředných neuronových sítí
- **str. 20**: v popisu Embedding vrstvy chybí základní vlastnost, kterou se liší od hustě propojené vrstvy: reprezentace slov jsou sdílené. To znamená, že stejné slovo je reprezentováno vždy stejným embeddingem bez ohledu na to, v které části sekvence slov se nachází.

- **str. 24**: střední kvadratická chyba se používa u (lineární) regrese, ale ne u logistické regrese
- **str. 31**: nerozumím možnosti použití kódu ISBN v odstavci „Obě metriky. . . “
- **str. 41**: v jazycích, pro něž chybí model pro MorphoDiTu, se používá český model a předpokládá se, že vrácí původní tvary slov místo lemmat. To ale není pravda, zejména když se používá morfologický guesser. I když se to týká jenom nízkého procenta dat, autor u nepodporovaných jazyků nemusel vůbec lemmatizovat, případně mohl použít nástroj UDPipe, který podporuje mnohem víc jazyků.
- **str. 44**: uvádí se počet jednotek u rekurentních sítí, i když v popisu této metody není vysvětleno, co to znamená. Je to velikost stavů?
- **str. 45**: v konvoluční architektuře je konvoluční vrstva následována sdružovací vrstvou (MaxPooling1D) a ukončena globální sdružovací vrstvou (GlobalMaxPooling1D). Není tak efekt prostřední sdružovací vrstvy neutralizován globální sdružovací vrstvou? Jinak řečeno, kdybychom vypustili prostřední sdružovací vrstvu, nedělala by síť to samé?
- **str. 45**: není jasné, zda rekurentní síť vrací výstup pro každé slovo, finální stav nebo oba druhy výstupů.

Závěr: Předloženou diplomovou prací autor ukázal, že je schopen samostatně provádět experimenty v oblasti zpracování přirozeného jazyka a rovněž tyto experimenty na velmi kvalitní úrovni popsat a analyzovat. Práce tak splňuje všechny požadavky stanovené pro diplomové práce a proto ji doporučuji k obhajobě.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 24. 1. 2022

Podpis: