Posudek disertační práce

Posudek vedoucího

Řešitel: Mgr. Dušan Variš
Název práce: Learning Capabilities of the Transformer Neural Networks

Vedoucí: doc. RNDr. Ondřej Bojar, Ph.D.

Dušan Variš se ve své disertační práci zabývá vybranými významnými charakteristikami trénování neuronové architektury Transformer s cílem prověřit, zda naše intuice ohledně toho, co se síť učí, odpovídá skutečnosti. Souhrnně lze prohlásit, že antropomorfizovat učení se u těchto typů sítí je předčasné, a že ve studovaných ohledech Transformery selhávají více či méně dramaticky. Hlavní studovanou doménou je doména strojového překladu, pro kterou byl Transformer ostatně navržen.

Práce je přehledně členěna do sedmi kapitol, motivaci a potřebnému úvodnímu přehledu jsou věnovány první tři (Introduction, Multi-task Learning, Neural Sequence-to-sequence Modelling). Následující tři kapitoly se zabývají postupně třemi oblastmi obecných vlastností učení se.

V kapitole 4, Generalization in NMT Transformers, Dušan na třech typech úloh prověřuje nakolik Transformer zobecňuje pozorování z dat. Podstatou pokusů je testovat natrénovaný model na třídě vstupů, které nám jako lidem připadají zcela analogické k příkladům trénovacím, ale nějakou poměrně jednoduchou charakteristikou se jejich distribuce od trénovacích dat zásadně liší. Dušan ukazuje, že Transformer neumí dobře pracovat s délkou věty. Výstupní věty základní verze Transformeru nikdy neopouštějí rozsah délek vět, které byly součástí trénování. Obdobné výsledky vykazovaly následně i některé pokročilejší verze Transformeru, které testovali moji další studenti. Dušan navrhuje zajímavou a jednoduchou techniku, jak problém empiricky obejít: trénovat na spojení několika vět za sebou, i když spolu věty nemusí jinak souviset. Obdobné pozorování představuje Dušan pro schopnost Transformeru kopírovat vstup: Transformer to umí jen v podobných délkách. Naproti tomu lexikální volba není v podobném smyslu přetrénována a kvalita výstupu Transformeru příliš neklesá, když uměle zvýšíme lexikální odlišnost trénovacích a testovacích dat.

V kapitole 5, Incremental Learning and Catastrophic Forgetting, se Dušan věnuje metodám trénování, které by dovolovaly během trénování přejít k jinému typu úlohy, aniž by výrazně poklesla kvalita v úloze původní. Dušan rozvíjí zavedenou metodu Elastic Weight Consolidation, která mu funguje obstojně pro účel kombinování překladových znalostí získaných z paralelních trénovacích dat a znalostí získaných z jednojazyčných dat. Další experimenty s uměle vytvořeným datasetem jednoduchých sekvencí a s mnohojazyčným paralelním korpusem ukazují limity metody EWC. Pro jednoduchou úlohu (kopírování sekvence) EWC dovoluje doučit se chodu na nové abecedě dobře, pro obtížnější úlohu (obrácení sekvence) souběžně se zlepšením se v nové abecedě dochází ke zhoršení v abecedě původní. Z praktického hlediska tedy EWC není užitečným řešením pro problematiku postupného učení se.

V kapitole 6, Transformer Modularization, Dušan usiluje o zlepšení kvality výstupu Transformerů explicitním oddělováním oddělitelných částí výpočtu. Zjištění, které části výpočtů (nebo řekněme typy chování) jsou na sobě natolik nezávislé, že lze některé z nich pro daný vstup apriori vylučit, aniž by kvalita utrpěla, má model také provést sám, trénovaným Controllerem. Výsledky ukazují, že je možné úspěšně model prořezávat (tj. části výpočtu nepoužít), ovšem toto prořezávání nesouvisí s nějak interpretovatelnými typy úloh, na nichž trénujeme. Metoda modularizace si tedy zaslouží další studium s cílem optimalizace, nevede však prozatím k lepší interpretovatelnosti chování Transformeru.

Závěrečná kapitola velmi srozumitelně výsledky práce shrnuje.

Práce je psána velmi dobrou angličtinou. Podobně i ze všech ostatních formálních hledisek je práce dle mého soudu vynikající, ať již se jedná o sazbu nebo práci s literaturou, citování vlastních i cizích prací.

Při práci s Dušanem jsem vždy oceňoval jeho samostatnost po celou dobu studia i při psaní samotného textu disertace. V pozdějších technických analýzách (EWC a modularizace) Dušan výrazně překročil moje standardní očekávání. Podobně bych rád vyzdvihl Dušanovu schopnost komplexní a tematicky poměrně široká pozorování představit formou stručných jasně formulovaných výzkumných otázek a odpovědí.

Závěrem jednoznačně konstatuji, že předložená práce Mgr. Dušana Variš splňuje nároky kladené na disertační práce, a proto ji doporučuji k přijetí.

Praha, 23. březen, 2023.

doc. RNDr. Ondřej Bojar, Ph.D.