Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	David Burian
Název práce	Document embedding using Transformers
Rok odevzdání	2024
Studijní program	Informatika **Studijní obor** Umělá inteligence

Autor posudku	Jindřich Libovický **Role** vedoucí
Pracoviště	ÚFAL MFF UK

Text posudku:

Diplomová práce Davida Buriana se věnuje problému, jak reprezentovat delší texty jedním vektorem, tak aby podobné texty měly podobné vektory. Takové reprezentace textu, tzv. document embeddings, se používá při vyhledávání podobných dokumentů

Hlavní myšlenka je práce je využití techniky, které se nazývá knowledge-distilation, jejímž cílem přenést schopnosti jednoho naučeného modelu do jiného modelu. V tomto případě se používá dva modely jako tzv. učitel a jeden model jako tzv. student. Učitelské modely jsou SentenceBERT, který poskytuje kvalitní sémantické reprezentace jednotlivých vět a textů kratších než přibližně 500 slov a Paragraph Vector, který poskytuje reprezentace delších textů, ale pouze pro dokumenty, které byly součástí trénování a navíc nebere v úvahu pořadí slov v textu. Studentský model je předtrénovaný model Longformer, který sice umí reprezentovat delší texty, ale jako posloupnost vektorů, ne jako jediný vektor.

Práce má celkem 66 stran, hlavní text začíná na straně 3 a končí na straně 61. Práce se skládá ze 7 kapitol včetně úvodu a závěru.

Text práce je velmi dobře strukturovaný. Práce je psaná dobrou angličtinou bez zjevných gramatických chyb. V práci je množství přehledných diagramů a grafů, které usnadňují pochopení textu. Experimenty jsou dobře navržené, provedené a popsané.

Kapitola 1 vysvětluje a definuje problém, kterému se je práce věnuje. Definuje dva stěžejní koncepty pro tuto práci: schopnost modelu zachytit strukturu dokumentu (kterou má model SentenceBERT) a schopnost zachytit dlouhý kontextu (kterou má model Paragraph Vector). Model student, Longformer, má obě tyto schopnosti, ale nereprezentuje dokumenty jedním vektorem.

Kapitola 2 popisuje existující přístupy k reprezentace dokumentů.

V kapitole 3 je detailně představena metoda trénování reprezentace dokumentů, kterou student v práci vyvinul, ovšem bez technických detailům, kterým se věnuje následující kapitola.

Kapitola 4 popisuje sérii předběžných experimentů, které směřují k výběru nejvhodnějších parametrů navržené metody. Cíle je zde vybrat nejvhodnější ztrátové funkce a hyperparamtery modelů. Tato část práce je velmi detailní a pečlivostí zpracování předčí velkou část v současnosti publikovaných odborných článků.

Kapitola 5 popisuje hlavní výsledky práce na úlohách, které zahrnují vyhledávání podobných dokumentů a klasifikaci dokumentů. Výsledky ukazují, že takto vyvinutý model překoná jak původní Longformer, tak oba učitelské modely v situacích, kdy je k dispozici jenom malé množství trénovacích dat (stovky až tisíce trénovacích instancí).

Autor předloženou prací prokázal, že se orientuje v problematice reprezentace textů pomocí neuronových modelů. Dále práce ukazuje, že autor dovede velmi dobře plánovat, implementovat a vyhodnocovat výpočetní experimenty s jazykovými modely.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 16. 5. 2024

Podpis: