Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Daniel Šipoš		
Název práce	Analýza a vizualizácia správania jazykového modelu GPT-2		
Rok odevzdání	2022		
Studijní program	Informatika	**Studijní obor**	Počítačová grafika a vývoj počítačových her

Autor posudku	David Mareček	**Role**	vedoucí
Pracoviště	ÚFAL MFF UK		

Text posudku:

Diplomová práce Daniela Šipoše se zabývá vizualizací závislostí mezi slovy v textu generovaným modelem GPT-2. Na rozdíl od předchozích prací, které se často zaměřují na tzv. self-attentions mezi jednotlivými vrstvami sítě Transformer, které jsou však těžko interpretovatelné, zde se do vnitřností sítě nezasahuje a měří se pouze pravděpodobnostní distribuce slov na výstupu modelu a obměňují se pouz e vstupy (prompty).

Práce je členěna na úvod, pět číslovaných kapitol, a závěr. Po motivaci a ukázce v úvodu následuje kapitola "Základní pojmy" popisující neuronovou síť Transformer a model GPT-2. Zabývá se i již existujícími nástroji pro vizualizaci Transformerů a dalšími nástroji, které byly v této práci použity. V následující kapitole "Metody vizualizácie" jsou popsány vlastní metody pro vizualizaci modelu GPT-2 a jsou vysvětleny na příkladech. Třetí kapitola popisuje architekturu vizualizátoru, čtvrtá kapitola pak obsahuje uživatelskou dokumentaci, kde jsou popsány všechny módy a nastavení. Další kapitola analyzuje modelu GPT-2, kde autor na různých vygenerovaných textech hodnotí jejich vlastnosti a ukazuje statistiky jak daleko do historie se model dívá. Následuje závěr.

Hodnocení práce:

Daniel Šipoš vytvořil přehledný nástroj pro vizualizaci modelu GPT-2 pro generování textu. Implementoval celkem čtyři módy, které pro daný text a model ukazují jak kandidátská slova na každé pozici a jejich pravděpodobnosti, tak dříve vygenerovaná slova, která vygenerování slova na dané pozici nejvíce ovlivnila. Tento vizualizátor bude užitečný pro analýzu vygenerovaných textů a zjišťování, proč se daný model rozhodl vygenerovat právě tento text.

Z důvodu nedostatku času se bohužel nepodařilo, aby program podporoval libovolný jazyk a vstupní model. Původním cílem bylo, že uživatel zadá vygenerovaný text a svůj model a program ho zanalyzuje. To se podařilo jenom částečně, vlastní model sice je podporován, ale chybí podpora

modelů natrénovaných v prostředí PyTorch a podpora jiných jazyků. S tím souvisí i méně zajímavá analýza v kapitole 5, která analyzuje pouze jeden model "GPT-2 Medium". Zajímavější by bylo porovnat různé modely mezi sebou, například jestli se větší modely dívají dál, nebo jestli se například modely generující básně dívají při rýmování na konec příslušného předchozího verše nebo na předchozí srofu.

Práce je psaná slovensky a obsahuje 65 stran čistého textu. Je přehledně členěna, psána srozumitelně, snad jen pátá kapitola "Analýza", která byla psaná už v časové tísni, je místy méně srozumitelná a některé odkazované obrázky v ní jsou posunuty o několik stránek dále.

I přes dříve popsané nedostatky ale celkově hodnotím práci kladně. Daniel ukázal, že umí pracovat samostatně, některé metody analýzy sám navrhnul a většinu vyskytnuvších se problémů dokázal sám bez problému vyřešit. Množství odvedené práce splňuje požadavky na diplomovou práci.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 2. 6. 2022

Podpis: