# 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** | Rudolf Rosa | **Role** | oponent |
| **Pracoviště** | Ústav formální a aplikované lingvistiky | | |

**Text posudku:**

Student se ve své diplomové práci věnuje aktuálnímu tématu analýzy chování předtrénovaného jazykového modelu GPT-2 užívaného pro generování textu, přičemž práce má podstatnou jak softwarovou tak výzkumnou složku. Softwarovou složku, která v práci převažuje, představuje návrh a vývoj nástroje pro analýzu a vizualizaci závislosti generovaného slova na předchozích slovech. Výzkumná složka se projevuje v návrhu čtyř metod pro analýzu (v kapitole 2) a v analýze jedné varianty modelu GPT-2 na středně velkých datech (v kapitole 5). Osobně bych více ocenil, pokud by důraz byl spíše na výzkumnou složku a zejména na vlastní analýzu existujících jazykových modelů, kterou považuji jednoznačně za nejzajímavější součást práce; uznávám však, že důraz spíše na softwarovou část práce je v souladu se zadáním práce i studijním zaměřením studenta.

Metody analýzy jsou navržené vhodným způsobem, jsou přímočaré a nepracují s příliš silnými předpoklady (narozdíl od mnohých předchozích metod), a tedy jsou snadno a důvěryhodně interpretovatelné, přitom přinášejí užitečné vhledy do fungování zkoumaných modelů. Oceňuji diskuzi navržených metod včetně motivace i omezení.

Navrženou architekturu nástroje i uživatelské rozhraní považuji za vhodně zvolené a provedené, odpovídající reálným potřebám i zadání práce. Oceňuji modulární návrh, který je zde vhodně motivovaný a umožňuje dva různé módy používání software, a to graficky i terminálově, což vzhledem k časové náročnosti analýzy větších jazykových modelů je velmi užitečné. I při provedení analýzy dávkově v konzoli je přitom možné výsledky analýzy exportovat do souboru a ten následně načíst do aplikace v grafickém módu, s vizualizacemi je tedy možné pracovat nezávisle na způsobu získání analýz, což je skvělé. Pochybnosti mám ale ohledně formátu exportu, kde autor používá textový soubor obsahující speciální řídící tokeny a následně podrobně diskutuje nepravděpodobnost náhodného výskytu těchto řídích tokenů v textu vygenerovaném jazykovým modelem (což by vedlo k nevalidnímu exportu); domnívám se, že by si mohl autor tyto problémy a diskuze ušetřit, použil-li by pro export nějaký standardní serializační formát (např. JSON nebo XML), navíc by se i citelně zjednodušila implementace exportu a importu, kde místo vývoje vlastního serializátoru a deserializátoru by postačovalo volání základních funkcí příslušných existujících knihoven (tato volba není v práci zdůvodněna).

Součástí textu práce je podrobná dokumentace (kapitola 3) a podrobný uživatelský manuál (kapitola 4); na můj vkus až příliš podrobné, u diplomových prací očekávám spíše menší podíl takovéhoto popisného obsahu bez složitějších myšlenek a naopak větší důraz na odbornější obsah (je možné, že v rámci daného studijního oboru je toto u softwarových diplomových prací očekáváno; pokud ano, pak je to v pořádku).

Nástroj jsem bez problémů zprovoznil i bez detailního zkoumání návodu a i na Linuxu (který není explicitně podporován), nástroj se chová poměrně standardně, je uživatelsky přívětivý a stabilní. Porozumění jednotlivým analýzám a vizualizacím již pochopitelně vyžaduje seznámení s textem práce, avšak po přečtení práce je považuji za dobře srozumitelné. Používání nástroje považuji za příjemné a samotný nástroj za užitečný. Umím si například velmi dobře představit jeho využití v projektu THEaiTRE, jejž vedu a ke kterému autor referuje. V tomto projektu si často při generování textů klademe otázku, proč v dané situaci model vygeneroval právě toto slovo; vyvinutý nástroj by nám právě v takových okamžicích mohl poskytnout velmi cenné odpovědi,

které pro nás bez takového nástroje jsou převážně či zcela nedostupné. Integrace tohoto nástroje do zmíněného projektu by proto měla potenciál významně usnadnit práci na projektu a tím umožnit zisk znatelně kvalitnějších výsledků. Oceňuji tedy zveřejnění gitového repozitáře nástroje; pro reálnou možnost dalšího využití nástroje bylo by vhodné ještě upřesnit licenci, a také nástroj alespoň stručně zdokumentovat v angličtině.

Práce je psaná velmi srozumitelně a téměř bez chyb. Oceňuji velké množství příkladů, které činí prácí velmi snadno pochopitelnou. Některé části jsou ale možná až příliš rozvláčné, samotného odborného obsahu není tolik, a mám místy dojem, že se autor snaží uměle natáhnout délku práce, aniž by měl dostatečné množství relevantního obsahu. Toto platí například o kapitole 2, kde jsou popisovány jednotlivé metody analýzy, přičemž některé jsou si v mnohém podobné, čtenář zde proto opakovaně čte velmi podobné texty, toto by šlo napsat stručněji a zároveň přehledněji; některé metody pak jsou ilustrovány na dvou různých příkladech, jednom textově a jiném vizuálně, přičemž by jistě postačoval jeden příklad.

Za nejzajímavější část považuji kapitolu 5, kde autor analyzuje jednu vybranou variantu modelu GPT-2 na autorem připravených datech. Analýzy dobře ukazují užitečnost nástroje, vhodně na sebe navazují a postupně vykreslují obrázek fungování zkoumaného jazykového modelu; zároveň slouží i jako zhodnocení užitečnosti jednotlivých metod analýzy. Autor často nejprve nabídne základní přímočarou analýzu, která naznačí směr dalšího zkoumání, který autor úspěšně sleduje a postupně nabízí další a další logicky navazující analýzy, až se dobere podstaty zkoumané situace a vhodným způsobem interpretuje získané poznatky. Některá zjištění jsou přitom velmi zajímavá, například nezanedbatelná závislost generátoru i na velmi vzdálených slovech, či vysoká závislost prvního a třetího (ale nikoliv druhého) slova ve větě na předchozích větách; toto by si jistě zasloužilo další zkoumání, a to i v kontextu jiných jazyků než je angličtina. Některé interpretace jsou trochu vágní, což je někdy možná nevyhnutelné, ale například tvrzení *Graf [5.8] má úplne rovnaké črty, ako graf 5.1* by jistě šlo exatně kvantifikovat. Oceňuji, že výsledkem analýz jsou jak poznatky o zkoumaném modelu GPT-2, tak i kritické zhodnocení užitečnosti navržených metod analýz, kde autor zjišťuje a otevřeně konstatuje, že metody *Missing word dependency* a *Replacing subword dependency* dávají velmi podobné výsledky (a tedy stačí používat jednu z nich), a metoda *Missing sentence dependency* se obecně neukazuje být příliš užitečnou; schopnost takto střízlivě zhodnotit svou vlastní práci často chybí i mnohem zkušenějším autorům. Analýzy by bylo možné dále rozvíjet mnoha zajímavými směry, například zkoumat závislost na slovním druhu, četnosti slova, délce věty, syntaktické funkci, jazyce, a podobně; práce tedy otevírá mnohé další zajímavé výzkumné cesty (vzhledem k výpočetní náročnosti analýz velkých jazykových modelů ale zcela chápu, že v rámci diplomové práce není možné provést analýzy do podstatně větší hloubky či šířky, proto provedené analýzy považuji za zcela postačující).

Za nedostatek práce považuji nedostatečné srovnání s existujícím výzkumem ve výzkumné části práce. Práce obsahuje stručnou rešerši podobných existujících nástrojů pro vizualizaci neuronových modelů (1.4), chybí však přímé srovnání navržených metod analýzy (kapitola 2) s existujícím výzkumem v oboru (případně konstatování, že navržené metody jsou zcela nové a v ničem se nepodobají předchozím přístupům, což ale považuji za nepravděpodobné). Zcela pak chybí jakékoliv srovnání s existujícím výzkumem v kapitole 5, kde autor provádí analýzy modelu GPT-2 a dochází k různým zajímavým zjištěním; hodnotu těchto zjištění přitom nelze posoudit bez porovnání se zjištěními předchozích autorů, tj. zda student novým způsobem potvrzuje již dříve představené poznatky, či zda tvrzení předchozích autorů vyvrací, případně zda představuje poznatky zcela nové, v existující oborové literatuře doposud neznámé. I přes zjevné primárně softwarové zaměření práce se domnívám, že úroveň srovnání s existujícími pracemi nedosahuje úrovně očekávané u diplomové práce a odpovídá spíše práci bakalářské.

Zjevným omezením nástroje (které je v práci zmíněno jen letmo) je fakt, že zkoumá pouze přímou závislost na jednom předcházejícím slovu či větě, ale nedokáže zohlednit nepřímé závislosti, kdy na daném předcházejícím slově (A) může záviset dále vygenerované slovo (B), a na něm pak teprve závisí zkoumané slovo (C); vynecháním či změnou slova A se tato závislost nemusí odhalit, neboť C závisí na B (a při zkoumání slova C za vynechání slova A nezkoumáme závislosti slova B); avšak ani vynecháním B se toto nemusí odhalit, neboť model v tomto případě má stále k dispozici slovo A a tedy může vygenerovat C nyní v závislosti na A. Nejde přitom o

nijak teoretickou situaci, toto pravděpodobně nastává opravdu velmi často, například i již v první motivační ukázce v práci (Obrázek 1), kde slovu *wildlife* (C) předchází nejprve slovo *bear* (A) a následně *bears* (B); analýza zde odhalí závislost vygenerování slova C na slovech A i B, avšak skutečná závislost C na A a B pravděpodobně bude ještě mnohem silnější, odhalila by se však až v případě vynechání obou těchto slov (A a B) zároveň, s čímž program nepracuje. Nabízelo by se zde právě toto, tedy umožnit vynechání více slov naráz, případně návrh nějaké agregační metody, která by propočítávala i tranzitivní závislosti mezi jednotlivými slovy. Řešení tohoto problému by práci značně zkomplikovalo, proto chápu, že tento problém nebyl v práci vyřešen; protože se však dle mého názoru jedná o zásadní omezení, očekával bych v práci k tomuto jasnou podrobnější diskuzi, nikoliv jen letmou zmínku (jeden krátký odstavec v 5.2).

Celkově jde o kvalitní a užitečnou softwarovou práci, s velmi zajímavou avšak poněkud nedůslednou výzkumnou složkou.

Práci navrhuji hodnotit známkou 2.

**Práci doporučuji k obhajobě.**

**Práci nenavrhuji na zvláštní ocenění.**

V Praze dne 30. 5. 2022

Podpis: