Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Lukáš Chaloupský		
Název práce	Automatic generation of medical from chest X-rays in Czech		
Rok odevzdání	2022		
Studijní program	Informatika	**Studijní obor**	Softwarové a datové inženýrství

Autor posudku	Jindřich Libovický	**Role**	oponent
Pracoviště	ÚFAL, MFF UK		

Text posudku:

Posuzovaná diplomová práce si klade za cíl vyvinout nástroj pro automatické popisování radiologických snímků plic v českém jazyce. Cíl práce je především aplikační, chce vyvinout funkční systém, proto ve velké míře využívá existují již natrénované modely a tam, kde připravuje vlastní modely, používá osvědčené metody z literatury.

Práce má celkem 60 stran včetně příloh, hlavní obsah začíná na straně 3 a končí na straně 51. Práce je psaná v anglickém jazyce s jen minimem stylistických a gramatických nedostatků. Práce má celkově velmi dobrou logickou strukturu: první kapitola popisuje problém, který se práce snaží řešit a základní koncepty, které se v průběhu práce používají (s. 5–13). Druhá kapitola (s. 16–23) popisuje metodologie, kterou se autor rozhodl postupovat. Třetí kapitola popisuje implementační detaily (s. 27–31), čtvrtá kapitola (s. 32–36) detaily provedených experimentů. V páté kapitole (s. 41–48) je hodnocení výsledků vyvinutých systémů.

Úvodní kapitole se kapitole se zcela nedaří definovat problém. Očekával bych, že autor uvede, jak popisy rentgenových snímků vypadají, jaké informace typicky obsahují včetně základních statistik, jak vypadá jazyk radiologických zpráv a jak se liší od běžného jazyka. Úloha je tak vlastně definována pouze nepřímo prostřednictvím trénovacích dat – cílem tedy je mít takové popisy snímků, jaké jsou v trénovacích datech. Důležitým faktem, který práce nezmiňuje, také je, jak se liší praxe při popisování snímků ve Spojených státech, odkud pochází většina datasetů a v České republice. Vysvětlení základních konceptů v první kapitole působí trochu nesystematickým dojmem: věnuje se zde prostor metodám, které se práci nepoužívají (např. rekurentní neuronové sítě), naopak jiné důležité otázky jsou vysvětleny velmi povrchně (např. jak se zařídí, že architektura Transformer se může použít pro autoregresivní modelování). Text také obsahuje tvrzení, které je potřeba doložit citací, např. že existuje velké množství medicínských aplikaci počítačového vidění a že mají vyso-

kou přesnost. Na jiných místech, například při výčtu příkladů důležitých předtrénovaných modelů strojového vidění (které se ale v práci nepoužívají) jsou citace velmi detailní.

Problém chybějících českých dat pro úlohu popisování rengenových snímků plic se autor rozhodl řešit automatickým překladem pomocí systému CUBBITT. Ten se rozhodl upřednostnit před systém Google Translate a DeepL především kvůli dostupnému API. Na tomto místě mi chybí kvantitativní zhodnocení kvality výstupů uvažovaných překladačů. K tomu by šlo využít kvalitní testovací data pro medicínský strojový překlad ze soutěžního úkolu ve strojovém překlad WMT17. Výrazně lepší kvality překladu by šlo dosáhnout dotrénováním systému CUBBITT na medicínských datech s pomocí UFAL Medical Corpus (především pokud by se použila technika backtranslation a jednojazyčná data, která se dále používají pro trénování medicínské varianty GPT-2).

Největší prostor je v práci věnovaný české verze generativního jazykového modelu GPT-2 podle postupu, který byl v minulosti vyvinut a otestován na indoevropských jazycích a lze tedy předpokládat, že bude dobře fungovat i pro češtinu. Autor zde postupoval velmi pečlivě a podle uvedených výsledků se zdá, že se mu podařilo vytvořit kvalitní český jazykový model.

Při trénování samotného systému pro popisování rentgenových snímku autor využil stejný postup jako u existujících anglických modelů: jako enkodér využil předtrénovaný model pro získání reprezentace obrázků a ten zkombinoval se systémem GPT-2, který plní roli dekodéru.

Výsledky systému jsou hodnoceny automatickými metrikami pro podobnost textu a ručním hodnocením. Autor zvolil širokou řadu automatických metrik, nediskutuje ale dostatečně jejich vhodnost pro tuto úlohu. Chybí také porovnání s tím, jaké metriky a proč používají modely pro angličtinu. Pro tento druh analýzy by bylo lepší použít ručně vytvořené české popisy snímků, spíše než automatické překlady (hrozí zde riziko přeučení na chybné překlady). V práci chybí srovnání se s nejdůležitější baseline, tj. generování popisu v angličtině a jeho následným automatickým překladem do češtiny.

Velmi kladně hodnotím, že práce zahrnuje i manuální evaluaci. V té ale opět chybí srovnání s překladem anglického systému a s původním anglickým systémem. Většina zdravotníků pravděpodobně mluví anglicky a používat anglicky mluvící systém by pro ně neměl být problém. Srovnání s původním systémem by tedy mělo být hlavním kritériem použitelnosti vyvinutého systému.

V úvodu si práce vytyčuje tři postupné cíle, které jsou potřeba k vyvinutí nástroj, který si autor klade za cíl. Závěr bohužel nekomentuje, jak se tyto dílčí cíle podařilo splnit. Z textu práce je ale zřejmé, že všechny tyto cíle se podařilo splnit více méně uspokojivě. Hlavním výsledkem analýzy výsledků je, že modely mají sklony mít vyšší přesnost (nálezy jsou obvykle správně), ale nízkou úplnost (nejčastějším druhem chyby je chybějící nález).

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 5. srpna 2022

Podpis: