

Posudek oponenta bakalářské práce

Student: Tvarožný Matuš
Téma: Zvyšování konzistence v datových sadách pro rozpoznávání textu (id 25170)
Oponent: Hradiš Michal, Ing., Ph.D., UPGM FIT VUT

1. **Náročnost zadání** průměrně obtížné zadání
2. **Splnění požadavků zadání** zadání splněno
3. **Rozsah technické zprávy** je v obvyklém rozmezí
4. **Prezentační úroveň předložené práce** 70 b. (C)

Ke struktuře práce mám několik výhrad. Celkově sice dává smysl, ale část 2.2 Transformer není v práci nijak využita, očekával bych, že v části 4.1 budou konkrétní ukázky problémů a že dále v kapitole 4 budou přesně definovány a algoritmy, které jsou dále používány. Algoritmy jsou dále popsány až v kapitole 5, ale to je už i s implementačními detaily a celkově není jednoduché dané algoritmy přesně pochopit. V textu jsem se několikrát ztrácel a bylo náročnější pochopit přesné záměry autora nebo prezentovaná fakta.
5. **Formální úprava technické zprávy** 77 b. (C)

Práce je napsaná slovensky, takže se mi špatně hodnotí jazyková úroveň. Při čtení jsem nenarazil na nějaké jasné problémy. Po formální stránce práce obsahuje jen menší nedostatky. Oceňuji, že se student snažil reprezentovat výsledky experimentů smysluplnými grafy a tabulkami, i když je otázka, jak je vhodné používat absolutní počty TP, FP, TN, FN a určitě není potřeba uvádět F1 score na pět desetinných míst. Za rovnicemi chybí interpunkce a vždy za nimi začíná nový odstavec.
6. **Práce s literaturou** 75 b. (C)

Práce se odkazuje celkově na 38 zdrojů, které jsou většinou kvalitní a dobře pokrývají řešenou úlohu. V textu jsou zdroje používány vhodně. Slabší je využití zdrojů v samotném řešení, kdy navržené přístupy a vytvořený software téměř nevyužívají vlastnosti použitých sítí pro přepis textu.
7. **Realizační výstup** 70 b. (C)

Student vytvořil postupy, jak odhalovat řádky s potencionálně chybnými přepisy a jak upravovat pozici řádků podle odezvy OCR sítě. Metody vhodně vyhodnotil. Vytvořené zdrojové kódy odpovídají textu práce, jsou částečně komentované (slovensky!) a jejich kvalita je rozumná. Do nápovědy každého scriptu by měl být doplněn popis jeho funkce. V některých případech jsou parametry součástí zdrojového kódu. Nejhorší je asi `detect_incorrect_lines.py`, ve kterém se použítá metoda musí měnit přepsáním zdrojového kódu a je tam i definovaný parametr `N`. Podobně nechápu, že v `calculator_of_incorrect_lines.py` jsou pevně definované konstanty `SHOULD_BE_PICKED`, `SHOULD_NOT_BE_PICKED` a cesty k souborům.

Přístup úpravy pozice řádků není výpočetně efektivní a ani není použitelný. Odhalování chybných přepisů pracuje čistě na úrovni zarovnání textu, čímž ignoruje informaci o jistotách na výstupu neuronové sítě.
8. **Využitelnost výsledků**

Některé informace z vyhodnocení jsou zajímavé, ale samotné metody a vytvořený software s největší pravděpodobností v projektu PERO nevyužijeme.
9. **Otázky k obhajobě**
 - Máte odhad skutečné chybovosti přepisů v použité datové sadě?
 - Jak ovlivňuje detekci chybných přepisů fakt, že využíváte neuronovou síť, která byla na těchto přepisech trénována?
 - Jestli to chápu dobře, výsledky Tabulky 6.3 jsou na řádcích již filtrovaných na rozdílnosti výstupů neuronové sítě vůči přepisům. Jak tento fakt ovlivňuje interpretaci těchto výsledků.
10. **Souhrnné hodnocení** 71 b. dobře (C)

Student se zvládl seznámit s novou oblastí a výzkumným software, navrhnout postupy řešící zadaný problém a jejich vlastnosti rozumným způsobem vyhodnotit. Navržené postupy ale jsou spíš základní, jejich realizace není efektivní a v některých ohledech jsou nevhodné. Vytvořené zdrojové kódy jsou sice použitelné, ale bylo by vhodné, aby s nimi bylo možné replikovat provedené experimenty bez nutnosti do nich zasahovat.

Prohlášení: Uděluji VUT v Brně souhlas ke zveřejnění tohoto posudku v listinné i elektronické formě.

V Brně dne: 3. června 2022

Hradiš Michal, Ing., Ph.D.
oponent