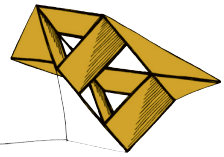

Význam syntetických dat v úloze rozpoznávání ručně psaného textu

Lukáš Soukup[1]

1 Úvod

Značná část naší historie je zaznamenána v ručně psaných dokumentech. Pro zajištění zachování těchto dokumentů a jejich sdílení s veřejností je nutná jejich digitalizace. Při digitalizaci je jedním z nejdůležitější a zároveň nejtěžších kroků rozpoznání ručně psaného textu.

Nízká výkonnost OCR (optical character recognition) systémů, které jsou standardně založeny na hlubokých neuronových sítích, může být vylepšena dotrénováních na specifickém typu dokumentů, ale pro tento krok je nezbytné mít k dispozici anotavaná data, jejichž získání je často časově i finančně náročné.

Potřeba anotovaných dat může být z části kompenzována vytvořením syntetických dat. Tvorba syntetických dat je rychlejší a levnější, než anotace reálných dat a jejich využití vylepšuje výkonnost OCR systému, což demonstruji na veřejném HKR (Handwritten Kazakh and Russian) datasetu Nurseitov et al. (2021) a na interním českém datasetu.

2 Data

Tato kapitola obsahuje přehled použitých datasetů s reálnými daty HKR dataset a interní český dataset a popis vygenerovaných syntetických dat.

2.1 HKR dataset

Handwritten Kazakh and Russian (HKR) dataset Nurseitov et al. (2021) je kolekce obrázků ručně psaných textů v azbuce a k nim odpovídající přepis strojového textu v ruském (95%) a kazašském (5%) jazyce. Tato databáze vznikla z 1500 vyplněných formulářů od 200 pisatelů a obsahuje zhruba 63 000 vět, 715 699 znaků a 106 718 slov. Autoři poskytují rozdělení dat na trénovací (70%), validační (15%) a testovací (15%) množinu. Testovací množina je dále rovnoměrně rozdělena na dvě podmnožiny, test1 je složen ze slov, které nejsou v trénovací ani validační množině a test2 obsahuje slova, která jsou i v trénovací množině, ale napsané jinými autory.

2.2 Interní dataset

Tento interní dataset byl v rámci projektu poskynut zákazníkem. Dataset byl vytvořen ze 112 historických kronik českých obcí a celkem obsahuje zhruba 126 000 řádek textu pro trénování a zhruba 16 00 řádek pro validaci.

[1] student doktorského studijního programu Aplikované vědy a informatika, obor Kybernetika, e-mail: lsoukup@kky.zcu.cz

Dataet	synth	test1	test2
HKR		51.39 / 73.42	11.55 / 17.24
HKR	y	**26.86 / 49.1**	**7.12 / 16.01**
CS		4.2 / 14.74	
CS	y	**3.08 / 12.83**	

Tabulka 1: CER / WER na testovacích sadách datasetů reálných dat HKR Nurseitov et al. (2021) a interního CS datasetu.

2.3 Syntetická data

Syntetická data byla vytvořena naším vlastním generátorem. Tento software dokáže vytvořit věrohodně vypadající ručně psaný text pomocí fontů pro ručně psané písmo. Text se generuje s využitím lokálních i globálních augmentaci, aby výsledný obrázek vypadal realisticky.

3 OCR metoda

Jako OCR metoda byla implementována stejná architektura jako v Kang et al. (2020). Tento model se skládá z konvoluční sítě, která z obrazů extrahuje příznaky, které jsou poté zpracovány transformerem enkodérem. Poté transformer dekodér z transformovaného příznakového vektoru dekóduje text.

4 Výsledky

Experimentální výsledky byly vyhodnoceny pomocí testovacích/validační sady dat. Vyhodnocujícím kritériem pro rozpoznávání ručně psaných textů byly zvoleno standarně character error rate (CER) a word error rate (WER), obojí spočteno jako levenstheinova vzdálenost od anotace.

Z výsledků v tabulce 1 můžeme vidět, že využití syntetických dat při trénování konzistentně zlepšuje výsledky této OCR metody.

5 Závěr

Přidáním syntetických dat při trénování OCR pro rozpoznání růčně psaných textů se konzistentně zlepší výkonnost. Tento jev byl experimentálně ověřen s využitím dvou datasetů, na jejichž testovacích sadách bylo dosaženo lepších výsledků s využitím syntetických dat při trénování.

Poděkování

Příspěvek byl podpořen grantovým projektem SVK1-2022-017.

Literatura

Nurseitov, D., Bostanbekov, K., Kurmankhojayev, D., Alimova, A., Abdallah, A., a Tolegenov, R. (2021) Handwritten Kazakh and Russian (HKR) database for text recognition. *Multimedia Tools and Applications*. Springer, pp. 1–23.

Kang, L., Riba, P., Rusiñol, M., Fornés, A., a Villegas, M. (2020) Pay attention to what you read: Non-recurrent handwritten text-line recognition. *arXiv preprint arXiv:2005.13044*.