Automatické rozpoznání ručně psaného textu pomocí Sequence2Sequence modelu

Lukáš Soukup[1]

1 Úvod

Automatické rozpoznání ručně psaného textu je velice náročná disciplína z oblasti umělé inteligence. Transforamce obrázků ručně psaného textu do strojově čitelného formátu umožňuje velké množství aplikací mezi něž patří například digitalizace historických, nebo administrativních dokumentů apod.

Tato práce se zabývá automatickým rozpoznáním ručně psaných historických knih v českém jazyce. Automatické zpracování je založeno na state-of-the-art modelu Seq2seq Kang et al. (2019).

2 Dataset

Dataset se skládá z historických česky psaných textů. Konkrétně se jedná ručně psané kroniky obcí, které byly v rámci projektu poskytnuty zákazníkem. Ke skenům dokumentů byly poskutnuty anotace ve formě bounding boxů a přepsaných textů na úrovni řádek textu.

3 Metoda

Pro řešení problému automatického čtení českého ručně psaného textu byl zvolen state-of-the-art model Seq2seq Kang et al. (2019). Implementace tohoto modelu byla přizpůsobena českému jazyku a poskytnutým datům.

Tento tzv. attention-based model se skládá ze tří částí: enkodér, attention mechanismus a dekodér. Enkodér začíná konvoluční neuronovou sítí (CNN), která extrahuje příznaky z obrázku. Extrahované příznaky jdou dále do Bi-directional Gated Recurrent Unit (BGRU), která zakodovává sekvenci ručně psaného textu. V dalším kroce je z zakodovných příznaků spočítána attention založená na lokaci. Dekodér je založen na jednosměrném vícevrstvém GRU. V každém kroce se provede konkatenace embedding vektoru z předešlého kroku a kontextového vektoru z attention modulu a výsledný vektor je dále zpracován pomocí GRU jednotky, která dekóduje výstupní text.

3.1 Příprava dat

Poskytnuté anotace k dokumentům byly v rámci toho projektu pouze na úrovni jednotlivých řádek. Seq2seq model ale pracuje na úrovni jednotlivých slov. Nejprve tedy bylo nutné provést segmentaci celé řádky tak, abychom dostali jednotlivá slova. K tomu bylo využito standardních metod počítačového vidění s využitím apriorních informací z anotací jako například počet slov na řádce, počty znaků v jednotlivých slovech apod. Celkově bylo vytěženo 7687

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

obrázků slov s příslušnými přepisy textu. Tato data byla rozdělena na trénovací a validační množinu náhodně v poměru 90% trénovací a 10% validační.

4 Výsledky

Experimentální výsledky byly vyhodnoceny pomocí validační sady dat. Vyhodnocujícím kritériem pro rozpoznávání ručně psaných textů byly zvoleno standarně character error rate (CER) a word error rate (WER):

$$CER = \frac{S + I + D}{N} \qquad WER = \frac{S_w + I_w + D_w}{N_w}, \tag{1}$$

kde $S(S_w)$ značí počet substituovaných znaků (slov), $I(I_w)$ značí počet vložených znaků (slov), $D(D_w)$ značí počet smazaných znaků (slov) a $N(N_w)$ značí celkový počet znaků (slov).

Metoda	**CER**	**WER**
Seq2seq + CE	17.72	43.39
Seq2seq + CE + augm	17.33	41.28
Seq2seq + CTC	13.89	38.2
Seq2seq + CTC + augm	12.82	34.9

Tabulka 1: Výsledky experimentů na validační sadě. CE - cross entropy loss, CTC - CTC loss, augm - augmentace.

5 Závěr

Přestože bylo v této úloze k dispozici prozatím pouze malé množství dat, výsledky experimentů ukazují vysoký potenciál tohoto modelu a slouží jako tkz. proof-of-concept pro další spolupráci na této problematice.

6 Seznam literatury a citace

Poděkování

Příspěvek byl podpořen grantovým projektem číslo SGS-2019-027. Výpočetní zdroje byly poskytnuty v rámci projektu ”e-Infrastruktura CZ”(e-INFRA LM2018140).

Literatura

Kang, L., Toledo, J. I., Riba, P., Villegas, M., Fornés, A., Rusiñol, M. (2019) Convolve, Attend and Spell: An Attention-based Sequence-to-Sequence Model for Handwritten Word Recognition. *Springer International Publishing*. Available from: `https://link.springer.com/chapter/10.1007/978-3-030-12939-2_32` [Accessed 22nd May 2021].