VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Identifikace popisů obrázků v tištěných dokumentech

KNN - Konvoluční neuronové sítě

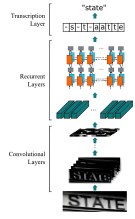
2024/2025

Radomír Bábek, Petr Volf, Přemek Janda

Úvod

OCR (Optical character recognition) je metoda, která rozpoznává tištěný či ručně psaný text a vrací jeho digitální přepis. Pro tuto techniku se již celkem standardně využívají neuronové sítě. Starší architektury bývají postaveny na kombinaci konvolučních a rekurentních neuronových sítí. Konvoluční neuronové sítě obvykle zastávají roli extraktoru rysů. Díky jejich správné utilizaci je možné detekovat text na obrázcích s různým osvětlením a na odlišných pozadích. Rekurentní neuronové sítě se dále starají o překlad rysů na znaky. V kontextu neuronových sítí lze mluvit o architektuře encoder-decoder, kdy je encoder tvořen konvolučními vrstvami a dekodér vrstvami rekurentními.

V poslední době jsou prováděny experimenty s novými moderními architekturami. Konvoluční neuronové sítě jsou nahrazovány vision transformery postavenými na attention mechanismu. Experimentuje se také s úplným vyloučením encoderu: výsledné neuronové sítě tak bývají postaveny na principu **decoder-only**. Jako dekodery se využívají například generativní předtrénované transformery (GPT), které vstupní obrázky kódují na tokeny a následně iterativně generují znaky až do rozpoznání konce textové sekvence. Generativní transformery se ukázaly jako vysoce efektivní napříč různými typy úloh, aktuálně jsou nejvíce používané v rámci velkých jazykových modelů (LLM).



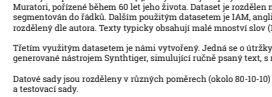
[4]

Datové sady

V práci chceme dosáhnout co nejlepších výsledků na některém z ručně psaných datasetů. Zvolili jsme proto dataset **LAM**[1], který obsahuje zápisy Italského autora Lodovico Antonio Muratori, pořízené během 60 let jeho života. Dataset je rozdělen na stránky a dále segmentován do řádků. Dalším použitým datasetem je IAM, anglicky ručně psaný dataset rozdělený dle autora. Texty typicky obsahují malé množství slov (1–2).

Třetím využitým datasetem je námi vytvořený. Jedná se o útržky z italského korpusu (oscar), generované nástrojem Synthtiger, simulující ručně psaný text, s množstvím perturbací.

Datové sady jsou rozděleny v různých poměrech (okolo 80-10-10) na trénovací, validační a testovací sady.



A — *Syntetický dataset*
1 257 000 vzorků

B — *LAM*
25 000 vzorků

C — *IAM*
115 000 vzorků

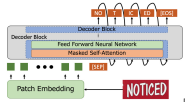
TrOCR

Model TrOCR[3] je postaven na architektuře encoder–decoder s využitím vizuálního transformeru **DeiT** pro zpracování obrazu a jazykového transformeru pro generování textu. Pro ručně psané texty jsme zvolili verzi **small** kvůli rychlosti a jednoduchosti. Při prvním finetuningu jsme dosáhli velmi špatných výsledků (CER kolem 0.9)

kvůli chybné dvojí normalizaci vstupních dat. Po nápravě se sice přesnost zlepšila, ale model stále selhával – rozpoznával jen polovinu znaků, patrně kvůli škálování vstupních obrázků na 384 × 384, což mohlo narušit čitelnost kurzívy. Kvůli těmto omezením jsme se rozhodli pro jiný přístup.

DTrOCR

Architektura DTrOCR[2], navržená Masatem Fujitakem, je postavena na decoder-only přístupu s využitím modelu **GPT-2** od OpenAI. Jelikož parametry původního modelu nejsou veřejně dostupné, bylo nutné ho v rámci práce natrénovat od začátku. Na rozdíl od počátečního tréninku TrOCR jsme k problému přistoupili systematičtěji a model nejprve rozsáhle předtrénovali na latinských textech, než jsme přistoupili k finetuningu.



[2]

Implementace

V rámci implementace jsme se rozhodli využít existující řešení DTrOCR a trénovat s využitím CTC loss a SAM optimizeru. Inspirace na zmíněné úpravy pochází z [5]. Váha našeho projektu je především v experimentech nad zmíněnou architekturou, která je poměrně nová (2023) a stojí za hlubší prozkoumání. Mnohá souvislost implementace je generování syntetických datasetů, které se odlišují dle cílové úlohy (ručně psaný text, tištěný text, umělecký text...).

Model byl nejprve předtrénován na 1,2 milionech vstupů ze syntetické datové sady průběhem jedné epochy. Pro generování Předtrénovaný model byl následně dotrénován nad LAM datasetem. Bylo provedeno 35 epoch nad trénovací datovou sadou, trénování bylo zastaveno, aby se předešlo neefektivnímu overfittingu. Z výsledků implementace lze odvodit závěry o důležitosti variability a velikosti datové sady pro předtrénování a výkonu testované architektury.

CTC loss
(Connectionist Temporal Classification)

Používána v seq-to-seq úlohách pro signálové úlohy. V našem případě byla použita pro predikci sekvence znaků bez nutnosti explicitního zarovnávání mezi vstupní a výstupní sekvencí.

SAM
(Sharpness Aware Minimization)

Myšlenkou použití SAM na trénování modelu DTrOCR je snaha o zvýšení odolnosti vůči malým změnám, které mohou při detekci znaků nastat. Model by měl být schpen lépe generalizovat napříč různými podmínkami, druhý textů, etc.

Metriky

Nejdůležitějšími sledovanými metrikami při trénování modelu byly hodnota ztrátové funkce při validaci a hodnoty *CER* (character Error Rate) a WER (Word Error Rate) zvláště během vyhodnocení modelu.

$$\text{CER} = \frac{S + D + I}{N}$$

$$\text{WER} = \frac{S + D + I}{W}$$

$$\text{Accuracy} = \frac{\text{Počet správně rozpoznaných slov}}{\text{Celkový počet slov}}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

S počet substitucí (špatně rozpoznaných znaků nebo slov)
D počet delecí (chybějících znaků nebo slov)
I počet insercí (nadbytečných znaků nebo slov)
N počet znaků v referenčním textu (pro CER)
W počet slov v referenčním textu (pro WER)

Zhodnocení

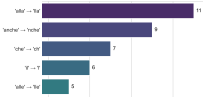
	fine-LAM	fine-IAM	fine-Synth	pre-LAM	pre-IAM	pre-Synth
↓ **Total CER**	0.0781	0.5251	0.1674	0.2865	0.1775	0.0309
↓ **Total WER**	0.1381	0.6509	0.3193	0.4224	0.2353	0.0678
↑ **Accuracy**	0.5819	0.3801	0.4685	0.1655	0.7631	0.8369
↑ **F1**	0.3926	0.1337	0.2830	0.0880	0.4624	0.7062
↓ **Average CER**	0.0582	0.4919	0.1796	0.2479	0.1175	0.0222
↓ **Median CER**	0.0000	0.5000	0.0400	0.2500	0.0000	0.0000
↓ **Std Dev CER**	0.1120	0.5542	0.3142	0.1799	0.2562	0.0700
↓ **Average WER**	0.1098	0.6595	0.3514	0.3638	0.2353	0.0595
↓ **Median WER**	0.0000	1.0000	0.1667	0.3750	0.0000	0.0000
↓ **Std Dev WER**	0.1736	0.5740	0.4972	0.2401	0.4303	0.1737

↓, ↑ označují výši respektive nižší hodnotu jako lepší

Předtrénování prokázalo vysoký výkon a nízkou chybovost na obou hlavních metrikách *CER* i *WER*. Následné doladění na reálné datové sadě LAM však vedlo ke zhoršení těchto metrik, přičemž celková *CER* vzrostla na *0,0781*, celková *WER* na *0,1381* a přesnost klesla na *0,5819*.

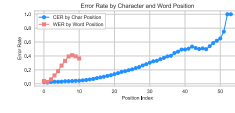
Vyhodnocení naznačuje, že krok doladění nezlepšil robustní základ vytvořený syntetickým předtrénováním napříč všemi datovými sadami. Nicméně lze tvrdit, že hlavní cíl byl splněn, neboť krok z předtrénovaného modelu na vyladěný je značný.

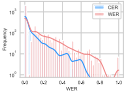
Z grafu nejčastějších chyb při rozpoznávání slov lze obecně pozorovat trend, kdy model vynechává počáteční samohlásky (např. 'a', 'i') nebo koncové hlásky u kratších, běžně užívaných italských slov.



Model má také tendenci dělat více chyb čím dále postupuje v generování nebo rozpoznávání sekvence. Začátky sekvencí jsou rozpoznávány s vyšší přesností než jejich konce. Chyby se akumulují a model má potíže s udržením kontextu na delší vzdálenosti (error propagation).

Pro obě metriky (*CER*, *WER*) jsou zastoupeny nejvyšší frekvence u nízkých hodnot chybovosti, vzhledem k logaritmickému měřítku osy *y*.





[1] Silvia Cascianelli, Vittorio Pippi, Maarand Martin, Marcella Cornia, Lorenzo Baraldi, Kermorvant Christopher, and Rita Cucchiara. The lam dataset: A novel benchmark for line-level handwritten text recognition. In International Conference on Pattern Recognition, 2022.

[2] Masato Fujitake. Dtrocr: Decoder-only transformer for optical character recognition, 2023.

[3] Minghao Li, Tengchao Lv, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models, 2021.

[4] Ritika Rai, Srushti Shitole, Pratiksha Sutar, Swapnali Kalbhone, and Jayashree Jadhav. Automatic license plate recognition using yolov4 and tesseract ocr. International Journal of Innovative Research in Computer and Communication Engineering, 10:1656, 01 2008.

[5] Yuting Li, Dexiong Chen, Tinglong Tang, and Xi Shen. Htr-vt: Handwritten text recognition with vision transformer. Pattern Recognition, 158:110967, 2025.