VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

http://fit.vut.cz

Konvoluční neuronové sítě 2023/2024

Handwritten text recognition (OCR)

Bc. Pavel Raur (xraurp00), Bc. Jiří Žilka (xzilka11), Bc. Marián Zimmermann (xzimme03)*

Abstrakt

Projekt se zabývá rozpoznáváním ručně psaného písma pomocí OCR (Optical Character Recognition) s použitím self-supervised učení. V rámci projektu byly porovnány různé postupy self-supervised trénování OCR a následného dolaďování takto natrénovaného modelu. Byla zvolena architektura sítě pro následné experimentování s těmito metodami a byli provedeny expirimenty. Následně tímto postupem bude vytvořeno několik kandidátních řešení pro porovnání.

Klíčová slova: OCR — self-supervised učení — strojové rozpoznávání ručně psaného textu

Přiložené materiály: Stáhnutelný Kód

*xraurp00@stud.fit.vut.cz, xzilka11@stud.fit.vut.cz, xzimme03@stud.fit.vut.cz,
Fakulta informačních technologií, Vysoké učení technické v Brně

1. Úvod

Tato práce se zabývá natrénováním OCR (Optical Character Recognition) modelu pro dataset historického ručně psaného písma. Většinová část datasetu je bez anotací. Projekt se zabývá zejména prozkoumáním metod self-supervised učení a provádením experimetů k nalezení optimální kombinace metod, které povedou k co nejlepšímu výsledku na daném datasetu.

2. Zvolený model a dataset

Pro experimenty byl vybrán model *TrOCR* [1], který používá transformerovou architekturu pro rozpoznávání textu na obrázcích a jeho přepis. Jak je popsáno v [3] model je inovativní *state-of-the-art* řešení využívající transformery v oblasti, kde tradičně byly použity zejména konvoluční neuronové sítě. Zároveň dosahuje kompetitivních výsledků s tradičními typy sítí. S ohledem na rostoucí popularitu transformerů, ne jen v oblasti zpracování textu a obrazu, jsme se rozhodli právě pro tento typ architektury.

2.1 Dataset

Použitý dataset je tvořený výřezy obrázků ručně psaného historického textu v anglickém jazyce. V každém výřezu je jediný řádek textu. Dataset obsahuje celkem 1153035 vzorků textu, z toho pouze 1% dat je anotovaných (11469 vzorků). Anotovaná data jsou rozdělena na trénovací (80%), testovací (7,5%) a validační sadu (12,5%).

[1] https://github.com/rsommerfeld/trocr

2.2 Augmentace dat

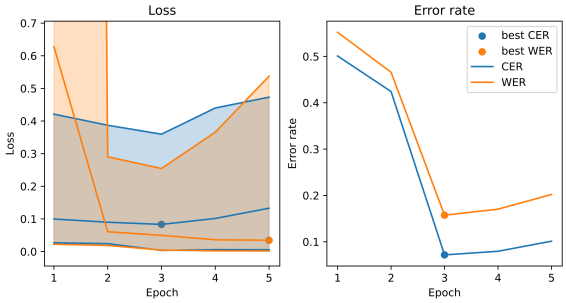
Pro zlepšení kvality modelu byla použita augmentace dat, což je běžný způsob využívaný při self-supervised učení. Byly zvoleny následujíci augmentace:

- Mixup
- Cutmix
- Horizontálni skosení
- Natáhnutí/smrštění řádku
- Gaussian noise
- Gaussian square - přidáni šumu v podobě čtverce, který zakrýva náhodné místo o velikosti max 40px
- Barevná maska

3. Postup trénovaní

Pro trénování byl použit model *microsoft/trocr-base-stage1*, který poskytuje základní váhy pro inicializaci TrOCR modelu. Váhy v tomto inicializačním modelu byly vybrány z modelů *BEiT* a *RoBERTa* pro inicializaci enkodéru pro detekci textu v obraze a dekodéru

pro přepis do textové podoby.[2] Tento model byl trénován 3 epochy na anotovaných datech. Ako vidno na 1 tak po 3 epochách dosahoval nejlepší výsledky.



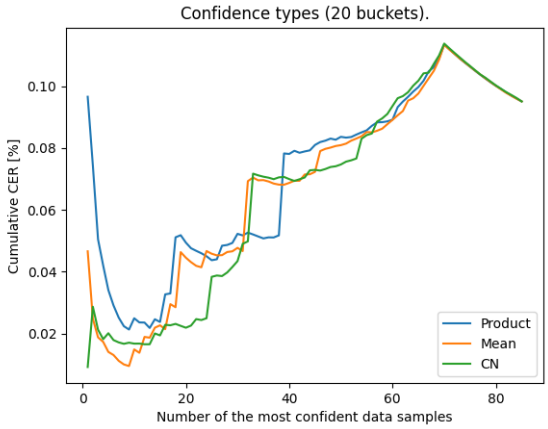
Obrázek 1. Vývoj CER a WER

3.1 Self-supervised učení

Tento model, který byl natrénován, byl následně použit k vytvoření anotací pro neanotovanou část datové sady. K tomu bylo nutné vybrat vhodnou metriku. Porovnali jsme tři metriky a získali jsme pro ně následující plochy pod křivkou:

- Confidence product - 5.937035306340725
- Confidence mean - 5.643832876624217
- Confidence confusion network - 5.480604598787505

Na základě těchto výsledků jsme zvolili confusion network jako metriku. Na grafu 2 můžeme vidět, jak moc jsou chybové pseudo-labely při výběru dat pomocí této metriky.



Obrázek 2. Porovnání metrik

Pomocí zvolené metriky confusion network[2] se postupně vybíralo určité procento nejlepších anotací(10%, 20% ... 100%) a přidali se do datasetu.

[2] https://github.com/DCGM/pero-ocr

Výchozí model	množství použitých labelů	počet epoch trénován

Tabulka 1. Caption

4. Vyhodnocování kvality modelu

Pro vyhodnocování modelů jsme použili *CER* (Character Error Rate) a *WER* (Word Error Rate).[1] Vzorec pro výpočet:

$$CER = \frac{S+D+I}{S+D+I+C} \quad (1)$$

Kde:

- S(Substitutions) = počet správných znaků zaměněných za chybné,
- D(Deletions)= počet chybějících znaků,
- I(Insertions) = počet znaků vložených navíc,
- C(Correct) = počet správných znaků.
- (toto je normalizovaná podoba CER)

Analogicky, WER je vypočítán nad slovy, na místo znaků. Někdy se udávají v podobě přesnosti (Accuracy), konkrétně *CAR* (Character Accuracy Rate) a *WAR* (Word Accuracy Rate).

$$CAR = 1 - CER \quad (2)$$

$$WAR = 1 - WER \quad (3)$$

5. Porovnání výsledků

Baseline model TrOCR Stage 1 byl před trénováním vyhodnocen na všech anotovaných datech a dosáhl následujících hodnot:

- CER = 0.3048
- WER = 0.5001

Model TrOCR Handwritten dosáhl hodnot:

- CER = 0.1768
- WER = 0.3524

Náš model, natrénovaný supervised příspupem z TrOCR stage1 dosáhl následujících výsledků:

- CER = 0.0717
- WER = 0.1573

Tento model byl následně použit pro generování pseudo-labelů pro další etapu trénování.

[[Doplnit data do tabulky!]] Modely natrénované na strojově anotovaných datech pak dosáhly následujících výsledků:

6. Kdo co udělal

Pavel Raur

- Výběr základního modelu pro trénování a použité architektury OCR.
- Základ kódů pro trénování modelu.
- Porovnání a selekce metrik pro výběr pseudolabelů.

Marián Zimmerman

- Výběr typů augmentací dat.
- Implementace augmentací.
- Dokumentace projektu.

Jiří Žilka

- Trénování OCR.
- Ukládání statistik.
- Dokumentace projektu.

Literatura

[1] LEUNG, K. Evaluate OCR Output Quality with Character Error Rate (CER) and Word Error Rate (WER). *Towards Data Science*. 2021. Dostupné z: https://towardsdatascience.com/evaluating-ocr-output-quality-with-character-

[2] LI, M., LV, T., CUI, L., LU, Y., FLORENCIO, D. et al. *TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models*. 2021. Dostupné z: https://huggingface.co/microsoft/trocr-base-stage1.

[3] LI, M., LV, T., CUI, L., LU, Y., FLORÊNCIO, D. A. F. et al. TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models. *CoRR*. 2021, abs/2109.10282. Dostupné z: https://arxiv.org/abs/2109.10282.