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 [5] 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%).

### 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:

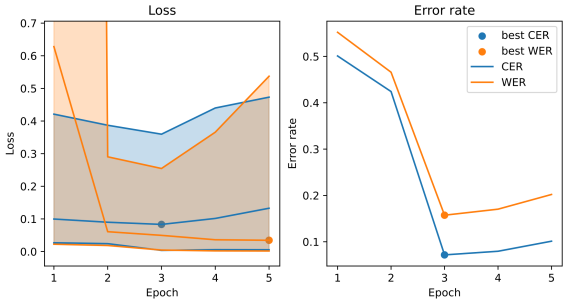
- Mixup
- Cutmix
- Rotace řádku o malý počet stupňů
- Horizontálni skosení
- Natáhnutí/smrštění řádku
- Gaussian noise
- Gaussian square - přidáni šumu v podobě čtverce, který zakrýva náhodné místo o velikosti max 40px
- Barevná maska

Mixup, cutmix a rotace byly určeny pro samostatný experiment. Ten měl zjistit, o kolik se síť zlepší při použití těchto typů augmentací.

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

# 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.[4] Tento model byl trénován 3 epochy na anotovaných datech. Jak je vidět na obrázku 1 tak po 3 epochách dosahoval nejlepší výsledky.




**Obrázek 1.** Vývoj CER a WER

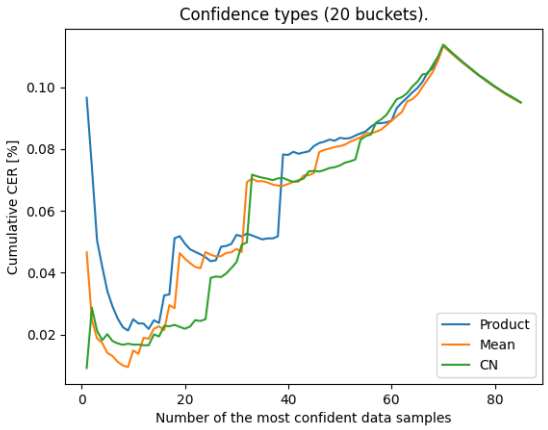
## 3.1 Self-supervised učení

Tento model, který byl natrénován, byl následně použit k vytvoření anotací pro neanotovanou část datové sady. K tomu bylo nutné vybrat vhodnou metriku[1]. Porovnali jsme tři metriky a získali jsme pro ně následující plochy pod křivkou:

- Confidence product - 5.9370
- Confidence mean - 5.6438
- Confidence confusion network - 5.4806

Na základě těchto výsledků jsme zvolili confusion networks pro výběr dat pro další trénování. Popis fungování confusion networks je např. zde [2]. Konkrétně jsme použili $1/confusion$, kde confusion je množství různých přepisů znaku na dané pozici textu, získaných z původního a augmentovaných verzí textu. Augmentace mixup, cutmix a rotace zde nebyly použity.

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 se postupně vybíralo určité procento nejlepších anotací(10%, 20% ... 100%) a přidali se do datasetu. Kód pro výpočet confusion networks byl převzat z repozitáře pero-ocr[2].

Pro prevenci výběru jen krátkých labelů, které jsou v průměru méně chybné, byla data rozdělena do bucketů podle délky. Celkem bylo vytvořeno 20 bucketů. Metriky byly počítány nad těmito rozděleními. Např. první hodnota v grafu odpovídá součtu prvních hodnot v každém bucketu, děleném 20.

# 4. Vyhodnocování kvality modelu

Pro vyhodnocování modelů jsme použili *CER* (Character Error Rate) a *WER* (Word Error Rate).[3] 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).

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

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

| Model | použité pseudo-labely [%] | počet epoch | learning rate | CER | WER |
|---|---|---|---|---|---|
| trocr-base-stage1 | - | - | - | 0.29820 | 0.49670 |
| trocr-base-handwritten | - | - | - | 0.17661 | 0.35244 |
| stage1-supervised | 0 | 3 | $5e-5$ | 0.06330 | 0.14670 |
| stage1-self-supervised | 10 | 13 | $1e-6$ | 0.04997 | 0.12471 |

**Tabulka 1.** Tabulka ukazující výsledky trénování – první 2 modely v tabulce jsou originální modely TrOCR sloužící ke srování.

## 5. Porovnání výsledků

Baseline model byl TrOCR Stage 1 a TrOCR handwritten. Stage 1 souží jako základ pro další trénování a byl použit pro dotrénování na našem datasetu. TrOCR handwritten je uveden pro srovnání, jde o model natrénovaný pro obdobnou úlohu.

Model stage1-supervised je model, který byl supervised metodou natrénován na anotovaných datech, následně byl použit pro přepis neanotovaného datasetu pro další etapu trénování. Celkové výsledky trénování jsou vidět v tabulce 1.

Bohužel se nepodařilo dokončit všechny experimenty včas kvůli nutnosti upravit konfigurace learning rate, aby se modely rychle nepřetrénovaly na vyším počtu labelů.

## 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 pseudo-labelů.
- Dokumentace projektu.

Marián Zimmerman

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

Jiří Žilka

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

## Literatura


[1] KIŠŠ, M., BENEŠ, K. a HRADIŠ, M. AT-ST: Self-training Adaptation Strategy for OCR in Domains with Limited Transcriptions. In: *Lecture Notes in Computer Science*. Springer International Publishing, 2021, s. 463–477. Dostupné z: http://dx.doi.org/10.1007/978-3-030-86337-1_31. ISBN 9783030863371.

[2] KIŠŠ, M., HRADIŠ, M., BENEŠ, K., BUCHAL, P. a KULA, M. *SoftCTC – Semi-Supervised Learning for Text Recognition using Soft Pseudo-Labels*. 2023.

[3] 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-

[4] 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.

[5] 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.


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