Identifikace popisů obrázků v tištěných dokumentech

2024/2025

Patrik Gáfrik (xgafri00)
Adrián Horváth (xhorva14)
Ondřej Bahounek (xbahou00)

1 Definícia úlohy

Cieľom je identifikovať časti textu na stránke, ktoré súvisia s obrázkami. Môže ísť o priame popisky alebo textové pasáže, ktoré obrázky opisujú. Máme k dispozícii rozsiahly dataset stránok s OCR výstupmi, ktorý obsahuje text aj detegované obrázky. Rozhodli sme sa túto úlohu riešiť dvoma krokmi:

1. **Detekcia popiskov** – Natrénujeme neurónovú sieť na rozpoznávanie popiskov obrázkov. Vstupom bude obrázok stránky, výstupom obrázok s anotovanými popiskami.

2. **Hľadanie popiskov v texte** – Pomocou modelu CLIP, ktorý už je natrénovaný na prepojovanie textu s obrázkami, budeme v OCR výstupe vyhľadávať texty súvisiace s obrázkami. Vstupom bude obrázok a text z OCR, výstupom zoznam textových úsekov, ktoré sa vzťahujú na daný obrázok.

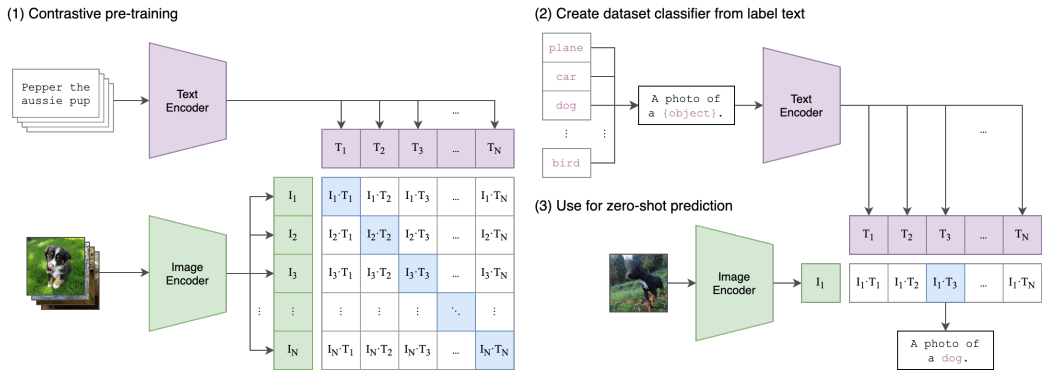
2 Prehľad existujúcich riešení

Existuje viacero prístupov na riešenie problému identifikácie častí textu na stránkach, ktoré súvisia s obrázkami. Tieto prístupy kombinujú techniky spracovania prirodzeného jazyka (NLP), počítačového videnia a hlbokého učenia.

2.1 OpenAI CLIP

CLIP (Contrastive Language-Image Pre-Training) je neurónová sieť trénovaná na rôznych pároch (obrázok, text). Môže byť inštruovaná v prirodzenom jazyku, aby predpovedala najrelevantnejší textový úryvok pre daný obrázok, bez priamej optimalizácie na túto úlohu, podobne ako schopnosti zero-shot modelov GPT-2 a GPT-3.

CLIP dosahuje výkonnosť pôvodného ResNet50 na ImageNet „zero-shot“ bez použitia ktoréhokoľvek z pôvodných 1,28 milióna označených príkladov, čím prekonáva niekoľko hlavných výziev v počítačovom videní.



Obr. 1: OpenAI CLIP Architecture

2.2 Ultralytics YOLO

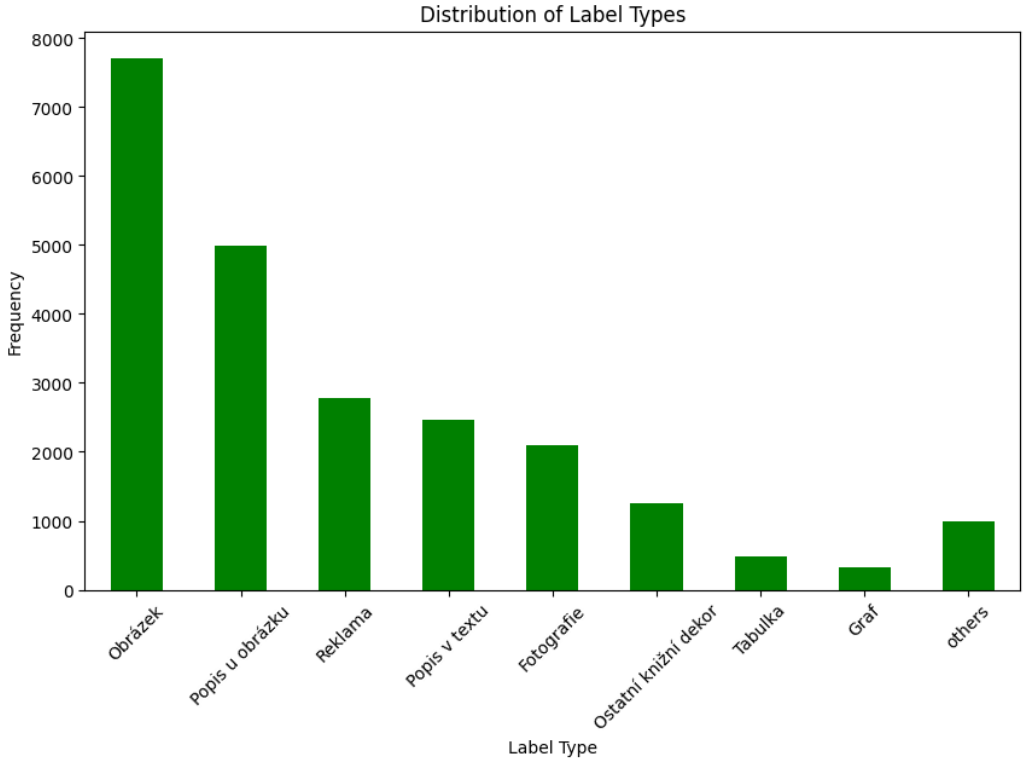
Ultralytics YOLO je model trénovaný na detekciu objektov a segmentáciu obrazu v reálnom čase, postavený na hlbokom učení. Umožňuje trénovanie a nasadzovanie modelov na rôznych hardvérových platformách, od edge zariadení po cloudové riešenia. Knižnica poskytuje optimalizované algoritmy na rýchlu a presnú inferenciu, čím je vhodná pre široké spektrum aplikácií, ako sú autonómne systémy, analýza videa alebo priemyselná automatizácia.

3 Dataset

Pre trénovanie siete bol využitý dataset `https://label-studio.semant.cz/projects/16/data?tab=363`, ktorý obsahuje 5765 rôznych článkov s textom a obrázkami, kde 80% datasetu bolo využitých pre trénovanie, 10% pre validáciu a 10% pre testovanie.

3.1 Rozloženie labels

Obrázky v datasete obsahujú niekoľko tried (labels), ktoré označujú rôzne objekty na obrázku. Pre zjednodušenie baseline riešenia a kvôli potrebám modelu YOLO budeme používať len jednu triedu - "Popis u obrázku", ktorá označuje malé popisky vedľa obrázkov.



Obr. 2: Početnosť labels

3.2 Príprava

Pre trénovanie baseline riešenia bolo nutné extrahovať labels z datasetu a previesť ich do formátu, ktorý akceptuje model YOLOv8. To znamená znormalizovať hodnoty súradníc stredu, šírky a výšky do rozsahu 0,1 a pridať CLASS_ID, ktoré označuje label - v našom prípade má vždy hodnotu 0 a označuje label "Popis u obrázku".

CLASS_ID	x_center	y_center	width_norm	height_norm
0	0.80634	0.47793	0.30839	0.04391

4 Trénovanie modelu

Trénovaný model využíval predtrénované váhy (yolov8m.pt), trénoval sa na 60 epoch s použitím adaptívneho optimizéra a jeho vstupom bola trénovacia časť datasetu. Pri trénovaní nás zaujímali metriky:

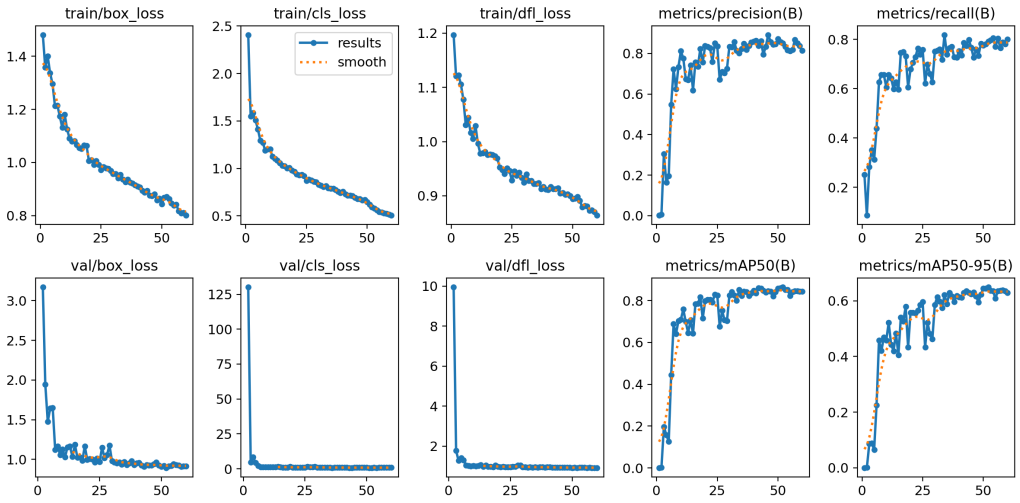
1. **Train Loss** - predstavuje chybu modelu na tréningovej množine počas učenia. Tieto hodnoty ukazujú, ako dobre model dokáže predpovedať ohraničujúce boxy, klasifikovať objekty a spresniť predikcie. Hodnoty train/box_loss, train/cls_loss a train/dfl_loss sa postupne znižujú, čo naznačuje, že model sa efektívne učí.

2. **Validation Loss** - predstavuje chybu modelu na validačnej množine, ktorá nebola použitá počas tréningu. Tieto hodnoty ukazujú, ako dobre model generalizuje na neznáme dáta. Hodnoty val/box_loss, val/cls_loss a val/dfl_loss sa taktiež znižujú, čo ukazuje, že model dobre generalizuje na neznáme dáta.

3. **Precision and recall** - Presnosť a odosah by sa mali postupne zvyšovať. Vysoká presnosť, ale nízky odosah: Model je konzervatívny a vynecháva veľa objektov (falošne negatívne predikcie). Vysoký odosah, ale nízka presnosť: Model deteguje veľa objektov, ale zahŕňa aj veľa falošne pozitívnych predikcií. Hodnoty metrics/precision(B) a metrics/recall(B) sa postupne zlepšujú, pričom presnosť dosahuje vysoké hodnoty, čo naznačuje menej falošne pozitívnych predikcií.

4. **mAP metriky** - mAP50 a mAP50-95 sú kľúčové ukazovatele výkonu modelu. Tieto hodnoty by sa mali postupne zvyšovať, pričom mAP50-95 je náročnejšie optimalizovať. Hodnoty metrics/mAP50(B) a metrics/mAP50-95(B) sa stabilne zvyšujú, pričom mAP50 dosahuje hodnoty nad 0.85, čo naznačuje silný výkon detekcie.

5. **Rýchlosti učenia** - Rýchlosti učenia (lr/pg0, lr/pg1, lr/pg2) sa postupne znižujú, čo pomáha modelu jemne doladiť svoje váhy počas tréningu.

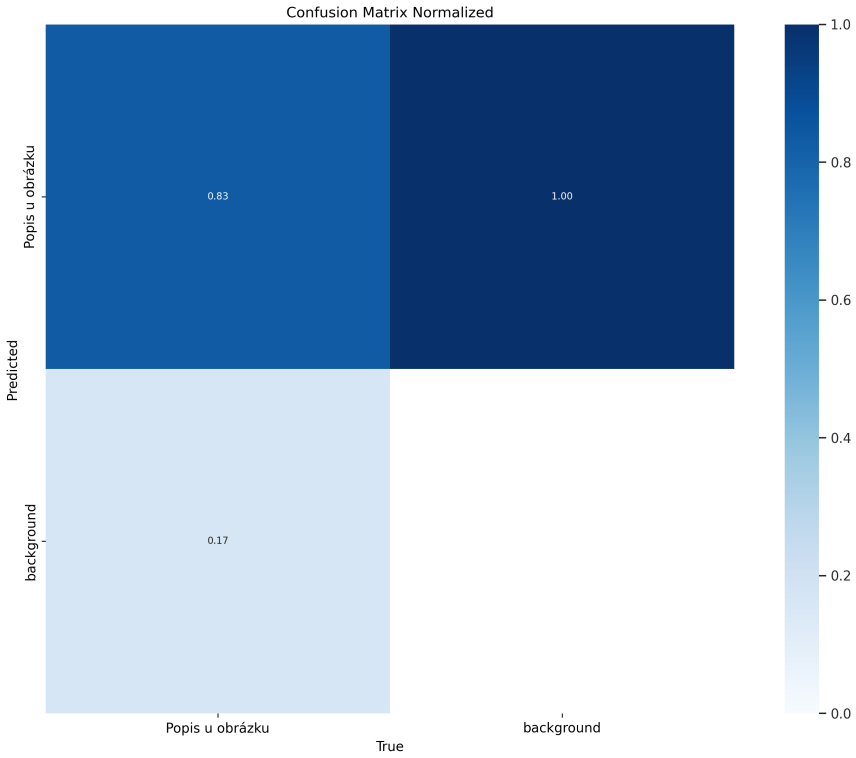


Obr. 3: Štatistiky trénovania

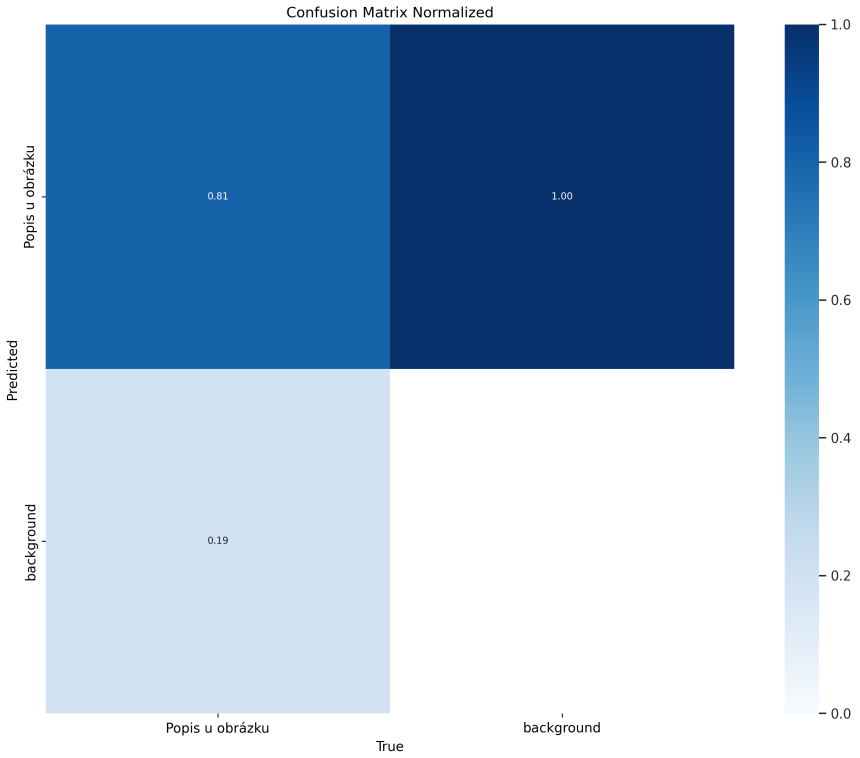
4.1 Zhodnotenie modelu

Po dokončení tréningu bol model YOLOv8 testovaný na samostatnej testovacej množine, ktorá nebola použitá počas tréningu ani validácie. Testovanie bolo vykonané s cieľom vyhodnotiť schopnosť modelu generalizovať na neznáme dáta a poskytnúť objektívne metriky jeho výkonu. Testovanie bolo vykonané pomocou metódy val z knižnice ultralytics, ktorá umožňuje vyhodnotenie modelu na testovacej množine. Vstupom testovania boli tiež nové natrénované váhy.

Pre zhodnotenie modelu sme porovnali matice zámen z trénovania oproti testovaniu, ktoré sú takmer identické, čo znamená, že natrénovaný model funguje rovnako aj na dátach, ktoré pri tréningu nevidel.



Obr. 4: Confusion Matrix (Training)



Obr. 5: Confusion Matrix (Test)

5 Ďalší postup

Súčasné riešenie využíva model YOLOv8 na detekciu popisov obrázkov v tlačených dokumentoch. Tento model je schopný efektívne detekovať popisy, ktoré sa nachádzajú priamo pri obrázkoch, napríklad v podobe textových polí alebo popisov umiestnených v blízkosti obrázkov. Výsledky testovania ukazujú, že model dosahuje vysoké hodnoty presnosti a odosahu pri detekcii takýchto popisov, čo naznačuje, že YOLOv8 je vhodný na riešenie úloh, kde je text jasne oddelený a vizuálne prepojený s obrázkami. Model YOLOv8 je navrhnutý na detekciu objektov na základe ich vizuálnych vlastností. Preto nie je schopný identifikovať popisy obrázkov, ktoré sú rozptýlené v texte alebo sa nachádzajú vo forme textových pasáží, ktoré obrázky opisujú. Napríklad, ak sa popis obrázka nachádza v odstavci textu vzdialenom od samotného obrázka, model YOLOv8 nedokáže vytvoriť spojenie medzi textom a obrázkom. Na prekonanie týchto limitácií navrhujeme použiť model OpenAI CLIP (Contrastive Language–Image Pretraining), ktorý je navrhnutý na multimodálne spracovanie obrazových a textových dát. CLIP

dokáže vytvárať spojenia medzi obrázkami a textom na základe ich významu, čo ho robí vhodným na identifikáciu popisov obrázkov v texte.

6 GIT

GitHub link: github.com/PoweredByAdrian/KNN