Identifikace popisů obrázků v tištěných dokumentech

Patrik Gáfrik, Adrián Horváth, Ondřej Bahounek

Abstrakt

Táto práca sa zaoberá identifikáciou textových častí na stránkach, ktoré súvisia s obrázkami – či už ide o priame popisky pod nimi, alebo textové pasáže roztrúsené v hlavnom obsahu. Navrhujeme dvojstupňový prístup. V prvom kroku detegujeme obrázky a ich bezprostredné popisky pomocou modelu YOLOv8, ktorý sme doladovali na vlastnom anotovanom datasete s dvoma triedami („Obrázok“ a „Popis pod obrázkom“). Dosiahnuté hodnoty mAP potvrdzujú, že model dokáže robustne lokalizovať dvojice obrázok – popis aj v dokumentoch s náročným rozvrhnutím.V druhom kroku vyhľadávame súvisiace popisky voľne v texte. Využívame multimodálny model CLIP: pre každý OCR riadok aj detegovaný obrázok počítame embedding a na základe kosínovej podobnosti hodnotíme relevanciu. Teksty priamo pod obrázkami filtrujeme, aby sme sa zamerali len na popisky v tele dokumentu, a spájame susedné riadky do koherentných blokov.

Úvod

V naskenovaných historických dokumentoch chýba strojovo čitateľné prepojenie obrázkov s textom, takže hľadanie „odseku, ktorý opisuje tento obrázok“ zostáva prácne. Automatické párovanie by uľahčilo vyhľadávanie, tvorbu znalostných grafov aj bádateľskú prácu. Cieľom je identifikovať časti textu na stránke, ktoré súvisia s obrázkami. Môže ísť o priame titulky alebo textové pasáže, ktoré obrázky opisujú či vysvetľujú. K dispozícii máme rozsiahly dataset fotografií stránok, z ktorých časť je označená štítkami identifikujúcimi objekty ako obrázky, diagramy a popisky. Ďalej máme OCR výstupy týchto stránok, ktoré obsahujú text. Medzi najpoužívanejšie prístupy patria vizuálny detektor YOLOv8 a multimodálny model OpenAI CLIP. YOLOv8 jedným prechodom obrázka lokalizuje objekty či rámiky s textom podľa ich vzhľadu, takže spoľahlivo nájde titulky priamo pri obrázkoch, no nedokáže priradiť rozptýlené textové pasáže. Naproti tomu CLIP mapuje obrázky a texty do spoločného vektorového priestoru a kontrastným trénovaním sa učí posudzovať ich sémantickú podobnosť, vďaka čomu dokáže bez ďalšieho ladenia (zero-shot) identifikovať popisy aj v odsekoch vzdialených od obrázka. Kombináciou oboch modelov tak možno pokryť titulky blízko obrázkov aj rozptýlené textové popisy.

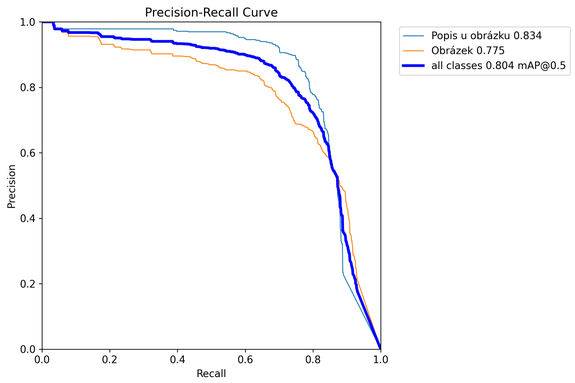
Detekce popisků u obrázků (YOLOv8m)

1. Príprava datasetu
   - ~6000 vzoriek
   - Dve triedy:
     - Popis u obrázku
     - Obrázok/Fotografia
   - Rozdelenie: train / val / test (≈ 80 / 10 / 10 %).
   - Normalizácia súradníc pre model YOLO

2. Tréning modelu
   - YOLOv8m (≈ 25,9 M parametrov) sme učili na clustre MetaCentrum nad dvojtriednym datasetom („Popisek pod obrázkem“, „Obrázek“)

- Pôvodne detegované len popisky, neskôr aj obrázky

3. Testovanie
   - Na nevidených dátach: mAP@0.5 = 0,804
   - Manuálne párovanie popiskov s obrázkami – 100% na testovacej vzorke



Detekce popisků v textu (OpenAI CLIP)

1. Vstupy – získať bbox obrázkov (YOLO/anotácie) a OCR riadky so súradnicami.

2. Filtrácia – odstrániť OCR riadky prekrývajúce popisky pod obrázkami (IoU).

3. Príprava dát pre CLIP
   - Orezať každý obrázok → samostatný obrazový vstup.
   - Každý OCR riadok spracovať samostatne (≈ 70 tokenov limit modelu).

4. Výpočet embeddingov
   - Obrazové embeddingy: pôvodná obrazová časť CLIP
   - Textové embeddingy: viacjazyčný transformer Multilingual-CLIP.

5. Porovnanie podobnosti
   - Pre každý riadok sa vypočíta kosínová podobnosť s obrázkom
   -Ak je podobnosť riadku vyššia než nastavený prah, riadok je označný ako kandidát na popisok

6. Zoskupovanie do blokov
   - Zoradiť kandidátov podľa skóre, nastaviť parameter pre výber k najlepších na stránku.
   - Ku každému vybranému riadku pridať susedné riadky nad/pod, ak tiež presiahnu prah → vytvoriť súvislý textový blok.

7. Výstup a validácia
   - Nakresliť bloky textu na pôvodné obrázky ako detegované popisky
   - Vizualizovať bounding boxy pre manuálnu kontrolu a automatizované porovnanie prienikov bounding boxov s anotovanými dátami.

8. Ladenie
   - Optimalizovať prah podobnosti, počet top kandidátov a pravidlá zoskupovania

Výsledky práce

- CLIP často správne identifikuje tematicky súvisiace texty, aj keď nie vždy presne.
- Na 100 stránkach model detegoval 109 popisov, anotácie mali 49. Prekrývanie s anotáciami bolo 21 % detekcií a 17 % stránok.
- Problémy s historickými alebo málo známymi objektmi kvôli slabšej reprezentácii v CLIP
- Manuálna kontrola potvrdila, že systém dobre zachytáva vzťah medzi obrazom a textom napriek obmedzeniam a nekonzistentným anotáciám.