Integrácia systému na distribúciu datasetov do HCPortal

Autor: Michal Vrbovsky
Obdobie: 24.10 - 8.11

Obsah

1. Charakteristika VEGA Datasetov

Pre kvalitnejší návrh systému a jeho use case som analyzoval VEGA datasety. Aktuálne máme prístup k datasetom s glyfmi a číslicami.
Oba datasety majú pôvodnú veľkosť obrázkov. Tieto využívajú COCO formát.
Oba datasety sú taktiež rozdelené na menšie bloky 640x640. Tieto využívajú formát YOLOv8.
Datasety s blokmi majú mnohé augmentácie.

1.1 Číslice

Original size:

Format: COCO Pocet obrazkov: 42
Kategorie:

```
_background_, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9
```

Kategória `_background_` nebola využitá ani raz naprieč celým datasetom.
COCO formát umožňuje aj tzv. superkategóriu, ktorá ale nebola definovaná (nie je povinná).

Bloky:

Formát: YOLOv8
Počet obrázkov: 3100 bez augmentácie
S augmentáciou sú tam mierne odchýlky pri počte obrázkov
Obsahuje originálne bloky a mnoho augmentácií
Kategórie:

```
names: ['0', '1', '2', '3', '4', '5', '6', '7', '8', '9']
```

Tieto datasety neobsahujú kategóriu `_background_`

1.2 Glyfy

Original size:

Formát: COCO
Počet obrázkov: 511
Kategórie: Názvy kategórií sú Unicode kódy, ktoré reprezentujú špecifické znaky (glyfy) čo najpresnejšie zodpovedajúce ich vizuálnej podobe. Tieto kódy sú doplnené o ďalšie symboly, ktoré presnejšie odzrkadľujú vzhľad a charakteristiku jednotlivých glyfov.
(Názvy kategórií ukážeme v Blokoch, keďže v YOLO formáte sú identické a zaberajú menej miesta)

Bloky

Formát: YOLOv8
Počet obrázkov: 10845 bez augmentácie
S augmentáciou sú tam mierne odchýlky pri počte obrázkov
Obsahuje originálne bloky a mnoho augmentácií
Kategórie:

```
nc: 162
names: ['N', 'Q', 'R', 'W', 'Z - -', 'Z - I', 'Z', 'u0023', 'u002a', 'u003d',
'u0043', 'u0054', 'u006d', 'u00d8', 'u0186', 'u01c2 - -', 'u01c2', 'u01c3',
'u0223', 'u0236 - - -', 'u0236 - -', 'u0236', 'u0255', 'u0266', 'u0271 - I',
'u0271', 'u0273', 'u0293', 'u0294', 'u0295', 'u0391', 'u039b', 'u03a5', 'u03a9',
'u03b1', 'u03b2', 'u03b8', 'u03ba', 'u03bb', 'u03c0 D _', 'u03c0', 'u03c6', 'u03c7
- -', 'u03c7', 'u03c9', 'u03f4', 'u03fe', 'u03ff', 'u04ba', 'u0564', 'u06ba - -',
'u06ba', 'u07c2', 'u0ba3', 'u0ce7 - -', 'u0ce7', 'u0e87 - - - -', 'u0e87 - - -',
'u0e87 - -', 'u0e87', 'u0ebd - -', 'u0ebd', 'u10c5', 'u10dd', 'u13ce', 'u1433',
'u146b', 'u1472 U -', 'u1472', 'u1542', 'u1543', 'u1546', 'u166d', 'u1687',
'u1691', 'u1692', 'u1722', 'u1723', 'u1d13', 'u1d133', 'u1d15f R -', 'u1d48',
'u1d78f', 'u1dc9', 'u1f74b', 'u1f75e', 'u1f761', 'u1f76a', 'u2020', 'u2026',
'u20df D I', 'u20df U I', 'u20df', 'u2113', 'u2118', 'u2125', 'u2164', 'u2200',
'u221e - -', 'u221e', 'u2290', 'u2293', 'u2295', 'u22a1', 'u22a5', 'u22b8 L I',
'u231c', 'u231d', 'u2571 R - - -', 'u2571 R - -', 'u2571 R -', 'u25a1', 'u25b3 -
_', 'u25b3', 'u25eb', 'u25ec', 'u25ef', 'u2609 D I I I', 'u2609 L - R -', 'u2609 U
I I I', 'u2609 U I _ I', 'u2609', 'u260a', 'u263E', 'u263f', 'u2640', 'u2641',
'u2642', 'u2644', 'u2648', 'u264b', 'u264d', 'u2650', 'u2651', 'u2652', 'u26bb',
'u26db', 'u2723', 'u2733', 'u27c6', 'u27d2', 'u27dc R I', 'u27dc', 'u2909',
```

```
'u29b5', 'u29b6', 'u29df - I', 'u29df', 'u2a4b', 'u2a68', 'u2ad8', 'u2aef - I',
'u2aef', 'u2af0', 'u2b35', 'u2caf', 'u2cc0 - -', 'u2cc0', 'u3059', 'ua609',
'ua72b', 'ufeea']
```

2. Research hostingu

Na účely nasadenia systému sme skúmali rôzne možnosti hostingu. Zvažovali sme riešenia na vlastnom serveri, ktorý by bol prevádzkovaný lokálne, ale aj využitie virtuálneho privátneho servera (VPS), ktorý by nám poskytol potrebnú flexibilitu a výpočtový výkon za rozumnú cenu.
Vlastný hosting by nám síce umožnil úplnú kontrolu nad systémom a hardvérom, ale kritická je práve bezpečnosť, kde by bolo nutné zabezpečiť ochranné mechanizmy ako firewall, ochrana proti DDoS útokom, prípadne šifrovanie dát možno?

Výber domény:

Pre jednoduché prepojenie s HCPortalom sme zvažovali aj kúpu domény, ktorá by používateľom zjednodušila prístup k aplikácii. Pre túto úlohu sme preskúmali možnosti u poskytovateľov domén, napríklad:

Godaddy: https://www.godaddy.com/en-uk/domainsearch/find?domainToCheck=historic-datasets.com
Kde kúpa domény stojí do 1 eur na mesiac.

Možnosti VPS hostingu:

Pre VPS hosting sme skúmali rôznych poskytovateľov, ktorí ponúkajú flexibilné možnosti a dostatočné zdroje na ukladanie a distribúciu veľkých datasetov. Zvažované možnosti zahŕňajú:

Netcup: poskytuje škálovateľné riešenia VPS za priaznivé ceny. Možnosť navýšiť úložisko za poplatok
https://www.netcup.com/en/server/vps
Contabo: ďalší poskytovateľ, ktorý ponúka veľké úložiská za nízke ceny
https://contabo.com/en/storage-vps/

3. Analýza skriptov na konverziu anotačných formátov

Aby sme si mohli vybrať anotačné formáty, ktoré chceme v systéme podporovať, musíme si nájsť a otestovať mnohé konvertory. A nájsť taký, ktorý najviac vyhovuje nášmu use case.

Príprava

Na prípravu testovania sme si vytvorili jednoduché datasety (3 fotky) s využitím len:

- bbox
- segmentácia
- bbox + segmentácia

Každý dataset sme stiahli vo formátoch: COCO, VOC, YOLOv5, YOLOv5obb, YOLOv5PyTorch, YOLOv7PyTorch, YOLOv8, YOLOv8obb, YOLOv9, YOLOv11

S touto prípravou sme išli otestovať open-source konvertory, ktoré sme našli na Githube.

3.1 Ultralytics

Ultralytics má vlastný konvertor, ktorý ale konvertuje len JSON anotácie do formátu YOLO.

- Podporuje len konverziu JSON formátov do YOLO, jednosmerne
- podporuje: COCO, infolks, vott, athm, labelbox
- Pri COCO je možné nastaviť, či má brať do úvahy bbox alebo segmentáciu. Segmentačné pole nie je povinné v COCO. Ideálny scenár je, že užívateľ si vyberie, či output chce mať v bbox alebo segmentácii.
    - Ak si vyberie segmentáciu a skript nájde anotáciu, ktorá nemá seg, tak by mohol vziať bbox, ktorý prepočíta na segmentáciu.
    - Ak si vyberie bbox, tak automaticky berie len bb (tie sú required v COCO formáte). Žiaľ, takto to v Ultralytics nefunguje a keď si užívateľ vyberie, že chce konvertovať segmentáciu a nastane situácia, že tam nie je, tak program spadne.
- Neexistuje žiadna dokumentácia. Tento skript je využívaný primárne in-house v Ultralytics. Neposkytujú podporu pre tento "produkt".

3.2 Taeyoung96/Yolo-to-COCO-format-converter

Našli sme veľmi málo konvertorov na YOLO -> COCO.
Toto bol jediný konvertor s väčším počtom hviezdičiek na Githube.
Každopádne testovanie skončilo prevažne rýchlo, keď prvá konverzia bola neúspešná a bboxy zmenili pozíciu.
Taktiež podporuje len konverziu bboxov.

3.3 PyLabel

Posledná testovaná, už knižnica PyLabel. Veľmi nádejná, nakoľko podporuje naše zvolené formáty a to COCO, YOLO, PascalVOC a konverzie navzájom medzi nimi.
Nevýhody:

- Treba definovať pred konverziou, či chceme bboxy alebo segmentáciu.
- Segmentácia je podporovaná len pre COCO -> YOLO.

4. Vlastný konvertor

Nakoľko každý konvertor mal nejaké drawbacks a bolo by nutné pospájať viacero konvertorov a asi ich aj poupravovať. Taktiež samotné testovanie zabralo veľa času. Jeden deň som testoval len 3 konvertory a to len pre COCO-YOLO medzi bbox a segmentáciou. Po týchto zlých skúsenostiach sme začali uvažovať nad implementáciou vlastného konvertora. Najviac sa budeme inšpirovať python knižnicou PyLabel, ktorá šikovne rieši problémy medzi konverziami viacerých formátov a to tak, že:

- **Import module**: pripraví anotácie a uloží ich do interného formátu, ktorý je špecificky upravený na naše potreby.
- **Export module:** Pri exporte si užívateľ zvolí formát exportovania. Tým, že všetky anotácie sú nahrané v rovnakom formáte, je tento proces o to jednoduchší.

Návrh interného formátu:

Tento návrh podporuje bboxy a polygóny. Taktiež ukladá veľkosti obrázku, čo je dôležité pri formáte YOLO, kde sa využívajú na normalizáciu súradníc.

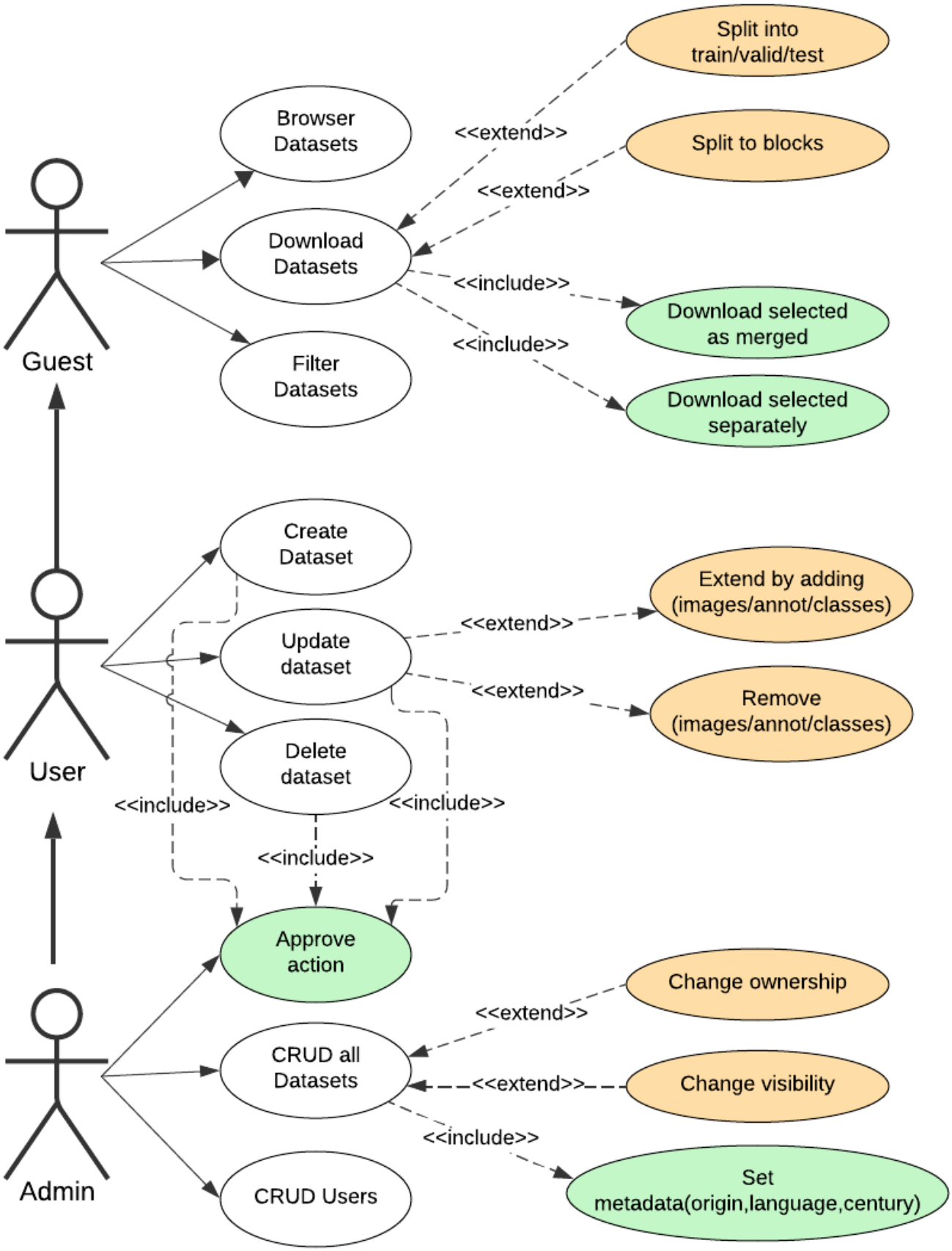
```
columns = [
    'img_folder', 'img_filename', 'img_width', 'img_height',
    'segmented', 'bbox_xmin', 'bbox_ymin', 'bbox_xmax', 'bbox_ymax',
    'segmentation', 'cat_name', 'cat_supercategory',
]
```

Požiadavky na náš konvertor:

- **Import:**
  - Naimplementovať zatiaľ len YOLO, COCO ako najpoužívanejšie formáty.
  - Import prekonvertuje anotácie do interného formátu.
  - Pri importe nie je nutné špecifikovať, či sa bude parsovať bbox alebo segmentácia.
  - Automaticky naplní bbox hodnoty a aj segmentáciu.
  - V prípade, že bbox alebo segmentácia chýba, automaticky sa dopočíta.
- **Export:**
  - Naimplementovať zatiaľ len YOLO ako najpoužívanejší formát.

5. Rework use case

Na podnet z posledného seminára sme prerobili use case diagram, aby lepšie odzrkadľoval naše požiadavky. Z veľkej časti sme zmenili Prihláseného používateľa, kde sme mu odstránili klonovanie a mergovanie datasetov.
Používateľa sme rozšírili o špecifikáciu, ako môže aktualizovať dataset.
Admin teraz musí pridať rôzne meta údaje o datasete, ako napríklad, z ktorého storočia sú texty, aký jazyk sa tam využíva atď. A taktiež možnosť potvrdiť akciu od používateľa, ako je nahranie datasetu a jeho zmeny.



6. Rework databázového návrhu

Zmenou use case diagramu a návrhom na vlastný konvertor formátov prišla aj zmena databázového návrhu.
Tento formát oproti minulému návrhu:

- Ukladá všetky anotácie v databáze vo vlastnom "formáte".
- Umožňuje datasetu priradiť rôzne metadata ako krajina pôvodu, jazyk a iné.
- Ukladá názvy tried, čo nám umožňuje jednoduché dotazovanie a filtrovanie.