Technical documentation

Contents

1 Lumen Geoguesser

1.1 Notices:

We assume you are located at the `.lumen-geoguesser` directory when running the scripts. All global variables are defined in `src/config.py` and `src/paths.py`

1.2 Setup

Create and populate the virtual environment. Simply put, the virtual environment allows you to install Python packages only for this project (which you can easily delete later). This way, we won't clutter your global Python packages.

```
[ ! -d "venv" ] && (echo "Creating python3 virtual environment"; python3 -m venv venv)

pip install -r requirements.txt
```

1.3 Dataset setup

The original dataset strucutre has a directory `data` with images and `data.csv` at the top level:

```
├── data
│   ├── 6bde8efe-a565-4f05-8c60-ae2ffb32ee9b
│   │   ├── 0.jpg
│   │   ├── 180.jpg
│   │   ├── 270.jpg
│   │   └── 90.jpg
│   ├── 6c0ed2ea-b31b-4cfd-9828-4aec22bc0b37
│   │   ├── 0.jpg
│   │   ├── 180.jpg
│   │   ...
│   ...
└── data.csv
```

Before running other scripts you have to properly setup new dataset structure using the `src/preprocess_setup_datasets.py` file. It's important to note that this file accepts multiple dataset directories as an argument and it will make sure to merge the datasets correctly. No changes will be done to your original directories.

```
usage: preprocess_csv_create_rich_static.py [-h] [--csv CSV] [--out dir] [--spacing SPACING] [--out-dir-fig dir] [--fig-format {eps,jpg,jpeg,pdf,pgf,png,ps,raw,rgba,svg,svgz,tif,tiff}] [--no-out]

optional arguments:
  -h, --help            show this help message and exit
```

```
  --csv CSV             Dataframe you want to enrich (default: None)
  --out dir             Directory where the enriched dataframe will be saved (default: None)
  --spacing SPACING     Spacing that will be used to create a grid of polygons.
                        Different spacings produce different number of classes
                        0.7 spacing => ~31 classes
                        0.5 spacing => ~55 classes (default: 0.7)
    --out-dir-fig  dir         Directory  where  the  figure  will  be  saved  (default:  /home/matej/projects/lumen-
geoguesser/figures)
  --fig-format {eps,jpg,jpeg,pdf,pgf,png,ps,raw,rgba,svg,svgz,tif,tiff}
                        Supported file formats for matplotlib savefig (default: png)
  --no-out           Disable any dataframe or figure saving. Useful when calling inside other scripts (default: False)
```

Example:

```
python3 src/preprocess_setup_datasets.py --dataset-dirs data/original_subset data/external_subset --out-dir data/complete_
```

To run scripts later, you must transform this structure to the following structure:

```
├── images
│   ├── test
│   │   ├── e75a769c-4193-491f-9062-c074d8cb80ab
│   │   │   ├── 0.jpg
│   │   │   ├── 180.jpg
│   │   │   ├── 270.jpg
│   │   │   └── 90.jpg
│   │   ├── e75b992b-5606-498b-ade5-0ac7f72b492e
│   │   │   ├── 0.jpg
│   │   │   ├── 180.jpg
│   │   │   ...
│   ├── val
│   │   ...
│   ├── test
│   │   ...
├── data_rich_static__spacing_0.5_classes_55.csv
└── data.csv
```

1. The dataset is split into train, val and test directories
2. `data.csv` is csv has concaternated rows of all `data.csv`s
3. *Rich static CSV* contains region information, which locations (images) are valid etc, centroids...

1.3.1 I have the directory `images` that looks like this: Creating enriched dataframe with centroids and regions:

1.4 Evaluate:

```
curl -X POST lumen.photomath.net/evaluate \
-F 'file=@mapped_to_country_pred-Mike_41-2022-05-06-10-01-15.csv' \
-F "team_code=<INSERT CODE HERE>"
```

Stats: 33.37094934360599 - mapped_to_country_pred-Mike_41-2022-05-06-10-01-15.csv

1.5 Directory structure

Directory	Description
data	dataset
models	saved and trained models
references	research papers and guidelines
reports	model stat's, figures
src	python source code

1.5.1 Developer notes:

To create `requirements.txt` use the following steps:

```
pip install pipreqs
cp requirements.txt requirements.txt.backup
pipreqs --force .
```

```
run python3 src/train.py --accelerator gpu --devices 1 --num-workers 32 --batch-size 8 --dataset-dir data/raw/ data/external/ --cached-df  data/complete/data_huge_spacing_0.21_num_class_211.csv  --image-size  224  --lr  0.00002  --unfreeze-at-epoch 1 --scheduler plateau --val_check_interval 0.25 --limit_val_batches 0.4
```

Merging PDFs:

```
pdfunite in-1.pdf in-2.pdf in-n.pdf out.pdf
```