Przygotowanie danych do dalszych badań w projekcie

mgr inż. Grzegorz Kossakowski

16.08.2024

1. Wprowadzenie

Celem tego notebook jest pobranie danych i przygotowanie ich do dalszej pracy. Dane zostaną podzielone na trzy podstawowe części:

- Dane uczące
- Dane walidacyjne
- Dane testowe

Każda z tych części zostanie zapisana w oddzielnym pliku w formacie fits, wraz z klasyfikacją. Ułatwi to późniejsze badania i pozwoli mieć pewność, że kolejne Notebook będą operowały na dokładnie tych samych zbiorach, co pozwoli porównać otrzymane wyniki.

Źródłem danych dla tego projektu jest AstroNN DECals. Są to obrazy o znacznie lepszej rozdzielczości i jakości, w porównaniu do wersji pierwszej projektu [1]. Dzięki temu będę w stanie przeprowadzić analizę na bardziej rozbudowanych modelach CNN. Jednak samo skorzystanie z większych obrazów powoduje zmianę podejścia do projektu. W pierwszym projekcie mogłem w jednym notebook wykonać wszytkie niezbędne operacje teraz ze względu na dużo większe zdjęcia (256 * 256 * 3), każdy notebook został podzielony na mniejsze części, aby mój system miał szanse je przetworzyć w rozsądnym czasie.

W projekcie wykorzystywane są obrazy z projektu AstroNN DECals [2]. Są tam zgromadzone zdjęcia galaktyk w rozmiarze 256 * 256 * 3 w ilości 17 736 kolorowych zdjęć. Zdjęcia pochodzą z Data Release 10 (DR10) to dziesiąta publiczna wersja danych z badań Legacy Surveys. Do każdego zdjęcia przypisano klasyfikację pochodzącą z projektu Galaxy Zoo [3]. W tym projekcie naukowcy i amatorzy klasyfikują galaktyki według 10 klas:

1. Galaktyki zaburzone 1 081 zdjęć (Disturbed Galaxies)
2. Łączące się galaktyki 1 853 zdjęć (Merging Galaxies)
3. Galaktyki okrągłe gładkie 2 645 zdjęć (Round Smooth Galaxies)
4. Galaktyki okrągłe gładkie pośrednie 2 027 zdjęć (In-between Round Smooth Galaxies)
5. Galaktyki gładkie w kształcie cygara 334 zdjęć (Cigar Shaped Smooth Galaxies)
6. Galaktyki spiralne z poprzeczką 2 043 zdjęć (Barred Spiral Galaxies)
7. Galaktyki spiralne bez poprzeczki 1 829 zdjęć (Unbarred Tight Spiral Galaxies)
8. Galaktyki spiralne bez poprzeczki 2 628 zdjęć (Unbarred Loose Spiral Galaxies)

9. Galaktyki krawędziowe bez wybrzuszenia 1 423 zdjęć (Edge-on Galaxies without Bulge)
10. Galaktyki krawędziowe z wybrzuszeniem 1 873 zdjęć (Edge-on Galaxies with Bulge)

2. Pobranie potrzebnych bibliotek

Pierwszym krokiem jest dodatnie wszystkich potrzebnych bibliotek, aby Notebook mógł zadziałać prawidłowo.

[2]:
```
TF_ENABLE_ONEDNN_OPTS=0
import os
import numpy as np
from astropy.io import fits
from astroNN.datasets import galaxy10
from sklearn.model_selection import train_test_split
```

3. Pobranie danych

W tym miejscu pobieramy dane, na których będziemy pracować w dalszej części artykułu.

[3]:
```
images, labels = galaxy10.load_data()
```

4. Tworzenie katalogu

Jednak aby zachować porządek, Wygenerowane dane w tym notebook będziemy przechowywać w oddzielnym katalogu. Jednak tego katalogu nie ma w repozytorium, dlatego po pobraniu repozytorium należy uruchomić notebook, aby odpowiednie pliki się wygenerowały. Jest to spowodowane wielkością utworzonych plików.

[4]:
```
path = 'Data'
isExist = os.path.exists(path)
if not isExist:
  os.makedirs(path)
```

5. Podział danych na trzy pod zbiory

Celem tego kroku jest podział całego pobranego zbioru na trzy mniejsze zbiory. Pierwszy z tych zbirów to będą dane uczące. Jest ich najwięcej i za ich pomocą każdy model będzie uczony. Drugim zbiorem będzie zbiór walidacyjny. Po każdym wykonanym kroku następuje proces walidacji. Pozwala to ocenić postępy nauki już podczas uczenia. Trzeci zbiór to zbiór testowy. Na którego podstawie będziemy testować modele.

[5]:
```
x_train, x_test, y_train, y_test = train_test_split(images, labels, test_size=0.
↪2)
x_train, x_valid, y_train, y_valid = train_test_split(x_train, y_train,␣
↪test_size=0.2)
```

Jeszcze sprawdzam rozmiar paczek

[6]:
```
x_train.shape, x_valid.shape, x_test.shape
```

[6]:
```
((11350, 256, 256, 3), (2838, 256, 256, 3), (3548, 256, 256, 3))
```

6. Generowanie plików fits

Został ostatni krok w tym notebook. Czyli wygenerowanie potrzebnych plików fits.

Opis plików:

train.fits

Dane w tym pliku posłużą nam do uczenia kolejnych modeli.

[7]:
```
hdu_train_image = fits.PrimaryHDU(x_train)
hdu_train_label = fits.ImageHDU(y_train)
hdu_train = fits.HDUList([hdu_train_image, hdu_train_label])
hdu_train.writeto('Data/train.fits',overwrite=True)
```

valid.fits

Dane w tym pliku posłużą do walidacji podczas uczenia.

[8]:
```
hdu_valid_image = fits.PrimaryHDU(x_valid)
hdu_valid_label = fits.ImageHDU(y_valid)
hdu_valid = fits.HDUList([hdu_valid_image, hdu_valid_label])
hdu_valid.writeto('Data/valid.fits',overwrite=True)
```

test.fits

Dane posłużą do ostatecznego testowania poprawności działania otrzymanych modeli.

[9]:
```
hdu_test_image = fits.PrimaryHDU(x_test)
hdu_test_label = fits.ImageHDU(y_test)
hdu_test = fits.HDUList([hdu_test_image, hdu_test_label])
hdu_test.writeto('Data/test.fits',overwrite=True)
```

Literatura

1. https://www.linkedin.com/pulse/por%C3%B3wnanie-klasyfikacji-obraz%C3%B3w-galaktyk-z-r%C3%B3%C5%BCnych-cnn-kossakowski-adctf/?trackingId=X%2BrVrE25A4JwO8RB4XP3Tg%3D%3D
2. https://astronn.readthedocs.io/en/latest/galaxy10.html
3. https://docs.astropy.org/en/stable/io/fits/