Strojové učenie a neurónové siete

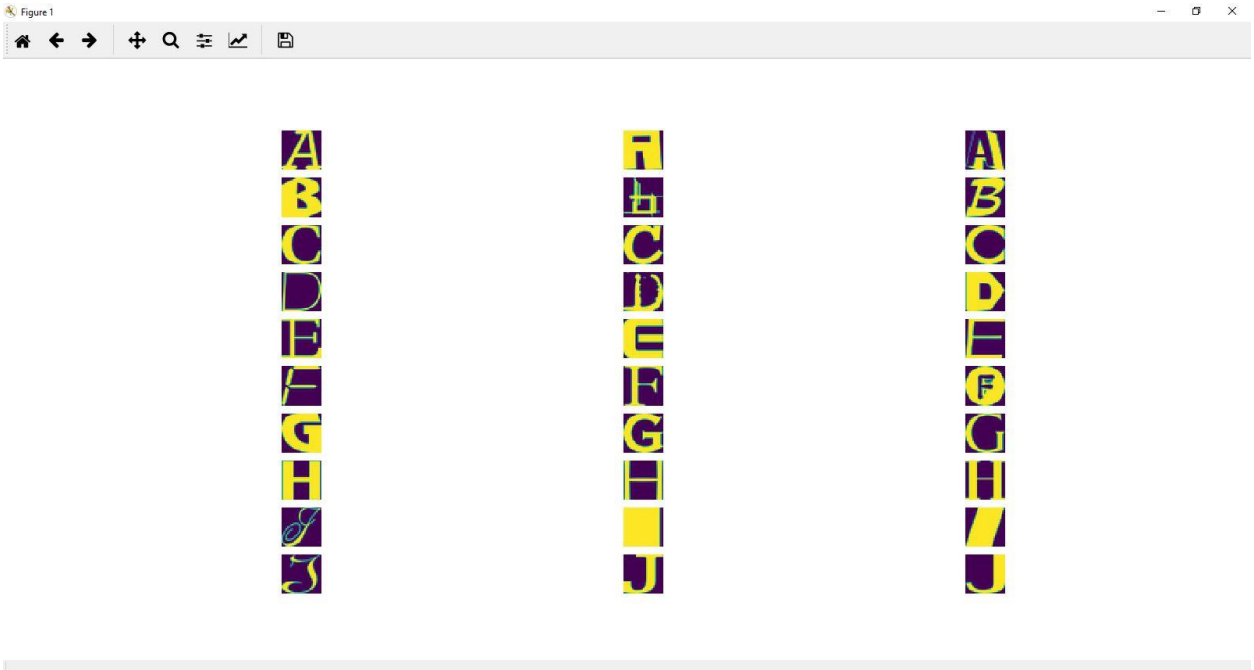
Zadanie číslo 1

Andrej Dubeň

74657

Cieľom prvého zadania je naučiť sa ako pracovať s dátami tak, aby bolo možné jednoducho pracovať s veľkým množstvom dát. Pracujeme s databázou notMNIST, ktorá obsahuje znaky od A-J veľkosti 28*28.

Prvým krokom bolo nahliadnutie do dát či sú zmysľuplné a vykresliť po 3 znakoch z každého. Tento krok je implementovaný v classe LoadFiles.py v ktorej pomocou metódy load_folders() načitame priečinky A-J. Táto metóda v sebe volá metódu load_pictures(), ktorá načíta 3 obrázky z priečinku zadaného vo vstupných parametroch a zapíše ich do „matice“ na vykreslenie.



Druhým krokom bolo pripraviť dáta na manipuláciu a to tak, že pre každý znak sme vytvorili vlastný súbor (.pickle) z normalizovaných dát, nekorektné obrázky boli vyradené.

To je realizované v classe DataCompression.py pomocou metódy ToPickle(), ktorá skontroluje či už neexistujú dané .pickle súbory, ak nie na každý priečinok sa zavolá metóda letterToDataset(), ktorá normalizuje jednotlivé obrázky(28*28, typ float32) a vráti 3D polia(index, x,y), ktoré sú uložené do súborov A-J.pickle. Na konzolu je vypísaný postup transformácie dát a to tak, že najprv sú vypísané nekorektné dáta, ktoré nenačítame a po vytvorení súborov pickle je rovno aj vypísaný počet normalizovaných obrázkov v jednotlivých súboroch.

```
C:\Users\andrej.duben\Programs\Programs\Anaconda\python.exe C:/Users/andrej.duben/PycharmProjects/SUNS1/__init__.py
('Could not read --> skipping', IOError("cannot identify image file 'C:/Users/andrej.duben/PycharmProjects/SUNS1/notMNIST_small/A\\\\RGVtb2NyYXRpY2FCb2xkT2xkc3R5bGUgQm9sZC50dGY=.png'",))
('Could not read --> skipping', IOError("cannot identify image file 'C:/Users/andrej.duben/PycharmProjects/SUNS1/notMNIST_small/F\\\\Q3Jvc3NvdmVyIEJvbGRPYmxpcXVlLnR0Zg==.png'",))
Created: C:/Users/andrej.duben/PycharmProjects/SUNS1/notMNIST_small/A.pickle with image count of 1873
Created: C:/Users/andrej.duben/PycharmProjects/SUNS1/notMNIST_small/B.pickle with image count of 1873
Created: C:/Users/andrej.duben/PycharmProjects/SUNS1/notMNIST_small/C.pickle with image count of 1873
Created: C:/Users/andrej.duben/PycharmProjects/SUNS1/notMNIST_small/D.pickle with image count of 1873
Created: C:/Users/andrej.duben/PycharmProjects/SUNS1/notMNIST_small/E.pickle with image count of 1873
Created: C:/Users/andrej.duben/PycharmProjects/SUNS1/notMNIST_small/F.pickle with image count of 1873
Created: C:/Users/andrej.duben/PycharmProjects/SUNS1/notMNIST_small/G.pickle with image count of 1872
Created: C:/Users/andrej.duben/PycharmProjects/SUNS1/notMNIST_small/H.pickle with image count of 1872
Created: C:/Users/andrej.duben/PycharmProjects/SUNS1/notMNIST_small/I.pickle with image count of 1872
Created: C:/Users/andrej.duben/PycharmProjects/SUNS1/notMNIST_small/J.pickle with image count of 1872
```

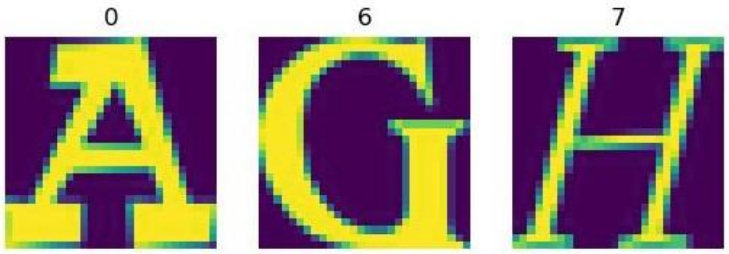
Tretím krokom je verifikovať či dáta sú korektné aj po transformácií, to je realizované v classe VerifyQuality.py, ktorá rozbalí pickle súbory a z každého zobrazí jeden obrázok.

Ďalsím krokom je rozdeliť dáta na trénovacie, testovacie a validačné a spojiť ich do jedného súbora. Nato slúži class DataDivision.py, ktorá v metóde SplitToSets() rozdelí jednotlivé sety písmen na trénovaciu, validačnú a trénovaciu množinu podla pomerov určených v class a priradí im kódovanie od 0-9. Pomery sú nastavené na 15% validačné, 15% testovacie a 70% trénovacie.

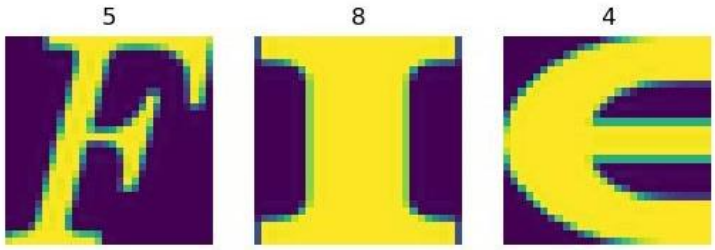
```
7
8   validation_ratio = 0.15
9   testing_ratio = 0.15
10  imageSize = 28
```

Tieto sety sú spravené tak, že sa vnich neprekrývajú žiadne dáta a sú spojené do jedného súbora .pickle v metóde toOneFile(). Dáta je potrebné overiť po uložení do jedného súbora, táto funkcionalita je v classe FinalShowImg(), ktorá rozbalí finálny súbor a obsahuje metódy na zobrazenie 3 obrázkov z každej kategórie setov.

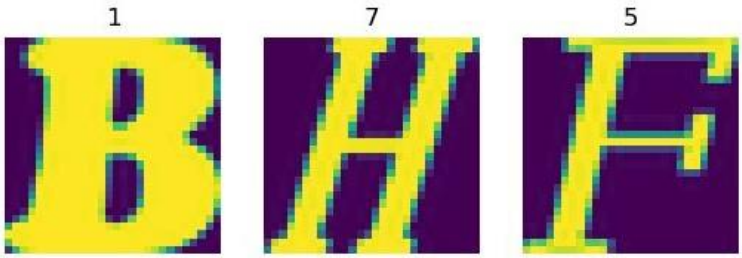
Training data



Testing data



Validation data



```
C:\Users\andrej.duben\Programs\Programs\Anaconda\python.exe C:/Users/andrej.duben/PycharmProjects/SUNS1/__init__.py
('Could not read --> skipping', IOError("cannot identify image file 'C:/Users/andrej.duben/PycharmProjects/SUNS1/notMNIST_small/A\\
('Could not read --> skipping', IOError("cannot identify image file 'C:/Users/andrej.duben/PycharmProjects/SUNS1/notMNIST_small/F\\
Created: C:/Users/andrej.duben/PycharmProjects/SUNS1/notMNIST_small/A.pickle with image count of 1873
Created: C:/Users/andrej.duben/PycharmProjects/SUNS1/notMNIST_small/B.pickle with image count of 1873
Created: C:/Users/andrej.duben/PycharmProjects/SUNS1/notMNIST_small/C.pickle with image count of 1873
Created: C:/Users/andrej.duben/PycharmProjects/SUNS1/notMNIST_small/D.pickle with image count of 1873
Created: C:/Users/andrej.duben/PycharmProjects/SUNS1/notMNIST_small/E.pickle with image count of 1873
Created: C:/Users/andrej.duben/PycharmProjects/SUNS1/notMNIST_small/F.pickle with image count of 1873
Created: C:/Users/andrej.duben/PycharmProjects/SUNS1/notMNIST_small/G.pickle with image count of 1872
Created: C:/Users/andrej.duben/PycharmProjects/SUNS1/notMNIST_small/H.pickle with image count of 1872
Created: C:/Users/andrej.duben/PycharmProjects/SUNS1/notMNIST_small/I.pickle with image count of 1872
Created: C:/Users/andrej.duben/PycharmProjects/SUNS1/notMNIST_small/J.pickle with image count of 1872
Total images in training dataset:  13126   13126
Total images in testing dataset  2800
Total images in validation dataset  2800
Total number of images in final pickle file   18726

Process finished with exit code 0
```