2. Druga laboratorijska vježba iz kolegija Duboko učenje

Cilj i opis laboratorijske vježbe

Cilj laboratorijske vježbe je upoznavanje s osnovnim konceptima proširivanja slikovnog seta podataka za treniranje neuronske mreže koja vrši klasifikaciju i obradom i vizualizacijom podataka numeričkih setova podataka.

Upute za rješavanje prvog djela laboratorijske vježbe (slikovni set podataka)

Potrebno je učitati podatke nasumično odabranog slikovnog seta podataka nekih slika koje volite. To može biti drveće, biljke, gljive, auti, umjetničke slike i slično. Velike količine slikovnih setova podataka dostupne su na stranicama https://www.kaggle.com/ i https://paperswithcode.com/datasets?mod=images&task=image-classification. Naravno, možete pronaći i na nekim drugim stranicama set podataka i to nije nikakav problem. Bitno je za napomenuti da set podataka treba biti namijenjen za klasifikaciju slika jer ćemo na ovoj pripremi podataka u jednoj od sljedećih vježbi raditi klasifikaciju slika, pa ćete kroz ovu vježbu odmah imati spreman set podataka.

Primjer nekih slikovnih setova podataka (**ne smiju se koristiti za izradu vježbe, već je potrebno pronaći vlastiti set podataka**):

- https://www.kaggle.com/gverzea/edible-wild-plants
- https://www.kaggle.com/maysee/mushrooms-classification-common-genuss-images
- https://paperswithcode.com/dataset/imagenet-sketch
- https://www.kaggle.com/jutrera/stanford-car-dataset-by-classes-folder

Nakon odabira seta podataka isti je potrebno podesiti na veličinu treniranja (npr. 224 x 224 za ResNet arhitekturu). Set podataka je potom potrebno uvećati kao što je to prikazano na predavanjima. Minimalno je potrebno napraviti promjenu boja, kuta slike i nasumično izrezivanje slike i jednu promjenu na slici koja se NE nalazi u primjeru s predavanja. Kako bi set podataka bio što brže obrađen, potrebno je uzeti 20% seta podataka i nad njime napraviti uvećanje. Naravno, kada će doći do treniranja neuronske mreže, povećanje će trebati napraviti nad cijelim trening setom podataka. **Programski kod je potrebno pohraniti za danje korištenje jer će se koristiti u budućim laboratorijskim vježbama.**

Upute za rješavanje drugog djela laboratorijske vježbe (CSV set podataka)

Slično kao i u prvom djelu vježbe, potrebno je učitati podatke nasumično odabranog CSV seta podataka dok god nije onaj iz predavanja. Velike količine CSV setova podataka moguće je pronaći na:

- https://www.kaggle.com/datasets?fileType=csv&datasetsOnly=true
- https://paperswithcode.com

Podatke je potrebno obraditi na način da se nepostojeće brojčane vrijednosti zamjene srednjim ili nekom adekvatnom (potrebno je samostalno razmisliti o tome). Nakon uklanjanja nepostojećih vrijednosti potrebno je ukloniti moguće duplikate po željenim kolonama. Duplikate je moguće ukloniti pomoću grupiranja (npr. da se preostale vrijednosti uprosječi) ili pomoću „drop_duplicates“ metode. S obzirom na odabir, korištenu metodu je potrebno obrazložiti. Sve podatke (ukoliko takvi postoje) potrebno je prebaciti u metrički sustav, a ukoliko postoje cijene koje nisu u Eurima, potrebno ih je prebaciti u Eure.

Osim obrade podataka, iz podataka je potrebno izraditi barem tri grafikona koristeći matplotlib biblioteku. Od grafikona je potrebno izraditi barem jedan boxplot, jedan histogram i treći po izboru koji se ne nalazi unutar predavanja. Slobodno koristiti https://matplotlib.org/cheatsheets/.