Klasifikacija morfologije galaksija

Dušan Komadinović, SV65/2022

Definicija problema

Problem koji se razmatra jeste klasifikacija galaksija na osnovu njihovih vizuelnih karakteristika pomoću mašinskog učenja upotrebom neuronskih mreža. Galaksije mogu imati različite oblike (osnovna klasifikacija na spiralne, eliptične, i nepravilne, tokom godina znatno proširena), koji odražavaju njihovu starost, sastav i pružaju uvid u istoriju formiranja i evoluciju kosmosa. Razumevanje morfologije galaksija je ključno za astrofiziku.

Motivacija za problem

Ranije se klasifikacija vršila ručno, što je izuzetno vremenski zahtevno i podložno ljudskim greškama, a sada i gotovo nemoguće zbog ogromnog broja podataka koji prikupljaju moderni teleskopi za znatno kraće vreme. Samim tim korišćenje mašinskog učenja omogućava bržu, tačniju i skalabilnu analizu tih podataka. Praktična primena ovakvog rešenja je u podršci astronomima i istraživačima, kao i u optimizaciji analize velikih astronomskih kataloga.

Skup podataka

Za realizaciju projekta koristiću Galaxy10 DECals skup podataka. Sadrži 17736 slika galaksija u boji dimenzija 256x256 piskela koje su podeljene na 10 klasa.

Skup podataka je sačuvan u vidu binarnog HDF5 (.h5) fajla, koji se koristi za strukturirano čuvanje velikih skupova podataka i omogućava efikasno učitavanje istih.

Klasa	Broj slika
Cigar Shaped Smooth Galaxies	334
Disturbed Galaxies	1081
Edge-on Galaxies without Bulge	1423
Unbarred Tight Spiral Galaxies	1829
Merging Galaxies	1853
Edge-on Galaxies with Bulge	1873
In-between Round Smooth Galaxies	2027
Barred Spiral Galaxies	2043
Unbarred Loose Spiral Galaxies	2628
Round Smooth Galaxies	2645

Statistički podaci	
Prosek	1774
Medijana	1863
Minimum	334
Maksimum	2645

Način pretprocesiranja podataka

- Slike su u formatu 256×256px, što je dobar format za CNN. ResNet, međutim, zahteva format slike 224x224px, tako da ću ih smanjiti na te dimenzije da bi treniranje bilo ravnopravno između ta dva modela
- Normalizacija piksela: vrednosti piksela se prebacuju u raspon [0,1] za CNN, a za ResNet [-1, 1]
- Augmentacija trening slika: tehnike rotiranja, premeštanja i horizontalnog okretanja (kao u ogledalu) radi smanjenja prenaučenosti (overfitting)
- Pošto skup nije uravnotežen, pored augmentacije dodeliću težine klasama kako bi model obratio veću pažnju na slike iz klasa sa manjim brojem slika. To bi u kombinaciji sa augmentacijom trebalo da znatno poboljša model, praviću modele sa i bez ovih tehnika i uporediti ih međusobno

Metodologija

Proces rešavanja problema uključuje sledeće korake:

1. **Učitavanje i podela podataka** na train/validation/test skupove
2. **Pretprocesiranje skupa podataka** (normalizacija, augmentacija, class weighting)
3. **Treniranje modela**:

   Radiću ga na GPU (CUDA), pošto posedujem NVIDIA grafičku karticu, što će značajno ubrzati treniranje i omogućava eksperimentisanje sa većim batch size-om i većim modelima.

   - CNN (Convolutional Neural Network) kao osnovni model
   - ResNet (Residual Neural Network) transfer learning za poređenje sa CNN
   - Poređenje različitih arhitektura (plića i dublja mreža, manji i veći broj epoha) radi procene performansi
   - Batch training: koristiću batch size od 32, što omogućava efikasno korišćenje grafičke kartice i stabilno treniranje modela
   - Early stopping: trening će se prekinuti ako validaciona metrika (loss) ne pokazuje poboljšanje tokom određenog broja epoha, čime se smanjuje rizik od overfittinga

- Model checkpointing: tokom treniranja čuvaju se najbolji parametri modela (prema performansama na validation skupu), što omogućava vraćanje najboljeg modela po završetku treninga
- Reduce Learning Rate on Plateau: stopa učenja će se smanjiti ako se validaciona metrika ne poboljšava

Ulazni podaci: slika galaksije (224×224×3).
Izlazni podaci: klasa kojoj galaksija pripada (jedna od 10 klasa).

4. **Evaluacija modela** korišćenjem test skupa i metrika
5. **Vizuelizacija rezultata** – grafik promena tačnosti i gubitka kroz epohe, "Classification Report" za prikaz metrika po klasama , matrica konfuzije

Način evaluacije

Podela skupa na train/validation/test skupove (70%/15%/15%)

Metrike performansi:

- **Accuracy** – opšta tačnost,
- **F1-score** – za neuravnotežene klase,
- **Confusion matrix** – za prikaz performansi po klasama

Tehnologije

- **Programski jezik: Python**
- **Glavne biblioteke: TensorFlow / Keras** (pravljenje i treniranje CNN i ResNet modela), **h5py** (učitavanje skupa podataka), **NumPy** (pretprocesiranje), **scikit-learn** (evaluacija i metrike), i **Matplotlib** (vizuelizacija rezultata)

Relevantna literatura

- Radovi koji su takođe koristili Galaxy10 skup su na astroGG sajtu
- Guruprasad, A. (2023). Galaxy Classification: A machine learning approach for classifying shapes using numerical data. ArXiv. https://arxiv.org/abs/2312.00184
- AstroDave, AstroTom, Christopher Read @ Winton, joycenv, and Kyle Willett. Galaxy Zoo - The Galaxy Challenge. https://kaggle.com/competitions/galaxy-zoo-the-galaxy-challenge, 2013. Kaggle.

- Willett, K. W., et al. "Galaxy Zoo 2: detailed morphological classifications for 304,122 galaxies from the Sloan Digital Sky Survey." *Monthly Notices of the Royal Astronomical Society* 435.4 (2013): 2835-2860. https://arxiv.org/abs/1308.3496