Proiect – Etapa 1

Deadline: 17.11.2024

Data postării: 30 Octombrie 2024

Contents

1 Introducere

1.1 Scopul proiectului

Scopul principal al proiectului nostru este dezvoltarea unei soluții pentru clasificarea tumorilor cerebrale utilizând imagini de rezonanță magnetică (MRI). Tumorile cerebrale sunt considerate unele dintre cele mai agresive afecțiuni, afectând atât copii, cât și adulți. Acestea reprezintă aproximativ $85 - 90\%$ din toate tumorile primare ale sistemului nervos central (SNC). Anual, în jur de 11.700 de persoane sunt diagnosticate cu o tumoră cerebrală, iar rata de supraviețuire pe cinci ani pentru pacienții cu tumori cerebrale canceroase sau SNC este de aproximativ 34% pentru bărbați și 36% pentru femei. Tumorile cerebrale pot fi clasificate în funcție de natura lor: tumori benigne, tumori maligne, tumori pituitare etc. O diagnosticare corectă, împreună cu un plan de tratament adecvat,

este esențială pentru a îmbunătăți așteptările de viață ale pacienților. Cea mai eficientă metodă de detectare a tumorilor cerebrale este imagistica prin rezonanță magnetică (MRI), care generează o cantitate imensă de date imagistice ce sunt examinate de către radiologi. Cu toate acestea, examinarea manuală poate fi predispusă la erori, din cauza complexității implicate în evaluarea tumorilor cerebrale și a proprietăților acestora.

Imaginile medicale, în special cele provenite din rezonanța magnetică, sunt esențiale în diagnosticarea și tratamentul tumorilor cerebrale, iar utilizarea inteligenței artificiale pentru a clasifica aceste imagini poate îmbunătăți semnificativ acuratețea diagnosticării și eficiența procesului decizional. Analiza datelor ne va permite să identificăm tiparele și variabilitățile care pot influența clasificarea.

Prin utilizarea unui set de date propus pentru clasificarea tumorilor cerebrale, ne propunem să investigăm cele mai bune practici de preprocesare a imaginilor medicale, inclusiv normalizarea, redimensionarea și augmentarea, pentru a ne asigura că modelul nostru este robust și capabil să generalizeze bine pe date noi.

2 Setul de date

Seturile de date joacă un rol crucial în dezvoltarea soluțiilor bazate pe inteligență artificială (AI) și învățare automată (ML) în domeniul medical, având un impact semnificativ asupra diagnosticării, tratamentului și gestionării pacienților. Aceste date permit antrenarea modelelor pentru a recunoaște tipare complexe în imagini, facilitând identificarea timpurie a afecțiunilor. În această etapă, ne vom concentra atenția asupra înțelegerii modului în care putem utiliza setul de date pentru a dezvolta un model robust de clasificare a tumorilor cerebrale. O parte esențială a acestui proces este analiza exploratorie a datelor (EDA), care ne va ajuta să identificăm tiparele, distribuțiile și posibilele anomalii din setul de date. EDA este crucială pentru înțelegerea caracteristicilor datelor, permițându-ne să ne ajustăm strategiile de preprocesare și să determinăm cele mai eficiente tehnici de augmentare a imaginilor. Printr-o examinare detaliată a datelor, putem descoperi relații semnificative și informații care pot influența performanța modelului, ceea ce va conduce la obținerea unor rezultate mai precise și mai de încredere.

2.1 Brain Tumor Classification (MRI)

Setul de date Brain Tumor Classification (MRI) conține imagini de tip MRI provenite de la pacienți și este destinat clasificării tumorilor cerebrale în patru categorii distincte:

1. **Glioma Tumor:** Imagini ale tumorilor gliomatoase, care afectează celulele gliale din creier.
2. **Meningioma Tumor:** Imagini care ilustrează tumorile meningiomatoase, localizate în meningele creierului.
3. **No Tumor:** Imagini de la pacienți care nu au tumori cerebrale, utilizate ca grup de control.

4. **Pituitary Tumor:** Imagini ale tumorilor pituitare, care se dezvoltă în glanda pituitară.

Setul de date include un număr semnificativ de imagini pentru fiecare categorie, facilitând astfel antrenarea și validarea algoritmilor de clasificare bazați pe învățare supervizată. Această diversitate de imagini ajută la îmbunătățirea acurateței modelului și la evaluarea performanței în condiții variate, reflectând astfel complexitatea reală a diagnosticării tumorilor cerebrale. Imaginile sunt etichetate corespunzător, ceea ce permite cercetătorilor și dezvoltatorilor să dezvolte soluții automate eficiente pentru detectarea și clasificarea tumorilor cerebrale.

Setul de date Brain Tumor Classification (MRI) poate fi descărcat de pe platforma Kaggle și vine deja împărțit în subseturi de antrenare și testare. Din setul de date destinat antrenării, vom realiza o împărțire suplimentară, alocând 80% din date pentru antrenare efectivă și 20% pentru validare. Această împărțire este esențială pentru dezvoltarea unui model robust.

- **Setul de antrenare (**80%**)** va fi utilizat pentru a învăța modelul să recunoască diferitele tipuri de tumori pe baza imaginilor MRI. Aceste date vor ajuta la ajustarea ponderilor în timpul procesului de antrenare. Atunci când realizăm împărțirea, vom avea grijă să selectăm 80% din imaginile disponibile în fiecare categorie, asigurându-ne astfel că distribuția datelor rămâne echilibrată și reprezentativă.

- **Setul de validare (**20%**)** va fi folosit pentru a evalua performanța modelului în timpul procesului de antrenare. Acesta permite monitorizarea overfitting-ului, asigurându-se că modelul generalizează bine la datele pe care nu le-a văzut anterior și ajută la selectarea variantei optime de model (vom alege varianta pentru care atingem cele mai bune performanțe pentru setul de validare).

- **Setul de testare**, care nu este utilizat în procesul de antrenare, va fi rezervat pentru evaluarea finală a modelului. Aceasta va oferi o măsură obiectivă a capacității modelului de a clasifica corect imagini noi, asemănătoare celor din practică, și de a identifica tipurile de tumori pornind de la imagini pe care nu le-a văzut anterior în timpul antrenării.

3 Cerințe

3.1 Cerința 1

Pentru a facilita lucrul cu setul de date Brain Tumor Classification (MRI) în cadrul acestui proiect, implementați o clasă personalizată în PyTorch, derivată din `torch.utils.data.Dataset`, utilizând un mod de încărcare lazy (doar în momentul accesării datelor). Aceasta va asigura că datele sunt încărcate doar atunci când este necesar, economisind astfel memorie. Această clasă ar trebui să:

- Primească ca argument o cale de unde va prelua imaginile (`train` / `test`).

- Clasa ar trebui să încarce imaginile pe baza numelor fișierelor și a etichetelor asociate, fără a le păstra în memorie până la momentul accesării, pentru fiecare categorie (`glioma_tumor`, `meningioma_tumor`, `no_tumor`, `pituitary_tumor`) din directorul furnizat ca argument în constructor.
- Conține implementarea metodelor `__len__` și `__getitem__`:
  - `__len__` – trebuie să returneze numărul total de imagini din dataset.
  - `__getitem__` – trebuie să încarce și să returneze imaginea și eticheta corespunzătoare pe baza poziției în lista de fișiere. De asemenea, dacă avem specificată o transformare, o va aplica înainte de a returna imaginea.

Această clasă va permite utilizarea setului de date împreună cu `DataLoader` pentru încărcarea optimizată a datelor în batch-uri, necesară pentru antrenarea și evaluarea modelului.

3.2 Cerința 2

Pentru a permite o evaluare robustă și generalizabilă a modelului, implementați o metodă de împărțire a setului de date de antrenare în două subseturi: 80% pentru antrenare și 20% pentru validare. Această metodă ar trebui să fie flexibilă, pentru a putea fi utilizată ulterior într-o strategie de tip cross-validation.

3.3 Cerința 3

Realizați o vizualizare a distribuției claselor (ex. histograme sau grafice de bare) pentru a determina dacă setul de date este echilibrat sau dezechilibrat între categoriile de tumori. Veți realiza acest lucru pentru fiecare dintre cele 3 seturi de date: antrenare, validare și testare.

Dacă sunt observate clase sub-reprezentate, documentați aceste observații în raportul realizat pentru această etapă. De asemenea, propuneți variante pe baza cărora putem remedia aceste situații, cum ar fi augmentarea datelor sau alte tehnici utilizate pentru echilibrarea claselor.

3.4 Cerința 4

Pentru a înțelege mai bine variabilitatea și tiparele vizuale ale fiecărei categorii de tumori cerebrale din setul de date Brain Tumor Classification (MRI), este necesară o analiză vizuală detaliată a imaginilor.

Extrageți și afișați un set de imagini din fiecare categorie prezentă în dataset (`glioma_tumor`, `meningioma_tumor`, `no_tumor`, `pituitary_tumor`), de preferat 5-10 imagini per categorie. Astfel, putem încerca să observăm în mod direct caracteristicile fiecărui tip de tumoră și variațiile între imagini (dacă ele există).

În urma analizei vizuale, documentați în raportul realizat pentru această etapă variabilitatea internă din cadrul fiecărei clase (de exemplu, tumorile gliomice pot varia ca

formă și textură). Notați și cazurile în care imaginile sunt similare între clase, ceea ce ar putea indica o dificultate mai mare pentru modelul de clasificare.

Examinați imaginile selectate pentru a identifica diferențele vizuale între categorii (unele tumori pot apărea mai luminoase sau mai întunecate în funcție de localizare și natura țesutului afectat).

3.5 Cerința 5

Pentru a asigura consistența și integritatea setului de date, implementați o verificare detaliată pentru fiecare imagine din setul de date, având în vedere următoarele criterii:

- Verificați dacă toate imaginile au același număr de canale (de exemplu, imagini în alb-negru vor avea un singur canal, în timp ce imaginile color vor avea 3 canale RGB). În cazul în care identificați imagini cu număr diferit de canale, găsiți o modalitate prin care să asigurați o uniformizare din acest aspect.
- Asigurați-vă că toate imaginile au dimensiuni uniforme (de exemplu, 256×256 pixeli). Imaginile de dimensiuni diferite pot introduce dificultăți în procesul de antrenare a modelului. Dacă identificați că există imagini de dimensiuni diferite, realizați o statistică și apoi aplicați o modalitate pentru uniformizarea dimensiunilor.
- Verificați dacă valorile pixelilor sunt exprimate pe aceleași scale sau unități (de exemplu, densitatea de protoni în imaginile MRI). Asigurați-vă că imaginile sunt normalizate corespunzător, astfel încât modelele de învățare automată să poată interpreta valorile corect.

În raportul final, includeți concluziile la care ați ajuns după analizarea aspectelor legate de consistența setului de date. Descrieți problemele identificate și detaliați procesările aplicate pentru a rezolva aceste probleme.

3.6 Cerința 6

Pentru a crește calitatea și consistența datelor de intrare, implementați sau selectați o serie de **minimum 5** operații de preprocesare și normalizare care pot fi aplicate imaginilor MRI din acest set de date. După aplicarea fiecărei operații, comparați imaginile rezultate cu cele originale pentru a înțelege impactul vizual al fiecărei transformări.

Alegeți tehnici precum filtrul Gaussian, normalizarea intensității, ajustarea contrastului (ex. CLAHE) și filtrele de detecție a marginilor (ex. Sobel). Pentru fiecare tehnică aleasă precizați în raport care este necesitatea sa pentru acest tip de imagistică și efectul așteptat asupra datelor.

3.7 Cerință 7

Implementați un pipeline de antrenare a unei rețele neurale folosind Pytorch / MONAI pentru clasificarea imaginilor din setul de date Brain Tumor Classification (MRI). Pipeline-ul ar trebui să includă următoarele etape:

1. **Încărcarea datelor:** Utilizați clasa derivată din Dataset pentru a citi și prelucra imaginile din setul de date, asigurându-vă că aplicați operațiile de preprocesare necesare.
2. **Definirea modelului:** Alegeți sau construiți o arhitectură de rețea neurală adecvată pentru sarcina de clasificare, precum rețeaua convoluțională (CNN) prezentată la curs.
3. **Definirea funcției de pierdere și a optimizatorului:** Selectați o funcție de pierdere potrivită și un optimizator pentru antrenarea modelului.
4. **Antrenarea modelului:** Implementați un ciclu de antrenare care să efectueze epoci multiple, să actualizeze ponderile modelului și să evalueze performanța pe setul de validare după fiecare epocă. Salvați modelul care obține cele mai bune performanțe pe setul de validare în timpul antrenării.
5. **Evaluarea modelului:** După antrenare, evaluați modelul pe setul de test pentru a determina: acuratețea, precizie, recall, F1-score. De asemenea, construiți și plotați matricea de confuzie.

Implementați și antrenați un model de bază (baseline) pentru clasificarea tumorilor cerebrale folosind setul de date propus. Un baseline este esențial deoarece ne oferă un punct de plecare pentru evaluarea performanței modelului, ajutându-ne să înțelegem cât de bine funcționează soluția noastră inițială. În etapa următoare, ne vom concentra pe îmbunătățirea acestui model de bază, explorând tehnici avansate de optimizare și ajustare a hiperparametrilor pentru a spori acuratețea și eficiența clasificării.

În raport, veți include detaliile legate de procesul de antrenare (modelul neural folosit, funcția de eroare, numărul de epoci, optimizatorul folosit, rata de învățare și alte aspecte considerate relevante). De asemenea, trebuie să realizați un grafic în care să evidențiați care este evoluția erorii pentru setul de antrenare și pentru setul de validare de la o epocă la alta. Un grafic similar va trebui să realizați și pentru a arăta cum evoluează acuratețea de la o epocă la alta pentru setul de date de antrenare și pentru setul de date de validare.

4 Format soluție

La final, va trebui să încărcați o arhivă în format zip care să conțină următoarele:

- **Raport.pdf** este fișierul în care veți include explicațiile, rezultatele, grafice pentru fiecare cerință în parte.

- Unul sau mai multe fișiere cu codul implementat pentru rezolvarea cerințelor (puteți alege dacă lucrați în fișiere `.py` sau în jupyter notebook).

5 Punctaj

Punctajul pentru cerințele propuse este următorul:

Cerința	Punctaj
Cerința 1	20 puncte
Cerința 2	5 puncte
Cerința 3	5 puncte
Cerința 4	5 puncte
Cerința 5	10 puncte
Cerința 6	15 puncte
Cerința 7	40 puncte

Atenție!

- Proiectul este individual!
- **Deadline-ul pentru această etapă este HARD!**
- Pentru a primi punctajul pentru proiect este obligatorie prezentarea lui în săptămâna dedicată.
- Punctajul pentru cerință 7 este acordat doar dacă acuratețea obținută este minimum 25% (cu alte cuvinte, aveți o soluție care funcționează mai bine decât o strategie random).
- Dacă realizați doar implementare fără a documenta în raport ceea ce vi se cere, nu veți primi punctajul pentru cerință respectivă.
- Proiectele vor fi verificate anti-plagiat, iar cele detectare drept copiate (colegi / internet) nu vor fi punctate.