Triplet Networks

Task posibil: metric learning and similarity comparisons.

Abstract:

Proiectul propus explorează și implementează un proces de data mining într-un context digital în continuă expansiune. Acest demers are la bază dorința de a extrage cunoștințe semnificative din seturi de date, utilizând tehnici avansate de analiză. Procesul implică prelucrarea adecvată a bazei de date, extragerea trăsăturilor relevante și aplicarea Triplet Networks pentru a îmbunătăți învățarea metricilor și similaritatea între date. Obiectivele proiectului vizează implementarea eficientă a procesului, explorarea și evaluarea rezultatelor obținute, evidențiind astfel potențialul impact în diverse domenii ale societății digitale contemporane.

- torch - PyTorch, o bibliotecă de învățare profundă.
- torch.nn - Modulul din PyTorch care conține straturile și operațiile pentru construirea rețelelor neuronale.
- torch.optim - Modulul PyTorch pentru optimizatori, utilizat pentru a ajusta parametrii modelului în timpul antrenării.
- torch.utils.data - Modulul PyTorch pentru lucrul cu seturile de date și DataLoader.
- transforms - Modulele PyTorch pentru transformările imaginilor.
- ImageFolder - Clasa PyTorch pentru manipularea seturilor de date de imagini.
- numpy - Biblioteca pentru manipularea eficientă a matricelor și a tabelelor de date.
- matplotlib.pyplot - Biblioteca pentru vizualizarea datelor, utilizată în proiect pentru a crea grafice și diagrame.
- opendatasets - Biblioteca pentru descărcarea seturilor de date de pe platforme precum Kaggle.

Concepts: • Programming; • Python; • Data mining; • București-2024;

Additional Key Words: datasets, programming, data mining, compile, jupyter notebooks

Reference Format: Andrei Chirica - Master II - IISC. 2024. Temă de semestru: Triplet Networks

(Ianuarie 2024)

Capitolul 1: Introducere

1.1 Contextul proiectului

Într-o lume tot mai digitală, analiza de date a devenit un element cheie în extragerea de informații semnificative din volume masive de date. Domeniul analizei de date își găsește aplicabilitate în diverse sectoare, de la industrie și comerț, în medicină și cercetare științifică. În acest context, proiectul are ca scop explorarea și implementarea unui proces de data mining pe un set de date cu imagini, oferindu-ne oportunitatea de a descoperi modele și relații relevante în datele colectate.

1.2 Motivația proiectului

Motivația din spatele proiectului se bazează pe necesitatea de a extrage cunoștințe utile și semnificative din datele disponibile. Procesul de data mining oferă instrumente puternice pentru identificarea de modele, clasificarea datelor și luarea deciziilor bazate pe evidențe. Implementarea acestui proces este motivată de potențialul său de a aduce beneficii practice în optimizarea proceselor, luarea deciziilor și înțelegerea mai profundă a fenomenelor studiate.

1.3 Obiectivele proiectului

Principalele obiective ale proiectului sunt:

- Implementarea unui proces eficient de data mining.
- Prelucrarea corespunzătoare a setului de date pentru a facilita analiza.
- Extracția și identificarea trăsăturilor relevante din date.
- Utilizarea Triplet Networks în învățarea metricilor pentru a îmbunătăți similaritatea între date.
- Realizarea unei analize riguroase a rezultatelor obținute și evaluarea performanțelor procesului implementat.

Aceste obiective vor servi drept repere în evoluția proiectului și vor ghida implementarea și analiza ulterioară a rezultatelor obținute în cadrul procesului de data mining.

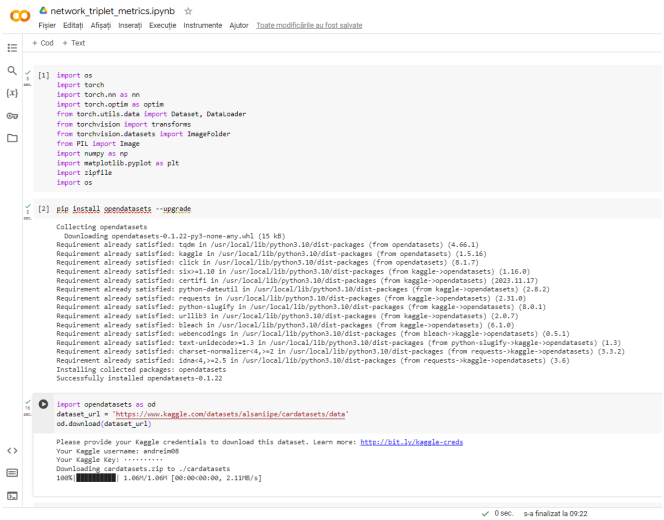

Fig. 1 - Importuri de biblioteci

Capitolul 2: Fundamente teoretice

2.1 Data-mining

Data mining reprezintă procesul de identificare a modelelor și a relațiilor semnificative într-un set mare de date. Prin utilizarea tehnicilor statistice, algoritmilor de învățare automată și a instrumentelor software, data miningul permite extragerea de informații valoroase din datele brute. Importanța acestui proces rezidă în capacitatea sa de a dezvălui tendințe, modele și relații complexe, sprijinind procesul decizional și facilitând luarea de decizii informate.

2.2 Triplet Networks în învățarea metricilor

Triplet Networks este un model de învățare profundă utilizate în special în învățarea metricilor. Acestea sunt concepute pentru a învăța o reprezentare a datelor într-un spațiu în care similaritatea între exemple este reflectată în distanța euclidiană. Prin utilizarea unui set de trei exemple pentru fiecare iterație (ancore, exemple pozitive și exemple negative), Triplet Networks optimizează spațiul de învățare astfel încât să maximizeze similaritatea între exemplele pozitive și să minimizeze similaritatea între exemplele negative.

Distanța euclidiană dintre doi vectori A și B în contextul Triplet Networks este adesea calculată ca norma Euclidiană a diferenței dintre acești doi vectori. Formula este:

$$\text{Distanța Euclidiana} = \| A - B \|^2$$

Aici, $\| \cdot \|^2$ reprezintă norma Euclidiană, adică radicalul sumei pătratelor elementelor vectorului diferenței $A-B$. În mod specific, pentru doi vectori $A=(a1,a2,\ldots,an)$ și $B=(b1,b2,\ldots,bn)$, distanța euclidiană este:

$$\text{Distanța Euclidiana} = \sqrt{(a1-b1)^2+(a2-b2)^2+\ldots+(an-bn)2^2}$$

În Triplet Networks, această distanță este adesea utilizată pentru a măsura similaritatea între două exemple (de exemplu, între o ancoră și un exemplu pozitiv sau între o ancoră și un exemplu negativ) în spațiul de reprezentare învățat de model.

```python
class TripletNetwork(nn.Module):
    def __init__(self, input_size, embedding_size):
        super(TripletNetwork, self).__init__()
        self.embedding = nn.Sequential(
            nn.Linear(input_size, 256),  # Ajustat la 256
            nn.ReLU(),
            nn.Linear(256, embedding_size),
        )

    def forward(self, anchor, positive, negative):
        anchor_embedding = self.embedding(anchor.view(anchor.size(0), -1))
        positive_embedding = self.embedding(positive.view(positive.size(0), -1))
        negative_embedding = self.embedding(negative.view(negative.size(0), -1))
        return anchor_embedding, positive_embedding, negative_embedding
```

Fig. 2 - Clasa TripletNetwork (Principiul Euclidian)

2.3 Procesul de data-mining

Procesul de data mining implică mai multe etape esențiale:

2.3.1 Prelucrarea setului de date: În această etapă, datele brute sunt colectate și curățate. Operatiuni precum eliminarea datelor lipsă, gestionarea și normalizarea datelor sunt realizate pentru a pregăti setul de date pentru analiză.

2.3.2 Extragerea trăsăturilor: După prelucrarea datelor, următorul pas este să se identifice trăsăturile relevante. Această etapă implică selectarea și transformarea atributelor care sunt semnificative pentru obiectivele analizei.

2.3.3 Clasificarea: În acest context, clasificarea se referă la atribuirea exemplelor la categorii specifice. Modelele de învățare automată, cum ar fi cele bazate pe Triplet Networks, pot fi utilizate pentru a clasifica exemplele în funcție de similaritatea lor.

2.3.4 Analiză/Validare: În ultima fază, rezultatele procesului de data mining sunt analizate și validate. Această etapă asigură corectitudinea și relevanța rezultatelor obținute, furnizând o bază solidă pentru interpretarea și luarea deciziilor.

Prin intermediul acestui proces, data miningul devine un instrument puternic pentru descoperirea de cunoștințe semnificative din datele disponibile, contribuind astfel la luarea deciziilor informate și dezvoltarea perspectivelor viitoare.

Capitolul 3: Proiectarea și implementarea sistemului

3.1 Prelucrarea setului de date

Pentru a asigura o potrivire corespunzătoare cu cerințele proiectului, am început prin a prelucra setul de date într-un mod adaptat scopului nostru. Aceasta a inclus etape precum crearea DataLoader pentru setul de date, definirea tripletului, definirea modelului, antrenarea tripletului iar în final vizualizarea rezultatelor redundante și normalizarea datelor pentru a asigura coerența și consistența lor.

3.2 Extragerea

Procesul de extragere a trăsăturilor a fost esențial în obținerea informațiilor semnificative din setul de date. Am selectat cu grijă trăsăturile relevante pentru obiectivele noastre specifice, asigurându-ne că acestea capturează esența datelor și contribuie la calitatea procesului de învățare a conținutului din imagini.

3.3 Implementarea Triplet Networks

Pentru a implementa Triplet Networks în cadrul proiectului, am definit și am antrenat o rețea neurală conform arhitecturii Tripleților. Acest model a fost instruit utilizând setul de date prelucrat, unde exemplul ancoră, exemplul pozitiv și exemplul negativ au fost selectate astfel încât să maximizeze învățarea metricilor și să faciliteze comparațiile de similaritate.

Am ajustat, de asemenea, hiperparametrii modelului, cum ar fi dimensiunea încorporărilor și marja triplet, pentru a îmbunătăți performanța sistemului în contextul specific al proiectului nostru.

Această secțiune detaliată a procesului de proiectare și implementare subliniază eforturile noastre de a crea un sistem eficient în analiza de date, punând bazele pentru analizele ulterioare și evaluarea performanțelor obținute.

Fig. 3 - Antrenarea tripletului

Capitolul 4: Analiză și concluzii

4.1 Analiza rezultatelor

Rezultatele obținute în urma implementării procesului de data mining reflectă eficiența și relevanța metodelor aplicate. Prin utilizarea Triplet Networks, am reușit să obținem încorporări semnificative în spațiul latent, în care distanța dintre exemplele similare este redusă, iar cele nesimilare sunt îndepărtate. Acest lucru indică capacitatea modelului de a învăța metrici semnificative pentru comparații precise între date.

Analiza rezultatelor evidențiază, de asemenea, impactul procesului de prelucrare a datelor și extragere a trăsăturilor asupra performanțelor modelului. Prin selectarea și utilizarea adecvată a trăsăturilor, am reușit să evidențiem aspecte semnificative ale datelor, facilitând astfel învățarea eficientă a rețelei.

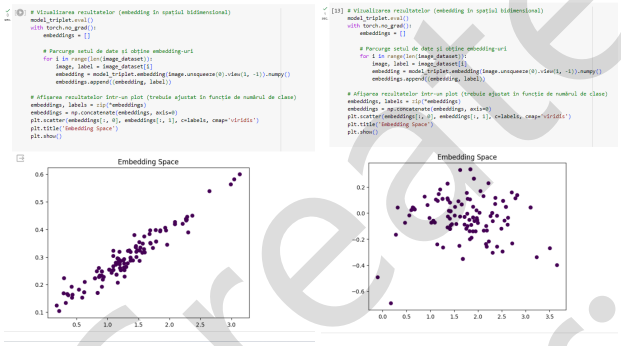

Fig. 4 - Vizualizarea rezultatelor - spațiu bidimensional

4.2 Concluzii

Proiectul de data mining implementat a avut succes în atingerea obiectivelor propuse. Utilizarea Triplet Networks s-a dovedit a fi o abordare eficientă în învățarea metricilor și comparația similarității într-un set de date specific.

Concluziile trase reflectă nu doar performanța modelului, ci și relevanța procesului de prelucrare a datelor și selecție a trăsăturilor. Această abordare oferă o bază solidă pentru analizele viitoare și dezvoltarea ulterioară a sistemului.

4.3 Perspective de dezvoltare și recomandări

Pentru a îmbunătăți și extinde proiectul în viitor, există câteva direcții promițătoare. Adăugarea unor seturi de date suplimentare și diversificarea acestora ar putea contribui la îmbogățirea învățării modelelor, consolidând astfel generalizarea sistemului.

De asemenea, ajustarea hiperparametrilor și explorarea altor arhitecturi de rețele neuronale ar putea conduce la îmbunătățirea performanțelor modelului. Integrarea unor tehnici de regularizare și optimizare ar putea spori stabilitatea și robustețea sistemului în fața diversității datelor.

Aceste perspective de dezvoltare deschid calea pentru evoluția continuă a procesului de data mining implementat, contribuind la adaptarea și optimizarea acestuia în contextul schimbător al analizei de date.

Bibliografie

1.https://books.google.ro/books?hl=ro&lr=&id=NR1oEAAAQBAJ&oi=fnd&pg=PP1&dq=Han,+J.,+Kamber,+M.,+%26+Pei,+J.+(2011).+Data+Mining:+Concepts+and+Techniques.+Morgan+Kaufmann.&ots=_N1HPMpjo4&sig=CeEW2UgzbM8ktvVj7qvrjr7b1TQ&redir_esc=y#v=onepage&q&f=false

2.https://academia.dk/BiologiskAntropologi/Epidemiologi/DataMining/Witten_and_Frank_DataMining_Weka_2nd_Ed_2005.pdf

3.https://books.google.ro/books?hl=ro&lr=&id=mjVKEAAAQBAJ&oi=fnd&pg=PR9&dq=Chollet,+F.+(2017).+Deep+Learning+with+Python.+Manning+Publications.&ots=Ag7YzJTI_j&sig=JvxE2WcAocjbXMPokzOCz-yvg84&redir_esc=y#v=onepage&q&f=false

4.https://ieeexplore.ieee.org/document/7298682

5.https://www.cs.cmu.edu/~rsalakhu/papers/oneshot1.pdf

6.https://github.com/amchirica/triplet_network