Aplicații Multimedia

Tema 2

AI în multimedia

Cuprins

Considerente generale

Inteligența artificială (AI – *Artificial Intelligence*) este un subiect intens discutat în ultimii ani, mai ales odată cu lansarea ChatGPT, chatbot bazat pe modelului lingvistic GPT4, dar și a modelelor care pot crea imagini sau videoclipuri, precum DALL-E al OpenAI sau Midjourney, toate acestea fiind rețele generative de mari dimensiuni.

Cu toate că rezultatele sunt impresionante, este normal să ne putem pune întrebarea cum ne vor afecta viețile, mai ales din punct de vedere profesional. Deja se observă o folosire intensă a ChatGPT în școli și universități, ceea ce înseamnă că în scurt timp va fi necesară o modificare a standardelor, verificărilor și cerințelor, astfel încât să se poată construi pe baza inteligenței artificiale, fără a diminua implicarea, efortul și creativitatea.

Având în vedere că sunteți viitori ingineri, aveți un privilegiu, dar și o responsabilitate, primul pas fiind acela de a înțelege cum se pot folosi aceste noi tehnologii în mod inteligent, dar și care sunt limitările și punctele lor slabe (pentru că există), iar ulterior să ajungeți să puteți contribui chiar voi la dezvoltarea lor și integrarea în sistemele la care veți lucra.

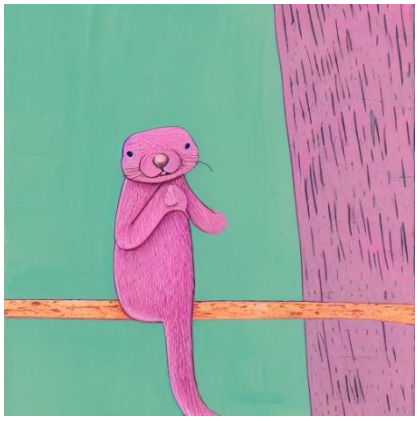
De aceea, în cadrul acestei teme, vom explora câteva modele și implementări care folosesc rețele neuronale și sintetizatoare de voce, pe care le putem folosi ca bază pentru procesări și aplicații ulterioare. Vom trece mai întâi prin cerințele temei, iar apoi prin configurările și instalările necesare.

Cerințe

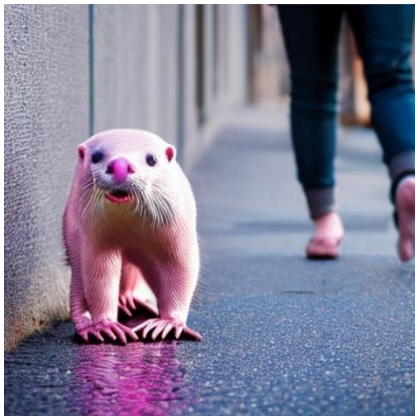
Ce aveți de făcut:

1. Folosind modelul generativ *Stable Diffusion* din *keras-cv* (puteți citi mai multe aici), generați o serie de imagini pe baza unui prompt (o propoziție). Aceste propoziții trebuie să conțină un obiect de o anumită culoare și un fundal, aceste trei caracteristici fiind atribuite fiecărui student în parte (coloanele ***COLOR***, ***OBJECT*** și ***BACKGROUND*** din resursa ***ASIGNARE.pdf***).
   Este recomandat să experimentați cu modul de formare al promptului, astfel încât imaginile generate să reprezinte bine atât obiectul asignat cât și fundalul.
   Spre exemplu, dacă v-a fost alocat obiectul *otter*, culoarea *pink*, și fundalul *city* câteva din rezultatele posibile sunt:

- Pentru promptul “Pink otter in a city”:

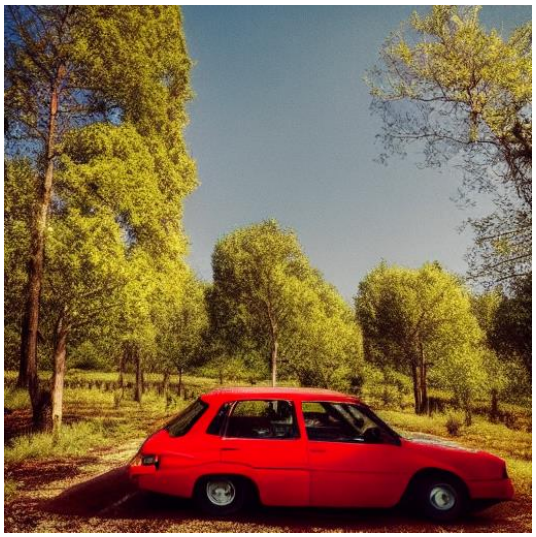


- Pentru promptul "Realistic image with a pink otter on a city street":



ATENȚIE: Generarea imaginilor rulează local, implicit folosindu-se procesorul, nu placa grafică. De aceea, timpul de generarea al unei imagini poate fi considerabil, în funcție de resursele disponibile.
Dacă sunteți interesați, puteți experimenta cu rularea rețelei pe GPU (dacă acest lucru este posibil pe sistemul vostru de operare).

2. Generați o serie de imagini pe baza imaginii de la punctul anterior:
    a. Desenați un dreptunghi vizibil în jurul obiectului de interes.
    b. Aplicați o mască de culoare pentru a păstra din imaginea inițială doar ce este specificat în coloana ***TO_EXTRACT*** din ***ASIGNARE.pdf***, și anume fie doar obiectul, fie doar fundalul.
    Spre exemplu, dacă am avut “*red*”, “*car*” și “*forest*”, iar imaginea generată a fost cea de mai jos:



dacă avem de extras obiectul, imaginea rezultată va fi următoarea:

iar dacă avem de extras fundalul, atunci imaginea rezultată va fi:

c. Obțineți separat masca binarizată aplicată mai sus, de exemplu:

 sau 

d. Transformați imaginea inițială și imaginea obținută la punctul b în spațiul de culoare asignat în coloana ***COLOR_SPACE*** din ***ASIGNARE.pdf***.

3. Pe baza imaginii inițiale și a imaginilor obținute la punctul 2, creați un video care să folosească codecul asignat în coloana ***CODEC*** din ***ASIGNARE.pdf***.
4. Pentru ficare imagine din video, plasați un text vizibil care să descrie imaginea afișată (de exemplu “Original image with a red car in a forest”, “Mask with the background”, etc.).
5. Pe baza titlurilor alese la punctul 4, folosind biblioteca de text-to-speech pyttsx3, generați un fișier audio care să conțină pronunția textului scris pe imagini. Trebuie să aveți atenție la sicronizarea timpului de pronunțare a textului cu timpul de afișare a respectivei imagini în video.

Configurări și instalări

1. Pentru generarea de imagini, puteți folosi environment-ul virtual al laboratorului, *am*, la care mai trebuie să adăugați modulele *tensorflow* și *keras_cv*
   pip install tensorflow
   pip install --upgrade keras_cv

   Puteți găsi un exemplu de utilizare al modelului generativ aici (de la pasul 2).
2. Pentru salvarea sau prelucrarea imaginilor puteți folosi modulul *opencv* (dacă nu aveți deja creat environment-ul *am*, atunci vă puteți crea un environment nou care să conțină doar modulele folosite în această temă).
3. Pentru sintetizarea fișierului audio, instalați biblioteca *pyttsx3*:
   pip install pyttsx3

   Pentru înțelegerea folosirii modulului, consultați documentația oficială aici.

De încărcat

O arhivă .zip cu numele vostru și grupa (*nume_prenume_grupa.zip*) care să conțină:

- Un fișier *generare.py*, cu codul folosit pentru generarea imaginilor, a fișierului audio, precum și a videoului final.
- Imaginea generată aleasă pentru procesare, cu denumirea *image.png*.
- Un fișier *readme.pdf* în care să explicați pe scurt (minim o pagină, maxim 2 pagini) observațiile voastre referitoare la modelul generativ și de sintetizare de voce, detalii despre spațiul de culoare (specificul său, cum se obține din RGB, unde este folosit, etc.), detalierea modului în care ați aplicat masca pentru extragerea culorii, precum și câteva informații despre codecul folosit.

Observații finale

- ✓ Dacă nu puteți rula unul din pași pe sistemele proprii, din cauza limitărilor hardware, vă rugăm să ne anunțați din timp pentru a găsi soluții.
- ✓ Respectați convențiile de nume și tipurile de fișiere cerute, pentru că se vor realiza verificări automate.
- ✓ Fișierele cu cod sursă încărcate vor fi testate anti-plagiat, și vor fi verificate împotriva generării cu modele lingvistice.
- ✓ Modulul de sintetizare de voce sugerat nu utilizează rețele neuronale, ci este bazat pe sintetizatoare de voce open-source pentru sistemele de operare. Puteți folosi alte metode de text-to-speech care să sune mai natural, cu observația că trebuie să specificați în fișierele de cod modul de configurare și folosire astfel încât să fie reproductibil.