Image Captioning with AI: Project Documentation

Introduction

This project demonstrates how to use a pre-trained model for image captioning. The task combines computer vision (image recognition) and natural language processing (caption generation) to generate descriptive captions for images. The model used is a combination of Vision Transformer (ViT) and GPT-2, available via the Hugging Face Transformers library.

Steps Overview

1. Install required libraries.
2. Load the pre-trained model, feature extractor, and tokenizer.
3. Load and preprocess the image.
4. Generate the caption.
5. Display the image with the caption.

Requirements

1. Python 3.8 or higher
2. Required Python packages:
   - torch
   - transformers
   - pillow
   - matplotlib

Installation

Install the necessary libraries using the following command:

```
pip install torch transformers pillow matplotlib
```

Code Explanation

1. Load Pre-trained Model

The VisionEncoderDecoderModel is loaded along with a ViT feature extractor and a GPT-2 tokenizer. These components work together to process the image and generate captions.

2. Preprocess the Image

The image is loaded using PIL (Python Imaging Library) and converted into pixel values compatible with the model using the ViT feature extractor.

3. Generate Caption

The preprocessed image is passed to the model's generate method, which produces token IDs for the caption. These token IDs are then decoded into a human-readable text caption using the tokenizer.

4. Display Image with Caption

The image is displayed using Matplotlib, with the generated caption shown as the title.

Code

```
import torch
from transformers import VisionEncoderDecoderModel,
ViTFeatureExtractor, AutoTokenizer
from PIL import Image
import matplotlib.pyplot as plt

# Step 1: Load pre-trained model, feature extractor, and tokenizer
model = VisionEncoderDecoderModel.from_pretrained("nlpconnect/vit-
gpt2-image-captioning")
feature_extractor =
ViTFeatureExtractor.from_pretrained("nlpconnect/vit-gpt2-image-
captioning")
tokenizer = AutoTokenizer.from_pretrained("nlpconnect/vit-gpt2-
image-captioning")

# Step 2: Load and preprocess the image
image = Image.open("your_image.jpg")
pixel_values = feature_extractor(images=image,
return_tensors="pt").pixel_values
```

```
# Step 3: Generate caption
caption_ids = model.generate(pixel_values, max_length=16,
num_beams=4)
caption = tokenizer.decode(caption_ids[0], skip_special_tokens=True)

# Step 4: Display the image with the caption
plt.imshow(image)
plt.title(caption)
plt.axis('off')
plt.show()
```

Output

For example, if the image is of a dog playing in a park, the generated caption might be:
"a dog playing in the grass."