**A demonstration of a machine leaning driven process**

**Supervised by:**

**Eng. Ehab Ibrahim**

**Team:  
1- Omar Abd Allah Omar**

**2- Yasmeen Mohamed Abd-Elradi**

**3- Moataz Hassan Ahmed**

**4- Rana Talaat Shalaby**

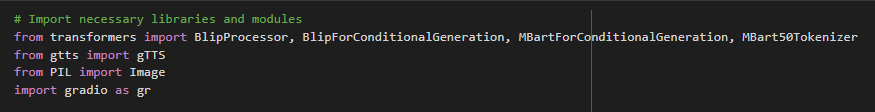
**Image Captioning, Arabic Translation and Text-to-Speech (TTS) Pipeline**

**Detailed Documentation: Image Captioning and Arabic Translation Pipeline**

**Overview**

This document provides an in-depth explanation of the code that implements a pipeline for processing an image, generating a caption, translating it to Arabic, and converting the text to speech (TTS) using machine learning models. The pipeline is made up of three main components: image captioning, Arabic translation, and text-to-speech, with Gradio used as the front-end interface to allow user interaction with the system.

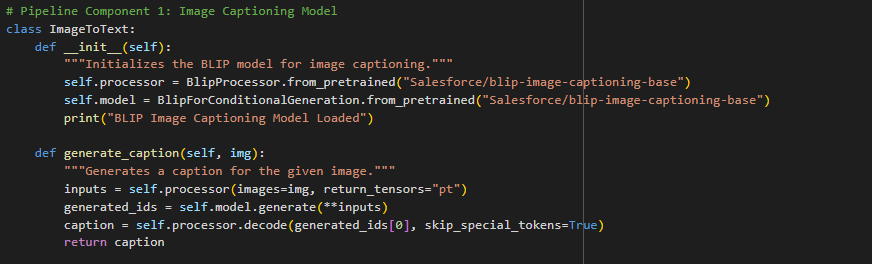
### **1. Importing Necessary Libraries and Modules**



#### **Explanation:**

* **BlipProcessor and BlipForConditionalGeneration**: Used for the image captioning task. The processor prepares the image input for the model, and the model generates the descriptive caption.
* **MBartForConditionalGeneration and MBart50Tokenizer**: Used for text translation. The mBART model can translate between multiple languages, and in this case, it translates from English to Arabic.
* **gTTS**: A Python library for interfacing with Google Text-to-Speech. It converts the generated Arabic text into audio.
* **PIL (Python Imaging Library)**: Used to open and manipulate images in Python.
* **Gradio**: A Python library for building user interfaces that can interact with machine learning models. It allows users to upload images and receive audio and text outputs through a simple web interface.

### **2. Pipeline Component 1: Image Captioning Model:**



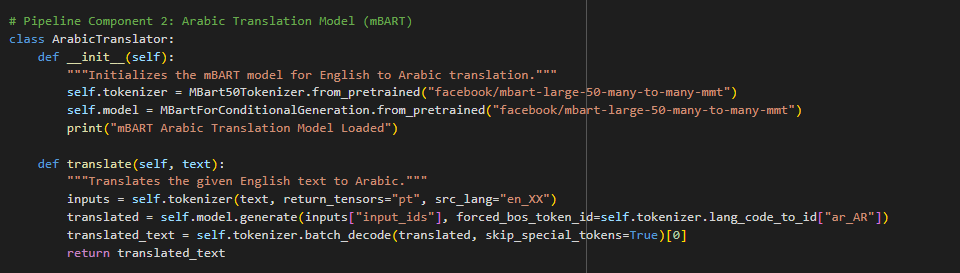
#### 

#### **Explanation:**

* **ImageToText class**: Initializes and loads the BLIP model for generating captions.
  + **BlipProcessor**: Prepares the image to be processed by the BLIP model.
  + **BlipForConditionalGeneration**: The model generates a descriptive caption for the image.
  + **generate\_caption(img)**: Accepts an image as input, processes it, and returns a caption.
  + **decode**: Converts the model's output from token IDs into human-readable text, skipping unnecessary tokens.

### 

### **3. Pipeline Component 2: Arabic Translation Model (mBART):**



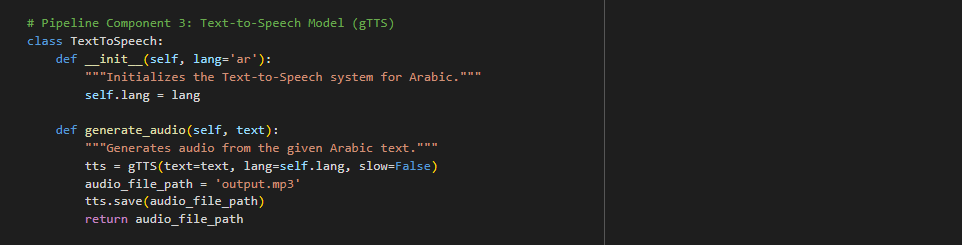
#### **Explanation:**

* **ArabicTranslator class**: Initializes and loads the mBART model for translating text from English to Arabic.
  + **MBart50Tokenizer**: Tokenizes the input English text.
  + **MBartForConditionalGeneration**: Translates the tokenized text into Arabic.
  + **translate(text)**: Takes an English text, tokenizes it, translates it, and decodes it into Arabic text.
  + **forced\_bos\_token\_id**: Forces the model to start generating tokens in the target language (Arabic).

### 

### 

### **4. Pipeline Component 3: Text-to-Speech (gTTS):**

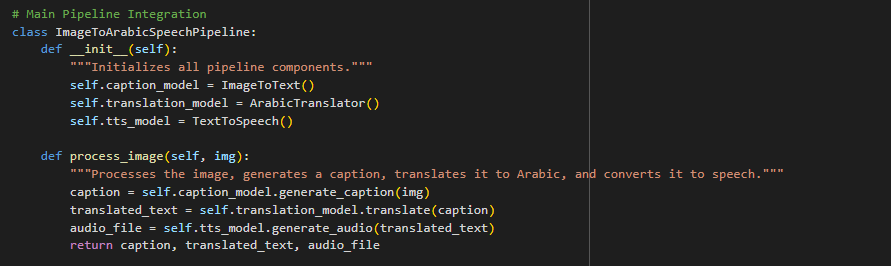


#### **Explanation:**

* **TextToSpeech class**: Uses Google Text-to-Speech (gTTS) to convert the Arabic text into spoken audio.
  + **gTTS**: Converts the provided text into an audio file in MP3 format.
  + **generate\_audio(text)**: Generates an MP3 file from the given text.
  + **lang**: Arabic ('ar') is set as the default language for TTS

### 

### **5. Main Pipeline Integration**



#### **Explanation:**

* **ImageToArabicSpeechPipeline class**: The main pipeline that integrates all the components (image captioning, translation, and TTS).
  + **process\_image(img)**: The function accepts an image, generates a caption, translates the caption into Arabic, and then converts the translated text into speech.

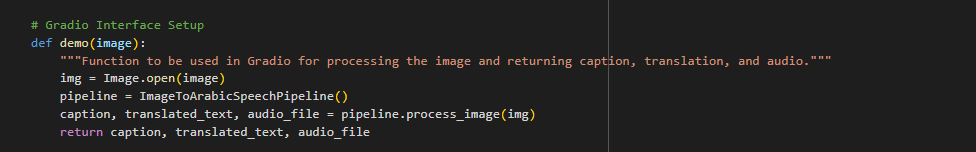
### 

### 

### 

### 

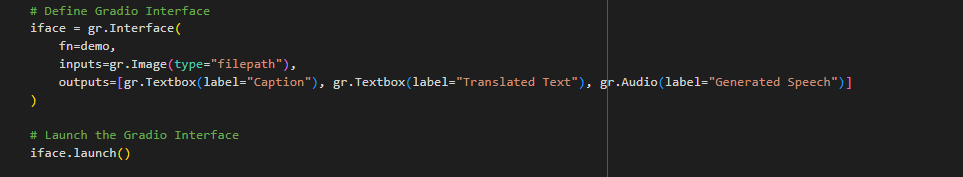
### **6. Gradio Interface Setup**



#### **Explanation:**

* **demo(image)**: This function handles the image uploaded by the user, passes it through the pipeline, and returns the generated caption, translated text, and audio file.

### **7. Gradio Interface Launch**



#### **Explanation:**

* **gr.Interface**: Sets up the Gradio interface to accept images and display captions, translated text, and generated audio.
  + **inputs**: Accepts an image file as input.
  + **outputs**: Returns a textbox for the image caption, a textbox for the Arabic translation, and an audio player for the generated speech.
  + **launch()**: Starts the Gradio web interface.

### **8. Conclusion**

This pipeline combines the power of machine learning models and user-friendly interfaces to transform images into spoken Arabic. By using the BLIP model for captioning, mBART for translation, and gTTS for text-to-speech, the system provides a seamless experience from image upload to audio output. Gradio simplifies the interaction with the models, allowing users to upload images and instantly receive text and audio outputs.

#### **Future Improvements:**

* **Additional Language Support**: Expand translation capabilities to other languages.
* **Alternative TTS Engines**: Integrate different TTS systems for more voice variety.
* **Enhanced Accuracy**: Implement a retrieval-augmented generation (RAG) model for improved context-based translation.