Day 30 Final Project: Fine-tuning GPT-2 for Product Description Generation

Objective:
Fine-tune a pre-trained GPT-2 model to generate product descriptions based on product names and categories from the Amazon Product Reviews Dataset

Dataset:
- Source: Amazon Product Reviews Dataset
- URL: : https://www.kaggle.com/datasets/PromptCloudHQ/amazon-reviews-unlocked-mobile-phones
- Description: The dataset contains various product reviews, including product names, categories, and descriptions.

Project Requirements:
- **Programming Language:** Python
- **Libraries:**
  - Hugging Face Transformers
  - Datasets
  - Tensorflow/PyTorch
  - Pandas
  - Scikit-learn

Project Steps:
- Set Up Environment:
  - Install necessary libraries:
  - !pip install transformers datasets torch huggingface_hub
- Load and Preprocess the Dataset
  - Load the dataset into a Pandas DataFrame.
  - Create a new column that combines product attributes (name, category) into a single input text.
  - Create a target column with the existing product descriptions.
  - Split the dataset into training and testing sets.
- Tokenization:
  - Load the pre-trained GPT-2 tokenizer and model.
  - Tokenize the input and target texts.
- Training Setup:
  - Define training arguments using TrainingArguments from Hugging Face, including:
    - Number of training epochs
    - Batch sizes
    - Evaluation strategy
    - Learning rate
    - Logging parameters
- Fine-tuning the Model:
  - Use the "Trainer" class to fine-tune the GPT-2 model on the training dataset.
  - Monitor the training process and evaluate the model's performance on the test dataset.
- Generate Product Descriptions:
  - Write a function to generate product descriptions based on new product names and categories using the fine-tuned model.

Expected Outcomes:
- A fine-tuned GPT-2 model capable of generating coherent and contextually relevant product descriptions.
- Understanding of the process of fine-tuning a pre-trained model and applying it to real-world tasks.

Hints and Tips:
- Use a maximum sequence length of 512 tokens for tokenization.
- Experiment with hyperparameters (learning rate, batch size) to improve model performance.
- Make use of model checkpoints to save the best version of your model during training.

Reference Materials:
- 30-Day GenAI Masterclass Material
- Hugging Face Transformers Documentation