Overview

This project fine-tunes the DistilBERT model to predict answers to questions given a context. It uses Hugging Face's transformers library for loading the pre-trained model, tokenization, and training pipeline.

Key features:

- Fine-tunes DistilBERT on SQuAD v2 dataset.
- Implements preprocessing to handle context truncation and map answers.
- Uses Hugging Face's Trainer for efficient training and evaluation.
- Supports mixed-precision training for faster computations.

Requirements

To run the code, you need the following libraries installed:

- torch
- datasets
- transformers

Dataset

This code uses the SQuAD v2 dataset. The dataset is automatically downloaded using the Hugging Face datasets library.

SQuAD v2: A collection of question-answer pairs with some unanswerable questions.

Code Explanation

1. Model and Tokenizer

We use the pre-trained distilbert-base-uncased model from Hugging Face. The tokenizer is used to prepare inputs for the model.

2. Preprocessing

The preprocess_data function tokenizes the question and context. It computes start_positions and end_positions for the answer within the context. For unanswerable questions, the positions default to 0.

Key parameters:

- max_length: Maximum length of input sequences (default: 512).
- doc_stride: Overlap between document splits for handling long contexts.

3. Training Arguments

The TrainingArguments class defines hyperparameters for training, such as batch size, learning rate, and number of epochs.

4. Trainer

The Hugging Face Trainer simplifies the training process by handling data batching, model optimization, and evaluation.