Name: Pravallika Cheekatimalla
Assignment: CPSC 8430 - Deep Learning - Homework 3
Git Hub: https://github.com/Pravallika-Cheekatimalla/DeepLearning/tree/main/Assignment3

Introduction

This project explores the implementation of an extractive question-answering system using **DistilBERT** on the **Spoken-SQuAD** (or SQuAD) dataset. The goal is to fine-tune DistilBERT for identifying answer spans from spoken or transcribed documents in response to user questions. Two models were developed:

1. A baseline DistilBERT model.
2. An improved DistilBERT model with performance optimizations.

Dataset

We used the **Spoken-SQuAD** dataset, a spoken question-answering dataset adapted from SQuAD, which includes question-answer pairs where the document is in spoken or transcribed form. Key steps involved:

- Converting spoken documents to text (using ASR if necessary).
- Filtering question-answer pairs where answers were missing in the transcriptions.

Dataset Details:

- **Training Set**: 37,111 question-answer pairs
- **Test Set**: 5,351 question-answer pairs
- **Word Error Rates (WER)**: Introduced varying WERs to simulate noisy real-world audio conditions.

Model Selection

We selected **DistilBERT** for its efficiency and suitability for question-answering tasks with limited computational resources. DistilBERT is a smaller, faster variant of BERT, retaining most of its capabilities for NLP tasks.

1. Baseline Model

The baseline model was fine-tuned with the following steps:

- **Data Tokenization**: Inputs were tokenized to a maximum length of 512 tokens, with a stride of 128 tokens to allow overlapping windows, accommodating longer documents.
- **Model Architecture**: Used the DistilBertForQuestionAnswering class from Hugging Face.

- **Training Setup**: Fine-tuning was conducted with:
  - **Batch size**: 8
  - **Epochs**: 3
  - **Learning rate**: 3e-5 with a linear decay scheduler.

2. Performance Improvement Model

Building on the baseline, this model incorporated additional optimizations:

- **Learning Rate Decay**: Implemented linear learning rate decay, reducing the learning rate incrementally at each step for more stable convergence.
- **Gradient Accumulation**: Applied gradient accumulation to manage larger effective batch sizes without increasing memory usage.

Results

Baseline Model Performance

Metric	Value
Exact Match (EM)	33.71
F1 Score	45.42

Performance Improvement Model

Metric	Value
Exact Match (EM)	48.26
F1 Score	54.33