Michael Elrod

# Implementation of BERT-Based Question Answering System on Spoken-SQuAD Dataset

## GitHub Repository Link

## 1 Introduction

This project implements a Question Answering (QA) system using BERT-based models on the Spoken-SQuAD dataset. The system is designed to handle extractive question answering tasks where the answer must be extracted as a span from a given context. Two implementations are presented: a baseline model and an improved version with enhanced training procedures and optimization techniques.

## 2 Dataset

The Spoken-SQuAD dataset is a modified version of the original SQuAD dataset, where documents are converted to spoken form. Key characteristics include:

- Training set: 37,111 question-answer pairs (WER 22.77%)
- Testing set: 5,351 pairs (WER 22.73%)
- Additional test sets with increased noise levels (WER 44.22% and 54.82%)

## 3 Methodology

### 3.1 Base Implementation (Simple Baseline)

The baseline model implements the basic BERT QA system using DistilBERT with:

- Model: DistilBertForQuestionAnswering
- Batch size: 16
- Learning rate: 2e-6
- Number of epochs: 10
- Basic tokenization and preprocessing

### 3.2 Medium Level Improvements

The following medium-level enhancements were implemented:

- Linear learning rate decay using get_linear_schedule_with_warmup
- Learning rate gradually decreases from 2e-6 to 0 over training period

### 3.3 Strong Level Improvements

At the strong level, preprocessing improvements were implemented:

- Enhanced token position handling for better answer span detection
- Improved character-to-token conversion process
- More robust handling of answer end indices

### 3.4 Boss Level Improvements

The final improvements focused on postprocessing:

- Implementation of AnswerEvaluator class for robust answer evaluation
- Enhanced answer normalization and text processing
- Improved F1 and exact match score calculations

## 4 Results

### 4.1 Performance Metrics

The implementations achieved:

- Base Model F1 Score: 52.54%
- Improved Model:
  - F1 Score: 52.54%
  - Exact Match Score: 39.02%

## 5 Analysis

### 5.1 Implementation of Required Improvements

The project successfully implemented several key requirements:

- Linear learning rate decay showing clear progression in training
- Enhanced preprocessing with improved token handling
- Robust postprocessing with comprehensive evaluation metrics

### 5.2 Training Dynamics

The improved model demonstrates more stable training dynamics with:

- Gradual learning rate decay from 2e-6 to 0

- More consistent loss reduction over epochs
- Better handling of position adjustments in preprocessing

### 5.3 Architecture Improvements

Key architectural enhancements include:

- Better token position handling
- Improved answer span extraction
- More robust evaluation metrics

## 6 Conclusion

While both implementations achieved similar F1 scores, the improved version offers several advantages:

- More robust training procedure
- Better code organization and maintainability
- Additional evaluation metrics
- Enhanced preprocessing pipeline

The project demonstrates the effectiveness of BERT-based models for spoken question answering tasks, while highlighting the importance of proper training procedures and preprocessing techniques.