Documentation Report: Question Answering System with BERT

1. Methodology and Steps Taken

Objective

The goal of this project is to develop a Question Answering (QA) system by fine-tuning a pre-trained BERT model using the Stanford Question Answering Dataset (SQuAD). The process involves data preprocessing, model adaptation for QA tasks, training, evaluation, and deployment using a Streamlit-based interface.

Steps Taken

1. **Setup and Initialization**
   - Verified CUDA availability for GPU acceleration.
   - Installed necessary dependencies including `datasets`, `transformers`, `torch`, and `streamlit`.
2. **Data Acquisition and Preprocessing**
   - Loaded the SQuAD dataset using the Hugging Face `datasets` library.
   - Tokenized questions and contexts using `DistilBertTokenizerFast`.
   - Mapped answer spans to their respective token positions.
   - Removed unnecessary columns from the dataset.
3. **Model Selection and Training**
   - Used the pre-trained `distilbert-base-cased-distilled-squad` model for fine-tuning.
   - Defined training arguments, including batch size, learning rate, and number of epochs.
   - Fine-tuned the model using the `Trainer` API.
   - Saved the trained model and tokenizer.
4. **Evaluation**
   - Implemented Exact Match (EM) and F1 score evaluation metrics.
   - Computed predictions using start and end token logits.
   - Conducted qualitative testing with sample inputs.
5. **Deployment with Streamlit**
   - Developed an interactive web app using Streamlit.
   - Allowed users to input a context and question to retrieve answers.
   - Implemented model inference with validation.

---

2. Experimentation Details

Hyperparameter Choices

Parameter	Value
Model Name	`distilbert-base-cased-distilled-squad`
Max Token Length	384
Batch Size (Train & Eval)	16
Learning Rate	5e-5
Number of Epochs	2
Weight Decay	0.01

Challenges Encountered

1. **Answer Alignment Issues**
   - The answer span did not always align with tokenized input.
   - Solution: Used offset mappings to correctly map answers.
2. **Limited Training Time and Resources**
   - Training deep learning models requires significant computational power.
   - Solution: Used DistilBERT, a lightweight alternative to BERT, for faster fine-tuning.
3. **Handling Incorrect Predictions**
   - The model sometimes returned an empty or irrelevant response.
   - Solution: Implemented logic to ensure start index precedes end index and adjusted token selections accordingly.

---

3. Evaluation Results and Insights

Qualitative Testing Examples

Example 1

- **Context**: "Paris is the capital and most populous city of France."
- **Question**: "What is the capital of France?"
- **Predicted Answer**: "Paris"

Example 2

- **Context**: "The Eiffel Tower is an iron lattice tower in Paris, France. It was built in 1889."
- **Question**: "When was the Eiffel Tower built?"
- **Predicted Answer**: "1889"

Insights

- The model performs well on direct fact-based questions.
- Performance declines when faced with complex or ambiguous questions.
- The lightweight nature of DistilBERT ensures faster inference, making it practical for real-time applications.

---

4. Conclusion

This project successfully implemented a fine-tuned QA system based on DistilBERT. By leveraging the SQuAD dataset and employing advanced token alignment techniques, we achieved a strong balance between performance and efficiency. Future improvements could include:

- Fine-tuning on domain-specific datasets for better accuracy.
- Implementing beam search or ensemble methods for more robust predictions.
- Exploring multi-turn question answering for conversational AI applications.

---

5. References

- SQuAD Dataset: https://rajpurkar.github.io/SQuAD-explorer/
- Streamlit Documentation: https://docs.streamlit.io/

Prepared by: Omar Esmat