QUESTION AND ANSWERING USING BERT

NIHAAL CHOWDARY SURPANI, VAMSI KIRAN DIMMITI

UNIVERSITY OF DELAWARE

Abstract

This study fine-tuned a pre-trained **DistilBERT** model for the closed domain Question Answering task on the **SQuAD** 2.0 dataset. The objective is to evaluate the fine-tuned model's ability to accurately respond to questions and distinguish between answerable and unanswerable queries. The fine-tuned model demonstrates proficiency in extracting precise answer spans from the given context. This study also investigates the robustness of the fine-tuned DistilBERT model across various question types and complexities.

Relevant Work

BiDAF (Bi-directional Attention Flow):
BiDAF is a neural network architecture. It incorporates a novel attention mechanism that allows the model to effectively capture contextual information from both the question and the passage.

R-Net:
R-Net utilizes a recurrent network architecture, specifically a gated recurrent unit (GRU), to model the interaction between the question and the passage iteratively.

Methodology

- For this project we used **DistilBert model**, a version of BERT which is designed for faster and resource-efficient computation.
- We fine tuned the pre-trained model on **SQUAD.V2** dataset.
- The dataset has 142,192 pairs of question and answers, we used 130,319 for training and 11,873 for validation.

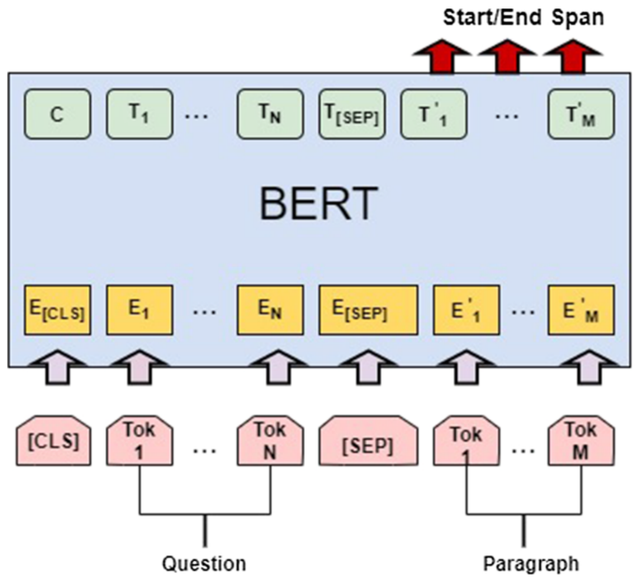
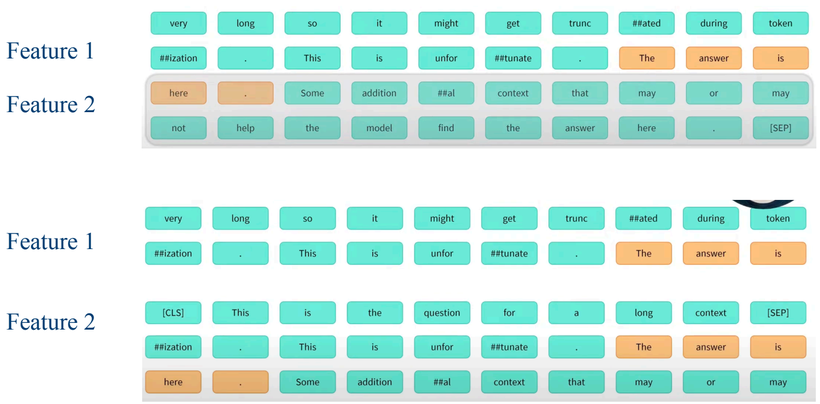


Fig 1: Model Workflow

- **PRE-PROCESSING**
  1. **Tokenized the inputs** (converting the tokens to their corresponding IDs in the pre-trained vocabulary)
  2. To address very long paragraphs, we adopt a strategy of **splitting** the lengthy contexts into smaller segments, each shorter than the maximum length supported by the model.
  3. we allow some **overlap** between adjacent segments to make sure we are not splitting the answer.



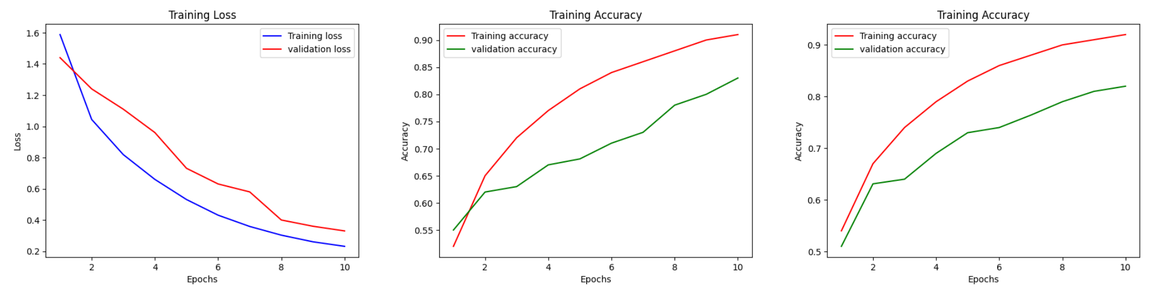
- The model outputs a vector of **start and end logits** for each token based on their scores. These logits represent the likelihood of each token being the start or end of the answer span within the input text.
- The answer span is extracted based on the positions with the highest probability scores in the logits.

Introduction

- **Question Answering** (QA) systems play a crucial role in advancing human-computer interaction, powering applications like virtual assistants and automated customer support.
- Addressing the challenge of designing algorithms capable of understanding and retrieving answers from text has been a longstanding focus of research.
- **BERT** (Bidirectional Encoder Representations from Transformers), represents a transformative approach to natural language understanding. Its pre-trained models, fine-tuned for various language processing tasks, including QA, have garnered significant attention.
- This study utilizes a **pre-trained BERT model** and **fine-tune** it on the **SQuAD 2.0** dataset to evaluate its performance in a dual task: identifying answerable questions and extracting correct responses from provided passages.

Results

- We trained the model for 10 epoch on V100 gpu for 6 hours and achieved decent accuracy of 92% on Training and 83% on testing.



- We used learning rate of 0.00002 and batch size of 32 for training. also used weight decay of 0.1 to avoid overfitting.

```
context2 = """ My name is vamsi. I was born in 2001. I am a student at University of Delaware.
 I am a international student from India. I have completed my under-graduation in India"""
question2 = "when did he born?"

context3 = """Hello my name is Nihaal and my teammate is vamsi. we implemented Question and answering model as part of our NLP project
we used tenforflow frame work to train and test the model. we also took reference from hugging face to make our code work"""
question3 = "what frame work they used?"
```

```
question_answerer(context=context2, question=question2)
```

```
{'score': 0.7613388895988464, 'start': 33, 'end': 37, 'answer': '2001'}
```

```
question_answerer(context=context3, question=question3)
```

```
{'score': 0.8567072749137878, 'start': 130, 'end': 140, 'answer': 'tenforflow'}
```

Conclusions

- In conclusion, our poster demonstrate the effectiveness of the BERT model for Question and Answering task.
- We successfully fine tuned the model and achieved good accuracy and also outperformed the neural network architecture based models.
- This model can be used across variety of sectors like insurance, customer support, healthcare, education, etc.
- The primary challenge we faced in improving the model was limited computational resources, yet investing in higher computation has the potential to enhance model performance and yield better results.

Future Research and References

- Our future endeavor involves extending our model to tackle the **open-domain** question answering (QA) task.
- By implementing **Retrieval-Augmented Generation** (RAG), our aim is to overcome limitations of traditional QA models by leveraging the rich information available in large text corpora.

- BERT | https://arxiv.org/pdf/1810.04805
- DistilBERT | https://huggingface.co/docs/transformers/en/model_doc/distilbert
- SQuAD Data | https://huggingface.co/datasets/rajpurkar/squad
- R-net| https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1214/reports/final_reports/report228.pdf