PDF Answering AI

Problem Statement and Objective

Problem Statement

The task is to develop an AI model capable of answering questions based on the content of a PDF document. This model should leverage natural language processing (NLP) techniques to understand the context and provide accurate responses to queries.

Objective

The objective of this project is to create a system that can interactively respond to questions about a PDF document, mimicking a scenario where the PDF "talks" back to the user. This involves fine-tuning a pre-trained model (the model is bert-large-uncased-whole-word-masking) on the SQuAD (Stanford Question Answering Dataset) dataset to ensure the model understands and processes the context of the text within the PDF to generate precise answers.

Methodology

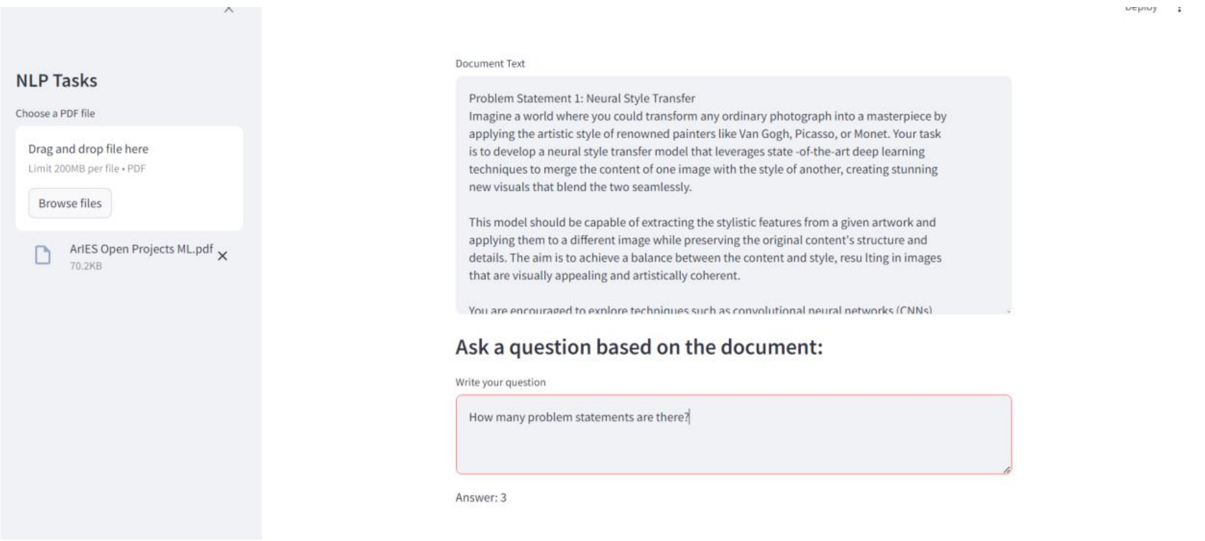
1. **Data Preparation**:
    - Collected relevant data by using the SQuAD dataset, which is a large dataset for training question-answering systems.
    - Pre-processed the data to ensure compatibility with the chosen model architecture.
2. **Model Selection and Fine-Tuning**:
    - Selected a pre-trained model, BERT (Bidirectional Encoder Representations from Transformers), due to its effectiveness in NLP tasks. [The specific model is bert-large-uncased-whole-word-masking].
    - Fine-tuned BERT on the SQuAD dataset to adapt it to the task of answering questions based on the context of the PDF content.
3. **Implementation**:
    - Developed a nlp pipeline from transformers to upload PDFs and extract text.
    - Implemented a mechanism to input questions and retrieve relevant text sections using the fine-tuned model.
    - Developed the frontend of the project using streamlit making it easy to interact.

Failed Approaches

1. **Vector Databases and Cosine Similarity**:
   - Attempted to convert PDF text into vectors using word2vec and GloVe embeddings.
   - Transformed questions into vector format and tried to find the closest match using cosine similarity.
   - This approach failed due to a lack of contextual understanding and difficulty in handling complex queries.
2. **Finding the correct pre-trained model and the best dataset for fine tuning:**
   - Some models were not up to the mark when it came to predicting answers.
   - Also in the fine tuning some datasets were very small and even fine tuning on them took a lot of time and yet they were not able to generate the correct or the expected answers.
3. **Other Attempts**:
   - **Using TF-IDF**: Tried using term frequency-inverse document frequency (TF-IDF) to identify important words and match queries, but it did not capture the context effectively.
   - **LSTM Networks**: Experimented with Long Short-Term Memory (LSTM) networks, but they struggled with the large size of text and required extensive computational resources.

Results:

I answered the questions from Aries Open Projects ML.pdf itself and below are some examples.



Problem Statement 1: Neural Style Transfer
Imagine a world where you could transform any ordinary photograph into a masterpiece by applying the artistic style of renowned painters like Van Gogh, Picasso, or Monet. Your task is to develop a neural style transfer model that leverages state -of-the-art deep learning techniques to merge the content of one image with the style of another, creating stunning new visuals that blend the two seamlessly.

This model should be capable of extracting the stylistic features from a given artwork and applying them to a different image while preserving the original content's structure and details. The aim is to achieve a balance between the content and style, resu lting in images that are visually appealing and artistically coherent.

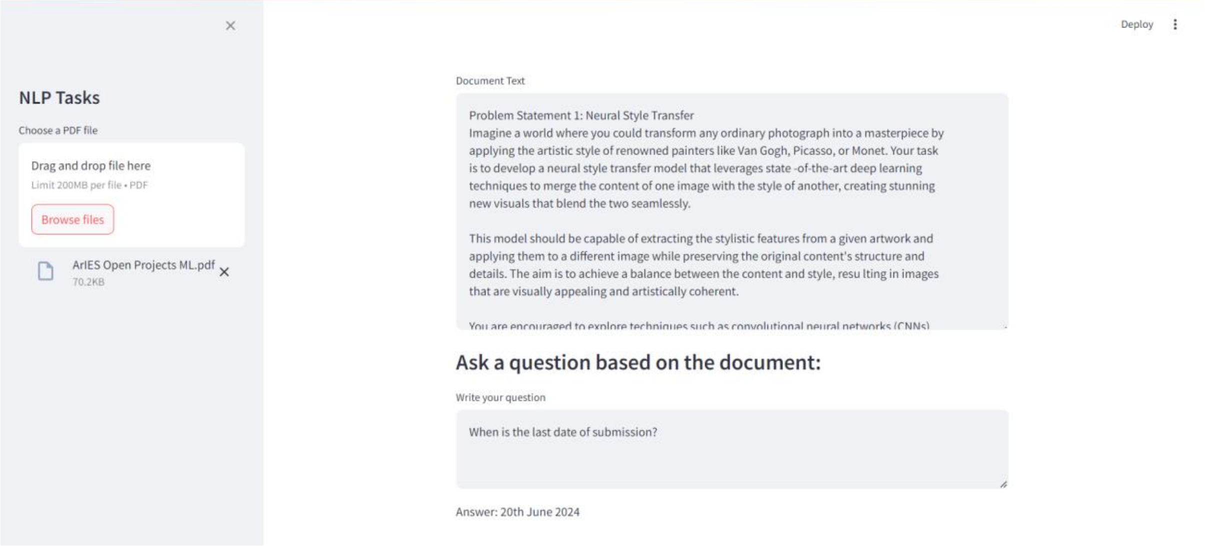
You are encouraged to explore techniques such as convolutional neural networks (CNNs)

Ask a question based on the document:

Write your question

How many problem statements are there?

Answer: 3



Problem Statement 1: Neural Style Transfer
Imagine a world where you could transform any ordinary photograph into a masterpiece by applying the artistic style of renowned painters like Van Gogh, Picasso, or Monet. Your task is to develop a neural style transfer model that leverages state -of-the-art deep learning techniques to merge the content of one image with the style of another, creating stunning new visuals that blend the two seamlessly.

This model should be capable of extracting the stylistic features from a given artwork and applying them to a different image while preserving the original content's structure and details. The aim is to achieve a balance between the content and style, resu lting in images that are visually appealing and artistically coherent.

You are encouraged to explore techniques such as convolutional neural networks (CNNs)

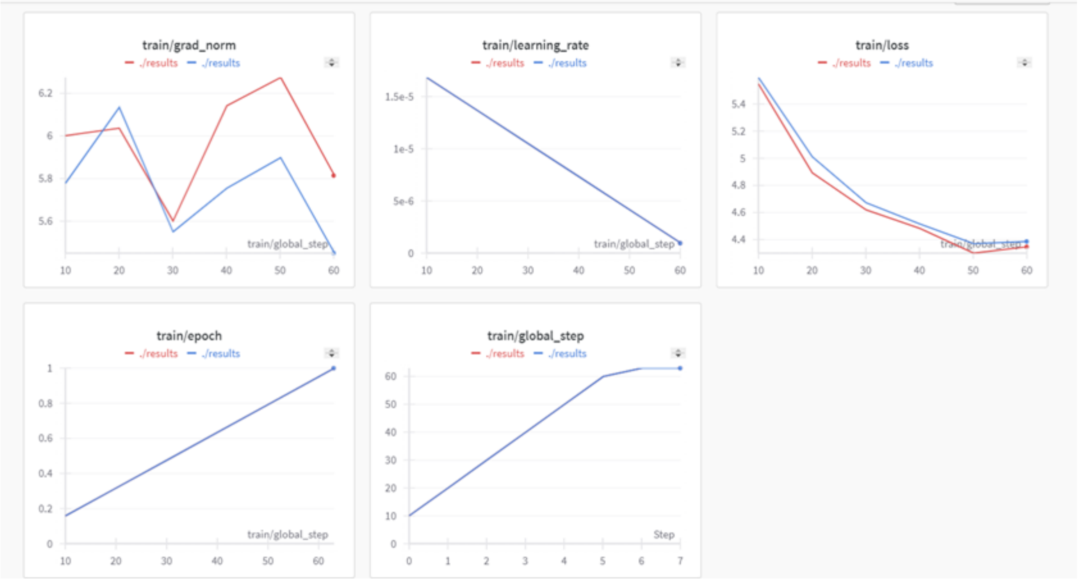
Ask a question based on the document:

Write your question

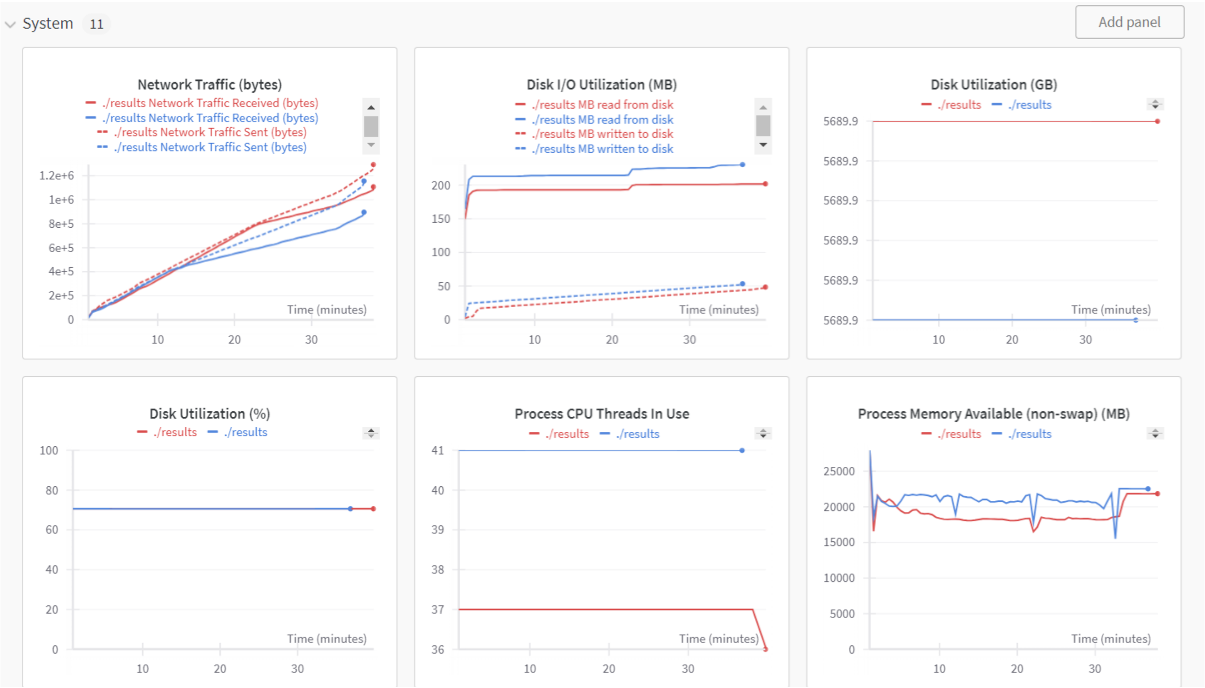
When is the last date of submission?

Answer: 20th June 2024

Also here are some results from training the model which I get from "**wandb**" which helps in visualising the training phase of the model and the system processing during that phase.



train/grad_norm
./results
./results
6.2
6
5.8
5.6
train/global_step
10
20
30
40
50
60
train/learning_rate
./results
./results
1.5e-5
1e-5
5e-6
0
train/global_step
10
20
30
40
50
60
train/loss
./results
./results
5.4
5.2
5
4.8
4.6
4.4
train/global_step
10
20
30
40
50
60
train/epoch
./results
./results
1
0.8
0.6
0.4
0.2
0
train/global_step
10
20
30
40
50
60
train/global_step
./results
./results
60
50
40
30
20
10
0
Step
0
1
2
3
4
5
6
7



System 11
Add panel
Network Traffic (bytes)
./results Network Traffic Received (bytes)
./results Network Traffic Received (bytes)
./results Network Traffic Sent (bytes)
./results Network Traffic Sent (bytes)
1.2e+6
1e+6
8e+5
6e+5
4e+5
2e+5
0
Time (minutes)
10
20
30
Disk I/O Utilization (MB)
./results MB read from disk
./results MB read from disk
./results MB written to disk
./results MB written to disk
200
150
100
50
0
Time (minutes)
10
20
30
Disk Utilization (GB)
./results
./results
5689.9
5689.9
5689.9
5689.9
5689.9
5689.9
Time (minutes)
10
20
30
Disk Utilization (%)
./results
./results
100
80
60
40
20
0
Time (minutes)
10
20
30
Process CPU Threads In Use
./results
./results
41
40
39
38
37
36
Time (minutes)
10
20
30
Process Memory Available (non-swap) (MB)
./results
./results
25000
20000
15000
10000
5000
0
Time (minutes)
10
20
30

Analysis and Insights

Learnings and Insights

1. **Understanding of Transformers**:
    - Gained a deep understanding of transformer models, particularly BERT, and their application in NLP tasks.
    - Learned about the architecture and functioning of transformers, including attention mechanisms and their advantages over traditional RNNs and LSTMs.
2. **Applications and Uses**:
    - Realized the broad applicability of transformer models in various NLP tasks beyond question answering, such as text summarization, translation, and sentiment analysis.
    - Understood the importance of fine-tuning pre-trained models on domain-specific data to improve performance.
3. **Challenges and Solutions**:
    - Encountered challenges with handling large volumes of text and ensuring the model's responses were accurate and contextually relevant.
    - Overcame these challenges by fine-tuning on a relevant dataset and optimizing the text extraction and processing pipeline.
    - Also, a major challenge was the time taken for fine tuning the model. SQuAD is a large dataset and fine tuning a large model like bert-large-uncased-whole-word-masking takes lot of time and patience.

Summary and Future Improvements

Summary

In this project, we developed an AI model that can answer questions based on the content of a PDF document. By fine-tuning a pre-trained BERT model on the SQuAD dataset, we achieved a system capable of understanding and responding to queries accurately. This project provided valuable insights into the workings of transformer models and their practical applications in NLP tasks.

Future Improvements

1. **Enhanced Contextual Understanding**:
    - Integrate more sophisticated text preprocessing techniques to improve the model's understanding of the document structure.
    - Experiment with other transformer models like RoBERTa for potentially better performance.
2. **User Interface Improvements**:
    - Develop a more user-friendly interface for interacting with the model, such as a browser extension or a mobile app.
    - Implement voice input for questions to make the system more accessible.
3. **Expand Dataset**:
    - Fine-tune the model on additional datasets to cover a wider range of topics and improve generalization.

References

1. **SQuAD Dataset:** https://rajpurkar.github.io/SQuAD-explorer/
2. **Attention is All You Need**: https://arxiv.org/abs/1706.03762
3. **BERT Playlist:** https://www.youtube.com/playlist?list=PLam9sigHPGwOBuH4_4fr-XvDbe5uneaf6
4. **Sequence model by Andrew Ng:** https://www.youtube.com/watch?v=S7oA5C43Rbc&t=5130s
5. **Illustrated Transformer**: http://jalammar.github.io/illustrated-transformer/

REPORT BY:

RAGHAV NYATI
22112082