Enhancing Research Paper Comprehension through BERT-Based Summarization and Retrieval

Group 1: Krishna Raketla, Rohith Chandra Kandambeth, Avinash Kasireddy

1. Description

The rapid growth in the volume of research papers necessitates efficient tools for navigating and understanding content swiftly. Our project aims to address this issue by developing a model that converts PDF research papers into embeddings, enabling quick retrieval of relevant information based on user queries. Furthermore, we intend to fine-tune BERT for abstractive summarization to provide concise and accurate summaries of the retrieved content. This approach will aid users in comprehending the material faster, ultimately facilitating improved accessibility and comprehension of research papers.
Reference papers: https://arxiv.org/pdf/1908.08345.pdf (Text Summarization with Pretrained Encoders)
https://arxiv.org/pdf/1508.06034.pdf (Better Summarization Evaluation with Word Embeddings for ROUGE)

2. Dataset

We will utilize the SCITLDR Dataset for training our BERT model in abstractive summarization. This dataset provides concise summaries of scientific papers, serving as a valuable resource for supervised learning. (https://metatext.io/datasets/scitldr)

3. Methodology and Expected Results

Our methodology encompasses three primary steps: PDF conversion, information retrieval, and text summarization.

- **PDF Conversion**: Convert PDF research papers into text and subsequently into embeddings. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
- **Information Retrieval**: Implement a cosine similarity search to retrieve sentences or paragraphs that closely match the user query.
- **Text Summarization**: Fine-tune BERT on the SCITLDR Dataset for abstractive summarization and summarize the retrieved content.
- **Tools and Libraries**: We plan to utilize TensorFlow, Scikit-learn, and various NLP libraries for text processing and model training. Additionally, we will employ matplotlib for visualizations.
- **Expected Results**: Our anticipated outcome is a system capable of efficiently retrieving and summarizing content from research papers, enhancing user accessibility and comprehension.

4. Timeline

Week 9: Data preparation and PDF to text conversion
Week 10: Embedding generation and implementation of cosine similarity search
Week 11: Implementation and training of the BERT model for abstractive summarization
Week 12: Testing, evaluation, and final adjustments
Week 13: Documentation, report writing, and final presentation preparation

5. Responsibilities

There are three member in our team, each member will have equal contribution towards the project. Krishna will focus on PDF conversion into text and generating embeds using all-MiniLM-L6-v2 and pre processing the SCITLDR dataset. Rohith will be implementing the BERT model and training on SCITLDR dataset. Avinash will be responsible for the evaluation of the model using ROUGE methodology. While these are assigned responsibilities, we will be collaborating and helping each other throughout all tasks.