Title: Text Summarization and Question-Answer generation using Gated Recurrent Units

Abstract:

Our project is a Text summarization and Question-Answer generation using Gated Recurrent Units (GRUs) which involves creating concise versions of longer documents while retaining essential information.GRUs, a type of recurrent neural network (RNN), are effective in capturing long-term dependencies in sequential data. The process employs an encoder-decoder architecture where the encoder reads and compresses the input text into a context vector, and the decoder generates the summary. Enhanced with attention mechanisms, GRUs focus on relevant parts of the text, improving summary quality. This approach offers a powerful method for generating coherent and relevant summaries from extensive texts, applicable in various natural language processing tasks.

Introduction:

Introduction to Text Summarization and Its Importance:

In the digital age, the vast amount of textual information generated daily necessitates efficient methods for content digestion. Text summarization, an essential task in the field of natural language processing (NLP), addresses this need by condensing large volumes of text into shorter, more manageable summaries without losing the core message. Effective summarization facilitates quicker information retrieval, enhances understanding, and aids in decision-making processes across various domains, including news media and academic research.

Overview of Neural Network Approaches:

Traditional methods of text summarization, such as extractive techniques, focus on selecting key sentences or phrases directly from the source text. However, these methods often fail to produce coherent and fluent summaries. The advent of neural network-based approaches, particularly those utilising recurrent neural networks (RNNs), has revolutionised the field by enabling the generation of more fluent and contextually relevant summaries. Among these, Gated Recurrent Units (GRUs) have gained prominence due to their capability to capture long-term dependencies and mitigate the vanishing gradient problem commonly associated with traditional RNNs.

Gated Recurrent Units (GRUs):

GRUs are a variant of RNNs designed to address the limitations of traditional RNNs. They introduce gating mechanisms that regulate the flow of information within the network, making them particularly effective for sequential data processing. GRUs have been successfully applied in various NLP tasks, including machine translation, text generation, and sentiment

analysis. Their relatively simple architecture, combined with robust performance, makes them an ideal choice for developing a text summarization model.

Problem Statement:

This project aims to develop a text summarization and question answer generation model using application of recurrent neural networks that is Gated recurrent units(GRU).

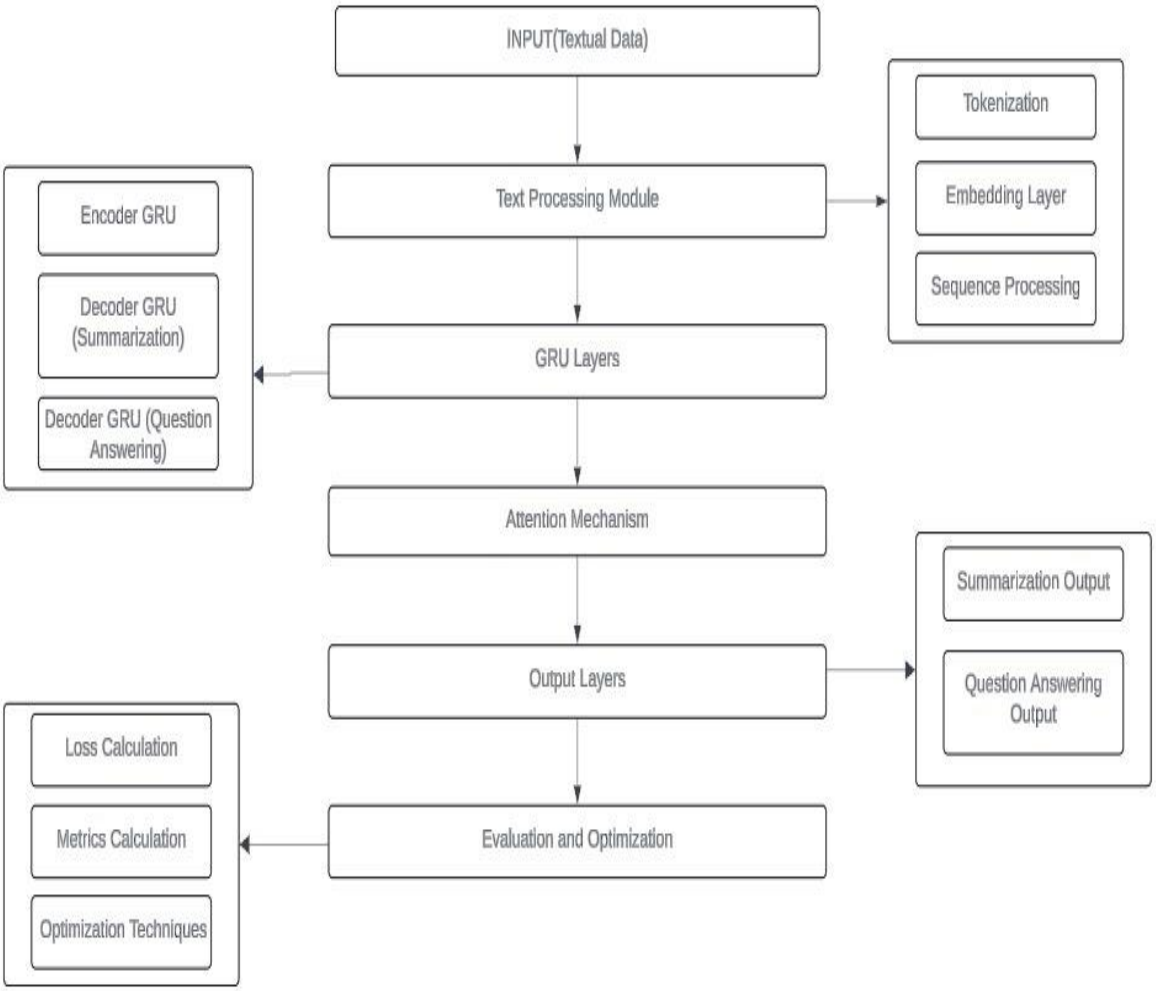
The primary research question is: How can a GRU-based model be effectively utilised to generate accurate and contextually relevant summaries and answers from large text?

This project addresses the dual challenge of developing automated systems for text summarization and question generation. The goal is to create a model that can produce concise, meaningful summaries and generate contextually relevant questions from extensive text. By leveraging Gated Recurrent Units (GRUs) within an encoder-decoder framework, this project aims to overcome the limitations of existing methods.

Objectives:

1. **Design and Implement a GRU-Based Neural Network Architecture:**
    - Develop a robust architecture using Gated Recurrent Units (GRUs) optimized for both text summarization and question-answer generation tasks.
2. **Data Preprocessing and Preparation:**
    - Clean, tokenize, and preprocess large-scale text datasets suitable for training the GRU model, ensuring data quality and relevance.
3. **Training and Evaluation of the GRU Model:**
    - Implementation of effective training strategies to optimize model performance for both text summarization (generating concise and coherent summaries) and question-answer generation (accurately answering user queries).
4. **Integration of Advanced NLP Techniques:**
    - Incorporate advanced techniques such as attention mechanisms and transfer learning to enhance the GRU model's ability to capture complex relationships and improve task-specific performance.
5. **Performance Evaluation and Comparison:**
    - Evaluate the developed GRU model against baseline methods and state-of-the-art models using standard evaluation metrics for text summarization and question answering.
6. **Optimization and Fine-Tuning:**
    - Fine-tune hyperparameters and optimize the GRU model's architecture to achieve optimal performance on diverse datasets and across varied tasks.

Architecture Diagram:



Modules Description:

Input Module:

- Responsible for handling input data, which could be textual documents or paragraphs for text summarization, and questions paired with context paragraphs for question answering.

Text Processing Module:

- **Tokenization:** Converts raw text into tokens (words or subwords) for further process.
- **Embedding Layer:** Maps tokens to dense vectors (word embeddings) that capture semantic meanings and relationships.
- **Sequence Processing:** Prepares tokenized sequences for input into the GRU layers.

Gated Recurrent Unit (GRU) Layers:

- **Encoder GRU:** Processes input sequences to capture contextual information and generate an encoded representation.
- **Decoder GRU (Summarization):** Utilises the encoded representation to generate a concise summary of the input text.
- **Decoder GRU (Question Answering):** Uses the encoded representation along with the question to generate an answer.

Attention Mechanism :

- **Attention Layer:** Enhances the model's ability to focus on relevant parts of the input during decoding for both summarization and question answering tasks.

Output Module:

- **Summarization Output:** Produces a summary of the input text based on the decoder's output.
- **Question Answering Output:** Generates an answer to the given question based on the decoder's output.

Evaluation Module:

- **Loss Calculation:** Computes the loss between predicted and actual outputs during training.
- **Metrics Calculation:** Evaluates model performance using metrics such as ROUGE for summarization and accuracy/F1 score for question answering.

Optimization Module:

- **Gradient Descent:** Updates model parameters to minimize the loss function.
- **Learning Rate Adjustment:** Optimises the learning process for better convergence.

Expected Output:

The expected outcome of the text summarization and question generation system is as follows:

Summarization Output:

The system should produce concise summaries that capture the key information and main ideas of the input text. These summaries should be coherent, contextually accurate, and free from irrelevant details or redundancies. The expected outcome includes summaries that are easily understandable and provide a comprehensive overview of the original text.

Question Generation Output:

Additionally, the system should generate contextually relevant questions based on the input text. These questions should cover various aspects of the text, probing different topics and points of interest. The questions should be meaningful, well-formed, and demonstrate a deep understanding of the text content. The expected outcome includes questions that prompt critical thinking and facilitate further exploration and analysis of the text.

References:

1.J. S. Ayyoubzadeh, M. A. Amolik, and A. Torabi, "Improving fake news detection on social media by adding network-level features," *J. Trust Manage.*, vol. 8, no. 1, pp. 1-15, May 2021. [Online]. Available: https://telrp.springeropen.com/articles/10.1186/s41039-021-00151-1

2.H. Lin, Y. Lin, L. Wu, and Z. Xiong, "Fact checkers' credibility assessment on social media: A triage model," *Int. J. Human-Comput. Interact.*, vol. 38, no. 6, pp. 701-712, Feb. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1110016822000138

3.M. Al-Ayyoub, A. Zobi, and O. B. H. Alhussein, "Detecting and categorizing rumors in microblogging platforms using deep learning techniques," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 5, pp. 524-531, Apr. 2021. [Online]. Available: https://link.springer.com/article/10.1007/s13748-023-00295-9

4.S. Mehmood, U. Qasim, and H. A. Khattak, "Fake news detection using machine learning: A systematic review," *Open J. Artif. Intell.*, vol. 5, no. 1, pp. 13-34, Jan. 2021. [Online]. Available: https://www.scirp.org/journal/paperinformation?paperid=100954

5.Y. Yang, X. Du, F. Meng, and Z. Liu, "Fake news detection with deep learning algorithms: A comprehensive review," *J. Inf. Sci.*, vol. 47, no. 3, pp. 320-340, Mar. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1389041721000826

6.M. A. Al-Ayyoub and O. B. H. Alhussein, "Automatic question-answer pairs generation and question similarity using deep learning," in *Proc. 2020 IEEE Jordan Int. Joint Conf. Electr. Eng. Inf. Technol. (JEEIT)*, Mar. 2020, pp. 272-277. [Online]. Available: https://researcher.manipal.edu/en/publications/automatic-question-answer-pairs-generation-and-question-similarit