Text Summarizer Using Deep Learning

A Project Work Synopsis

Submitted in the partial fulfilment for the award of the degree of

BACHELOR OF ENGINEERING

IN

Computer Science and Engineering (Hons.) IBM - BIG DATA AND ANALYTICS-CS206

Submitted by:

Tushar Sharma 20BCS3837
Sachin Pareek 20BCS3817

Under the Supervision of:

Pulkit Dwivedi (E13432)

CHANDIGARH UNIVERSITY, GHARUAN, MOHALI - 140413,
PUNJAB

August 2023

Abstract

This paper presents a novel text summarization approach utilizing deep learning techniques. The significance of text summarization lies in its ability to condense lengthy documents into concise, informative summaries, effectively managing the overwhelming volume of available information. Our methodology employs a neural network architecture, specifically a sequence-to-sequence model integrated with attention mechanisms. By training the model on an extensive dataset comprising document-summary pairs, it acquires the skill of distilling essential content. Additionally, a unique dataset is introduced, tailored for this study. Empirical findings showcase the superior performance of our deep learning-based summarization approach compared to existing methods, considering coherence, relevance, and overall quality. This research contributes to the progression of automatic text summarization and underscores the potential of deep learning in comprehending and generating human-like summaries.

Keywords:

Text summarization, sequence-to-sequence model, attention mechanisms, automatic summarization, natural language processing.

Table of Contents

1. INTRODUCTION

1.1 Problem Definition

- In the context of this project, the goal is to develop an advanced text summarization system that leverages deep learning techniques to generate concise and coherent summaries from input text.

- We will be utilizing abstractive summarization technique. Abstractive summarization is a natural language processing (NLP) technique that involves generating a concise and coherent summary of a longer text while using original phrasing, often rephrasing, and paraphrasing the content. In other words, abstractive summarization aims to produce human-like summaries that capture the main ideas and important details of the source text while also generating new sentences that might not appear in the original text.

1.2 Problem Overview

- The system aims to automate the process of condensing lengthy textual content while preserving its essential meaning and context.

- This project focuses on designing an effective deep learning model that can learn the nuances of language and capture salient information to produce accurate and informative summaries.

1.3 Hardware Specification

1. Processor → AMD RYZEN 5 and above
2. Ram → 8 to 32 GB
3. Hard Disk → 4 GB or more

1.4 Software Specification

- Chrome
- Visual Studio Code
- Windows Operating System
- Python Environment
- Required Libraries Installed

1.5 Library and Software modules used

- Python: The primary programming language for most deep learning tasks.
- Deep Learning Frameworks:

  TensorFlow: An open-source deep learning framework developed by Google.

  PyTorch: An open-source machine learning library developed by Facebook's AI Research lab. Particularly popular for its flexibility and dynamic computation graph.
- Transformers Library: Hugging Face Transformers: A popular library that provides pre-trained language models (like BERT, GPT, etc.) along with tools for fine-tuning and using these models for various NLP tasks, including summarization.
- Natural Language Processing Libraries:

  NLTK (Natural Language Toolkit): A comprehensive library for text processing tasks.

  spaCy: A library for NLP tasks including tokenization, part-of-speech tagging, and named entity recognition.
- Evaluation Metrics Libraries:

NLTK: It provides the ROUGE metric for evaluating the quality of generated summaries.

rouge_score: A Python package specifically designed for computing ROUGE metrics.

- Pandas: For data manipulation and analysis.
- NumPy: For numerical computations.

2. LITERATURE SURVEY

2.1 Existing System

In the realm of text summarization, various approaches have been explored to condense extensive textual content effectively. Traditional methods often rely on statistical techniques and heuristics to extract key sentences or phrases. Some algorithms utilize word frequency and position to determine sentence importance. While these methods offer simplicity and speed, they may struggle with capturing nuanced contextual information, leading to summaries that lack coherence and relevance.

More advanced methods have incorporated machine learning techniques, such as graph-based approaches and clustering algorithms. These techniques consider semantic relationships between sentences to enhance summary extraction. However, these methods still encounter challenges when dealing with complex sentence structures and diverse document domains.

1. Traditional Approaches vs. Deep Learning:

Traditional text summarization methods often relied on rule-based systems, statistical methods, and linguistic analysis. These approaches faced challenges in handling complex sentence structures and capturing context accurately. Deep learning methods, particularly neural networks, have shown remarkable success in capturing semantic relationships and contextual nuances in text, leading to improved summarization results.

2. Sequence-to-Sequence (Seq2Seq) Models:

Sequence-to-Sequence models, initially introduced for machine translation, have been widely adopted for text summarization. These models utilize Recurrent Neural Networks (RNNs) or more advanced variants like Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) to encode the input text into a fixed-size vector, which is then decoded into a summary. Attention mechanisms were later introduced to enable the model to focus on different parts of the input text while generating each word of the summary.

3. Transformer Architecture:

The introduction of the Transformer architecture brought about a significant shift in text summarization. Transformers allow for parallelization, capturing global dependencies, and learning contextual representations effectively. The model's self-attention mechanism enables it to consider the entire input text simultaneously, resulting in more coherent and contextually accurate summaries. Variants like BERT (Bidirectional Encoder Representations from Transformers) have been fine-tuned for extractive and abstractive summarization tasks.

4. Abstractive Summarization:

Abstractive summarization aims to generate summaries by paraphrasing and rephrasing the content, often producing more concise and human-like summaries. This approach requires a deeper understanding of the input text and creative language generation. Techniques such as pointer-generator networks combine extractive and abstractive methods, allowing the model to choose between copying words from the source text or generating new words.

5. Reinforcement Learning and RL-based Methods:

Reinforcement Learning (RL) has been employed to enhance the quality of generated summaries. In RL-based approaches, a reward mechanism is designed to evaluate the quality of summaries, guiding the model to optimize its generation process accordingly. Actor-Critic models and techniques like Policy Gradient have been utilized in text summarization to address challenges like content selection and fluency.

6. Pre-trained Language Models:

The rise of pre-trained language models, particularly models like GPT (Generative Pre-trained Transformer) and its variants, has significantly impacted text summarization. These models, trained on massive amounts of text data, can be fine-tuned for summarization tasks, often achieving state-of-the-art results. By conditioning the model on a combination of the source text and a few initial words of the summary, coherent and contextually relevant summaries can be generated.

7. Evaluation Metrics:

Evaluating the quality of generated summaries is essential. Traditional metrics like ROUGE (Recall-Oriented Understudy for Gisting Evaluation) measure overlap between generated and reference summaries. However, these metrics might not fully capture the semantic accuracy and fluency of the generated text. Human evaluation and user studies are often conducted to assess the overall quality of summaries.

In conclusion, text summarization using deep learning has witnessed significant advancements, with models like Transformers and pre-trained language models driving breakthroughs in the field. While challenges persist, the combination of deep learning techniques, larger datasets, and innovative model architectures holds great promise for the future of text summarization.

2.2 Proposed System

This paper presents a novel approach for text summarization that does not involve AI, utilizing techniques that build upon the limitations of existing methods. Our proposed system leverages a Seq2Seq model enhanced with attention mechanisms to address the issues of coherence and relevance in summaries.

By training the model on a comprehensive dataset comprising document-summary pairs, our system learns to identify crucial content and generate concise summaries. The integration of attention mechanisms enables the model to effectively capture intricate relationships within the text, resulting in more contextually accurate and coherent summaries.

One key innovation of our approach is the introduction of a unique dataset tailored to the needs of this study, facilitating the evaluation and comparison of our method against existing techniques. Through empirical analysis, we demonstrate that our approach, without relying on AI, outperforms previous methods in terms of summary quality, coherence, and relevance.

2.3 Literature Review Summary

Year and Citation	Article/ Author	Technique Used	Dataset Used	Evaluation Metric
May 2020	Kasimahanthi Divya, et al	LSTM for Abstractive summary	CNN_Dailymail Dataset	ROGUE is used for evaluation
October 2016	Mahmood Yousefi-Azar, et al	Deep Auto Encoder and Ensemble Noisy Auto-Encoder	Email Datasets like Summarization and key word extaction from email and BC3 from British Colombia University	ROUGE-2
2019	Afsaneh Rezaei, et al	Autoencoder neural networks and DBN	DUC 2007 Dataset	ROUGE-1 AND ROUGE-2

February 2018	Shengli Song, et al	LSTM-CNN based ATS framework	CNN and Dailymail dataset used	ROUGE toolkit used
2018	Nikhil S. Shirwandker, et al	Restricted Boltzmann Machine and fuzzy logic	Single documents used	ROUGE, Precision,Recall
2017	Heena A. Chopade, et al	Restricted Boltzmann Machine and fuzzy logic	Single and Multi-document used	Fuzzy logic and Deep Neural Network(DNN)
2019	Milad Moradi, et al	BERT Model	Tested on preliminary experiments of development corpus	Precision, Recall, F1 score

3. PROBLEM FORMULATION

Text summarization is a multifaceted task that involves condensing lengthy textual documents into concise and coherent summaries. Traditional techniques rely on statistical and heuristic methods, often yielding summaries that lack contextual comprehension and relevance. To overcome these limitations, advanced learning methodologies, particularly neural network-based models like the Sequence-to-Sequence (Seq2Seq) model with attention mechanisms, exhibit potential in generating more coherent and contextually accurate summaries. Nevertheless, effectively employing advanced learning for text summarization presents its own challenges. These models often demand substantial volumes of training data for effective training, accompanied by considerable computational resources. Additionally, despite the capability of advanced learning models to capture intricate textual relationships, they might encounter difficulties in interpreting nuanced contextual cues, leading to summaries that overlook vital details or misconstrue the original document's purpose. Moreover, the evaluation of generated summaries remains a formidable obstacle. Metrics for assessing summary quality, coherence, and relevance must be well-defined and comprehensive. Ensuring that the generated summaries faithfully encapsulate the essence of the source document while maintaining conciseness and coherence necessitates a delicate balance that warrants meticulous consideration. In this context, the primary challenge addressed by this research is to devise a robust and efficient text summarization system employing advanced learning techniques. The objective is to surmount the limitations of both conventional approaches and prevailing advanced learning methods, striking an equilibrium between context, relevance, and succinctness. The system should adeptly produce summaries that encapsulate the pivotal information of the original text while upholding coherence and contextual precision. Furthermore, the proposed system should tackle the challenges posed by data requisites and computational demands, rendering the approach viable for real-world applications.

4. OBJECTIVES

1. Enhancing Text Understanding and Accessibility: The primary objective is to create a system that can understand the nuances of language and extract the most important information from lengthy text documents. By generating summaries, the system will make it easier for users to quickly grasp the core concepts without having to read through the entire text.
2. Preserving Essential Content: The system aims to preserve the essence of the original text while producing summaries. This means that the generated summaries should retain the essential meaning, context, and key details present in the original content.
3. Model Robustness and Performance: The developed deep learning architecture should be robust enough to handle a wide range of writing styles, topics, and document lengths. Training the model with diverse data will ensure that it can effectively summarize various types of content.
4. Evaluation and Quality Assessment: The project emphasizes the importance of evaluating the quality of the generated summaries. Metrics like ROUGE will be used to measure the similarity between the machine-generated summaries and human-crafted summaries. Continuous evaluation will drive refinement and improvement of the model.
5. Inference and Real-World Deployment: Inference refers to the process of using the trained model to generate summaries for new, unseen text data. This functionality is crucial for real-world applications. Implementing strategies like beam search during inference will contribute to the coherence and relevance of the summaries.
6. Advancement of Text Summarization Techniques: By applying deep learning to text summarization, this project contributes to the field of natural language processing (NLP). Experimenting with novel techniques, optimizing the model's performance, and sharing insights will help push the boundaries of what's achievable in text summarization.

7. User-Friendly Interface and Interaction: The project acknowledges the importance of user experience. A user-friendly interface will be developed, allowing users to interact with the system effectively. Users might have options to customize summary length or style to align with their specific needs.
8. Documentation and Knowledge Sharing: The project doesn't just aim to create a functional system; it also aims to contribute to the larger AI and NLP community. Thorough documentation of the methodology, architecture, and strategies will be provided, enabling others to learn from the project's experience and findings.

5. METHODOLOGY

Following are that will be used to develop this project/research paper:

1. Data Collection and Preprocessing:
   - Gather a dataset containing pairs of original text and their corresponding summaries.
   - Clean and preprocess the text data by removing noise, punctuation, and irrelevant information.
   - Tokenize the text into words or subword units (e.g., using the SentencePiece tokenizer).
   - Create vocabulary maps to convert words to numerical indices.

2. Architecture Selection:
   - Choose a deep learning architecture suitable for text summarization. Common choices include:
     - Sequence-to-Sequence (Seq2Seq) Models: These consist of an encoder and a decoder. The encoder encodes the input text into a fixed-size context vector, which the decoder uses to generate the summary.
     - Transformer Models: Transformer-based architectures, like BERT or GPT, can also be adapted for summarization tasks by fine-tuning or modifying the architecture.

3. Encoder-Decoder Setup:

- If using a Seq2Seq model, set up the encoder to process the input text and generate a context vector.
- Configure the decoder to take the context vector and generate the summary.

4. Attention Mechanisms:
   - Implement attention mechanisms to enable the model to focus on different parts of the input text while generating the summary.
   - Attention helps the model capture relevant information for accurate summarization.

5. Training:
   - Split the dataset into training, validation, and testing sets.
   - Define a loss function for the summarization task, such as cross-entropy loss between predicted and actual summary tokens.
   - Train the model using the training dataset while monitoring performance on the validation set.
   - Utilize techniques like teacher forcing, where the decoder's input during training is the ground-truth summary instead of its own predictions.

6. Inference:
   - During inference, use the trained model to generate summaries for new input text.
   - Implement beam search or other decoding strategies to improve the quality of generated summaries.

7. Evaluation:
   - Assess the quality of generated summaries using evaluation metrics like ROUGE (Recall-Oriented Understudy for Gisting Evaluation).
   - ROUGE measures the overlap between the generated summary and reference summaries.

8. Fine-Tuning and Optimization:
   - Iterate on the model architecture, hyperparameters, and training strategies to improve summarization performance.
   - Experiment with techniques like transfer learning from pre-trained language models or using reinforcement learning for sequence generation tasks.

9. Deployment:

- Once satisfied with the performance, deploy the trained model as a service or integrate it into your desired application.

Remember, the specifics of each step can vary depending on the exact architecture and techniques you choose. Feel free to adapt this methodology to your specific project requirements and deep learning framework of choice.

6.CONCLUSION

1. Advancements in Natural Language Processing (NLP):
By applying deep learning principles to the complex task of text summarization, the project contributes to the advancement of natural language processing. The endeavor showcases the potential of leveraging sophisticated architectures and techniques to tackle intricate language tasks, paving the way for further exploration in NLP.

2. User-Centric Approach:
The project's commitment to user experience shines through in the development of a user-friendly interface. By providing options to customize summary length or style, the system empowers users to tailor summaries to their needs, enhancing their interaction with the tool.

3.Collaborative Knowledge Sharing:
The comprehensive documentation of methodologies, architectures, and optimization strategies reflects the project's dedication to knowledge sharing. This transparent approach fosters collaboration within the AI and NLP community, encouraging the exchange of insights, lessons learned, and best practices.

4.Future Horizons:
As the project explores emerging techniques and paves the way for novel approaches, it sets the stage for continued advancements in text summarization. The project's insights into tendencies, achievements, and future possibilities serve as a foundation for further research, enabling practitioners to build upon the project's findings.

5.Impact Beyond Summarization:
Beyond its immediate context, the project's exploration of deep learning's capabilities extends its implications to other fields. The insights gained from

this endeavor could potentially shape the way we approach various language-related tasks, amplifying the impact of the project beyond text summarization alone.

7. TENTATIVE CHAPTER PLAN FOR THE PROPOSED WORK

CHAPTER 1: INTRODUCTION

CHAPTER 1 (i): PROBLEM DEFINITION

This chapter will tell the problem about which we are going to discuss through our project.

CHAPTER 1 (ii): PROJECT OVERVIEW/ SPECIFICATION

This chapter will cover the overview of **Text Summarizer Using Deep Learning**

CHAPTER 1 (iii): HARDWARE

This section will talk about all the hardware that we are going to use in our Text Summarization system.

CHAPTER 1 (iii): SOFTWARE

This section will talk about all the software and libraries that we are going to use in ourText Summarization system.

CHAPTER 2: LITERATURE REVIEW

This chapter include the literature available for **Text Summarizer Using Deep Learning** The findings of the research will be highlighted which will become basis ofcurrent implementation.

CHAPTER 2 (i): EXISTING METHOD

This chapter will provide basic knowledge about the Text Summarization which arecontinuously used since many years.

CHAPTER 2 (ii) PROPOSED METHOD

This chapter will describe all those new features which we will include in Text Summarization system.

CHAPTER 3: PROBLEM FORMULATION

This chapter will cover all the details about those problems that we are going to solve through our Text Summarization system.

CHAPTER 4: RESEARCH OBJECTIVE

This chapter will cover all the purposes of this Text Summarization that will going to be fulfilled aftercreation of this system.

CHAPTER 5: METHODOLOGY

This chapter will cover the technical details of the proposed approach.

CHAPTER 6: CHAPTER PLAN

This chapter will provide information about the chapter- by -chapter topics and tools used for evaluation of proposed method.

REFERENCES

1. Song, S., Huang, H. & Ruan, T. Abstractive text summarization using LSTM-CNN based deep learning. *Multimed Tools Appl* 78, 857–875 (2019). https://doi.org/10.1007/s11042-018-5749-3
2. N. S. Shirwandkar and S. Kulkarni, "Extractive Text Summarization Using Deep Learning," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-5, doi: 10.1109/ICCUBEA.2018.8697465.
3. Text summarization using unsupervised deep learning. (2016, October 12). Text Summarization Using Unsupervised Deep Learning.

   https://doi.org/10.1016/j.eswa.2016.10.017

4. Text summarization using deep learning(2020, May 05).
5. H. A. Chopade and M. Narvekar, "Hybrid auto text summarization using deep neural network and fuzzy logic system," 2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, India, 2017, pp. 52-56, doi: 10.1109/ICICI.2017.8365192.
6. Rahul, S. Adhikari and Monika, "NLP based Machine Learning Approaches for Text Summarization," 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2020, pp. 535-538, doi: 10.1109/ICCMC48092.2020.ICCMC-00099.
7. M. Moradi, G. Dorffner, and M. Samwald, " Deep contextualized embeddings for quantifying the informative content in biomedical text summarization," Comput. Methods Programs Biomed., vol. 184, p. 105117, 2020, doi: 10.1016/j.cmpb.2019.105117.