Optimizing Scientific Paper Summarization with Fine-Tuned T5 on the ArXiv Dataset

Muhammad Ibtisam Afzal
FA22-BCS-073
13 Oct 2024

Abstract

In today's fast-paced scientific environment, the rapid growth of research publications has created a demand for efficient paper summarization tools. The motivation for this study stems from the need to help researchers quickly comprehend large volumes of scientific literature. A key challenge lies in developing summarization models that maintain both coherence and accuracy, despite the complexity of the source material. Previous approaches, including various transformer-based models like T5, have demonstrated potential, but they often struggle with domain-specific nuances and scalability.

In this work, we fine-tune the T5-small model on the ArXiv-summarization dataset to optimize scientific paper summarization. Our contributions include improving the model’s ability to recognize critical concepts and structure within scientific texts, enhancing the summarization quality while maintaining computational efficiency. We also employ the ROUGE score for a rigorous evaluation of model performance, highlighting areas for further refinement.

The significance of this research lies in its potential to substantially reduce the time researchers spend reviewing literature, fostering more efficient knowledge dissemination. By advancing summarization capabilities, our model can contribute to the development of scalable, automated tools that streamline the research process across various domains.

Keywords

Scientific paper summarization, Fine-tuned T5 model, Automated summarization, ArXiv Summarization dataset, ROUGE score evaluation, Research efficiency, Scientific Paper Summarization, Abstractive Summarization.

Introduction

The exponential growth of scientific research publications has created a significant information overload for researchers, making it challenging to keep up with the latest findings across various fields. Automatic text summarization, first introduced by [1] through his frequency-based approach, aimed to address this issue by condensing large texts into concise summaries. Since then, numerous advancements have been made, such as the introduction of vector space models by [2], which provided a robust framework for information retrieval and laid the foundation for statistical summarization techniques. Further developments by [3] in the 1990s introduced statistical methods to improve summary accuracy, and researchers like [4, 5] expanded the field by exploring machine learning and graph-based methods for summarization.

Despite these advancements, existing models still face challenges when applied to scientific texts, which are often dense, technical, and domain specific. The need for improved tools that can efficiently summarize such papers remains urgent as the number of publications continues to grow.

While many text summarization models have been developed, transformer-based models like the Text-to-Text Transfer Transformer (T5) have demonstrated considerable promise. However, these models often struggle with the specialized language and complex structure of scientific papers.

By building upon earlier works, such as the vector-based models of [2] and the statistical approaches introduced by [3], this study fine-tunes the T5-small model specifically for scientific literature. The fine-tuning process, conducted on the ccdv/arxiv-summarization dataset, enables the model to better capture scientific concepts and generate concise, coherent summaries. Our research aims to overcome these challenges, making it easier for professionals and researchers to stay informed.

1. Luhn, H.P., *The automatic creation of literature abstracts.* IBM Journal of research and development, 1958. **2**(2): p. 159-165.
2. Salton, G., *Modern information retrieval.* (No Title), 1983.

3. Kupiec, J., J. Pedersen, and F. Chen. *A trainable document summarizer*. in *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*. 1995.
4. Mani, I. and M. Maybury, *Automatic summarization John Benjamin's publishing Co*. 2001.
5. Mihalcea, R. and P. Tarau. *Textrank: Bringing order into text*. in *Proceedings of the 2004 conference on empirical methods in natural language processing*. 2004.