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

Muhammad Ibtisam Afzal

FA22-BCS-073

05 Oct 2024

**Abstract**

In today's fast-paced scientific world, the rapid increase in research publications has created a need for better paper summarization tools. The motivation for this study comes from the need to help researchers quickly understand large amounts of scientific information. A major challenge is developing summarization models that stay clear and accurate, even with the complexity of the original material. Earlier methods, including transformer-based models like T5, have shown promise but often struggle with specific subject details and handling large amounts of data.

In this work, we fine-tune the T5-small model on the ccdv/arxiv-summarization dataset to improve the summarization of scientific papers. Our contributions include enhancing the model’s ability to identify important ideas and structure within scientific texts, improving summary quality while keeping it efficient. We also use the ROUGE score to carefully evaluate how well the model performs and highlight areas for further improvement.

This research is important because it can significantly reduce the time researchers spend reading papers, making it easier to share and spread knowledge. By improving summarization abilities, our model can help create scalable, automated tools that make the research process faster and more efficient in different fields.

**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 rapid increase in scientific research publications has resulted in an overwhelming amount of information for researchers, making it difficult to stay updated with the latest discoveries across various disciplines. Automatic text summarization, first introduced by [1] in 1958 using a frequency-based method, sought to solve this problem by simplifying lengthy documents into shorter summaries. Since then, various improvements have been made, including the development of vector space models by [2], which created a strong system for retrieving information and formed the basis for statistical summarization methods. Further progress was made in the 1990s when [3] introduced statistical techniques to improve summary precision, while researchers like [4, 5] expanded the field with machine learning and graph-based approaches.

Despite these advancements, current models still face difficulties when dealing with scientific texts, which are often complex, highly technical, and specific to particular fields. As the volume of publications continues to rise, there is an increasing need for better tools that can effectively summarize such documents.

While several text summarization models have been created, transformer-based models like the Text-to-Text Transfer Transformer (T5) have shown great potential. However, these models often encounter challenges with the specialized vocabulary and detailed structure found in scientific papers.

Building on earlier works, such as [2] vector-based models and [3] statistical methods, this study fine-tunes the T5-small model specifically for scientific content. The fine-tuning, done on the ccdv/arxiv-summarization dataset, helps the model better understand scientific ideas and create clear, concise summaries. Our research addresses these issues, helping researchers and professionals stay up to date with the latest developments more easily.

**References**

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.