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

Muhammad Ibtisam Afzal

FA22-BCS-073
03 Oct, 2024

Abstract

Abstractive text summarization has gained significant attention due to its ability to generate concise summaries while maintaining the essence of the source text. Despite advancements in pre-trained transformer models, challenges remain in summarizing lengthy, technical documents like those found in scientific repositories. Existing models, while effective, often struggle with retaining factual accuracy and contextual coherence in abstractive summaries. To address these limitations, we fine-tuned the T5-small model using the ccdv/arxiv-summarization dataset, specifically focusing on enhancing the model's ability to distill complex scientific content. Our contribution lies in fine-tuning the T5-small model for generating high-quality, domain-specific abstractive summaries and deploying the model on Hugging Face to offer real-time summarization capabilities. We evaluated the model's performance using the Rouge score, demonstrating improved precision and recall over baseline models in summarizing technical documents. The successful deployment on Hugging Face provides a practical solution for researchers and industry professionals to generate accurate, meaningful summaries, potentially streamlining research workflows and enhancing information dissemination.

Keywords: Abstractive Summarization, T5-small Model, ArXiv Summarization Dataset, Rouge Score Evaluation, Transformer Models in NLP, Scientific Paper Summarization