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 a lot of attention because it can create short summaries while keeping the main ideas of the original text. Even with improvements in pre-trained transformer models, there are still difficulties in summarizing long, technical documents like those in scientific databases. Current models, though useful, often have trouble keeping the facts accurate and the summaries clear in terms of context. To overcome these issues, we fine-tuned the T5-small model using the ccdv/arxiv-summarization dataset, focusing on improving the model's ability to simplify complex scientific content. Our work involved adjusting the T5-small model to produce high-quality summaries specific to certain subjects and making them available on Hugging Face for real-time use. We used the Rouge score to measure the model's performance, showing better accuracy and recall than earlier models in summarizing technical documents. The successful launch of Hugging Face provides an efficient tool for researchers and professionals to create accurate and useful summaries, making research processes faster and improving the spread of information.

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