A Deep Dive into Transformer Models for Abstractive Summarization: A ROUGE Score Comparison

Ms. Maria Varghese
APJ Abdul Kalam Technological University, Kerala, India
mariathelappilly@gmail.com

January 08, 2025

Abstract

Abstractive summarization enables concise and meaningful summaries of long-form text by leveraging natural language generation. This project compares the performance of three state-of-the-art transformer models—BART, T5, and Pegasus—using the CNN/DailyMail dataset. ROUGE metrics are employed to evaluate the quality of the generated summaries. BART outperformed other models in overall ROUGE scores, while Pegasus demonstrated strong bigram overlap. A real-world deployment of the summarization models was achieved through interactive Gradio and Streamlit interfaces, enabling broader accessibility and usability.

1 Introduction

The growing volume of textual data necessitates efficient tools for condensing information into concise summaries. Unlike extractive summarization, which selects sentences directly from the text, abstractive summarization generates human-like summaries by rephrasing content. Recent advancements in deep learning, particularly through transformer architectures, have significantly improved the effectiveness and coherence of abstractive summarization. Models such as BART, T5, and Pegasus have transformed this field, achieving state-of-the-art performance by leveraging pretraining on extensive datasets and fine-tuning for specific tasks. This project evaluates these models for summarization and demonstrates their real-world deployment using Gradio and Streamlit interfaces.

2 Literature Review

Lewis et al. [1] introduced BART, a transformer-based model designed for sequence-to-sequence tasks such as summarization, machine translation, and text generation. BART employs a novel denoising objective during pretraining, which involves corrupting text input (e.g., through token masking, deletion, or reordering) and training the model to reconstruct the original sequence. This approach allows BART to excel at both text understanding (via its bidirectional encoder) and text generation (via its autoregressive

decoder). The paper demonstrates that BART achieves state-of-the-art performance on a variety of NLP tasks, including abstractive summarization, where its ability to generate coherent and fluent summaries is particularly highlighted.

Raffel et al. [2] proposed T5 (Text-to-Text Transfer Transformer), a model that treats all NLP tasks as a unified text-to-text problem. This framework simplifies task-specific modifications, as both the input and output are formatted as text sequences. By pretraining on the massive C4 dataset with a span-corruption objective, T5 demonstrates remarkable generalization capabilities across a wide range of tasks, including summarization, translation, and classification. The study reveals that T5's effectiveness stems from its flexibility and scalability, making it suitable for a diverse array of NLP challenges. However, the model's generalized nature may require additional fine-tuning for optimal performance on specific tasks like abstractive summarization.

Zhang et al. [3] introduced PEGASUS, a transformer model explicitly tailored for abstractive summarization. The unique pretraining objective, known as gap-sentence generation (GSG), involves masking entire sentences in a document and training the model to predict these sentences based on the remaining context. This task closely mimics the summarization process, making PEGASUS highly effective in generating concise and meaningful summaries. The authors evaluated PEGASUS on multiple summarization benchmarks, where it achieved state-of-the-art results, particularly excelling in capturing salient points and maintaining coherence in generated summaries. The pretraining strategy significantly reduces the need for extensive fine-tuning, demonstrating the potential of task-specific pretraining for NLP applications.

3 Methodology

3.1 Dataset

The **CNN/DailyMail dataset** consists of:

- **Articles:** News articles covering diverse topics, ranging from technology to politics.
- **Summaries:** Human-written highlights summarizing the main points of each article.

The dataset is preprocessed to tokenize the text and truncate inputs to the models' maximum token limits.

3.2 Models Evaluated

The following transformer-based models were used:

1. **BART (facebook/bart-large-cnn):** Pre-trained for summarization and fine-tuned on the CNN/DailyMail dataset.
2. **T5 (t5-large):** A general-purpose text-to-text model requiring specific fine-tuning for summarization tasks.
3. **Pegasus (google/pegasus-xsum):** Optimized for abstractive summarization, trained with gap-sentence compression tasks.

3.3 Experimental Setup

- **Preprocessing:** Articles were tokenized using each model's tokenizer. Input lengths were truncated to a maximum of **1024 tokens**.
- **Model Configuration:** Beam size: 4 (to explore multiple decoding paths for better summaries). Output length: Restricted to a maximum of 128 tokens.
- **Evaluation Metrics:** ROUGE-1, ROUGE-2, and ROUGE-L were computed to evaluate unigram, bigram, and longest subsequence overlaps.
- **Tools and Frameworks:** Models and tokenizers were implemented using the Hugging Face Transformers library. Experiments were conducted on **Google Colab** with GPU acceleration for faster computation.

3.4 Deployment

Gradio:

- A Gradio web interface was developed to enable users to input custom text and generate summaries in real time using any of the three models.

Streamlit:

- A Streamlit application was created to visualize ROUGE scores and allow users to interact with the summarization models.

4 Results and Analysis

4.1 ROUGE Scores

Model	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
BART	0.535	0.377	0.479	0.479
T5	0.284	0.092	0.209	0.235
Pegasus	0.368	0.349	0.368	0.368

Table 1: ROUGE Scores of Different Models

4.2 Analysis

- **BART:** Achieved the highest overall ROUGE scores, demonstrating its ability to generate accurate and coherent summaries.
- **Pegasus:** Performed well in ROUGE-2, indicating its strength in capturing bigram overlaps and abstract patterns.
- **T5:** Delivered significantly lower ROUGE scores, highlighting the need for additional fine-tuning to optimize performance for summarization tasks.

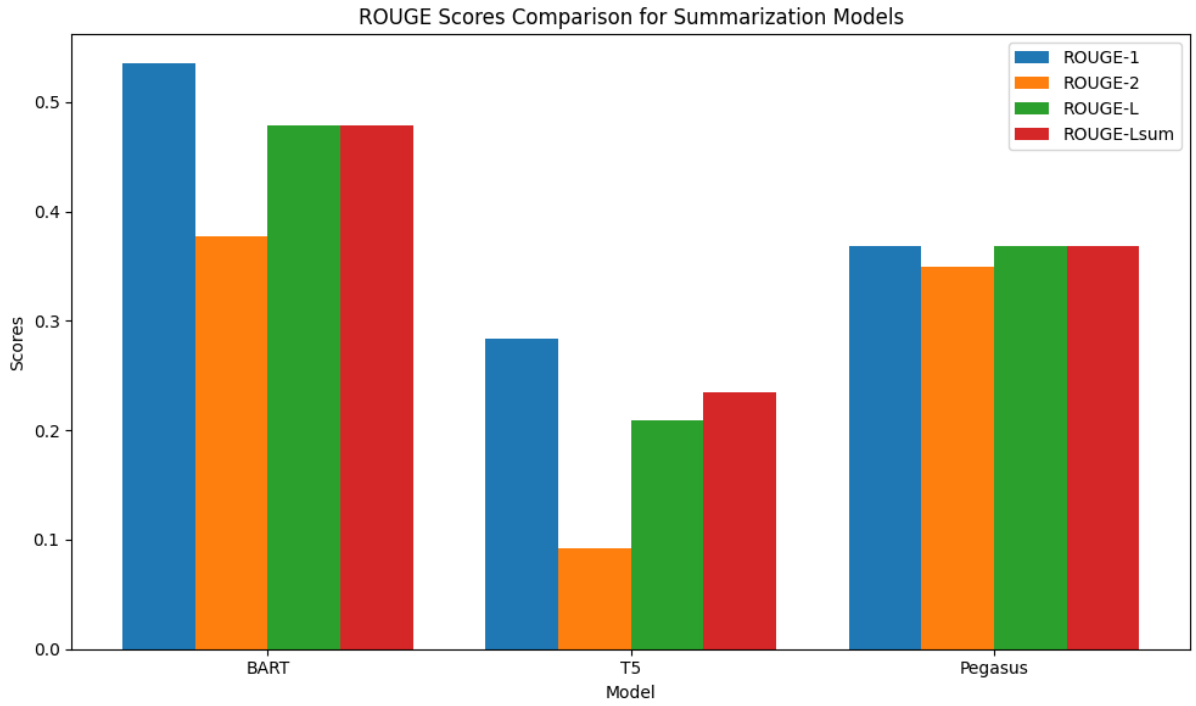

Figure 1: Graphical Representation of Results

5 Deployment

5.1 Gradio Implementation

An interactive **Gradio interface** was developed, allowing users to:

1. Input custom articles for summarization.
2. Select the desired summarization model (BART, T5, Pegasus).
3. Generate and view model-generated summaries in real time.

Gradio provided an easy-to-use interface, making the tool accessible to both technical and non-technical users.

5.2 Streamlit Implementation

A **Streamlit dashboard** was created to:

1. Visualize ROUGE scores for all three models.
2. Allow users to select models and compare their outputs interactively.
3. Provide a clean and user-friendly summarization platform.

Streamlit's flexibility enabled seamless integration of visualizations and interactivity.

6 Discussion

6.1 Key Findings

- **BART** is the most effective model for abstractive summarization, achieving the highest ROUGE scores across all metrics.
- **Pegasus** is a strong alternative for tasks requiring more abstract summaries.
- **T5**, while versatile, requires task-specific fine-tuning to perform well in summarization.

6.2 Deployment Benefits

The use of Gradio and Streamlit demonstrated the practical applicability of abstractive summarization models, enabling easy access and usability.

6.3 Challenges and Limitations

The results of this study are based on the CNN/DailyMail dataset, which may not generalize to other domains, such as scientific or medical texts. The models' performance might vary with different types of content, and further evaluation across diverse datasets is necessary. Additionally, ROUGE metrics, while widely used, primarily measure surface-level word overlap and may not fully capture the semantic quality or coherence of the generated summaries. Incorporating human evaluation or alternative metrics could provide a more comprehensive assessment. Lastly, the computational demands of large transformer models like BART, T5, and Pegasus can limit accessibility for researchers with limited hardware, highlighting the need for more efficient models or optimization techniques.

7 Future Scope

Future research could explore the models' performance across a wider range of datasets, especially from different domains, to assess their adaptability. Integrating human feedback into model training could improve the alignment with human expectations, enhancing summary quality. Additionally, addressing the computational challenges through techniques like model pruning or distillation would make these models more accessible and efficient for broader use.

8 Conclusion

This study evaluated the performance of three transformer-based summarization models—BART, T5, and Pegasus—using the CNN/DailyMail dataset. BART emerged as the best-performing model, achieving the highest ROUGE Pegasus offered competitive performance, while T5 required further fine-tuning for optimal results. The deployment of the summarization tool via Gradio and Streamlit showcased its real-world applicability, making abstractive summarization more accessible to users.

References

[1] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Zettlemoyer, L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, 7871–7880.

[2] Raffel, C., Shinn, E., Roberts, A., Lee, S., Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020).*

[3] Zhang, J., Zhao, Y., Saleh, M. (2020). PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, 11338–11349.