The Voice of Creative Research

Vol. 7 & Issue 2 (April 2025)

OPEN ACCESS PEER REVIEW

https://doi.org/10.53032/tvcr/2025.v7n2.43

Research Article

Comparative Analysis of Transformer-Based Models for Patent Summarization

Muhammad Ibrahim Khan
Student,
Department of Data Science,
Mumbai University, Kirti College, India
ibrahimincognito@gmail.com

Dr. Sampada Margaj
Assistant Professor,
Department of Computer Science,
Mumbai University, Kirti College, India
sampada.malhar@gmail.com

Abstract:

Patent documents contain critical technical and legal information but are often lengthy and complex, making it difficult for researchers and businesses to extract key information efficiently. This research paper compares the performance of four transformer-based models - PEGASUS, BART, LED, and BigBirdPegasus for abstractive summarization of patents using the BIGPATENT dataset (subset "d" - Textiles; Paper patents), explores the architectural differences among these models, discusses the training strategies used and examines their implications for improving automated patent summarization. Each model was fine-tuned under identical training conditions, using 10% of the training dataset as per the hardware constraints and their performance was evaluated using ROUGE metrics. The results provide insights into which model is best suited for summarizing patent documents efficiently.

Keyword: Patent Summarization, Abstractive Summarization, Transformer Models, BigBirdPegasus, PEGASUS, BART, Longformer (LED), Hugging Face Transformers, ROUGE Evaluation Metrics, Fine-Tuning Pretrained Models

1. Introduction

Patents documents are important for keeping intellectual property safe and encouraging new ideas. However, academics, patent examiners, and legal experts find it challenging to efficiently extract relevant information from long patents that are filled with technical and legal jargon. In order to improve accessibility and comprehension, patent summary provides a means of simplifying these papers while maintaining their key information. Patent summarization can be broadly categorized into two approaches [9]:

- **Extractive Patent Summarization:** Creates a summary by choosing key phrases or words verbatim from the patent. Despite its effectiveness, it frequently lacks consistency and might not accurately convey the document's message.
- **Abstractive Patent Summarization:** Generates new textual content that communicates the main ideas of the patent document in a more natural, human-like manner. This approach is more challenging but offers better readability and contextual knowledge.

This study employs recent advances in transformer-based models to investigate abstractive patent summarization. In general text summarizing tasks, models like PEGASUS, BART, BigBirdPegasus, and LED have shown excellent performance [1],[2],[3],[4]. However, research is still being done to determine how well they can manage long, domain-specific documents like patents. In this study, subject to the hardware constraints - System RAM: 12.7GB, GPU RAM: 15GB and Storage: 112.6GB, each model is fine-tuned using only 10% of the training dataset, under identical conditions to ensure a fair comparison and to identify the most effective model for abstractive patent summarization. The performance was assessed using ROUGE-1, ROUGE-2, ROUGE-L and ROUGE-Lsum metrics. This study provides a comparative evaluation of these models for abstractive patent summarization using the BIGPATENT dataset (subset "d" – Textiles; Paper patents).

2. Literature Review:

1. The Transformer model, a self-attention mechanism-based architecture first presented by Vaswani et al. (2017), transformed natural language processing (NLP) by facilitating parallelization and enhancing performance in tasks including text summarization and machine translation [6]. The groundwork for later advances in summarization models was established by this study.

2. For text summarizing, PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive summarizing) was created especially [1]. It presents a fresh pretraining idea in which crucial phrases are obscured by gap-sentence generation (GSG), requiring the model to make contextual predictions. PEGASUS is optimized for summarizing tasks and has been pre-trained on big corpora such as CNN/DailyMail and arXiv.

3. A denoising autoencoder trained to restore damaged input text is called BART (Bidirectional and Auto-Regressive Transformer) [2]. In contrast to PEGASUS, BART is useful for sequence-to-sequence tasks like summarization since it employs both left-to-right and bidirectional encoding. Prior to being optimized on summarization datasets, it was pretrained on Wikipedia and BookCorpus.

4. BigBird [3], a sparse-attention model that tackles the shortcomings of conventional transformers in processing lengthy texts, was presented by Zaheer et al. (2020). By lowering the quadratic complexity of attention processes, BigBird has demonstrated encouraging results in summarizing challenges for lengthy texts.

5. The n-gram overlap between system-generated summaries and human reference summaries is measured using ROUGE (Recall-Oriented Understudy for Gisting assessment), an automated assessment metric for summarizing tasks introduced by Lin (2004) [7]. The paper describes several variations of ROUGE, such as ROUGE-L (longest common subsequence), ROUGE-N (unigram and bigram overlap), and ROUGE-S (skip-bigram-based evaluation). The author shows that ROUGE is a common benchmarking technique in summarization research and has a good correlation with human evaluations.

6. To improve summary variety and relevance, Valli et al. (2023) suggest a BigBirdPegasus-based method for multi-document summarizing that incorporates Maximal Marginal Relevance (MMR). Their research emphasizes BigBirdPegasus' sparse attention mechanism, which makes it ideal for patent summarizing since it allows for the quick processing of lengthy documents. By choosing the most instructive phrases and eliminating repetition, they show how MMR enhances summary coherence. The model outperforms existing transformer-based models in preserving important information while guaranteeing conciseness, achieving 80% accuracy after being trained on a 70%–30% data split.

3. Transformer-Based Models for Summarization

Hugging Face is a leading platform in natural language processing (NLP), which offers a large collection of pretrained transformer-based models via the Hugging Face Model Hub. It makes it simple for academics and developers to acquire, optimize, and implement modern models for a range of natural language processing (NLP) activities, such as question-answering, translation, and text summarization.

For this study, we utilized four pretrained transformer models sourced directly from Hugging Face’s Transformers library.

i. PEGASUS (google/pegasus-cnn_dailymail)
ii. BART (facebook/bart-large-cnn)
iii. BigBirdPegasus (google/bigbird-pegasus-large-bigpatent)
iv. LED (allenai/led-large-16384)

These models have been trained on diverse corpora and optimized for text summarization tasks, making them suitable for our patent summarization experiments.

4. Dataset

For this study, we used the BIGPATENT dataset, a large-scale dataset for patent summarization created by Sharma et al. (2019) [5]. The dataset consists of over 1.3 million U.S. patent documents across nine technology categories, each corresponding to a different Cooperative Patent Classification (CPC) category. We specifically selected subset "d" (Textiles; Paper patents) for this study. It contains 10,164 training samples, 565 validation samples and 565 test samples.

5. Preprocessing and Fine-Tuning Methodology

5.1 Preprocessing and Sampling:

Since fine-tuning large transformer models requires significant computational resources, we used only 10% of the training dataset (1016 training samples) as per the hardware constraints (System RAM: 12.7GB, GPU RAM: 15GB and Storage: 112.6GB).

- Truncation: Inputs were truncated to 1024 tokens, except for BigBirdPegasus and LED, which support longer sequences.
- Tokenization: We used standard Hugging Face tokenizers for each model.

5.2 Training Configuration:

All models were fine-tuned using the same hyperparameters to ensure fair comparison:

- Epochs = 2
- Batch Size Per Device = 1
- Gradient Accumulation Steps = 4 to simulate larger batch size.
- Mixed Precision Training (fp16=True) to reduce memory usage.
- Gradient Checkpointing (gradient_checkpointing=True) to save memory by recomputing activations.
- Evaluation Steps = 500 to perform evalution after every 500 steps.
- Save Steps = 1000 to save models every 1000 steps.

6. Evaluation & Comparative Analysis

6.1 Understanding Evaluation Metrics

We employ ROUGE (Recall-Oriented Understudy for Gisting Evaluation), a commonly used measure in summarization tasks, to evaluate the quality of produced summaries. ROUGE provides information on material retention, fluency, and structural similarity by measuring the n-gram overlap between the produced summary and the reference summary. The key ROUGE metrics used in this study are:

- **ROUGE-1**: It determines the amount of important material that is kept by measuring the overlap of unigrams (single words) between the generated and reference summaries.
- **ROUGE-2**: It evaluates the summary's coherence and fluency by measuring the overlap of bigrams, or two-word sequences.
- **ROUGE-L**: It measures the similarity in sentence structure between the produced and reference summaries by measuring the longest common subsequence (LCS).
- **ROUGE-Lsum**: A variant of ROUGE-L that takes sentence-level structure and coherence into account and is tailored for multi-sentence summaries.

Additionally, we analyze training loss and validation loss:

- **Training Loss**: It measures how well the model fits the training data. Lower loss means better learning.
- **Validation Loss**: It measures performance on unseen data. A decreasing validation loss suggests improved generalization, while a high gap between training and validation loss may indicate overfitting.

6.2 ROUGE Score Comparison:

The table below presents the final ROUGE scores of all models after fine-tuning on the BIGPATENT dataset (subset "d" - Textiles; Paper patents):

Model	**Training Time**	**Training Loss**	**Validation Loss**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**ROUGE-Lsum**
PEGASUS	19:10	2.4415	2.19097	0.4061	0.2073	0.3077	0.3093
BigBirdPegasus	29:37	1.4108	1.67343	0.5076	0.3132	0.4017	0.4024
BART	10:21	1.7661	2.05088	0.4364	0.2330	0.3271	0.3282
LED	27:16	1.8681	2.14806	0.4372	0.2389	0.3319	0.3339

6.3 Performance Analysis:

1. BigBirdPegasus achieved the highest ROUGE scores, outperforming other models across ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum. This is likely due to its ability to handle longer sequences (4096 tokens), making it better suited for patent summarization, where documents are lengthy and technical.

2. PEGASUS and BART performed reasonably well, but their 1024-token input limit likely caused information loss when summarizing long patents. PEGASUS's gap-sentence generation (GSG) strategy is powerful, but it may be less effective for long-form domain-specific texts.
3. LED (Longformer Encoder-Decoder) showed moderate performance. While it supports up to 16,384 tokens, its ROUGE scores were slightly lower than BigBirdPegasus. This suggests that LED's extended attention span alone is not enough—pretraining and fine-tuning strategies also play a significant role.
4. Training & Validation Loss Trends:
    i. BigBirdPegasus has the lowest training loss (1.41), indicating stable learning and good convergence. Its validation loss (1.67) is also the lowest among the models, suggesting better generalization.
    ii. PEGASUS shows the highest training loss (2.44) and validation loss (2.19), indicating it struggled to learn efficiently, likely due to its pretraining on news articles (CNN/DailyMail), making it less domain-adaptive to patents.
    iii. BART and LED have slightly lower training losses than PEGASUS but show higher validation loss than BigBirdPegasus, suggesting they learned patterns but did not generalize as effectively.

7. Conclusion:

BigBirdPegasus is the most effective model for patent summarization, excelling in both performance and efficiency. BART and PEGASUS are strong general-purpose models but are limited by their 1024-token constraint for long documents. LED's extended attention span alone does not guarantee superior results—pretraining and optimization strategies play a crucial role.
This study demonstrates that BigBirdPegasus is the most effective model for abstractive patent summarization, followed by LED and BART. Future work should explore longer training durations, domain adaptation, and hybrid extractive-abstractive approaches.

References

1. Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020). PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. Proceedings of the 37th International Conference on Machine Learning (ICML), 113, 11328-11339.

2. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & 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), 7871-7880.

3. Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., & Ahmed, A. (2020). Big Bird: Transformers for Longer Sequences. Advances in Neural Information Processing Systems (NeurIPS), 33, 17283-17297.

4. Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The Long-Document Transformer. arXiv preprint arXiv:2004.05150.

5. Sharma, A., Li, Y., Wang, W. Y., & Mitamura, T. (2019). BIGPATENT: A Large-Scale Dataset for Abstractive and Coherent Summarization of Patent Documents. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), 2204-2213.

6. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, vol. 30, pp. 5998–6008.

7. Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

8. Valli, P. S., Satwika, N. U., Swaroop, S. V., Prasad, V. V. L., & Lakshmi, B. V. (2023). BigBirdPegasus-based Abstractive Multi-Document Summarization.

9. Eduard Hovy and Chin-Yew Lin. 1998. Automated Text Summarization and the Summarist System. In TIPSTER TEXT PROGRAM PHASE III: Proceedings of a Workshop held at Baltimore, Maryland, October 13-15, 1998, pages 197–214, Baltimore, Maryland, USA. Association for Computational Linguistics.