Beyond Size: Harnessing the Power of Small Models for Effective Dialogue Summarization

Utkarsh Avadhut Dabholkar
Data science
Rochester institute of technology
ud9701@rit.edu

***Abstract*—This study explores automated abstractive text summarization with transformer models (Pegasus, T5, BART), focusing on the SAMSum Dialogue Dataset. It investigates whether a fine-tuned T5-small model can outperform larger models and quantifies performance enhancement using the ROUGE metric. Results indicate that the fine-tuned T5-small surpasses its larger counterparts, highlighting the potential of smaller models in NLP and emphasizing computational sustainability.**

***Index Terms*—text summarization, transformer models, NLP, SAMSum Dialogue Dataset, T5-small, fine-tuning, ROUGE metric**

I. Introduction

This In an era marked by rapidly diminishing attention spans, particularly among newer generations, the ability to quickly process and comprehend large volumes of information has become increasingly valuable. Research indicates a trend towards shorter attention spans, influenced by the digital landscape's fast-paced nature and the proliferation of short-form content. This shift poses a challenge for traditional long-form texts, which can be overwhelming or unattractive to many, especially younger audiences.

The project "Text Summarization Using the SAMSum Dialogue Dataset" addresses this challenge by exploring automated abstractive text summarization. Unlike traditional summarization methods, which are time-consuming and rely heavily on human cognitive resources, abstractive summarization uses advanced AI models to generate concise and coherent summaries. These summaries encapsulate the essence of longer texts, making them ideal for quick consumption. Moreover, recent comprehensive reviews on automatic text summarization methods highlight the evolving techniques in this field [1].

This project focuses on fine-tuning transformer models, such as Pegasus, T5, and BART, specifically for the task of summarizing dialogue-based texts. By adapting these models to the unique characteristics of conversational data, the project aims to create summaries that are not only brief but also contextually relevant and engaging. This approach is particularly pertinent in catering to the preferences of an audience that values brevity and speed in content delivery, thereby aligning with the evolving patterns of information consumption in the digital age.

II. Research Questions

The primary objective of this project is to assess and compare the performance of pre-trained transformer models, specifically Pegasus, T5, and BART, with their fine-tuned versions in abstractive text summarization. The study, centered on the SAMSum Corpus, is structured around two key research questions:

1) **T5-Small-Finetuned vs. Larger Models:** Investigating whether a fine-tuned T5-small model, with about 60 million parameters, can outperform larger models like Pegasus-Large (568 million parameters), T5-Large (770 million parameters), and BART-Large (400 million parameters). This comparison is crucial to understand if a smaller model can match or exceed the capabilities of larger models, with significant implications for efficiency and resource utilization in NLP.
2) **Performance Enhancement Post-Fine-Tuning :** Measuring the improvement in the performance of these models post fine-tuning using the ROUGE metric, a standard in summarization assessment.

The project aims to elucidate the effectiveness of fine-tuning in enhancing transformer models for tasks such as abstractive text summarization, offering insights into optimizing NLP models for both efficiency and performance.

III. Methodology

The SAMSum Corpus [2], integral to this study, comprises around 15,000 pairs of dialogues and their summaries. Reflecting real-life interactions, it includes a broad range of topics in various conversational styles, from casual to structured.

A. Dataset Characteristics

- **Dialogue and Summary Length Distribution:** Analysis shows a right-skewed distribution in both dialogue and summary lengths, indicating most dialogues and summaries are short, with a few longer outliers in Fig1&2.
- **Correlation between Dialogue and Summary Lengths:** Positive correlation observed, where longer dialogues typically have longer summaries in Fig3.

B. Preprocessing Steps

- **Data Preprocessing:** Removal of irrelevant tags (non-verbal cues, metadata) to reduce noise and enhance focus on dialogue content.

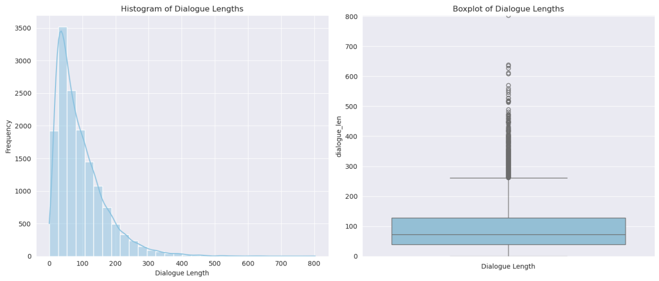

Fig. 1. Plots of Dialogue Lengths

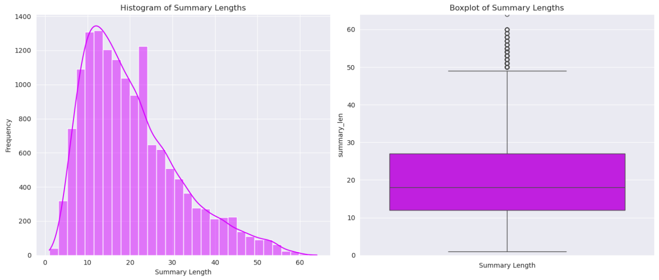

Fig. 2. Plots of summary Lengths

- **Tokenization and Segmentation:** Standard processing to split text into tokens, and segment dialogues by speakers to maintain conversational flow.
- **Ensuring Coherency:** Additional steps to handle interruptions and non-sequiturs, ensuring input data coherency for summarization models.

The dataset's meticulous preparation ensures clean, structured input, vital for effective model training in summarization tasks.

C. Model Fine-Tuning

The fine-tuning of the T5-small model was executed with GPU acceleration to optimize computational efficiency and shorten training duration.

Selection of Model:

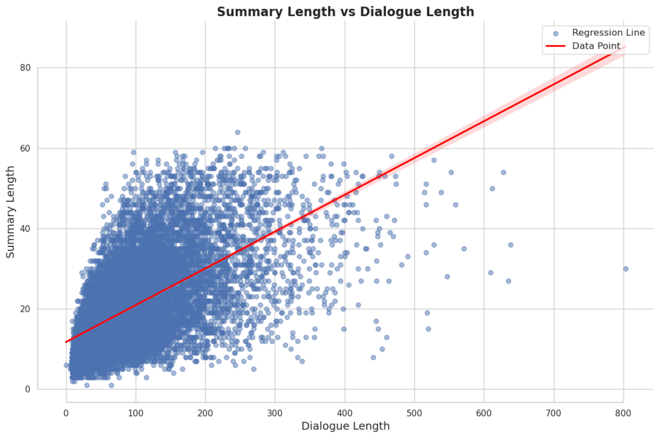

Fig. 3. Summary Length vs Dialogue Length

- Chosen for its efficiency-performance balance, the T5-small model, with roughly 60 million parameters, is ideal for fine-tuning with the SAMSum dataset.

Preprocessing for T5:

- A tailored preprocessing function formatted the dialogue data for the T5, with 'dialogue' as input and 'summary' as output. The T5 tokenizer ensured proper sequence lengths through truncation and padding.

Training Details:

- **Hyperparameters**: Fine-tuned over 4 epochs with a learning rate of 2e-5 and batch size of 2, employing gradient accumulation for effective error backpropagation.
- **Mixed Precision Training**: Utilized mixed precision (fp16) for resource-efficient training without compromising performance.
- **Optimizer**: Adopted PyTorch's AdamW for its gradient and learning rate management capabilities.

Training Infrastructure:

- Utilized the Hugging Face Trainer API for its robust fine-tuning capabilities, monitored with tensorboard for performance tracking and hyperparameter adjustments. The best model configuration was saved post-training based on evaluation metrics.

Dataset Mapping:

- The dataset was processed with a custom function to align inputs and labels correctly for the T5 model, facilitating the model's learning from the SAMSum dialogues.

This streamlined fine-tuning approach was designed to enhance the T5-small model's summarization capabilities, taking advantage of its ability to capture subtle patterns in conversational data.

D. Evaluation

The fine-tuned T5-small model's performance was evaluated using the ROUGE metric [3], with scores for ROUGE-1, ROUGE-2, and ROUGE-L in Table I, titled "Validation and Test Set Performance." reflecting the quality of content overlap and structural fluency in summaries.

TABLE I
VALIDATION AND TEST SET PERFORMANCE

Metric	Validation Set	Test Set
ROUGE-1 (%)	56.67	55.36
ROUGE-2 (%)	29.23	27.18
ROUGE-L (%)	53.06	52.10

The metrics provide a benchmark for the model's summarization capabilities and highlight areas for future enhancements. The T5-small's results are promising, indicating its potential for deployment in resource-constrained production environments, with room for further improvement.

IV. RESULTS

The study showcases a remarkable improvement in the summarization quality of the fine-tuned T5-small model, especially when compared to larger pre-trained models.

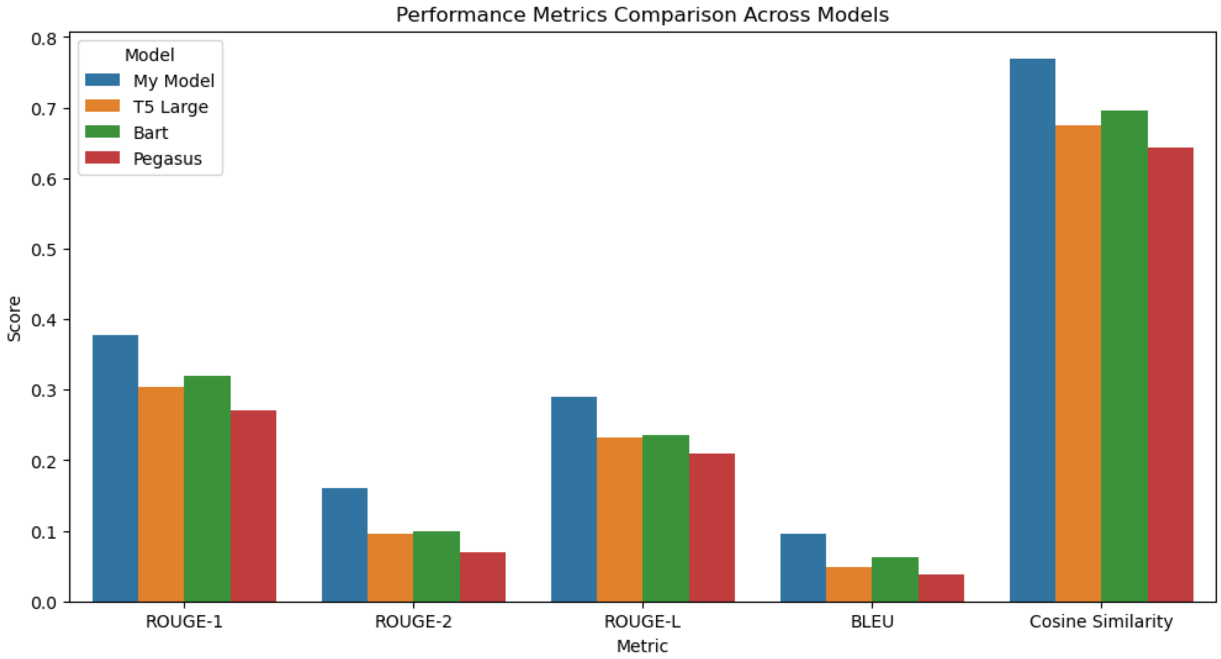

Fig. 4. Comparing all the models

TABLE II
PERFORMANCE METRICS COMPARISON ACROSS MODELS

Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	Cosine Similarity
My Model	0.3767	0.1596	0.2896	9.52	0.7698
T5 Large	0.3045	0.0960	0.2315	4.82	0.6745
Bart	0.3189	0.0989	0.2352	6.28	0.6961
Pegasus	0.2702	0.0703	0.2093	3.88	0.6432

This section can be expanded to include an experimental analysis of a subset of the dataset: The experiment involved analyzing 40 samples from the test set, leading to a comprehensive evaluation of each model across multiple metrics. The average scores for each metric were calculated, as presented in Table II, 'Performance Metrics Comparison Across Models.' These scores clearly illustrate that the fine-tuned T5-small model not only competes with but in several aspects surpasses the performance of larger models such as T5-Large, Bart, and Pegasus. This finding is particularly significant, suggesting that despite its smaller size, the T5-small model, post-fine-tuning, demonstrates enhanced efficiency and effectiveness in capturing the essence of the input text. This addresses our second research question and highlights the substantial benefits of fine-tuning in improving model performance, as evident in Table II & Fig 4.

V. CONCLUSION

The results of this study, particularly the analysis of 40 samples from the test set, highlight the impressive performance of the fine-tuned T5-small model in comparison to its larger counterparts. This is not only a significant finding in terms of model efficiency but also in the potential for smaller models to achieve high-quality text summarization.

In conclusion, this project underscores the efficacy of fine-tuning in enhancing transformer models for abstractive text summarization. The potential of fine-tuning to optimize smaller models without compromising on performance quality paves the way for computational sustainability in NLP. Future research directions include expanding the scope of fine-tuning across diverse NLP tasks, exploring its application in multilingual contexts, and integrating multimodal data to enrich the summarization process. These avenues hold promise for further advancing the field of NLP.

REFERENCES

[1] D. Yadav, J. Desai, and A. K. Yadav, "Automatic text summarization methods: A comprehensive review," *arXiv preprint arXiv:2204.01849*, 2022.

[2] B. Gliwa, I. Mochol, M. Biesek, and A. Wawer, "Samsum corpus: A human-annotated dialogue dataset for abstractive summarization," in *Proceedings of the 2nd Workshop on New Frontiers in Summarization*. Association for Computational Linguistics, 2019. [Online]. Available: http://dx.doi.org/10.18653/v1/D19-5409

[3] Y. Liu, A. R. Fabbri, J. Chen, Y. Zhao, S. Han, S. Joty, P. Liu, D. Radev, C.-S. Wu, and A. Cohan, "Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization," *arXiv preprint arXiv:2311.09184*, 2023.