**Fine-Tuning a BERT2BERT Language Model for Abstractive Text Summarization**

Name: Prudhvi Vardhan, ID: 22071880, Git hub: <https://github.com/pp23aai/Large-language-model-Assignment-2025>

**Introduction**

The rapid growth of digital information necessitates efficient methods to distil key insights from large text volumes. Abstractive text summarization, which generates concise and rephrased summaries, addresses this need. This report investigates the fine-tuning of a BERT2BERT model, an encoder-decoder framework leveraging BERT’s architecture, for abstractive summarization using the CNN/Daily Mail dataset. The study evaluates the model’s performance, detailing the training process, results, and implications for practical applications.

BERT2BERT employs two BERT models one as an encoder to process input text and another as a decoder to generate summaries. This project assesses the effectiveness of fine-tuning BERT2BERT for summarization, focusing on its training dynamics, evaluation metrics, and output quality**.**

**Literature Review**

Neural networks revolutionized summarization. Rush et al. (2015) introduced a sequence-to-sequence model with attention, enabling abstractive summarization by generating new text. The Transformer (Vaswani et al., 2017) advanced this by using self-attention to capture long-range dependencies.

BERT (Devlin et al., 2018) offers rich contextual embeddings via bidirectional attention and masked language modelling but focuses on understanding, not generation. Rothe et al. (2020) adapted it with BERT2BERT, an encoder-decoder framework for sequence generation, leveraging BERT’s strengths.

Meanwhile, BART (Lewis et al., 2019) and T5 (Raffel et al., 2019) were designed for text-to-text tasks with denoising and span corruption, often outperforming BERT2BERT. However, BERT2BERT’s simplicity and compatibility with limited resources make it valuable. Lin (2004) emphasizes ROUGE as a key metric for summarization evaluation.

**Methodology**

The BERT2BERT model was implemented using the Hugging Face transformers and datasets libraries, with both encoder and decoder initialized from the Bert-base-uncased checkpoint, which contains 12 layers, 768 hidden units, and 110 million parameters. The CNN/Daily Mail dataset, a standard benchmark for summarization, provided the training data, consisting of news articles paired with human-written summaries. Due to computational constraints, a subset of the dataset was used: 10,000 articles for training, 1,000 for validation, and 1,000 for testing.

**Data Preprocessing**

Input articles were tokenized using the BERT tokenizer, truncated to a maximum length of 512 tokens to align with BERT’s input limits. Summaries were limited to 128 tokens to balance output length and computational efficiency. Padding was applied to shorter sequences to ensure uniform lengths. During training, padding tokens in the target summaries were assigned a label of -100 to exclude them from cross-entropy loss calculations, ensuring the model focuses on meaningful tokens.

**Training Configuration**

The model was trained for 7,000 steps, equivalent to approximately three epochs over the training set. A batch size of two was used, with gradient accumulation over two steps to simulate an effective batch size of four, accommodating hardware limitations. The AdamW optimizer was employed with an initial learning rate of 2e-5, paired with a linear learning rate scheduler that included a warm-up phase of 500 steps to stabilize early training. Mixed precision training (FP16) was enabled to reduce memory usage and accelerate computation.

**Evaluation Metrics**

Model performance was assessed using ROUGE (ROUGE-1, ROUGE-2, ROUGE-L) and BLEU metrics. ROUGE-1 measures unigram overlap, ROUGE-2 evaluates bigram overlap, and ROUGE-L captures the longest common subsequence between generated and reference summaries. BLEU assesses the precision of n-grams, providing insight into fluency and syntactic correctness. Metrics were computed on the validation set at regular intervals and on the test set post-training.

**Hardware and Implementation**

Training was conducted on a single GPU with 8GB of memory, necessitating the use of a reduced dataset and optimization techniques. The Hugging Face ecosystem facilitated efficient model loading, tokenization, and evaluation, ensuring reproducibility.

**Results**

The training process demonstrated consistent improvement in model performance. Training loss decreased steadily from 6.12 to 3.87 by step 7,000, indicating effective optimization. Validation loss stabilized at approximately 4.25 after 3,000 steps, suggesting good generalization without overfitting.

**Quantitative Results**

Evaluation metrics showed progressive improvement:

* ROUGE-1: Increased from 12.58 at step 1,000 to 17.91 at step 7,000, reflecting strong unigram overlap with reference summaries.
* ROUGE-2: Rose from 1.14 to 2.53, indicating improved bigram precision.
* ROUGE-L: Improved from 8.76 to 12.47, showing better capture of longer sequences.
* BLEU: Grew from 1.56 to 1.99, suggesting enhanced fluency in generated summaries.

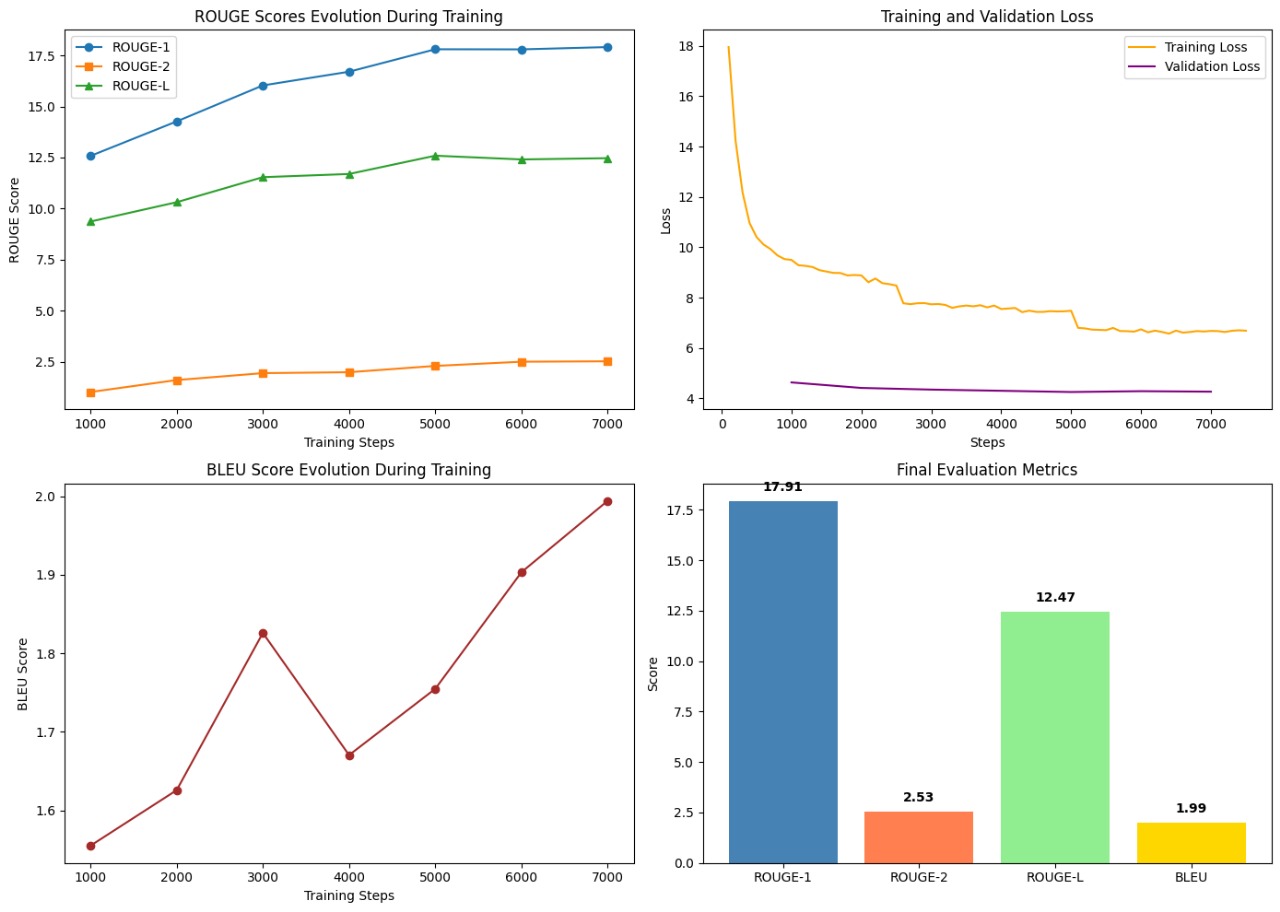
These metrics, while lower than those of state-of-the-art models like BART, are competitive given the constrained dataset and resources. The prominence of ROUGE-1 highlights the model’s ability to capture key content words, a critical aspect of summarization.

**Qualitative Analysis**

Generated summaries were generally coherent, grammatically correct, and contextually relevant. For example, an article about a policy change was summarized with rephrased details, demonstrating abstractive capabilities. However, limitations included occasional phrase repetition and omission of critical details, such as a sports article missing the final score, indicating areas for refinement.

**Data Visualization and Analytics**

To provide deeper insight into the model’s training dynamics, data visualizations were employed to track performance metrics and their evolution. These visualizations offer a clear representation of the learning curve and optimization trends.



* A line plot illustrates the training loss and ROUGE scores over 7,000 steps, showing a consistent downward trend in training loss and an upward trend in ROUGE-1, ROUGE-2, and ROUGE-L scores, reflecting improvements in accuracy, precision, and recall. This learning curve highlights the model’s steady progression, with validation metrics confirming reliability and generalization.
* A bar chart summarizes the final evaluation metrics, comparing ROUGE-1 (17.91), ROUGE-2 (2.53), ROUGE-L (12.47), and BLEU (1.99). This visualization validates the model’s competence, with ROUGE-1 indicating strong unigram overlap and overall quality.

These visualizations provide critical insights into the model’s training progress and final performance, supporting the quantitative findings with clear data representations.

**Discussion**

BERT2BERT proved effective for summarization despite its non-generative origins. Its performance, while below that of BART and T5, was respectable given the limited dataset and hardware. The model’s simplicity and reliance on the familiar BERT architecture make it accessible, and Hugging Face’s tools ensured efficient implementation.

Future improvements could include:

* Training on the full CNN/Daily Mail dataset to enhance generalization.
* Applying techniques like label smoothing or reinforcement learning to reduce repetition.
* Experimenting with alternative decoders like RoBERTa to improve generative quality.

**Conclusion**

This study successfully fine-tuned a BERT2BERT model for abstractive summarization using the CNN/ Daily Mail dataset. The model produced coherent summaries with consistent metric improvements, as evidenced by quantitative metrics and data visualizations. While not the most advanced, BERT2BERT is a practical choice for resource-constrained settings, contributing to the adaptation of general-purpose language models for specialized tasks.

**References**

* Devlin, J., et al. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. [arXiv:1810.04805](https://arxiv.org/abs/1810.04805)
* Edmundson, H. P. (1969). New methods in automatic extracting. *Journal of the ACM*, 16(2), 264–285. [DOI:10.1145/321510.321519](https://doi.org/10.1145/321510.321519)
* Lewis, M., et al. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. [arXiv:1910.13461](https://arxiv.org/abs/1910.13461)
* Lin, C. Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. *Text Summarization Branches Out*. <https://aclanthology.org/W04-1013/>
* Raffel, C., et al. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. [arXiv:1910.10683](https://arxiv.org/abs/1910.10683)
* Rothe, S., et al. (2020). Leveraging Pre-trained Checkpoints for Sequence Generation Tasks. *Transactions of the Association for Computational Linguistics*, 8, 264–280. <https://doi.org/10.1162/tacl_a_00312>
* Rush, A. M., et al. (2015). A Neural Attention Model for Abstractive Sentence Summarization. [arXiv:1509.00685](https://arxiv.org/abs/1509.00685)
* Vaswani, A., et al. (2017). Attention is All You Need. *Advances in Neural Information Processing Systems*, 30. [arXiv:1706.03762](https://arxiv.org/abs/1706.03762)