Automatic Text Summarization Using Transformer Models

1. Introduction

The goal of this project is to build an automatic text summarization system that can take lengthy articles and condense them into short, meaningful summaries without losing the core message. This tool can be useful for a variety of people: journalists can quickly grasp the key points in breaking news, businesses can extract actionable insights from long reports, and students can simplify academic articles or notes to save time while learning.

2. Dataset

We used a dataset of 9,000 articles, each paired with its headline. The dataset has two main columns:

- **Headlines**: The summarized headline of each article
- **Text**: The full article content

The dataset allowed us to train models to generate concise summaries while keeping the important details intact.

3. Approaches Tried

Approach 1: Custom Transformer Model

We first built our own Transformer from scratch, using positional embeddings, self-attention layers, and feed-forward networks. While this approach was a good learning experience, it had limitations: the summaries often lacked coherence, frequent `<OUT>` tokens appeared due to limited vocabulary, and ROUGE scores were low. This showed us that training a summarization model from scratch requires a large amount of data and fine-tuning to perform well.

Approach 2: BART Transformer

Next, we used BART, a pre-trained sequence-to-sequence model. It generated more coherent summaries, but some details were missed, and repetitive phrases appeared. Training for only three epochs with a small beam size limited diversity, and fixed hyperparameters prevented further optimization. While this was better than the custom model, it highlighted the importance of proper fine-tuning.

Approach 3: T5 Transformer (Fine-Tuned)

Finally, we fine-tuned the pre-trained T5-base model. This approach produced fluent, coherent summaries that captured context and details better than the previous models. ROUGE scores improved significantly, indicating that the summaries closely matched human-written ones. With additional fine-tuning and optimization, the model can handle even complex articles effectively.

4. Results and Key Observations

Aspect	Initial Models	T5 Fine-Tuned Model
Summary Quality	Incoherent, low ROUGE scores, frequent <OUT> tokens	Coherent, context-rich summaries
Model Performance	Underfitting, poor generalization	Better generalization and accuracy
Hyperparameter Tuning	Limited, suboptimal	Optimized using Optuna (learning rate: 3e-5, batch size: 4)
Training	Few epochs, insufficient data	Optimized epochs and training setup
Data Handling	No augmentation, noisy input	Cleaned and augmented data
Tokenization & Vocabulary	Frequent <OUT> tokens	Expanded vocabulary and better tokenization
Robustness	Low	Improved via better data handling
Training Efficiency	Slow, suboptimal	Faster with optimized setup
Outcome	Poor summaries	High-quality, human-aligned summaries

Example:

- Original headline: *South Korea urges US declare end Korean War*
- Generated summary (BART): *south Korean president Moon Jaein asked north Korea formally declare end 195053 Korean War Jaein added would imply US ended longstanding hostile relations.*
- Issue: The summary lacked context, coherence, and included irrelevant tokens.

The fine-tuned T5 model, however, generates concise summaries that retain essential details and maintain the overall context, closely mirroring human-written summaries.

5. Key Insights

- Pre-trained models like T5 outperform custom architectures for summarization tasks.
- Proper tokenization, vocabulary handling, and data augmentation are crucial for high-quality outputs.
- Hyperparameter tuning improves both accuracy and efficiency.
- Fine-tuning on the dataset significantly improves coherence and context retention.

6. Conclusion

This project demonstrates that transformer-based models, especially pre-trained ones, are highly effective for text summarization. By fine-tuning models like T5 and optimizing data handling, tokenization, and hyperparameters, we can generate summaries that are coherent, contextually rich, and closely aligned with human-written summaries. This system can save time and provide value for journalists, businesses, and students alike.

Model Files: Drive