News Summarizer

INTRODUCTION

Automated news summarization is the task of condensing long news articles into concise, meaningful summaries while preserving factual accuracy and key information. This project focuses on building an abstractive text summarization system using T5-Base, a transformer model designed for sequence-to-sequence natural language generation tasks.

The model is fine-tuned using supervised learning on a high-quality summarization dataset and evaluated using standard text-generation metrics. This report includes data preprocessing, model training, hyperparameters, evaluation results, and limitations.

DATASET

The dataset was sourced from Kaggle public dataset – CNN/DailyMail. Due to large dataset size and hardware limitations, it was slightly modified.

The dataset consisted of news articles with their respective summaries.

Each row included:

- Input: News article
- Output: News summary

Dataset details:

- Number of news article: 20886
- Average article length: 563 words

DATA PREPROCESSING

- Addition of prefix
    - Each article was formatted using:
      summarize: < news article >
- Tokenization
    - Tokenizer: T5-Base Tokenizer
    - Input token limit: 1024
    - Output token limit: 128
    - Each article Padded or Truncated to 1024 tokens.
    - Each summary Padded or Truncated to 128 tokens.
- Train-Test-Split
    - The standard 80/20 train-test split was applied.
    - Test dataset was used for evaluation metric calculation.

MODEL ARCHITECTURE

A custom model was build using pytorch library.

- Base model: T5-Base
- Model type: Encoder-Decoder
- Loss function: Cross-entropy
- Evaluation metric: Rouge scores (e.g.: RougeL, Rouge1, Rouge2)

GENERATION SETTINGS

Min New Tokens	30
Max New Tokens	128
No Repeat Ngram Size	3
Num Beams	6
Early Stopping	False

These settings ensure quality and adequate summary.

TRAINING

Model was trained using Hugging Face Trainer.

Training was performed in Kaggle Environment using:

- GPU T4 x 2

Model training took approximately 10 hours.

HYPERPARAMETERS

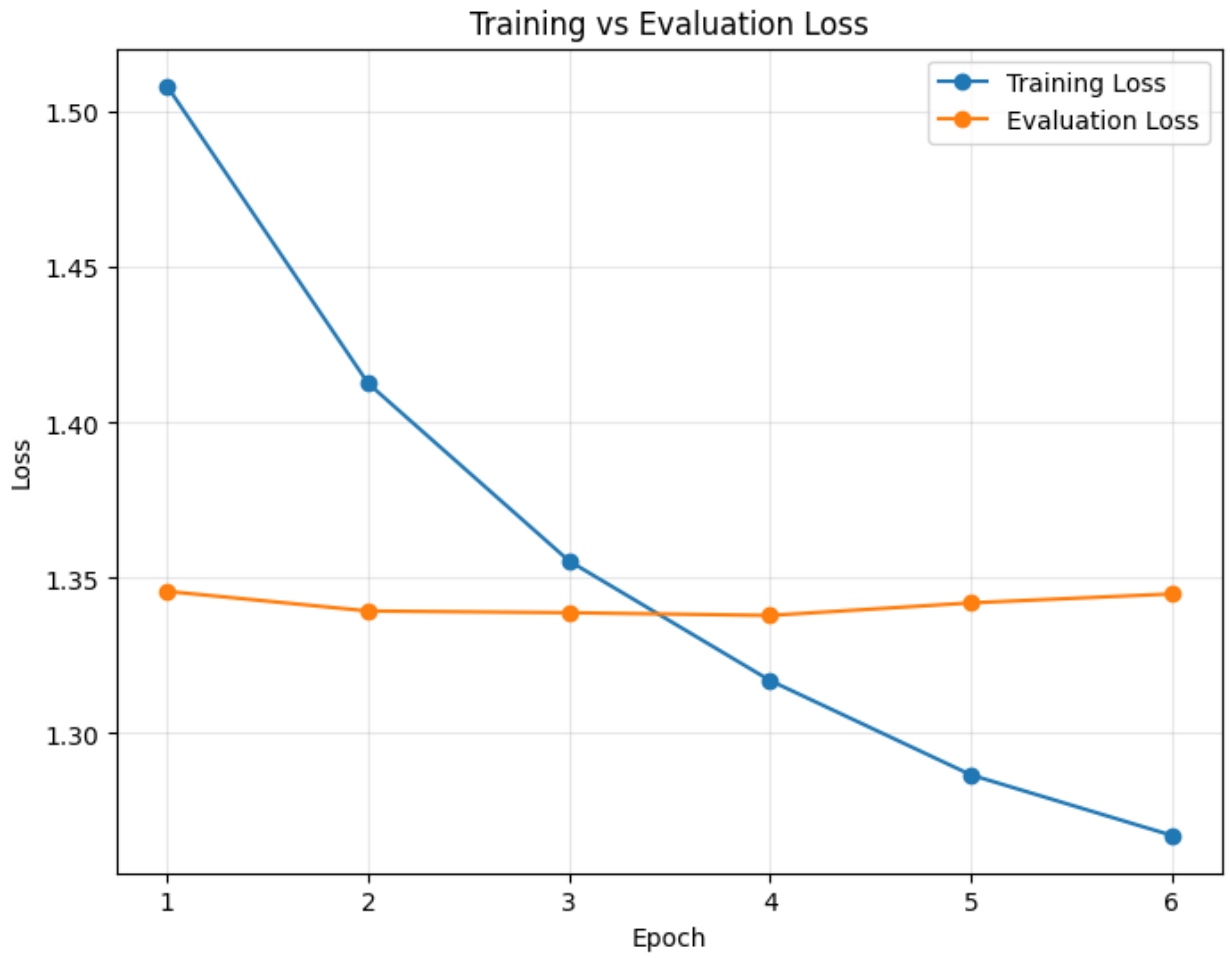
Learning Rate	4e-5
Optimizer	AdamW
Training Batch Size	4
Evaluation Batch Size	4
Weight Decay	0.05
Best Model Evaluation Metric	RougeL
Epochs	6
Predict With Generate	True

Model was trained for all 6 epochs, even after using Early Stopping.

EVALUATION

Evaluation table is available is “results” folder with filename “results.csv”

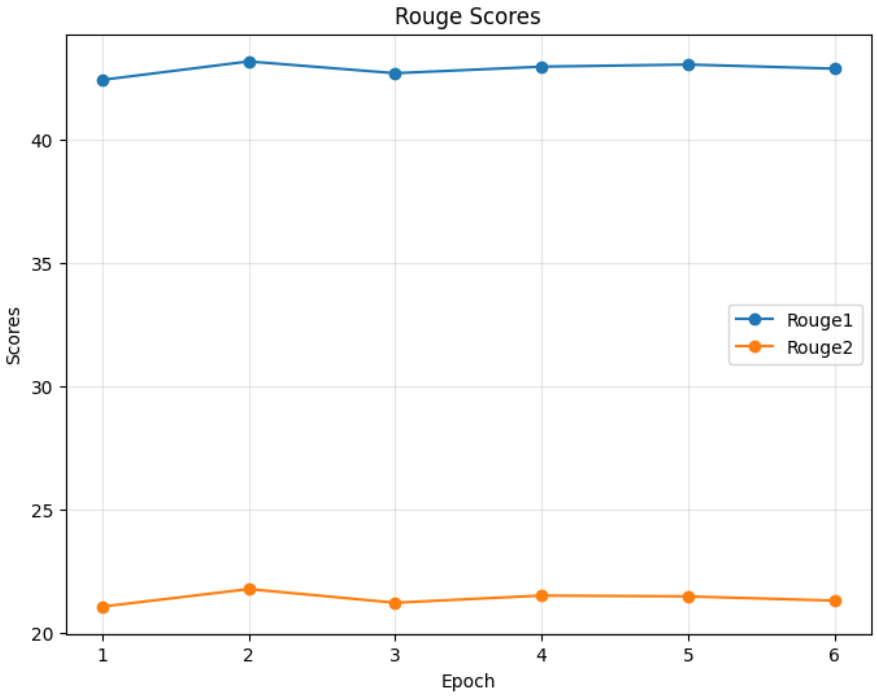
Training vs Evaluation loss:

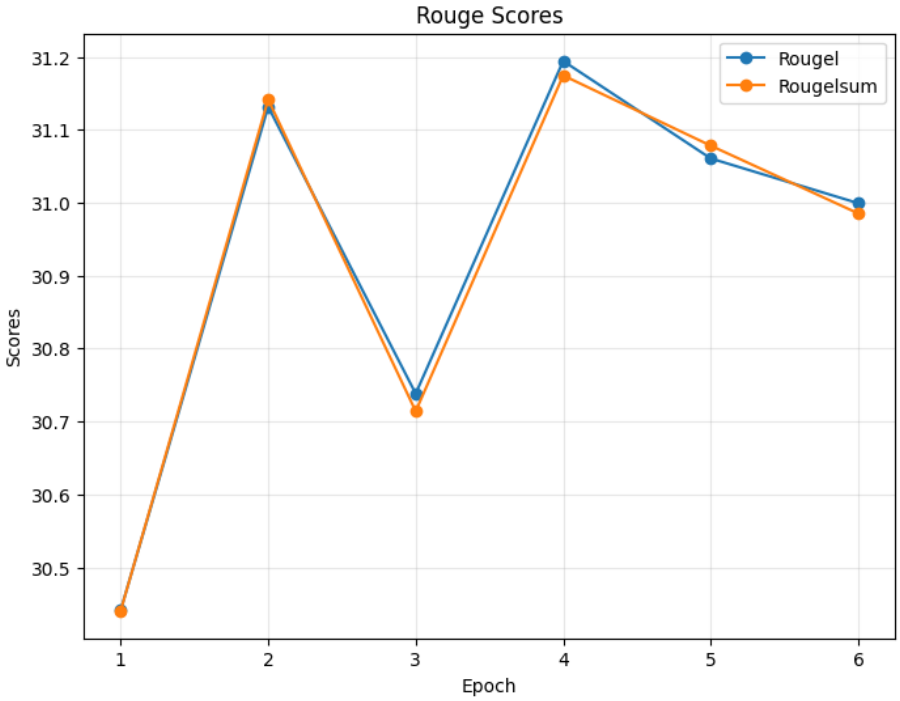


Observations:

- Training loss decreases across all 5 epochs.
- Validation loss across epochs remains identical

Rouge Scores:





Best model performance is at 4th epoch.

LIMITATIONS

- Limited input: Articles longer than 1024 tokens may lose crucial information.
- Domain bias: Cannot produce quality summary of other domains than news.
- Summary length: Sometimes outputs may be shorter than desired.
- Information loss: Model may omit important details due to output limit.
- Grammatical errors: May sometimes generate paragraphs that are grammatically incorrect.

CONCLUSION

This project successfully fine-tuned a T5-Base abstractive summarization model capable of generating concise, high-quality summaries for news articles. The system demonstrates strong ROUGE scores and produces readable summaries suitable for daily application.