Fine-tuning T5 for Text Summarization

May 2025

1 Introduction

Objective: Fine-tune T5-small for abstractive text summarization on XSum

Model: T5-small, a compact Transformer for text-to-text tasks

Dataset: XSum, news articles with single-sentence summaries

Process: Data preprocessing, model fine-tuning, ROUGE metric evaluation

2 Related Work

1 - Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

link: https://arxiv.org/abs/1910.10683

2-Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization.

Link: https://arxiv.org/abs/1808.08745v1

3-Evaluating LLMs and Pre-trained Models for Text Summarization Across Diverse Datasets

Link: https://arxiv.org/abs/2502.19339

4-A Survey of Recent Abstract Summarization Techniques

Link: https://arxiv.org/pdf/2105.00824

3 Methodology

3.1 Data Acquisition and Preprocessing

Dataset: XSum, sourced via Hugging Face datasets, stored in ./raw_data/

Preprocessing:

Tokenization with AutoTokenizer for T5-small

Added “summarize: ” prefix to inputs

Truncated inputs to 1024 tokens, summaries to 128 tokens

Saved tokenized data in ./processed_data/, JSON subset for inspection

3.2 Model Architecture

Model: T5-small, using AutoModelForSeq2SeqLM, suited for sequence-to-sequence tasks

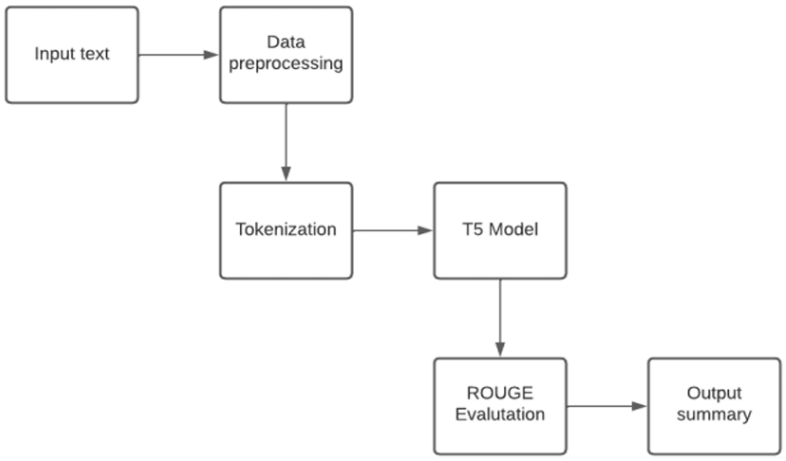

Fig-1: System Architecture

3.3 Training Procedure

Setup: Fine-tuned with Seq2SeqTrainer, GPU (CUDA) or CPU, FP16 on GPU

Hyperparameters:

- Learning Rate: 2×10^{-5}
- Batch Size: 4, with 4 gradient accumulation steps (effective: 16)
- Epochs: 1
- Optimizer: AdamW
- Weight Decay: 0.01

Optimization: Gradient checkpointing, checkpoints saved every 200 steps, max 5

retained in ./checkpoints/

Output: Best/final model saved in ./model_output/t5-small-finetuned-xsum

Data Limits: Optional 20,000 training, 1,000 validation samples

3.4 Evaluation Protocol

Metrics: ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-Lsum via evaluate

Process: Tested on XSum test set (up to 500 samples), beam search (4 beams, 64–128 tokens, length penalty 2.0)

Text Prep: Segmented with nltk.sent_tokenize for ROUGE

4 Experimental Setup

Dataset: XSum, limited to 20,000 training, 1,000 validation, 500 test samples

Model: T5-small, pre-trained from Hugging Face

Hardware: NVIDIA GPU (CUDA 12.1) or CPU

Software: Python, transformers (≥4.11.0), datasets, evaluate, nltk, rouge_score, torch (2.5.1+cu121), accelerate (≥0.26.0), numpy, pandas, huggingface-hub. TensorBoard for monitoring

5 Results and Analysis

5.1 Quantitative Results

ROUGE-1: 22.12

ROUGE-2: 4.88

ROUGE-L: 15.35

ROUGE-Lsum: 16.98

5.2 Qualitative Results

Example 1:

Original: Chronic housing need for prison leavers in Wales, says charity

Generated: Prison Link Cymru notes chronic housing need for Welsh prison leavers, suggesting one-bedroom flats could save costs

Example 2:

Original: Man in court after police seize firearms, ammunition, cash in Edinburgh

Generated: Edinburgh police recover firearms, ammunition, and cash; man charged with conspiracy

Example 3:

Original: Four denied bail for kidnapping, torturing man in racially motivated attack streamed on Facebook

Generated: Four charged with hate crimes, kidnapping in Chicago; court details violent acts

5.3 Analysis

Quantitative: ROUGE-1 (22.12) shows moderate unigram overlap; lower ROUGE- 2, ROUGE-L, ROUGE-Lsum typical for small model, single epoch

Qualitative: Captures key entities, fluent but repetitive (Example 2), factually inconsistent (Examples 1, 3), longer than references

Insights: TensorBoard logs could reveal convergence, overfitting trends

6 Conclusion

Summary: Fine-tuned T5-small on XSum, achieving ROUGE scores: 22.12 (ROUGE- 1), 4.88 (ROUGE-2), 15.35 (ROUGE-L), 16.98 (ROUGE-Lsum)

Findings: Produces relevant summaries but struggles with conciseness, accuracy, redundancy

Limitations:

- Small model size limits performance
- Single epoch, subsampled data
- Default hyperparameters
- ROUGE misses semantic quality