ASSIGNMENT- 3: Fine-Tuning a Large Language Model

Technical Report

Fine-Tuning FLAN-T5 for Biomedical Summarization

1. Executive Summary

This project focuses on fine-tuning Google's **FLAN-T5-Small** model on the **PubMed Summarization dataset (`ccdv/pubmed-summarization`)** to enhance its capability to produce concise and contextually accurate biomedical summaries. The objective is to adapt a general instruction-tuned language model to the biomedical domain, where text is often dense, technical, and requires high factual precision.

Rather than training a large language model from scratch—which would demand massive amounts of computational resources and time—the fine-tuning approach capitalizes on the pre-existing linguistic and reasoning capabilities of FLAN-T5. By further training it on biomedical data, the model learns to compress lengthy scientific texts into coherent abstracts that maintain essential information.

The project pipeline employed **Hugging Face's Transformers library** along with **Ray Tune** for hyperparameter optimization. The fine-tuned model demonstrated a significant improvement in key evaluation metrics such as **ROUGE-1**, **ROUGE-2**, and **ROUGE-L**, outperforming the baseline zero-shot model by a wide margin. The results validate that domain-specific fine-tuning can make relatively small models competitive for specialized tasks such as biomedical literature summarization, clinical evidence extraction, and automated review generation.

---

2. Methodology and Approach

2.1 Methodological Rationale

Fine-tuning was chosen as the preferred methodology because it offers the best balance between **computational efficiency** and **performance gain**. Instead of training a model from scratch on millions of biomedical texts, fine-tuning enables the adaptation of an already instruction-tuned model to the target domain. This approach leverages transfer learning principles, allowing the model to retain general language understanding while improving its ability to process domain-specific vocabulary and semantics.

The decision to use **FLAN-T5** stems from its architecture and prior instruction-tuning across multiple text generation tasks such as summarization, question answering, and translation. These characteristics make it inherently capable of following structured prompts such as "summarize this article," aligning perfectly with the PubMed summarization task. The **Small variant**, which has approximately 60 million parameters, was chosen to strike a balance between accuracy and resource constraints. Larger models like FLAN-T5-Base or Large would likely yield marginally higher performance but at substantially greater computational cost.

The methodology followed a linear progression starting with data preprocessing and cleaning, followed by tokenization and formatting, model initialization, fine-tuning using the Hugging Face Trainer API, and final evaluation using ROUGE metrics. Early stopping was implemented to avoid overfitting, and Ray Tune was used to search for the best hyperparameter configuration based on ROUGE-L performance.

2.2 Dataset Preparation

The dataset used for this project is **`ccdv/pubmed-summarization`**, a curated collection of biomedical articles paired with expert-written abstracts. The dataset provides long-form scientific text along with concise human summaries, making it ideal for developing summarization systems that mirror the structure of actual PubMed abstracts.

For experimental efficiency, a **random sample** of the dataset was used, resulting in approximately **10,000 training**, **1,500 validation**, and **1,500 testing** samples. Each entry contains two primary fields: "article" (the full text) and "abstract" (the summary). The goal was for the model to learn the mapping from the article's main content to its summary.

Extensive **text cleaning** was carried out to ensure input consistency. HTML tags and special characters were removed, and unusually short or long entries were filtered out. Articles shorter than 800 characters or abstracts shorter than 40 characters were excluded. The remaining text was trimmed to a maximum of 8,000 characters to avoid input overflow.

Tokenization was performed using **`T5TokenizerFast`**, with a **maximum input length of 512 tokens** for the article and **128 tokens** for the output summary. This limit was chosen to balance contextual completeness and GPU memory usage. The formatted input-output pairs followed the schema:

```
Input  :  article text
Target :  abstract summary
```

This direct text-to-text formulation matched FLAN-T5's pretraining format, enabling smooth domain adaptation.
The **PubMed Summarization** dataset was selected because it closely represents real-world biomedical writing, containing highly structured text with factual density. It trains the model not only to shorten text but also to preserve scientific meaning, an essential capability for real biomedical applications.

2.3 Model Architecture and Hyperparameters

The project used the **FLAN-T5-Small** architecture, which follows the standard encoder-decoder Transformer structure. The model was trained using Hugging Face's **Trainer API**, simplifying implementation of backpropagation, evaluation, and logging. Key hyperparameters were carefully tuned to achieve the best trade-off between training stability and performance.

Parameter	Value	Explanation
Base model	`google/flan-t5-small`	Pretrained, instruction-tuned transformer suitable for summarization
Batch size	4	Optimized for GPU memory constraints
Learning rate	2e-5 and 5e-5 (tested)	Tuned using Ray Tune; 5e-5 chosen as optimal
Epochs	3	Converged by epoch 3, preventing overfitting
Optimizer	AdamW	Efficient handling of sparse gradients
Scheduler	Linear	Smooth learning rate decay across epochs
Weight decay	0.01	Adds regularization to stabilize training
Metric	ROUGE-L	Primary metric for summarization quality

Training was performed on a **single NVIDIA T4 GPU** and completed in roughly **two hours**. Early stopping with a patience of two epochs prevented unnecessary computation once validation performance plateaued.
Hyperparameter optimization using **Ray Tune** explored learning rates of 2e-5 and 5e-5 with a fixed batch size and epoch count. The best configuration—**5e-5 learning rate, batch size of 4, and 3 epochs**—produced the highest ROUGE-L score, confirming this as the final setup.

This configuration ensures that the model learns effectively without requiring extensive computational resources, making it a reproducible and scalable solution for similar domain adaptation tasks.

---

3. Results and Analysis

3.1 Quantitative Evaluation

Performance was assessed using standard **ROUGE metrics**—ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum—which measure word-level and phrase-level overlap between generated summaries and human references.

The results clearly show that fine-tuning led to substantial improvements compared to the baseline zero-shot FLAN-T5 model:

Model	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
Baseline (FLAN-T5)	0.0526	0.0051	0.0420	0.0470
Fine-tuned (PubMed)	0.1981	0.0433	0.1353	0.1563

These numbers indicate a **near fourfold improvement** across all metrics. The most notable gain was observed in ROUGE-L, which measures sentence-level coherence and is particularly relevant for summarization tasks. The validation loss steadily decreased over the three training epochs, confirming that the model was learning effectively without overfitting.

3.2 Qualitative Evaluation

Beyond numerical metrics, qualitative inspection of model outputs provided deeper insights. The fine-tuned model produced summaries that captured the key findings and conclusions of biomedical papers in a clear and concise manner.
For instance, when summarizing long clinical trial reports, the model learned to prioritize the outcome statements ("X increased patient survival" or "Y improved response rates") rather than background context.

However, the summaries occasionally exhibited **truncation** in longer sentences and sometimes **paraphrased complex biomedical terms**. These minor issues stemmed primarily from input length constraints and the generic tokenizer used.
Overall, the model demonstrated a solid grasp of domain-specific writing patterns and was able to generate fluent, factual abstracts consistent with human-written references.

3.3 Visualization and Interpretation

The comparative visualization of baseline and fine-tuned model scores revealed that fine-tuning improved all four ROUGE categories substantially. The fine-tuned FLAN-T5 captured more relevant content and maintained better sentence structure, showing that domain exposure allows even smaller models to become strong summarizers.

These improvements are not merely numerical; they represent a tangible increase in the model's ability to preserve factual detail, maintain coherence, and avoid generic phrasing—three critical success factors for biomedical summarization tasks.

---

4. Error Analysis

While the fine-tuned model performed well overall, systematic error analysis exposed areas for potential improvement.
The most frequent issue was **sentence truncation**, caused by the 256-token limit during generation. In some cases, the model generated incomplete summaries or omitted supporting details from the original text. Increasing the input length or adopting hierarchical summarization strategies could mitigate this.

Another pattern observed was **minor factual drift**, particularly where biomedical entities were paraphrased. For instance, “tumor suppressor” might be replaced with “protein that inhibits tumor growth.” Though semantically similar, such variations can reduce precision in highly technical contexts.

Additionally, the model tended to produce **generic phrasing** when confronted with very long or complex inputs, suggesting a limit in contextual recall. Incorporating a **domain-specific tokenizer** or performing **Domain-Adaptive Pretraining (DAPT)** on biomedical corpora could

address this gap.
Finally, **numeric detail omission**—such as skipping specific dosage values or sample sizes—was another limitation, likely tied to token truncation during encoding.

---

5. Lessons Learned

One of the key lessons from this project is that **smaller, instruction-tuned models can perform competitively** when fine-tuned carefully on specialized data. FLAN-T5-Small demonstrated that size is not the sole determinant of model quality; rather, task alignment and high-quality data play equally critical roles.

Another insight is that **data quality outweighs quantity**. Even though only some percent of the PubMed dataset was used, its curated structure ensured effective learning. The experiment also highlighted the value of **comprehensive evaluation**, as metrics like ROUGE-L provided a more realistic measure of output quality than simple accuracy.

The experience further emphasized the power of **hyperparameter optimization** using tools like Ray Tune, which automated the process of finding the most effective learning rate and reduced manual trial-and-error. Finally, the project reaffirmed that practical trade-offs—such as shorter sequence lengths and modest batch sizes—can yield strong results without high computational overhead.

---

6. Limitations and Future Improvements

Despite its strong performance, the fine-tuned FLAN-T5 model has several limitations.
First, the **intrinsic ambiguity of biomedical text** sometimes led the model to produce cautious summaries with phrases like “may cause” or “appears to influence.” While this reflects the uncertainty present in many scientific articles, it can reduce clarity in deterministic summaries.

Second, the model’s **context window** is limited, preventing it from processing entire research papers. This restriction occasionally caused omission of results or conclusions from later sections. Increasing the token length to 720 or using transformer variants with extended context windows could help.

Third, the absence of **factual verification mechanisms** means that while the summaries read fluently, they are not guaranteed to be factually correct. Integrating factuality scoring or biomedical knowledge graph cross-checking could improve this aspect.

For future work, fine-tuning **larger FLAN-T5 variants** (Base or Large) could further enhance performance without excessive cost. Additionally, conducting **Domain-Adaptive Pretraining**

(DAPT) on unlabeled PubMed articles and integrating **factual consistency metrics** such as BERTScore or FactCC could help ensure greater reliability in generated summaries.

---

7. Research Best Practices

This project followed robust research and engineering practices to ensure reproducibility and transparency. All scripts and notebooks were version-controlled through Git, and training logs were stored using TensorBoard for performance tracking. Random seeds were fixed across libraries (PyTorch, NumPy, and Datasets) to maintain determinism.

Every hyperparameter choice was grounded in prior research on T5 fine-tuning, particularly in biomedical NLP contexts. The evaluation pipeline was fully reproducible using the **Evaluate** and **Rouge-Score** libraries. Model checkpoints and tokenizers were saved under `./finetuned_flan_t5_pubmed/` for deployment or future retraining.

This rigorous methodology highlights that with clear documentation, small models, and open-source tools, high-quality domain-specific language systems can be built efficiently. The resulting pipeline can serve as a blueprint for fine-tuning other instruction-tuned models on domain-specific corpora in medicine, law, or finance.

---

8. References

1. Cohan, A. et al. (2018). *Discourse-Aware Summarization of Scientific Documents.* NAACL.
2. Zhang, Y. et al. (2020). *PubMedQA: A Dataset for Biomedical Research Question Answering.* ACL 2020.
3. Raffel, C. et al. (2020). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer (T5).* JMLR.
4. Chung, H. W. et al. (2022). *Scaling Instruction-Finetuned Language Models.* arXiv preprint arXiv:2210.11416.
5. Hugging Face (2025). *Transformers Documentation.* https://huggingface.co/docs/transformers