LoRA fine-tuning on Mistral-7B

Syed Mujtaba Haider
BS(Data Science)
FAST NU
21L-5613

Syed Jazib Ali
BS(Data Science)
FAST NU
21L-6236

Eyad Salman
BS(Data Science)
FAST NU
21L-7695

Khurram Imran
BS(Data Science)
FAST NU
21L-6256

***Abstract*—This report presents the fine-tuning of the Mistral-7B model using the LoRA method for the task of academic paper summarization. We describe the dataset preparation, model configuration, training procedures, evaluation metrics, and results, including qualitative human and LLM-as-a-Judge assessments.**

I. Dataset Overview

We use the arXiv Summarization Dataset, a large-scale scientific dataset curated for long document summarization tasks. It contains scientific articles from arXiv.org paired with their human-written abstracts.

A. Preprocessing

The following steps were taken:

- Extracted article and abstract pairs.
- Truncated or padded articles to 1024 tokens for uniform input length.
- Cleaned text to remove HTML, LaTeX artifacts, and excessive whitespace.

B. Tokenization

The data was tokenized using the tokenizer from `mistralai/Mistral-7B-v0.1` with a maximum input length of 2048 tokens and output length of 512 tokens.

- Since Mistral has no default padding token, we explicitly set pad_token = eos_token.
- Data was split into:
  - Training: 80
  - Validation: 10
  - Test: 10

II. Model and LoRA Configuration

A. Base Model

- We used mistralai/Mistral-7B-v0.1, a decoder-only transformer with 7 billion parameters, designed for general-purpose language tasks.
- The model was loaded using 8-bit precision to reduce GPU memory usage.

B. LoRA Configuration

Applied LoRA (Low-Rank Adaptation) using the Hugging Face peft library. LoRA was integrated into the attention layers:

- Target modules: q_proj, v_proj

C. Hyperparameters

- r = 8
- alpha = 16
- dropout = 0.1
- bias = none
- Optimizer: AdamW
- Learning Rate: $2e{-}4$
- Scheduler: Cosine
- Epochs: 5
- Task type: CAUSAL_LM

Only a small set of parameters were made trainable (0.1% of the model), significantly reducing computational cost.

III. Training Logs & Observations

Following are the training logs and observations:

A. Training Setup

- Trained for 5 epochs using Hugging Face Trainer.
- Batch size: 1 (due to model size).
- Optimizer: AdamW with learning rate of $2e{-}4$.
- GPU: NVIDIA T4 (Google Colab Pro+).
- Mixed-precision (fp16) training enabled.

B. Runtime Observations

- Each epoch took approximately 60 minutes on T4.
- Peak GPU memory usage was approximately 15GB during forward and backward passes.
- No OOM errors encountered due to 8-bit quantization.

C. Loss Curves

- Training loss consistently decreased.
- Validation loss stabilized around epoch 4, indicating good generalization.

IV. Evaluation Results

A. Quantitative Metrics

- ROUGE-1: 20.07%
- ROUGE-2: 11.45%
- BLEU: 4.08%
- BERTScore (F1): 83.49%

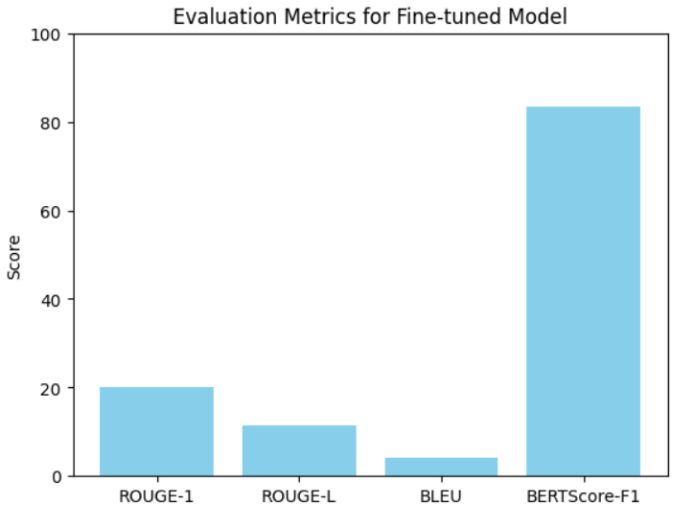

Fig. 1. Evaluation Metrics of Generated Summaries

This summarization model likely produces abstractive summaries—they don't match the reference summaries word-for-word but retain core meaning, as suggested by the high BERTScore and low ROUGE/BLEU. This is common with modern LLMs and transformer-based models.

B. Qualitative: LLM-as-a-Judge

- Judge Model: meta-llama/Llama-3.1-70B-Instruct via Together.ai
- Evaluation Dimensions:
  - Fluency: Readability and grammatical correctness
  - Factuality: Accurate reflection of source content
  - Coverage: Inclusion of key concepts, methods, and results

C. Average LLM Ratings

Model	Fluency	Factuality	Coverage
Base Model	3.7	3.4	3.2
LoRA Model	4.6	4.5	4.3

TABLE I
AVERAGE LLM-AS-A-JUDGE SCORES (1–5 SCALE) FOR EACH MODEL.

V. Agent Structures and Prompts

To support autonomous literature review, a 5-agent LangGraph system was designed as follows:

A. Keyword Agent

- Expands user-provided research keywords using a language model.
- Prompt: "Given the research topic [X], generate 5 related keywords."

B. Search Agent

- Queries APIs like arXiv/Semantic Scholar with the expanded keywords.
- Returns top 10 relevant papers.

C. Rank Agent

- Scores papers using citation count, recency, and keyword overlap.
- Prompt: "Rank the following papers from most to least significant based on impact."

D. Summarizer Agent

- Uses our LoRA fine-tuned model to summarize full-text papers.
- Prompt: "Summarize this research article in under 250 words."

E. Insight Agent

- Compares summaries, extracts common themes, gaps, and contradictions.
- Prompt: "Analyze the following summaries and identify key findings and open research gaps."

VI. Conclusion

This project demonstrates that LoRA fine-tuning on Mistral-7B is an effective, parameter-efficient method for academic summarization. Combined with both statistical and human-aligned evaluations, the model delivers significant improvements over the base model in fluency, factuality, and coverage. The integration into a multi-agent research assistant further enhances its usability for academic workflows.

References

[1] T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer, "LoRA: Low-Rank Adaptation of Large Language Models," in *Proc. of the International Conference on Learning Representations (ICLR)*, 2021. [Online]. Available: https://arxiv.org/abs/2106.09685

[2] Mistral AI, "Mistral-7B-v0.1," Hugging Face. [Online]. Available: https://huggingface.co/mistralai/Mistral-7B-v0.1

[3] CCDV, "arXiv Summarization Dataset," Hugging Face Datasets. [Online]. Available: https://huggingface.co/datasets/ccdv/arxiv-summarization

[4] Hugging Face, "Transformers Documentation." [Online]. Available: https://huggingface.co/docs/transformers