Medical Insight Extraction from Clinical Reports

---

1. Project Idea

This project applies a LoRA fine-tuned Large Language Model to extract key medical insights from clinical radiology reports.

Task Description:

- Converting lengthy medical reports into concise, structured impressions
- Identifying and extracting critical diagnostic findings
- Automating the generation of medical impressions from radiological observations

Real-World Application:

- Assists radiologists in report generation
- Increases efficiency in clinical workflows
- Improves consistency in medical documentation

---

2. Dataset Information

Dataset: MIMIC-III (Medical Information Mart for Intensive Care III)

- **Source:** Hugging Face dataset "hejazizo/mimic-iii"
- Link: https://huggingface.co/datasets/hejazizo/mimic-iii
  - **Size:**
  - Training: 59,320 samples
  - Validation: 7,413 samples
- Test: 13,057 samples
  - **Structure:**

  Each sample contains a clinical report ("prompt") and corresponding clinical impression ("impressions")

Example Data Sample:

```
Report: "the liver pancreas spleen adrenals and kidneys are normal the aorta is of
normal caliber no enlarged lymph node identified in the retroperitoneum there is
bilateral hydronephrosis right side greater than left"

Impression: "1 bilateral adenxal tumors with resultant bilateral hydronephrosis right
side greater than left most likely metastatic ___"
```

Preprocessing:

- Added "extract: " prefix to input text
- Converted dataset structure to input/output format
- Tokenized inputs and outputs with appropriate truncation
  - Input max length: 512 tokens
  - Output max length: 256 tokens

---

3. Base Model Information

Model: T5-Small (Text-to-Text Transfer Transformer)

- **Architecture:** Encoder-decoder transformer model
- **Size:** 60 million parameters (small variant)
- **Originally trained on:** C4 (Colossal Clean Crawled Corpus)

- **Capabilities:**
  - Text-to-text generation
  - Designed for transfer learning across NLP tasks
  - Handles tasks via text transformation

Why T5 for this task?

- Text-to-text framework naturally fits extraction/summarization
- Efficient parameter usage
- Strong foundation in general language understanding
- Suitable for fine-tuning with limited resources

---

4. Fine-Tuning with LoRA

Low-Rank Adaptation (LoRA) Overview:

- Efficient fine-tuning method that reduces trainable parameters
- Adds low-rank matrices to pre-trained weights rather than updating all parameters
- Significantly reduces memory requirements and training time

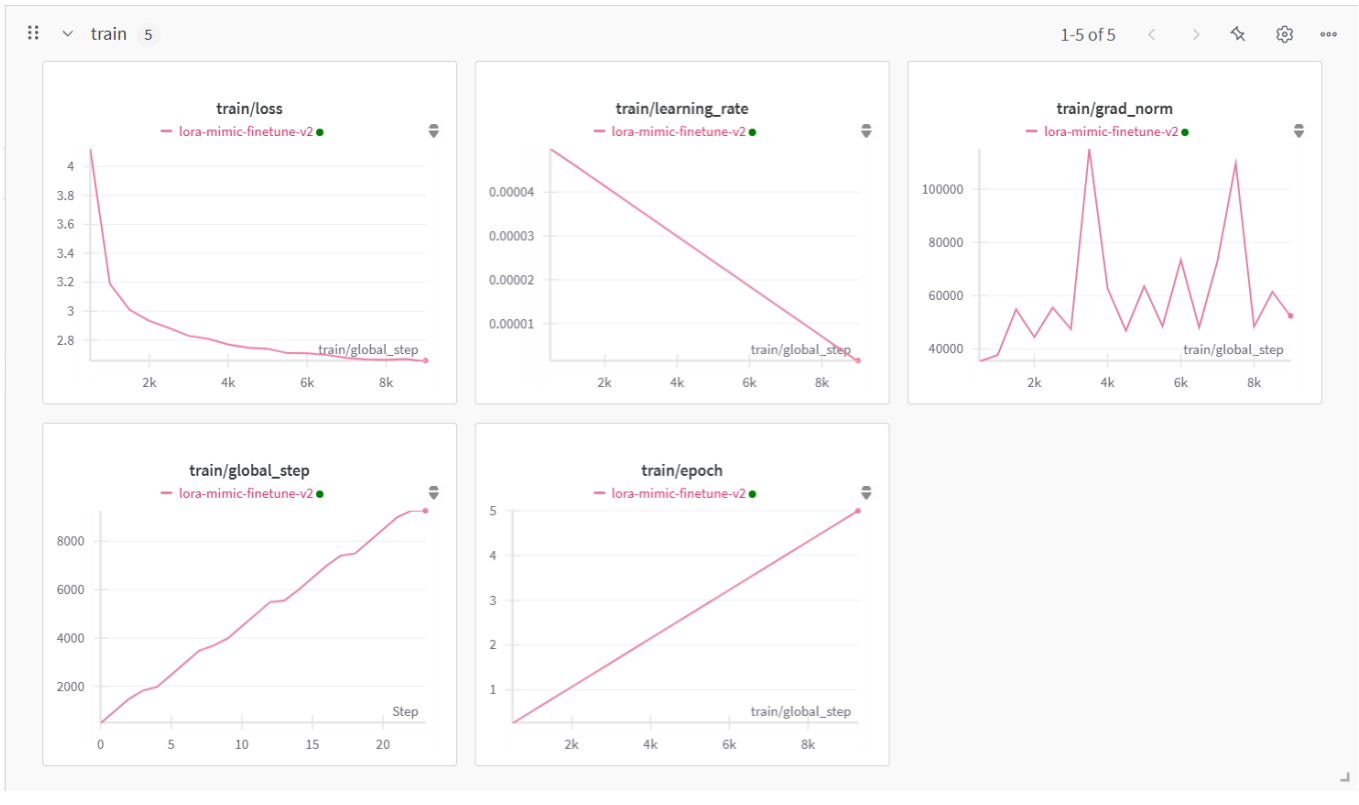
LoRA Configuration Used:

```python
lora_config = LoraConfig(
    r=16,
    lora_alpha=64,
    lora_dropout=0.1,
    bias="none",
    target_modules="all-linear",
    task_type=TaskType.SEQ_2_SEQ_LM
)
```

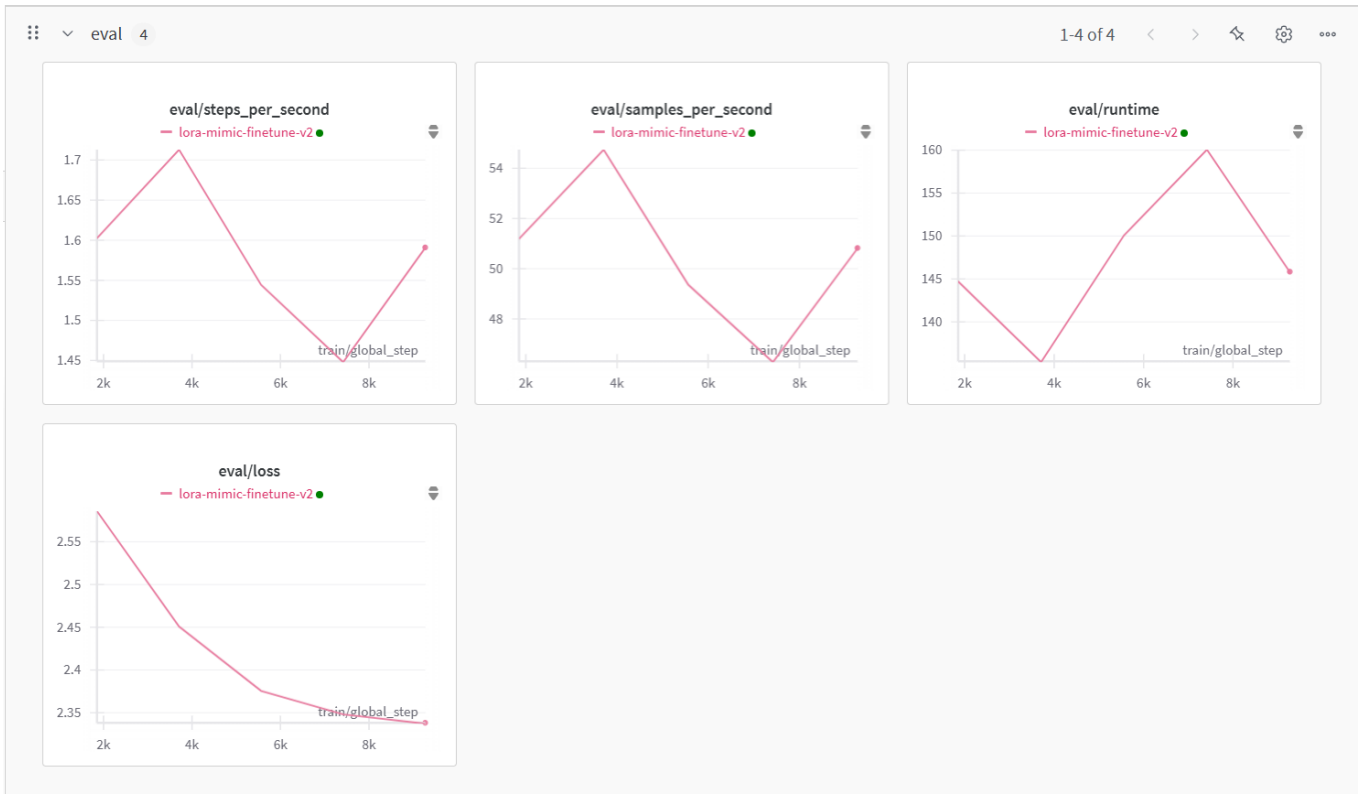
Training Configuration:

- Epochs: 5
- Batch size: 16
- Learning rate: 5e-5
- FP16 precision
- Weight decay: 0.01
- Warmup steps: 500

Training Metrics



Validation Metrics



5. Evaluation Methods

Metrics Used:

- **ROUGE Scores:**
  - ROUGE-1: Unigram overlap (word-level)
  - ROUGE-2: Bigram overlap
  - ROUGE-L: Longest common subsequence
  - ROUGE-Lsum: Variant for summary evaluation
- **BLEU Score:**
  - N-gram precision-based metric
  - Includes brevity penalty for length assessment

Evaluation Approach:

- Compared base T5-small model vs. LoRA fine-tuned model
- Evaluated on 1,000 random test samples
- Generated predictions with max_new_tokens=128
- Applied ROUGE and BLEU metrics to assess quality
- Qualitative comparison of generated outputs

---

6. Project Results

Training Progress

- Initial training loss: ~3.01
- Final training loss: ~2.66
- Validation loss improved from 2.59 to 2.34

[9270/9270 3:11:36, Epoch 5/5]

Epoch	Training Loss	Validation Loss
1	3.013500	2.585762
2	2.814000	2.451436
3	2.716400	2.376428
4	2.683400	2.349041
5	2.658200	2.338226

ROUGE Score Comparison

Metric	Base Model	LoRA Fine-tuned	Improvement
ROUGE-1	24.9%	32.3%	+7.4%
ROUGE-2	11.6%	14.9%	+3.3%
ROUGE-L	18.2%	25.5%	+7.3%
ROUGE-Lsum	18.2%	25.5%	+7.3%

What These Scores Mean

- ROUGE-1 (32.3%)
  This measures unigram (word-level) overlap. A score in this range indicates a decent level of basic content similarity between the generated and reference texts.
- ROUGE-2 (14.8%)
  This evaluates bigram overlap. The lower score suggests limited fluency or coherence in terms of consecutive word pairs.
- ROUGE-L / ROUGE-Lsum (~25.5%)
  These scores assess the longest common subsequence, making them effective for capturing sentence-level structural similarity.

Interpretation

Use Case: Feature Extraction
These scores fall within the typical range for early or baseline models.
For example, many feature extraction baselines score around 0.2–0.4 ROUGE-1 on datasets like CNN/DailyMail.

Use Case: Radiology Report Generation / Medical Domain
These ROUGE scores are considered quite reasonable in the medical domain, due to the following factors:

- Medical language often involves domain-specific vocabulary.
- There are usually multiple valid ways to phrase the same clinical content.
- ROUGE metrics frequently underrepresent semantic quality, especially in specialized or technical texts.

BLEU Score Analysis

- Base model: **10.3%**
- LoRA fine-tuned: **8.2%**
- Note: Lower BLEU but higher ROUGE suggests better semantic extraction despite different phrasing

Interpretation

- LoRA fine-tuning significantly improved ROUGE metrics
- Lower BLEU score with higher ROUGE indicates the model captures key content but with different phrasing
- Brevity penalty shows base model tends to be wordy while fine-tuned model is more concise

---

7. Structured Output Examples

Example 1:

Input Report:

```
 "the liver pancreas spleen adrenals and kidneys are normal the aorta is of normal caliber no enlarged lymph node identified in the retroperitoneum there is bilateral hydronephrosis right side greater than left"
```

Base Model Output:

": the liver pancreas spleen adrenals and kidneys are normal the aorta is of normal caliber no enlarged lymph node identified in the retroperitoneum there is bilateral hydronephrosis right side greater than left"

Fine-tuned Model Output:

"1 bilateral hydronephrosis right side greater than left"

Example 2:

Input Report:

"status post radical hysterectomy there is susceptibility artifact on the left lying between the bladder and the rectum -- are there surgical clips in this location adjacent to the artifact there is an ill-defined 16 x 29 mm area of abnormal soft tissue intensity hypointense on t1 with intermediate intensity on t2 no discrete mass is seen no enlarge pelvic lymph nodes are detected..."

Base Model Output:

": status post radical hysterectomy there is susceptibility artifact on the left lying between the bladder and the rectum -- are there surgical clips adjacent to the artifact there is an ill-defined 16 x 29 mm area of abnormal soft tissue intensity hypointense on t1 with intermediate intensity on t2 no discrete mass is seen no enlarge pelvic lymph nodes are detected the bladder wall is not thickened a small amount of free fluid is present within the pelvis there"

Fine-tuned Model Output:

"1 no evidence of a stranding or hydronephrosis

2 ill-defined stranding on the left lying between the bladder and the rectum"

Conclusions

- **LoRA fine-tuning significantly improved extraction quality**
- **Structured outputs follow medical report conventions:**
  - Numbered findings
  - Concise language
  - Focus on critical information
- **Business Impact:**
  - Reduces radiologist time spent on documentation
  - Improves report standardization
  - Enhances clinical workflow efficiency
- **Next Steps:**
  - Experiment with larger base models
  - Incorporate medical-specific pre-training
  - Expand to other medical specialties beyond radiology