NLP for Healthcare - CL3.411
Assignment 3

X-ray to Report

Deadline: 7th October, 2025 11:59 PM

1 Objectives:

Task: Given a Chest X-ray, generate:

1. **Findings:** A detailed, descriptive account of what is directly observed on the image.
2. **Impression:** A concise interpretation of the findings.

Approach: You will experiment with three different approaches, compare their performance, and critically analyze the results.

2 Dataset:

Link: https://huggingface.co/datasets/itsanmolgupta/mimic-cxr-dataset

- 80% – Training set
- 10% – Validation set
- 10% – Test set

3 Task 1: Baseline (Pre-trained Models): [5 marks]

Evaluate at least three pre-trained models (trained on medical domain data) for:

- Image → Findings
- Findings → Impression
- Image + Findings → Impression

Evaluation: **ROUGE (1, 2, L)**, **BLEU**, **METEOR**.

4 Task 2 (Experiments):

4.1 Task 2.1 Findings Generation (Image → Findings): [10 marks]

Fine-tune any image-to-text model to generate findings from the chest X-ray.

- Models: Experiment with at least 3 different architectures (e.g., BLIP, GIT, CNN+Transformer).
- Training: Report hyperparameters (learning rate, batch size, optimizer, epochs).
- Evaluation: **ROUGE (1, 2, L)**, **BLEU**, **METEOR**.

4.2 Task 2.2 Impression Generation (Findings → Impression): [10 marks]

Fine-tune sequence-to-sequence models to generate the impression from the findings.

- Models: Experiment with at least 5 model configurations (e.g., T5, BART, Pegasus, LLaMA-based fine-tuning).
- Evaluation: **ROUGE (1, 2, L)**, **BLEU**, **METEOR**.

5 Task 3: Multi-Modal Impression Generation (Image + Findings → Impression) [15 marks]

Fine-tune image-text-to-text models for generating an impression using both image + findings.

- Models: Experiment with at least 5 model configurations
- Evaluation: **ROUGE (1, 2, L)**, **BLEU**, **METEOR**.

6 Comparison and Analysis: [25 marks]

6.1 Quantitative Comparison

Create a summary table comparing ROUGE/BLEU/METEOR/clinical term accuracy across:

- Task 1 (Baseline)
- Task 2.1 + Task 2.2 (Two-stage pipeline)
- Task 3 (Multi-modal model)

6.2 Qualitative Comparison

Select 5–10 representative case studies and show:

- Chest X-ray image
- Ground truth findings + impression
- Model outputs for all three tasks

Analyze which model produces the most clinically accurate impression and which errors could be harmful in a real clinical workflow.

7 Analysis and Report: [5 marks]

1. Analyse which of the methods gives a better score & output on the test set and why?
2. Compare 3 methods and elaborate your findings using graphs and visualizations.
3. Discuss whether combining image + findings reduces error rate.

8 Viva: [30 Marks]

Good Luck