MedTeller: Vision-Language Transformer for Automated Radiology Report Generation

Team #2

Members:

Harsh Shrishrimal [121331366]

Nikhil Roy [117936216]

Priyam Chhonkar [121118301]

Rishabh Ranka [121084965]

Rishika Thakre [121331403]

Deep Learning Project Proposal

University of Maryland

October 16, 2025

1. Background

Radiology is a cornerstone of modern medical diagnosis, yet the process of interpreting medical images and composing reports is time-consuming and prone to human variability. Radiologists often review hundreds of X-ray images per day, leading to fatigue and inconsistencies in reporting. Recent advancements in deep learning, particularly in **Transformer architectures**, have demonstrated superior performance in both computer vision and natural language generation tasks.

By integrating image understanding (Vision Transformers) and language generation (GPT-style models), it is now possible to build a system that can automatically generate clinically coherent radiology reports. Such systems can support radiologists by providing draft reports, flagging abnormalities, and ensuring consistency in clinical documentation.

2. Significance and Problem Statement

Traditional AI models for medical imaging are task-specific, focusing on single-disease classification (e.g., pneumonia detection). However, real-world clinical practice requires **com-**

prehensive report generation that describes multiple findings, impressions, and contextual observations.

3. Problem Statement

Develop a Vision-Language Transformer system, named **MedTeller**, that takes a chest X-ray as input and generates a complete radiology report as output. The model should identify anatomical features, detect abnormalities, and produce medically accurate descriptions similar to those written by human radiologists.

3. Training Data Acquisition

We plan to use publicly available medical datasets containing paired X-ray images and textual reports:

- **IU X-Ray Dataset (Indiana University)** – 7,470 chest X-ray images paired with 3,955 radiology reports. Available via the NIH OpenI repository and Kaggle.

Preprocessing Steps:

- Resize all images to 224×224 and normalize intensity values.
- Extract "Findings" and "Impression" sections from each report.
- Tokenize textual reports using Byte Pair Encoding (BPE) or WordPiece tokenizer.
- Split data into training (80%), validation (10%), and test (10%) subsets.

4. Deep Learning Framework and Model Architecture

We will implement our model using the **PyTorch** deep learning framework and the **Hugging Face Transformers** library.

Architecture Overview

- **Vision Encoder:** Pretrained **Vision Transformer (ViT-Base)** for extracting patch embeddings from chest X-rays.
- **Text Decoder: GPT-2** or **BART Transformer** fine-tuned on medical text to generate radiology-style language.
- **Cross-Attention Layer:** Connects image features to text tokens, ensuring alignment between visual patterns and generated sentences.
- **scikit-learn / NLTK**: For metric evaluation and scoring

Training Objective:

Minimize cross-entropy loss for report generation, with optional auxiliary loss for disease tag prediction. AdamW optimizer will be used with a learning rate scheduler for stability.

5. Validation and Verification Metrics

We will evaluate the generated reports using both **textual similarity metrics** and **clinical correctness scores**:

- **BLEU-1/2/4** – Measures n-gram overlap between generated and reference reports.
- **ROUGE-L** – Captures recall-oriented structural similarity.
- **METEOR** – Considers synonym and semantic alignment.
- **CheXbert Score / RadGraph F1** – Evaluates medical accuracy and disease mention consistency.

Additionally, qualitative verification will include:

- Visualizing cross-attention heatmaps showing regions influencing each sentence.
- Manual inspection by human evaluators for fluency and accuracy.

6. Expected Outcomes

- A trained Vision-Language Transformer capable of generating radiology reports from unseen X-rays.
- Achieve BLEU-4 score > 0.25 and ROUGE-L > 0.3, consistent with current benchmarks.
- A web-based demo (built with Streamlit) allowing users to upload an X-ray and receive an auto-generated report.

7. Impact:

MedTeller can reduce reporting time by **40–50%,** assist radiologists in preliminary diagnosis, and improve clinical documentation efficiency. The project also lays groundwork for future multimodal medical AI systems capable of handling CT or MRI scans.