Multimodal Medical Assistant — Prototype (HPPCS[04])

Abstract

This project presents a prototype Multimodal Medical Assistant that integrates clinical text and chest X-ray images to assist healthcare professionals with diagnostic insights. The system fuses natural-language understanding and image feature extraction using lightweight transformer-based encoders. A small projection head aligns the two modalities via contrastive learning. Large Language Models (LLMs) — Flan-T5-Small for explanation generation and DistilBART for diagnostic tagging — produce concise clinical interpretations. Synthetic patient cases were generated to simulate real-world multimodal inputs. The prototype outputs a similarity score, a diagnostic tag, and a brief medical explanation per case, stored as JSON conversations. Results demonstrate the feasibility of multimodal reasoning even on CPU/GPU-limited environments using open-source models. This work serves as a foundation for future fine-tuning on clinical datasets to achieve higher diagnostic accuracy and reliability.

1. Introduction

Medical diagnostics increasingly require intelligent systems capable of combining visual and textual data. Radiologists often correlate imaging results with patient notes, but such interpretation is time-consuming. This project explores a fusion-based AI assistant that processes both clinical notes and medical images to provide interpretable outputs. The motivation arises from the growing accessibility of open-source multimodal transformers (e.g., CLIP, MiniLM) and LLMs capable of summarization and reasoning. Leveraging these advances enables creation of assistive diagnostic tools even on modest computational resources.

2. Problem Statement

Existing AI diagnostic systems are typically unimodal — either image-based or text-based. There is a lack of lightweight frameworks that can jointly interpret textual and imaging data for medical reasoning while maintaining explainability. The problem addressed here is designing a compact, interpretable multimodal pipeline that can correlate chest X-rays with clinical descriptions to suggest diagnostic insights.

3. Objectives

- • Develop an end-to-end multimodal pipeline combining medical text and image data.
- • Implement and align text and image embeddings using contrastive projection.
- • Generate human-readable explanations and diagnostic tags using LLMs.
- • Demonstrate feasibility on synthetic datasets simulating clinical inputs.
- • Produce structured outputs and an automated project report in compliance with Capstone submission standards.

4. Methodology

Tools & Technologies: Python 3.11, PyTorch, Transformers v4, Sentence-Transformers, ReportLab. Models: sentence-transformers/all-MiniLM-L6-v2 (text), openai/clip-vit-base-patch32 (image), google/flan-t5-small (explanation), and sshleifer/distilbart-cnn-12-6 (tagging). Environment: Kaggle Notebook (GPU/CPU fallback).

Workflow: (1) Data Generation: Synthetic pairs of clinical notes and sample X-rays. (2) Feature Extraction: Text embeddings via MiniLM; image embeddings via CLIP. (3) Fusion Module: Projection heads trained with InfoNCE contrastive loss. (4) LLM Reasoning: Flan-T5-Small produces explanations; DistilBART yields diagnostic tags. (5) Reporting: Results saved as JSONs and summarized automatically into Report.pdf.

5. System Design / Implementation

Architecture Overview:

[Clinical Note] → [MiniLM Encoder] → [Projection Head + Similarity Computation] → [Fusion Score]

[Chest X-ray] → [CLIP Encoder] → [Projection Head] → [Fusion Score → Flan-T5 (Explanation), DistilBART (Tag)]

Modules: text_pipeline.py (text embedding), image_pipeline.py (image encoding), fusion_pipeline.py (fusion and similarity), main.py (orchestration and reporting), demo_data.py (synthetic data). All modules reside in Codebase/ to ensure compatibility with Capstone evaluation.

6. Results and Analysis

Five demo cases were processed, generating JSON conversation and report summary files. Key diagnostic results are summarized below:

Case	Similarity	Diagnostic Tag	Key Explanation
1	-0.011	Pneumonia	Focal consolidation consistent with lobar pneumonia.
2	-0.068	COPD	Hyperinflation and chronic changes suggestive of COPD.
3	-0.051	Pneumothorax	Small apical pneumothorax; recommend urgent evaluation.
4	-0.090	Pulmonary Embolism	Non-specific on X-ray; suggest CT angiography.
5	0.056	CHF	Cardiomegaly and congestion indicating heart failure.

The system successfully combined multimodal features to produce interpretable diagnostic summaries. Training executed efficiently on Kaggle GPU within 6 epochs. Similarity values were near zero due to synthetic data, which is expected for a prototype.

7. Conclusion and Future Work

This prototype demonstrates that multimodal fusion with open-source LLMs can generate concise, clinically relevant summaries from text and image data. Contributions include modular integration of MiniLM, CLIP, and LLMs with automated reporting. Future work involves fine-tuning encoders on real paired datasets, integrating attention-based fusion, and validating outputs with clinician feedback.

9. References

1. Radford A. et al., 'Learning Transferable Visual Models from Natural Language Supervision', ICML 2021 (CLIP).

2. Reimers N. et al., 'Sentence-Transformers: Sentence Embeddings using Siamese BERT-Networks', 2020.

3. Raffel C. et al., 'Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer', JMLR 2020 (T5).

4. Hugging Face Transformers Documentation (https://huggingface.co/docs).

5. PhysioNet MIMIC-CXR Database (https://physionet.org/content/mimic-cxr/).

Acknowledgement

I would like to express sincere gratitude to the course instructors and evaluators for their guidance, and to the open-source AI community for providing models and tools that made this project feasible. Special thanks to Kaggle for enabling accessible GPU-based experimentation.