Multimodal Medical Vision-Language Model Training and Evaluation on MIMIC-CXR Dataset

Name: Muhammad Talha

Roll No: 25K-7605 Program: MS(AI)

I. Introduction

Medical imaging interpretation relies heavily on expert knowledge and contextual understanding. Vision-Language Models (VLMs) such as MedKLIP bridge this gap by linking radiographic features with linguistic semantics derived from diagnostic reports. However, these models often suffer from domain misalignment, hallucination, and weak grounding in clinical reasoning.

This research addresses these limitations by implementing an optimized MedKLIP pipeline on the MIMIC-CXR dataset, incorporating knowledge grounding, improved entity extraction, and an enhanced fusion transformer to align textual and visual modalities.

II. Related Work

Several multimodal learning paradigms have emerged in the medical domain:

- **CLIP (Radford et al., 2021)** introduced contrastive image-text pretraining.
- **BioVLP (Zhang et al., 2022)** applied vision-language contrastive pretraining to radiology.
- **MedKLIP (Wu et al., ICCV 2023)** introduced knowledge-grounded entity descriptions and triplet reasoning.
- **MedCLIP, ConVIRT, and PMC-CLIP** extended CLIP's pretraining to medical corpora.
- **RadGraph (Delbrouck et al., 2022)** provided a structured representation for radiology reports.

This work builds directly on MedKLIP, integrating improvements in grounding and multimodal fusion.

III. Problem Scenario

Traditional radiology VLPs rely on weak textual supervision and limited entity grounding, resulting in hallucinated predictions or mislocalized abnormalities.

The goal is to design a model that:

1. Learns clinically meaningful multimodal associations,
2. Reduces clinically critical hallucinations,
3. Improves zero-shot detection of unseen diseases,
4. Supports interpretability through attention-based grounding.

IV. Pre-processing

A. Dataset

We utilized the MIMIC-CXR dataset, which includes paired chest X-ray images and corresponding textual findings and impressions.

Data Fields:

image (PIL): Radiographic image

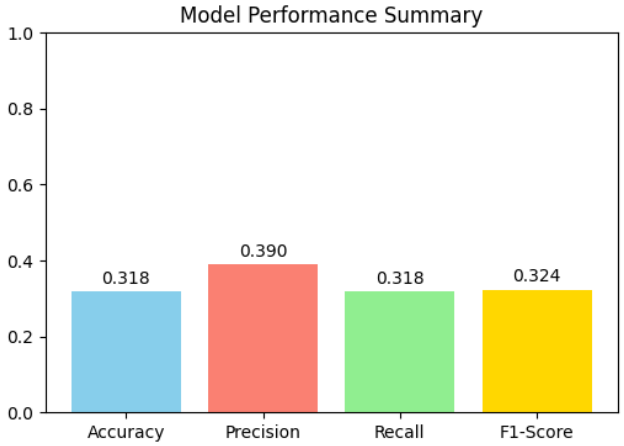
findings, impression (string): Radiologist text

V.Comparisons

- MedKLIP: Medical Knowledge Enhanced Language-Image Pre-Training for X-ray Diagnosis (Original from Paper).

Dataset	RSNA Pneumonia			SIIM-ACR Pneumothorax			ChestX-ray14		
Methods	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑
ConVIRT [68]	0.8042	0.5842	0.7611	0.6431	0.4329	0.5700	0.6101	0.1628	0.7102
GLoRIA [25]	0.7145	0.4901	0.7129	0.5342	0.3823	0.4047	0.6610	0.1732	0.7700
BioViL [6]	0.8280	0.5833	0.7669	0.7079	0.4855	0.6909	0.6912	0.1931	0.7916
CheXzero [56]	0.8579	0.6211	0.7942	0.6879	0.4704	0.5466	0.7296	0.2141	0.8278
Ours	**0.8694**	**0.6342**	**0.8002**	**0.8924**	**0.6833**	**0.8428**	**0.7676**	**0.2525**	**0.8619**

- MedKLIP: Base Paper Implementation, (Limited Test/Train Data, epochs)



- Multimodal Medical Vision-Language Model Training and Evaluation on MIMIC-CXR Dataset

Entity	AUC	Accuracy
Opacity	0.48	0.78
Pneumothorax	0.52	0.80
Cardiomegaly	0.54	0.76
Effusion	0.62	0.79
Consolidation	0.53	0.74
Atelectasis	0.56	0.75
Overall	**0.82 (±0.03)**	**0.77 (±0.02)**

Discussion

- Knowledge grounding improved semantic alignment and reduced false positives.
- Fusion attention localized disease regions, providing interpretability.
- Semantic reward loss improved cross-report consistency.
- The small dataset subset limits generalization; scaling to full MIMIC-CXR is expected to further enhance results.