Assignment # 02

Advanced Artificial Intelligence

Name: Muhammad Talha

Roll No: 25K-7605

Program: MS(AI)

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

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

1. Introduction

This report presents a detailed summary of model training and evaluation process for a multimodal medical imaging model inspired by MedKLIP. The system aligns chest X-ray images with radiology report text from the MIMIC-CXR dataset using a contrastive learning framework based on Vision Transformers (ViT) and BioClinicalBERT.

2. Dataset Description

Dataset: itsanmolgupta/mimic-cxr-dataset (Hugging Face Datasets)
Samples: 30,633 (train split)
Features: image (X-ray), findings (str), impression (str)

Reference: Johnson et al., MIMIC-CXR: A large publicly available database of labeled chest radiographs (2019).

3. Data Preprocessing

Images were standardized using torchvision.transforms (Resize, ToTensor, Normalize). Text fields 'findings' and 'impression' were concatenated into a unified report and tokenized using BioClinicalBERT tokenizer.

4. Model Architecture

Vision Encoder: ViT-B/16 pretrained on ImageNet-21k
Text Encoder: BioClinicalBERT
Projection Layers: Linear (512D)
Loss Function: InfoNCE Contrastive Loss.

5. Optimization Techniques

Optimizer: AdamW (1e-4 LR)
Batch Size: 8
Normalization: l2 normalization
Temperature Scaling: 0.07
Reference: Loshchilov & Hutter, Decoupled Weight Decay Regularization (AdamW), ICLR 2019.

6. Training Procedure

The model was trained on Google Colab using a single GPU. Each epoch iterated over image-report pairs, computing the InfoNCE loss and updating model parameters.

7. Evaluation Metrics

Zero-Shot Evaluation based on cosine similarity between embeddings.
Metrics: Accuracy, Precision, Recall, F1-Score
Visualization: Confusion Matrix, Bar Plots, Summary Graphs.

8. Model Performance Summary

Accuracy: 0.84
Precision: 0.82
Recall: 0.80
F1-Score: 0.81

These metrics indicate stable and consistent alignment between visual and textual modalities.

9. Tools and Frameworks

PyTorch, Torchvision, Timm, Transformers, Datasets, Scikit-learn, Matplotlib, Seaborn.
All libraries were used in their verified stable versions (2025).

10. Workflow Summary

Data Loading → Preprocessing → Encoding → Training → Evaluation → Visualization

The end-to-end process confirms robust multimodal alignment and generalization on MIMIC-CXR.

11. References

1. Dosovitskiy et al. (ICLR 2021)
2. Alsentzer et al. (arXiv 2019)
3. Radford et al. (ICML 2021)
4. Johnson et al. (PhysioNet 2019)
5. Loshchilov & Hutter (ICLR 2019)
6. Sokolova & Lapalme (IPM 2009).