Statement of Purpose: Region-Aware MedCLIP for Medical Image–Text Understanding

1. Project Details

Project Title: Region-Aware MedCLIP for Medical Image–Text Understanding
Code Repo Link: To be shared upon completion
If Own Idea: No

2. Problem Statement

Traditional multimodal models such as CLIP only align whole images with text, limiting their effectiveness in domains like medical imaging where fine-grained regional details (e.g., lesions, tumors) are critical for diagnosis. Our goal is to design a region-aware CLIP variant for medical imaging that jointly aligns global image–report pairs and region-of-interest (ROI)–finding pairs. This will enhance performance in zero-shot classification, cross-modal retrieval, and diagnostic report understanding.

3. Methodology

- **Base Model:** Pretrained CLIP (ViT backbone).
- **Region Extraction:** Use pretrained segmentation models (e.g., U-Net, CheXDet) to extract lung or lesion regions.
- **Text Processing:** Decompose reports into multi-level text units (disease labels, short findings, full reports).
- **Alignment:**
  - Global alignment: image $\leftrightarrow$ full report.
  - Regional alignment: ROI $\leftrightarrow$ phrase-level finding.
- **Training:** Contrastive loss inspired by RegionCLIP (CVPR 2022) for both global and regional pairs.
- **Evaluation:** Compare against baseline CLIP and MedCLIP on retrieval accuracy and classification performance.

4. Dataset Details

We will use publicly available chest X-ray datasets:

- **MIMIC-CXR:** 377,110 chest X-rays with corresponding radiology reports.
- **CheXpert:** 224,316 chest radiographs with 14 labeled observations.

Preprocessing includes normalization, ROI extraction, and report parsing. These datasets are well-suited due to their scale and detailed report annotations.

5. Required Resources

- **Hardware:** 16 gb ram rtx 4050 4gb-6gb.
- **Software:** Python, PyTorch, HuggingFace Transformers, OpenAI CLIP.
- **Tools:** Pretrained segmentation models (U-Net, CheXDet).

6. Novelty of Approach

Our novelty lies in extending MedCLIP with **region-aware alignment**,implement the given architure in the paper and also borrowing from RegionCLIP (CVPR 2022). Unlike global-only alignment, we explicitly align ROIs with localized clinical findings, enabling fine-grained reasoning in medical imaging. This adaptation introduces region–text contrastive supervision into medical vision–language learning, a capability unexplored in current MedCLIP variants.

7. Team Composition and Contributions

- Member 1: [Shashank Yadav, 12342010] – all work will be done by me

8. Expected Outcomes

- A trained region-aware MedCLIP model
- Visualization of attention/ROI grounding in medical images.
- A research-style report documenting findings.

9. References

- Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," ICML 2021.
- Zhong et al., "RegionCLIP: Region-based Language-Image Pretraining," CVPR 2022.
- Johnson et al., "MIMIC-CXR: A Large Public Dataset of Chest Radiographs," arXiv 2019.

- Irvin et al., "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels," AAAI 2019.