Quick Summer Summary 😅

I read an article from Korean Journal of Radiology which had some nice things to say about multimodal models on radiology and health care.

The issues it mostly talks about were :

- achieving generalizability of AI models
- establishing the explainability of the decision-making process

-> https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10613849/

> To overcome these issues, some of the various possible solutions are as follows:
>
> inclusion of training with longitudinal and multimodal datasets, dense training with multitask and multimodal learning, new generative models including anomaly detection, XAI
>
> --- From Summary and conclusion

I Mostly read on Collecting data and tokenization , since that will be our first steps.

For Data:

- Medpix2.0
  - 12,000 patient case scenarios,9,000 topics, and nearly 59,000 images.
  - But Far less than what 4m pretrains on which is 12M (CC12M).
  - But Medpix claims : `CLIP succeeds in achieving competitive performance in zero-shotcontexts on a wide range of classification datasets by learning the relationshipsbetween images and their textual descriptions` entirely trained on their own dataset.

For tokenization

- Here 4m does different than almost every other paper
- 4m has 1 unified embedding for all modalities
- But Llama , Medpix and others have different image and text encoder. [so as many embeddings and cross attention as modalities]
- Having different embedding would allow us to use specailized tokenizer like BioBert.[https://doi.org/10.1093/bioinformatics/btz682, arXiv:1901.08746 [cs]]
- But 1 unified embedding would be great to plot after pretraining
- So I am not really sure of what tokenization technique i should use for now.

Extra Ideas

Just so that its not complete replication of any 1 paper. I was thinking of making some changes by taking pieces of others architectures so that i could do ablation studies [Thats a research gap i think].

- like 4m does not use grouped query attention (which is supposedly faster with same resoulution as Multihead attention) (unlike Llama 3.1).
- 4m does a lot of masking (Massively Masked Multiple Modality) but does not talk about efficiency of this masking. A lot of other paper CLIP [https://arxiv.org/pdf/2103.00020] and SIGLIP [https://arxiv.org/abs/2303.15343] which newer models are trained on ; go in a lot of detail.
  - I did some triton and cuda programming over the summer . Maybe i can try writing their code in cuda/triton to show if their methods are as efficient and fast as others or not. (But this seems slightly out of scope for the internship plan we submitted!)