Research Internship Notes

Chang

her email

me

my email

A compilation of approaches; there are some claims from aritcles/paper that i have not been able to reproduce yet or seems contradicting

1. Dataset Generation

1.1. Strategy of Dataset Generation

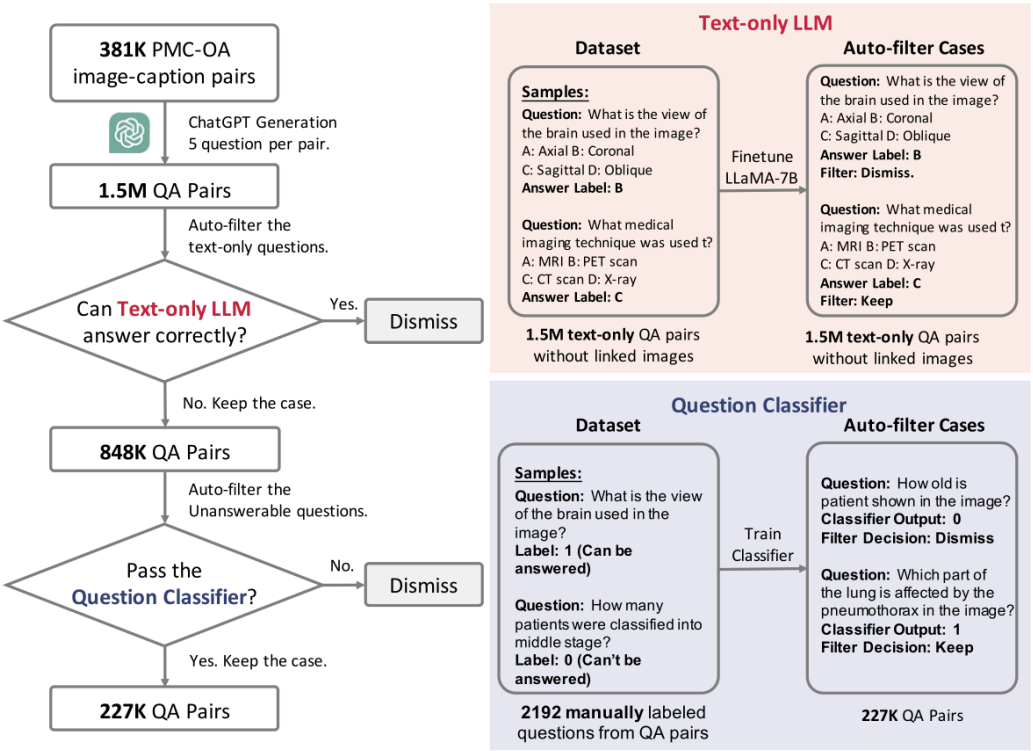

Figure 4 | The whole flowchart demonstrating how we build up our PMC-VQA dataset. In left, we show the general progress and in right we show how we build up the two auto-filter models used in our data collection.

Figure 1: Got the idea for generating QA pairs form here

- this has been trained on larger volume.
- I pass in the prompt to deal with filtering unanswerable questions from Text only llm
- current modified prompt : (subject has information about plain modality and photographed region)

```py
prompt = f"""
Based on the following medical image caption and case information,
Caption Information:
{caption}
Plane and Location Information of the Image:
{subject}
Generate question-answer pairs (exhaustive of the information given).
Assume I am going to use this to train a Visual Question Answering model for a medical dataset.
Keep the QA pairs such that they can only be answered when the image is in context.
Do not ask questions about measurements (avoid numericals in question answer pairs) , history of patient , or if information is unavailable.
Question on 1 line and answer on the new line. Please! Don't use any filler text.
"""
```

- Their prompt (costlier per example) generates MCQ Based

```py
prompt = f"""
Ask 5 questions about the content and generate four options for each question. The questions should be
answerable with the information provided in the caption, and the four options should include one correct
and three incorrect options, with the position of the correct option randomized. The output should use
the following template: i:'the question index' question:'the generate question' choice: 'A:option content
B:option content C:option content D:option content' answer: The correct option(A\B\C\D).
"""
```

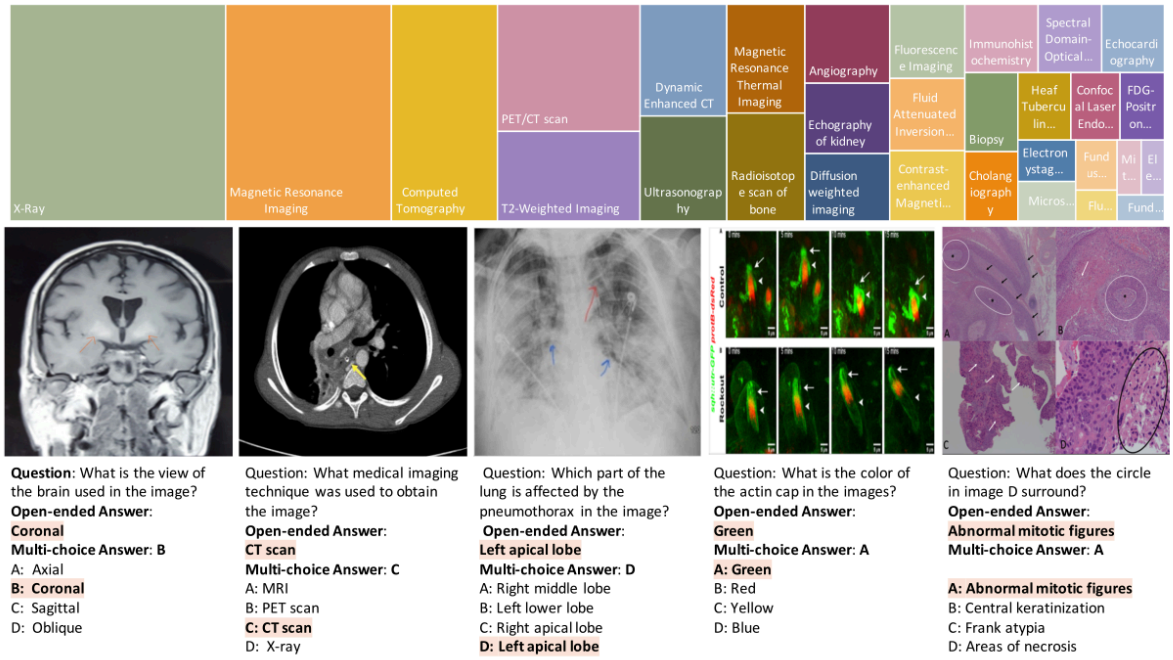

Figure 1 | (a) Several examples of challenging questions and answers along with their respective images. To answer questions related to these images, the network must acquire sufficient medical knowledge, for example, for the first two images, it is essential to recognize the anatomy structure and modalities; for the third image, recognizing the X-ray image pattern of pathologies is necessary; for the final two images, apart from the basic biomedical knowledge, the model is also required to discern colors, differentiate subfigures, and perform Optical Character Recognition (OCR). (b) The top 20 figure types in PMC-VQA, cover a wide range of diagnostic procedures.

Figure 2: Exemplar Question and dataset Distribution of their prompt!

Figure 3: [Left] Diversity in Medpix

Figure 4: sample image

"caption": "The prostate is enlarged with several calcifications noted within. No dominant prostate mass is evident.", "subject": "CT - noncontrast • Coronal • Genitourinary • Reproductive and Urinary System"

Q: Is the prostate enlarged? **A**: Yes **Q**: Are there calcifications noted within the prostate? **A**: Yes **Q**: Is there a dominant prostate mass evident? **A**: No **Q**: What type of imaging modality was used to obtain this image? **A**: CT - noncontrast **Q**: What plane and location is the image in? **A**: Coronal, Genitourinary, Reproductive and Urinary System"

Figure 5: QA Pairs form our prompt

2. Modelling

2.1. MLLM's

> We prove that medical LLM should be first pretrained with domain corpus, and then tuned with instructions following dataset.
>
> – https://github.com/chaoyi-wu/PMC-LLaMA

• This is too too heavy needs multi gpu even for inference https://huggingface.co/katielink/llava-med-7b-vqarad-delta https://huggingface.co/photonmz/llava-roco-8bit

• we tried on paligemma 224 and paligemma 448. https://huggingface.co/google/paligemma-3b-pt-224

• This is famous for transfer task

• But not pretrained on medical corpus

2.2. Based on Clip

https://huggingface.co/kaushalya/medclip

• training script :

```sh
python src/medclip/run_medclip.py \
    --output_dir ./snapshots/vision_augmented_biobert \
    --text_model_name_or_path="allenai/scibert_scivocab_uncased" \
    --vision_model_name_or_path="openai/clip-vit-base-patch32" \
    --tokenizer_name="allenai/scibert_scivocab_uncased" \
    --train_file="data/train_dataset.json" \
    --validation_file="data/valid_dataset.json" \
    --do_train --do_eval \
    --num_train_epochs="40" --max_seq_length 128 \
    --per_device_train_batch_size="64" \
    --per_device_eval_batch_size="64" \
    --learning_rate="5e-5" --warmup_steps="0" --weight_decay 0.1 \
    --overwrite_output_dir \
    --preprocessing_num_workers 32 \
#    --push_to_hub
```

• Finetunes on ROCO Dataset https://huggingface.co/datasets/MedIR/roco?row=33 with clip vit base [not pretrained on medical corpus]

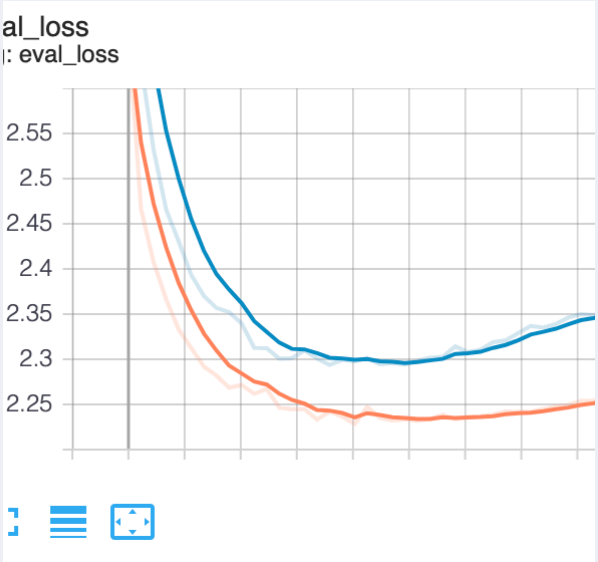

Figure 6: Their Loss on ROCO

• Their model did not have any evaluation metrics

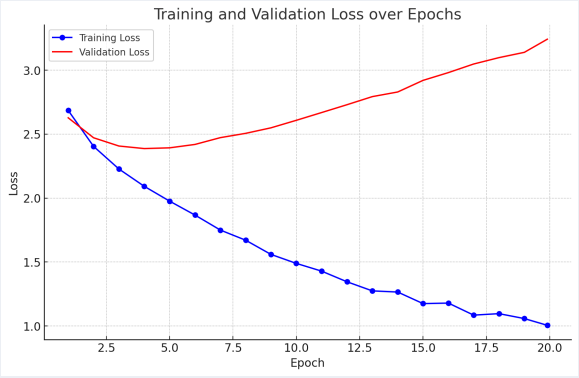

Figure 7: Our loss on Medpix with Paligemma 448 on dataset https://huggingface.co/datasets/adishourya/MEDPIX-ShortQA

2.3. Based on Blip

```
######################################################################
Question:  what does this image show?
Predicted Answer:  rocky mountain
Actual Answer:  typical excellent pinworm
```

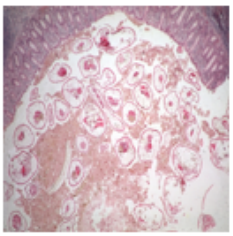

Figure 8: This was trained on pathalogy images

- This is a popular notebook [trains on short QA] sometimes gives bad answers
  https://www.kaggle.com/code/basu369victor/blip-medical-visual-question-answering
- Based on llama :
  https://github.com/aldraus/quilt-llava

2.4. Runs :

- full Roco on paligemma
  - does not work then:
    - delete the full vision tower
    - https://github.com/photomz/BabyDoctor
    - https://huggingface.co/photonmz/llava-roco-8bit
    - Full Roco Dataset : https://huggingface.co/datasets/mdwiratathya/ROCO-radiology/viewer/default/train?p=0
    - Roco Instrict : photonmz/roco-instruct-65k
- when trained from scratch
  - some instruct dataset : liuhaotian/LLaVA-Instruct-150K
  - visual : cc12m
  - medical : photonmz/roco-instruct-65k
- medpix on llava med

|<End of Document>|