Finetune a Multimodal Model for X-Ray Radiology Report Generation

Task Description

Deep learning has revolutionized traditional medical image analysis tasks such as image segmentation, classification, and detection. However, generating radiology reports remains an unsolved yet important clinical task. Recently, multimodal and LLM models have made significant improvements, which has shown great potential for automatic report generation. This automation can not only can enhance diagnostic accuracy and reduce human error but also speed up the reporting process, allowing radiologists to focus on more critical tasks which inturn improves overall patient care.

In this quiz, we provide 2900+ chest X-Ray images from the public IU X-Ray dataset for model development and validation. The quiz aims to test two important skills on LLM and multimodal models: prompt engineering and efficient fine-tuning.

Task 1. Prompt engineering: reorganize the X-Ray report findings into predefined anatomical regions

Please write a prompt to use LLaMA or GPT4 to separate the findings on the **validation set (296 patients)** into the four predifined anatomical regions: lung, heart, mediastinal, and bone. If the model cannot assign the sentence to any anatomical region, please put it in others. Here is an example:

- Input: a report of a typical chest X-Ray radiology findings.

> The cardiomediastinal silhouette and pulmonary vasculature are within normal limits in size. The lungs are mildly hypoinflated but grossly clear of focal airspace disease, pneumothorax, or pleural effusion. There are mild degenerative endplate changes in the thoracic spine. There are no acute bony findings.

- Expected output:

```
{
    "lung": "Lungs are mildly hypoinflated but grossly clear of focal airspace disease, pneumothorax, or pleural effusion. Pulmonary vasculature are within normal limits in size.",
    "heart": "Cardiac silhouette within normal limits in size.",
    "mediastinal": "Mediastinal contours within normal limits in size.",
    "bone": "Mild degenerative endplate changes in the thoracic spine. No acute bony findings.",
    "others": ""
}
```

> Note: We have re-organized the findings in the training and testing sets.

Task 2. Efficient Model Fine-tuning

Please fine-tune the well-known LLaVA-Next on the provided training set for report generation. Candidates should use parameter efficient fine-tuning methods (e.g., LoRA) instead of fully fine-tuning for better compute efficiency.

Notes:

- The objective of this task is to test the applicants' ability of using advanced code repositories without costing too much compute. Thus, the task is designed to be done on freely provided compute resources, such as Google Colab. If you have access to better computing, please feel free to train the larger models.
- Different LLMs can be easily incorporated into LLaVA's framework. You can also use smaller LLMs instead of the default 7B+ LLMs.
- LLaVA-NeXT blog: Improved reasoning, OCR, and world knowledge
- The official codebase does not provide the training scripts but there are some third-party implementations. We leave this as a simple searching task.
- Useful package: https://huggingface.co/docs/peft/en/index

Dataset Folder Structure

```
data/
├── annotation.json # includes data split, patient id and report findings. Ignore the findings in "others"
└── images # each patient contains 1-4 images
    ├── CXR1000_IM-0003
        ├── 0.png
        ├── 1.png
        └── 2.png
    ├── ...
    └── CXR9_IM-2407
        ├── 0.png
        └── 1.png
```

Evaluation Metric

GREEN (Generative Radiology Report Evaluation and Error Notation) score is used to evaluate the report generation results. Please refer to the paper for more details. Candidates should compute the the GREEN metric for each anatomical region (if the corresponding report findings are available in ground truth).

Submission: Report and Code

Please submit a report that describes the model developments and evaluaiton results (on the testing set) for each anatomical regions (lung, heart, mediastinal, and bone). The report should also include a Github link for reproducing the results. Please follow this checklist to prepare the Github.

Result Table format: Average GREEN scores on the validation and testing set.

Data Split	Lung	Heart	Mediastinal	Bone
Validation				

Data Split	Lung	Heart	Mediastinal	Bone
Testing				

Additional Resources

Here are some existing studies on multimodal medical AI:

- MedGemini: https://research.google/blog/advancing-medical-ai-with-med-gemini/
- Consensus, dissensus and synergy between clinicians and specialist foundation models in radiology report generation: https://arxiv.org/abs/2311.18260
- RadFM: https://chaoyi-wu.github.io/RadFM/
- M3D: https://github.com/BAAI-DCAI/M3D