Finetune a Multimodal Model for X-Ray Radiology Report Generation

Adibvafa Fallahpour – WangLab

This technical report summarizes the procedures used to reorganize the X-Ray report findings into predefined anatomical regions and to efficiently finetune a LLaVA-Next[2] model on IU X-Ray dataset[1].

GitHub Repository: https://github.com/Adibvafa/WangLab-OA

TASK1

The objective of this task was to break down a radiology report into findings of lung, heart, mediastinal, bone, and others. Here, the OpenAI API for "gpt-4o-2024-08-06" model with structured outputs was used. The prompt given to model was:

``` *You are an expert at structured data extraction and medicine. You will be given unstructured findings of a x-ray radiology report and should separate the findings into four predefined anatomical regions: lung, heart, mediastinal, and bone. If you cannot assign the sentence to any anatomical region, put it in others.* ```

Inference on the 296 datapoints of the validation dataset was completed in about 6 minutes.

TASK2

This task aimed to finetune the LLaVA-Next model (llava-v1.6-mistral-7b-hf)[2] on the IU X-Ray dataset[1]. The main libraries used were PyTorch Lightning, Transformers, Deepspeed, Peft, and GREEN. The report for this task is broken down into 3 parts.

Dataset

A PyTorch dataset class was developed which would load the radiology report and 2 images associated with each datapoint. The report was prepared in a conversation-style structure suitable for finetuning LLaVA-Next[2]. The user prompt associated with each datapoint was ```*Generate an X-ray report for Lung, Heart, Mediastinal, and Bone.*``` A collate function was developed to create a list of prompts and images of a given batch, and use Hugging Face's processor for LLaVA-Next to tokenize input report, preprocess images, and add labels to feed into the model.

Finetune

To enable efficient finetuning, only the linear layers of the model were finetuned in a quantized (4-bit) setting using QLoRA. The chosen parameters were about 0.3% of all model parameters. A LLaVA Finetuner class was developed using PyTorch Lightning that would load and train the model. The AdamW optimizer with cosine schedule was used which bumped learning rate to 5e-5 in the first 10% of training and decreased it in the next 90%. The training ran for 4 epochs and the model with lowest validation loss from epoch 3 was chosen. Training strategy was Deepspeed stage2 and used 1 NVIDIA A100-80GB GPU with a gradient accumulation of 8 and batch size of 1. Further details are provided on `*finetune.py*`.

Inference

To evaluate the model, radiology reports were generated for validation and test datasets. Then, the reports were broken down into findings of each anatomical region using regex, and evaluated with the GREEN package[3]. The resulting evaluation along with the generated reports were saved to disk in the `*results*` directory and a summary of results is provided.

Data Split	Lung	Heart	Mediastinal	Bone
Test	0.70	0.89	0.66	0.32
Valid	0.69	0.72	0.50	0.31

References

1. V. Kougia, J. Pavlopoulos, and I. Androutsopoulos, "A Survey on Biomedical Image Captioning," arXiv preprint arXiv:1905.13302, 2019.

2. H. Liu et al., "LLaVA-NeXT: Improved reasoning, OCR, and world knowledge," Jan. 2024. [Online]. Available: https://llava-vl.github.io/blog/2024-01-30-llava-next/

3. S. Ostmeier et al., "GREEN: Generative Radiology Report Evaluation and Error Notation," arXiv preprint arXiv:2405.03595, 2024.
```