Examining Applicability of MM-LLM-RO to Lung Cancer Segmentation

By: Aadit Juneja (W, SEAS '27)
Department of Radiation Oncology, University of Pennsylvania School of Medicine
Mentors: Sang Ho Lee, Ying Xiao
Program: Penn Undergraduate Research Mentoring Program (PURM)

Penn CURF
CENTER for UNDERGRADUATE RESEARCH & FELLOWSHIPS

Perelman School of Medicine
UNIVERSITY of PENNSYLVANIA

Abstract

Tumor segmentation from CT scans is a task typically performed by trained professionals in practice. However, the advent of vision language models (VLMs) present an alternative, machine-learning based approach to tackling this task. We explored the usage of MM-LLM-RO, a VLM, in the context of volume contouring for lung cancer. Although no state-of-the-art performances were matched, we discovered modest performances after training the model over a span of 1000 epochs, which took approximately 18 hours on 2 NVIDIA A40 GPUs. The training jobs on these NVIDIA GPUs were submitted largely with the SLURM job submission system on Penn Med's CBICA cluster, though were initially submitted with the SGE method before Penn Med completed the migration.

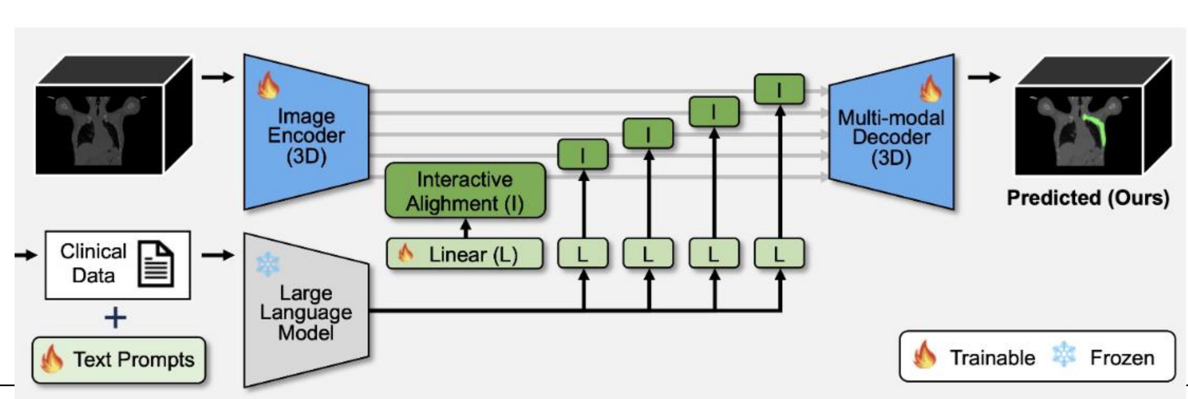
Data Preparation

Data preparation took place largely in two parts: text data preparation and CT scan data preparation. Firstly, text data preparation incorporated cleaning a large patient database including information like age, previous conditions, BMI, and other information that would be included by a doctor in a clinical note. I wrote a Python script to turn this information, which was maintained in a spreadsheet-like format, into a natural language clinical note for each patient. I then wrote a new column into the spreadsheet including this note. For the next step, I had to convert the CT scan files, which were provided in Dicom format, to the Nifti format, which proved to be very challenging. Dicom is a 2D format representing cross sections of the scan, meaning 20-25 Dicom files are required to represent one CT scan. However, the Nifti format is one file that represents a 3D image, and thus is more commonly used for inputs to vision language models such as MM-LLM-RO. The script I wrote to convert the file took approximately 4 seconds per sample to run.

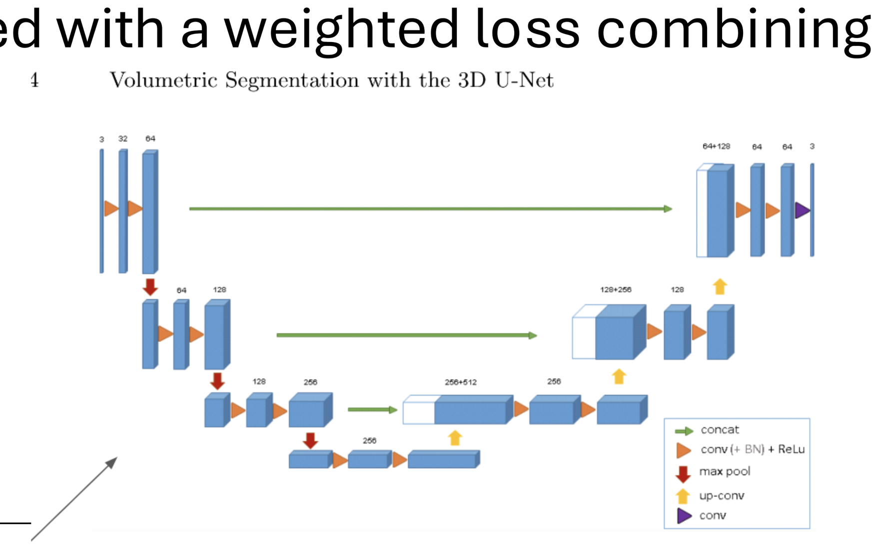
ID	HupMrn	Age	Clinical Note
1	147074469	43.50719	Smoking Pack Years: 62.5,BMI: 28.7,ECOG Performance Status Prior to First RT Infraction:
3	144465112	81.39357	Smoking Pack Years: 0.0,BMI: 36.04,ECOG Performance Status Prior to First RT Infraction:
4	147022085	66.31622	Smoking Pack Years: 45.0,BMI: 17.98,ECOG Performance Status Prior to First RT Infraction:
5	155477327	59.12115	Smoking Pack Years: 42.0,BMI: 19.63,ECOG Performance Status Prior to First RT Infraction:
6	147097234	61.11431	Smoking Pack Years: 0.0,BMI: 27.35,ECOG Performance Status Prior to First RT Infraction:
8	147085046	69.62081	Smoking Pack Years: 20.0,BMI: 14.88,ECOG Performance Status Prior to First RT Infraction:
9	147076253	79.28268	Smoking Pack Years: 0.0,BMI: 27.76,ECOG Performance Status Prior to First RT Infraction:
10	161048153	77.51677	Smoking Pack Years: 20.0,BMI: 24.81,ECOG Performance Status Prior to First RT Infraction:
12	148038616	58.94867	Smoking Pack Years: 40.0,BMI: 26.89,ECOG Performance Status Prior to First RT Infraction:
13	112082186	66.15469	Smoking Pack Years: 40.0,BMI: 40.25,ECOG Performance Status Prior to First RT Infraction:
14	144065392	84.44079	Smoking Pack Years: 50.0,BMI: 25.6,ECOG Performance Status Prior to First RT Infraction:
15	144091029	75.83299	Smoking Pack Years: 17.5,BMI: 21.14,ECOG Performance Status Prior to First RT Infraction:
16	155834980	78.5681	Smoking Pack Years: 5.0,BMI: 19.16,ECOG Performance Status Prior to First RT Infraction:
17	146597038	72.80767	Smoking Pack Years: 0.0,BMI: 25.84,ECOG Performance Status Prior to First RT Infraction:
18	147031181	73.03217	Smoking Pack Years: 104.0,BMI: 24.97,ECOG Performance Status Prior to First RT Infraction
19	14531141	46.99521	Smoking Pack Years: 9.5,BMI: 20.36,ECOG Performance Status Prior to First RT Infraction:
20	144128256	74.00137	Smoking Pack Years: 0.0,BMI: 24.87,ECOG Performance Status Prior to First RT Infraction:

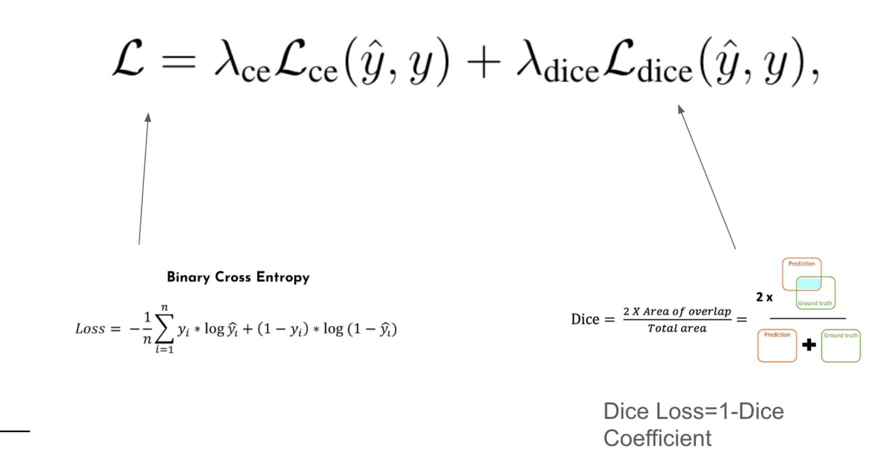
Background and Architecture

In November 2023, a paper titled MM-LLM-RO was published by Oh et. al. The paper introduced a convolutional U-Net with cross-attention mechanisms to output a segmentation mask on a CT scan predicting the location of a tumor. The model first utilizes downsamples the initial CT scan to a lower dimensional space to learn more meaningful features. Simultaneously, a natural language doctors note is converted into a sequence of tokens via a Llama encoder, with learnable artificial tokens prepending the actual sequence via a prompt tuning mechanism. At each downsampling step, the self-attention adjusted text sequence is projected to the same dimensionality as the current dimensionality of the image sequence via a linear layer, after which cross-attention is conducted between the image and text sequences, which are now of the same dimensionality. This is performed until the CT scan representation reaches the "bottom" of the U-Net architecture in its lowest dimensionality. Afterwards, the representation is upsampled back up to original dimensionality, with the initial representation during the downsampling phase pre-attention is concatenated with this representation as part of a residual layer. Once mapped back to the CT scan representation, softmax is applied to arrive at per-voxel probabilities. The model is trained with a weighted loss combining binary cross entropy and a DICE loss.



(c)





Results

ID	Avr	Dice	IoU	H
100	0.888			
129	0.589			
140	0.788			
177	0.845			
183	0.855			
201	0.854			
365	0.723			
366	0.87			
378	0.801			
444	0.272			
468	0.311			
50	0.691			
524	0.74			
542	0.752			
773	0.461			
791	0.557			
792	0.47			
877	0.712			
878	0.728			
882	0.655			

```
Mon Aug  5 06:50:34 2024
epoch, loss, mean, ptv, ctv, gtv
60, 1.025, 0.000019, 0.000
65, 0.982, 0.031315, 0.031
70, 0.935, 0.094633, 0.095
75, 0.917, 0.126797, 0.127
80, 0.852, 0.228411, 0.228
85, 0.828, 0.263805, 0.264
90, 0.792, 0.290622, 0.291
95, 0.830, 0.227073, 0.227
100, 0.835, 0.247232, 0.247
105, 0.693, 0.427618, 0.428
110, 0.755, 0.362875, 0.363
115, 0.819, 0.252012, 0.252
120, 0.706, 0.406319, 0.406
125, 0.706, 0.414761, 0.415
130, 0.729, 0.391443, 0.391
135, 0.712, 0.360358, 0.360
140, 0.701, 0.372080, 0.372
145, 0.734, 0.325806, 0.326
150, 0.659, 0.429564, 0.430
155, 0.639, 0.438234, 0.438
160, 0.672, 0.420778, 0.421
165, 0.685, 0.375816, 0.376
170, 0.623, 0.440226, 0.440
175, 0.679, 0.389830, 0.390
180, 0.658, 0.403954, 0.404
185, 0.632, 0.426750, 0.427
190, 0.628, 0.450067, 0.450
195, 0.587, 0.502042, 0.502
200, 0.537, 0.534581, 0.535
```

The left image appears to represent the average IoU and Dice loss score for each of the images in the validation dataset after the model is trained. The right image is a picture of the loss logs over the model's training horizon, indicating decreasing losses over the 1000 epoch training period. Note that not all 1000 epochs are included in this picture due to sizing limitations.

Acknowledgements

I'd like to thank my mentors, Ying Xiao and Sang Ho Lee, for advising me throughout the project, as well as the entire medical physics research group within the Department. I'd also like to thank my family and friends for supporting me throughout the completion of the project. Lastly, I'd like to thank CURF for the opportunity to conduct such stimulating research to advance my horizons and my intellectual curiosity; the experience was invaluable.