GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-efficient Medical Image Recognition

Shih-Cheng Huang* Liyue Shen* Matthew P. Lungren Serena Yeung

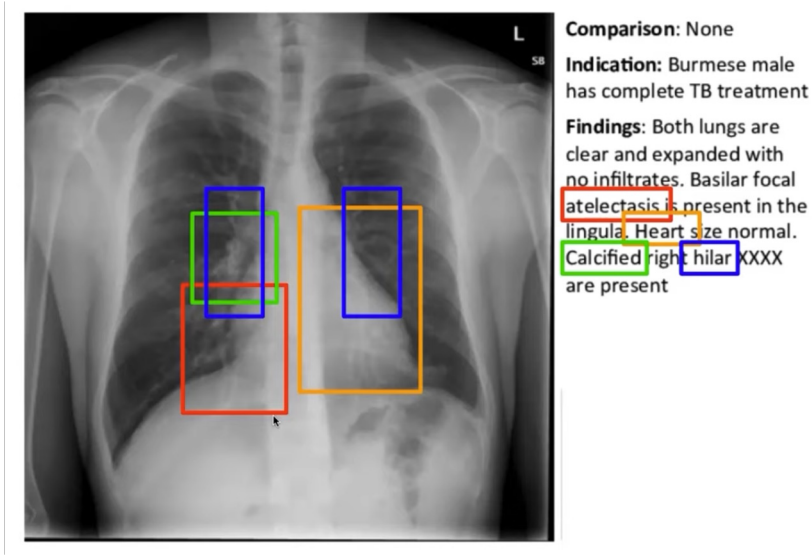
Presenter: Andi Xu

Motivation

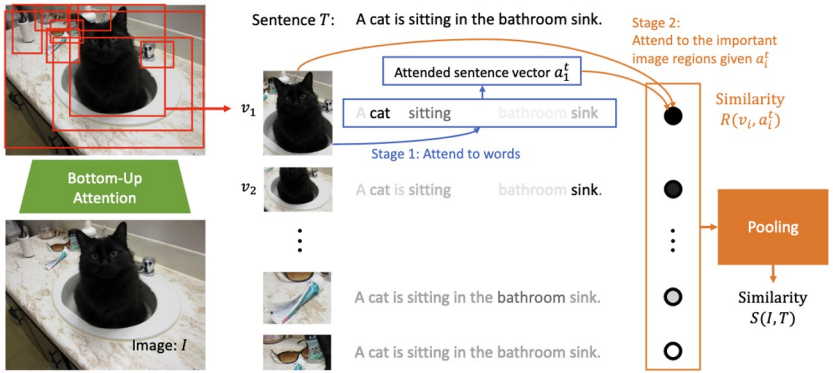
- Supervised learning for biomedical imaging requires large-scale manually labeled datasets
- Annotating medical imaging datasets requires domain expertise and is cost-prohibitive at scale
- Solution: self-supervised learning that can be fine-tuned for downstream tasks + use medical reports as supervision signal

Motivation (continue)

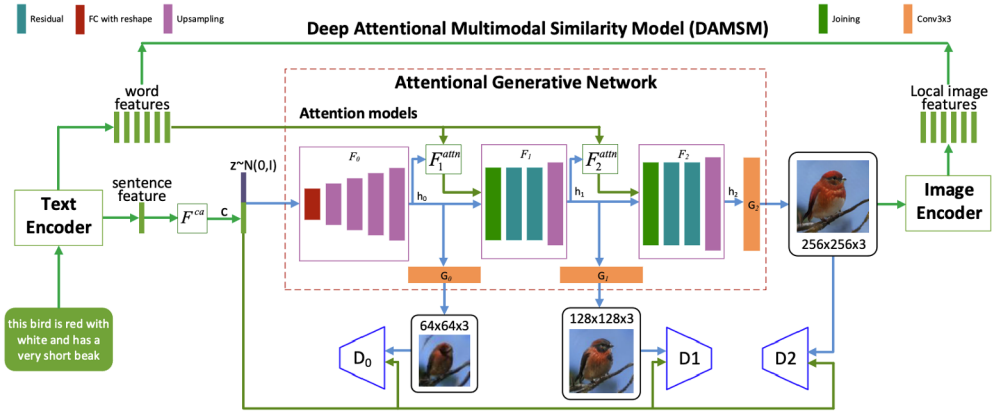
- Prior work learn multimodal representations by maximising mutual information between the global representations of the paired image and report
- However, pathology usually occupies only small proportions of the medical image
- Solution: Local features!



Localized representation as a solution



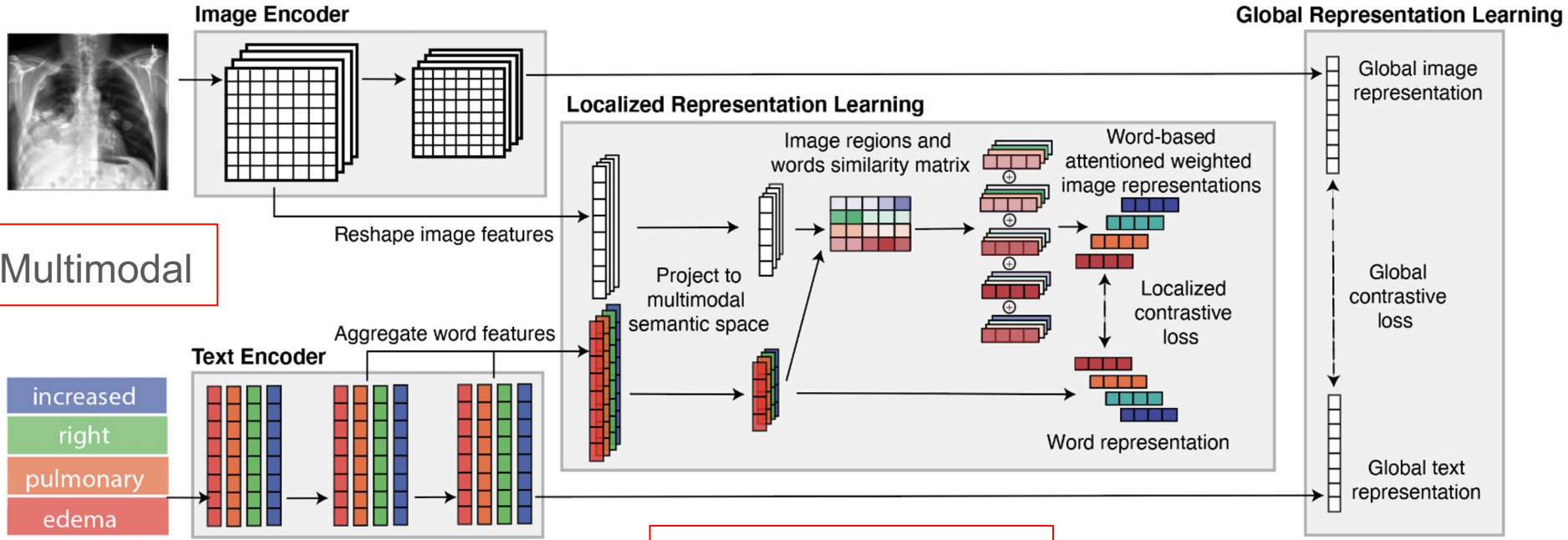
Problem: rely on object detection models that are pre-trained using natural image datasets to extract image region features



Problem: only demonstrated effectiveness for specific natural image tasks.
But medical reports often contain typographical errors and long-range context dependencies

GLoRIA

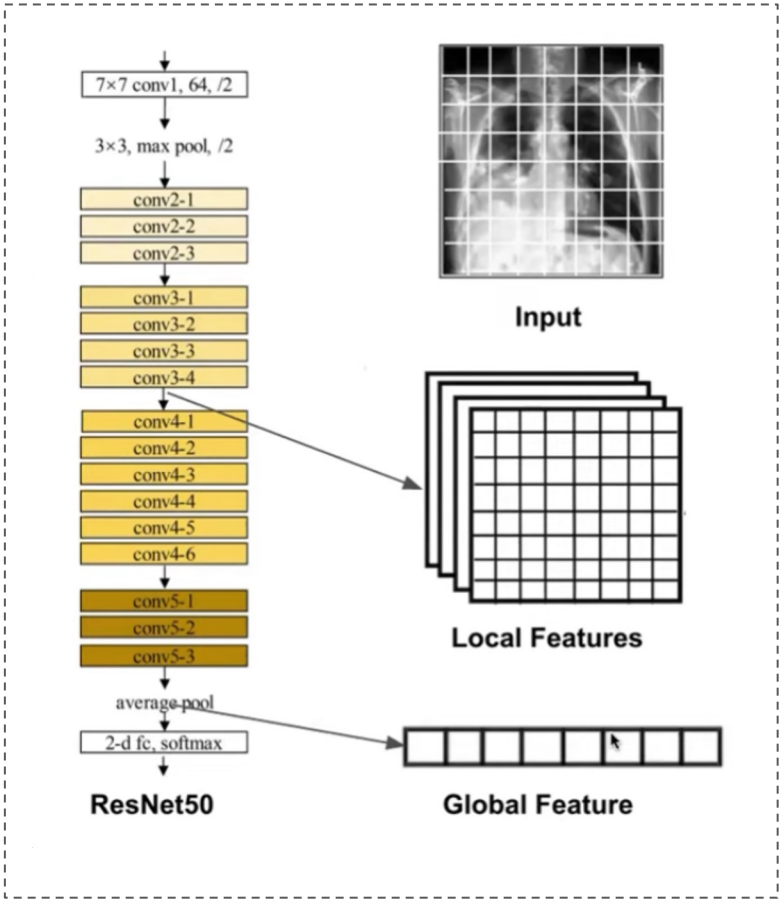
Global + Local



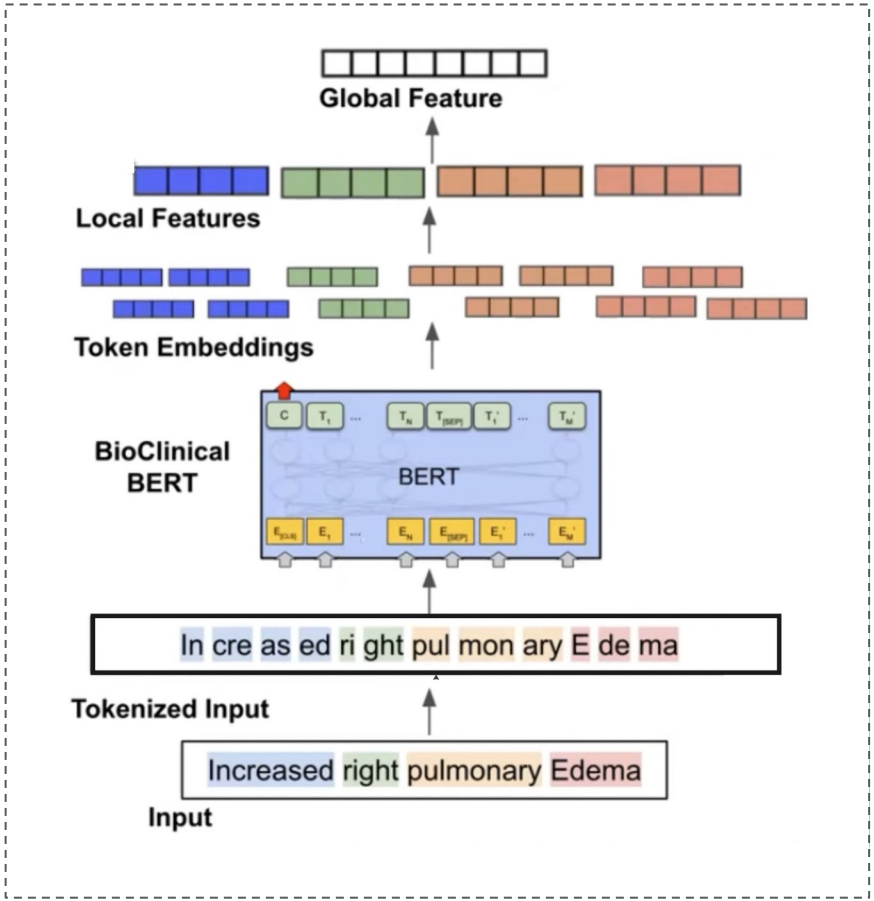
Multimodal

Self-attention-based

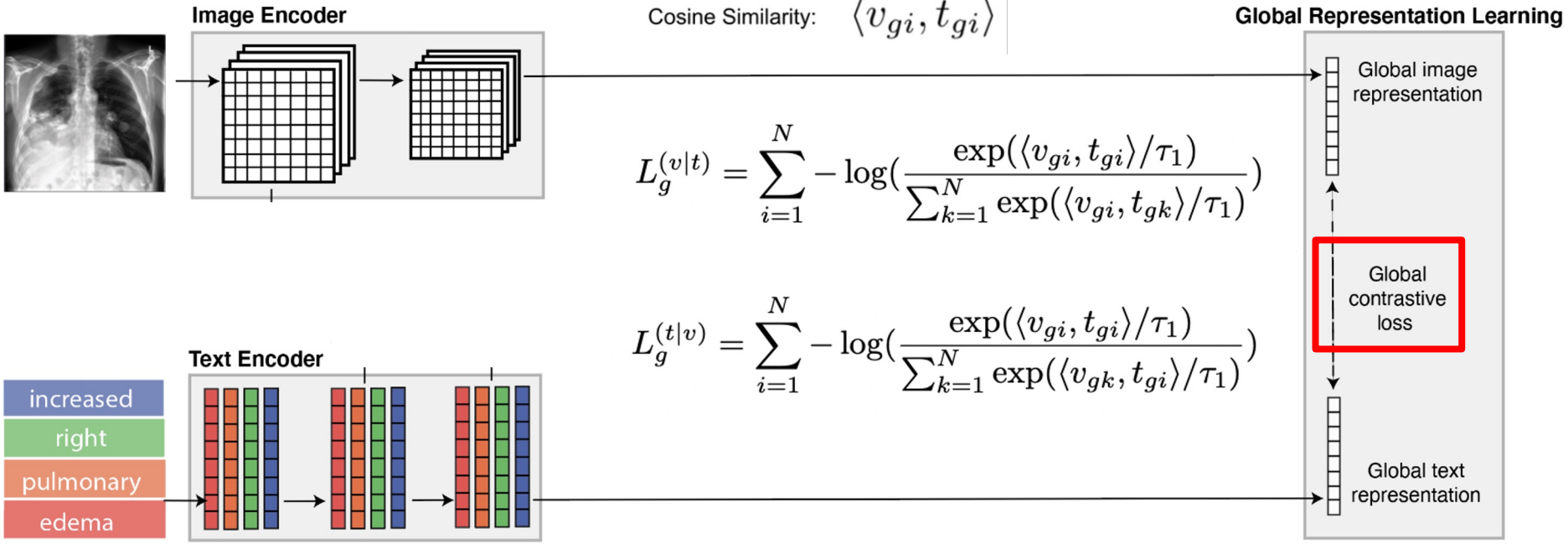
Image Encoder



Text Encoder



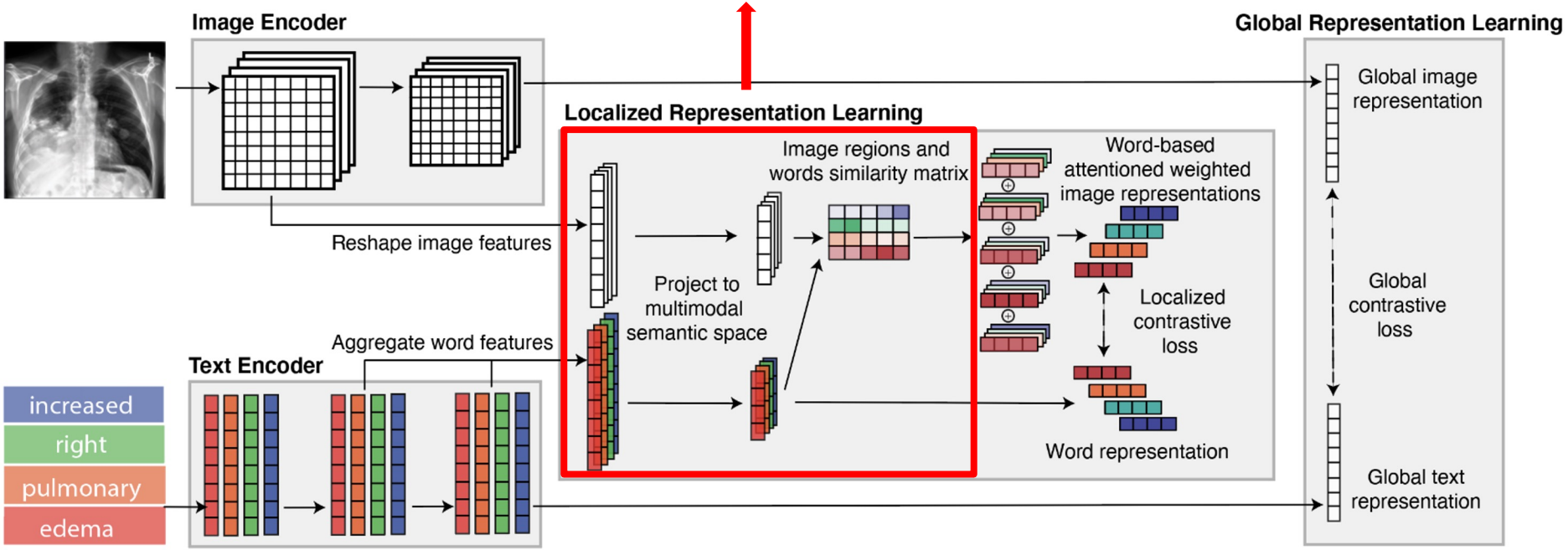
Learning objectives for global representation learning



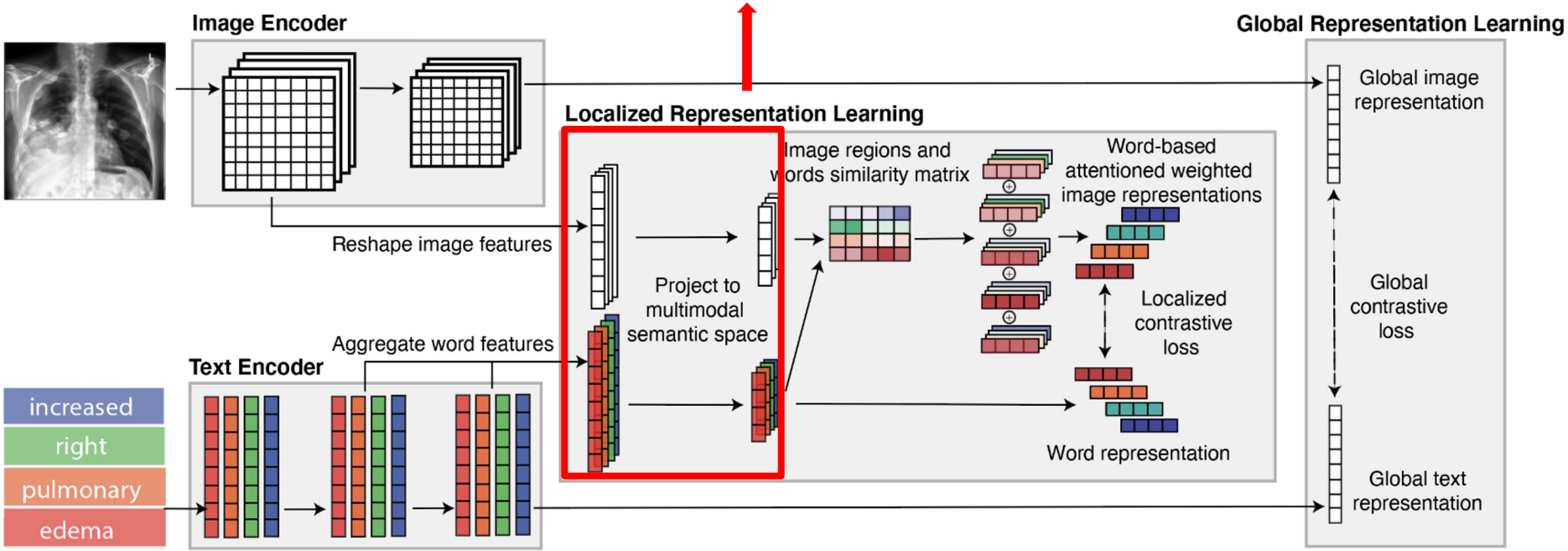
Similarity matrix: $s = v_l^T t_l$

$s_{i,j}$ = similarity between the word i and sub-region j in the image.

row = # words
col = # image features



A Neural Network for dimension matching



Similarity matrix:

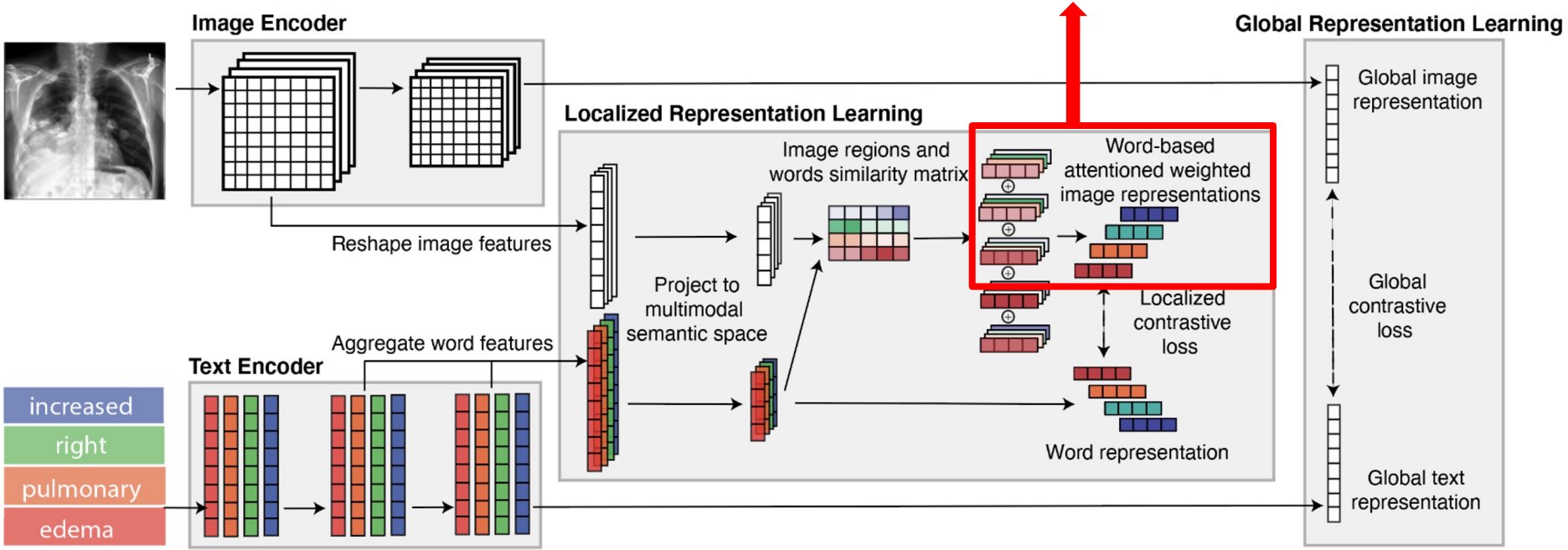
$$s = v_l^T t_l$$

Attention weights:

$$a_{ij} = \frac{\exp(s_{ij}/\tau_2)}{\sum_{k=1}^{M} exp(s_{ik}/\tau_2)}$$

Attention weighted image region representation

$$c_i = \sum_{j=0}^{M} a_{ij} v_j$$

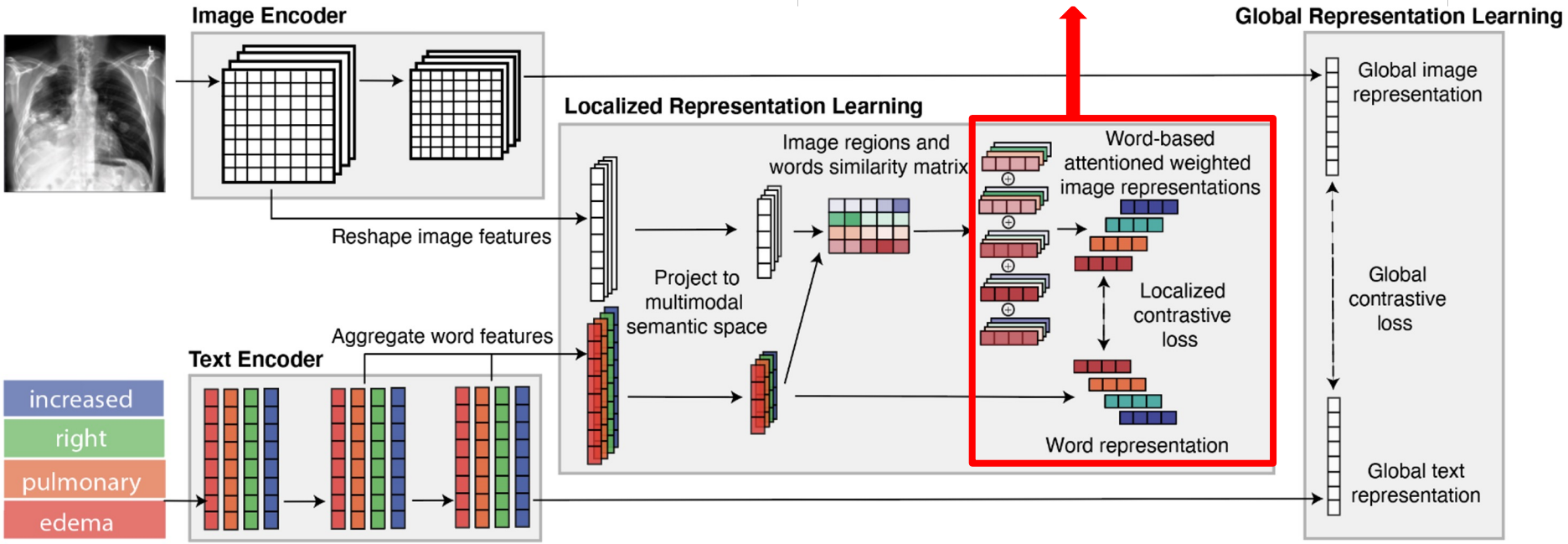


Matching functions for weighted image features & text features

$$Z(x_t, x_v) = \log(\sum_{i=1}^{W} \exp(\langle c_i, t_i \rangle / \tau_3))^{\tau_3}$$

$$L_l^{(v|t)} = \sum_{i=1}^{N} -\log(\frac{\exp(Z(x_{vi}, x_{ti})/\tau_2)}{\sum_{k=1}^{N} \exp(Z(x_{vi}, x_{tk})/\tau_2)})$$

$$L_l^{(t|v)} = \sum_{i=1}^{N} -\log(\frac{\exp(Z(x_{vi}, x_{ti})/\tau_2)}{\sum_{k=1}^{N} \exp(Z(x_{vk}, x_{ti})/\tau_2)})$$



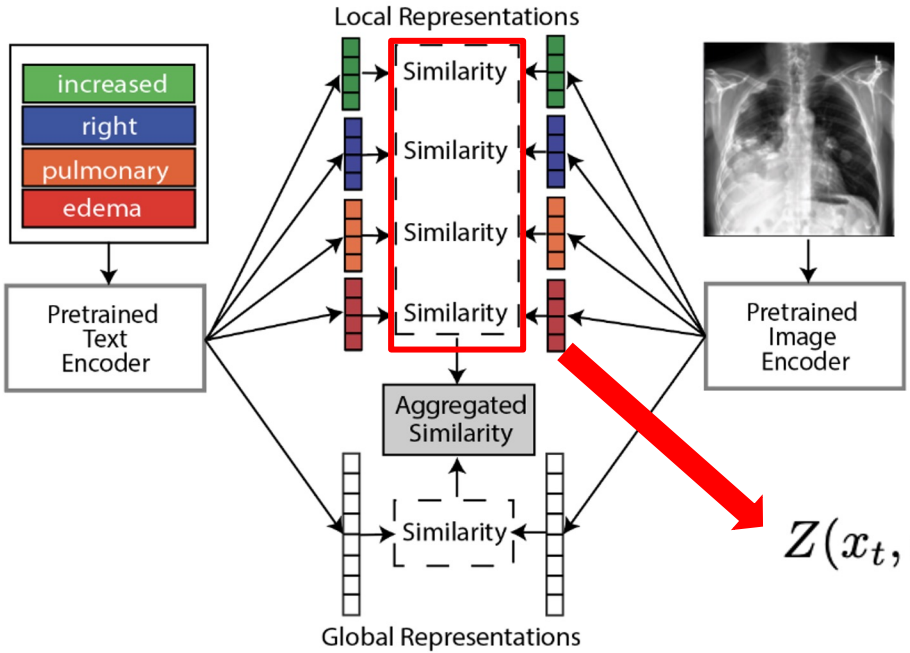
Total Loss

$$L = \boxed{L_g^{(t->v)} + L_g^{(v->t)}} + \boxed{L_l^{(t->v)} + L_l^{(v->t)}}$$

Global Loss

Local Loss

Downstream Task 1: Image-Text Retrieval

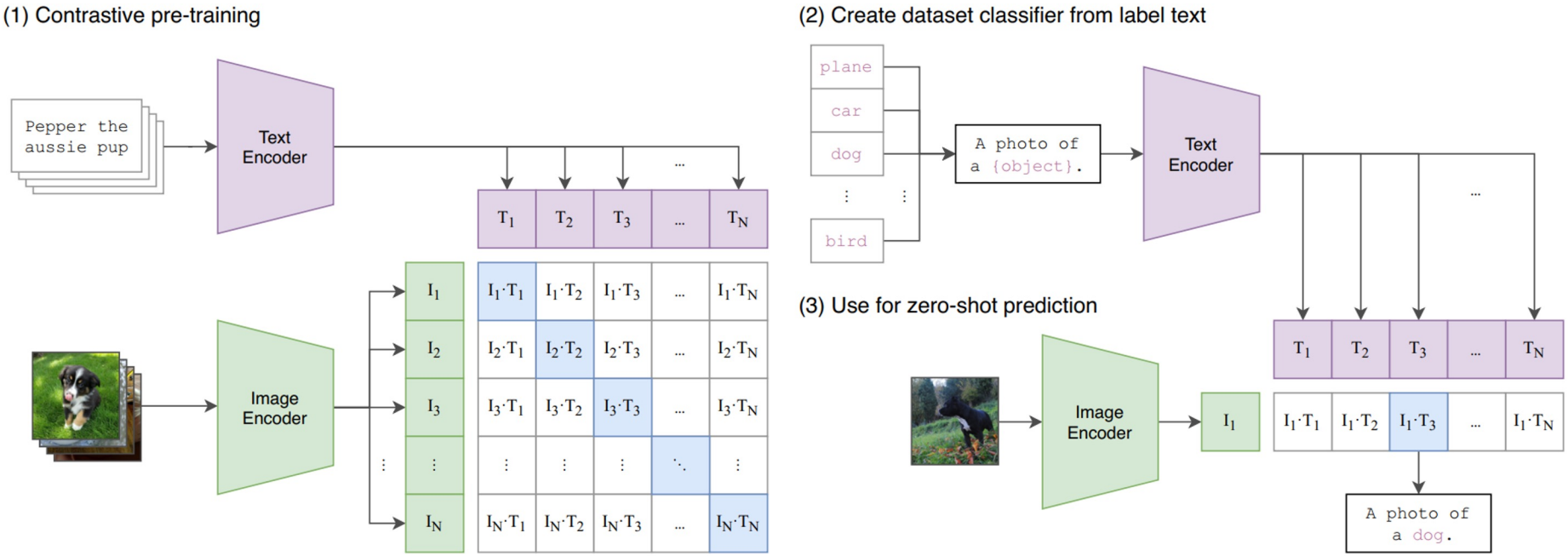


$$Z(x_t, x_v) = \log(\sum_{i=1}^{W} \exp(\langle c_i, t_i \rangle / \tau_3))^{\tau_3}$$

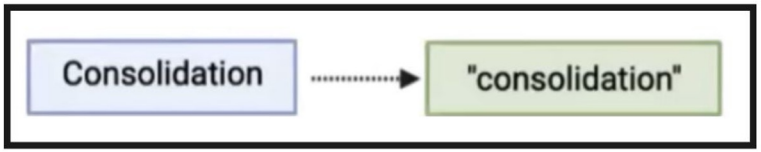
Retrieval Results

Method	Prec@5	Prec@10	Prec@100
DSVE [8]	40.64	32.77	24.74
VSE++ [9]	44.28	36.81	26.89
ConVIRT [40]	66.98	63.06	49.03
GLoRIA (Ours) - global only	67.02	64.68	49.55
GLoRIA (Ours) - local only	68.22	64.58	48.17
GLoRIA (Ours)	**69.24**	**67.22**	**53.78**

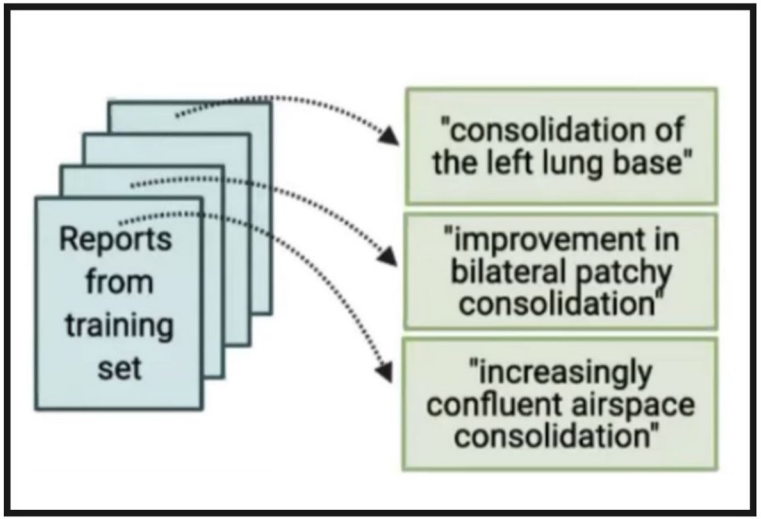
Downstream Task 2: Zero-shot Classification



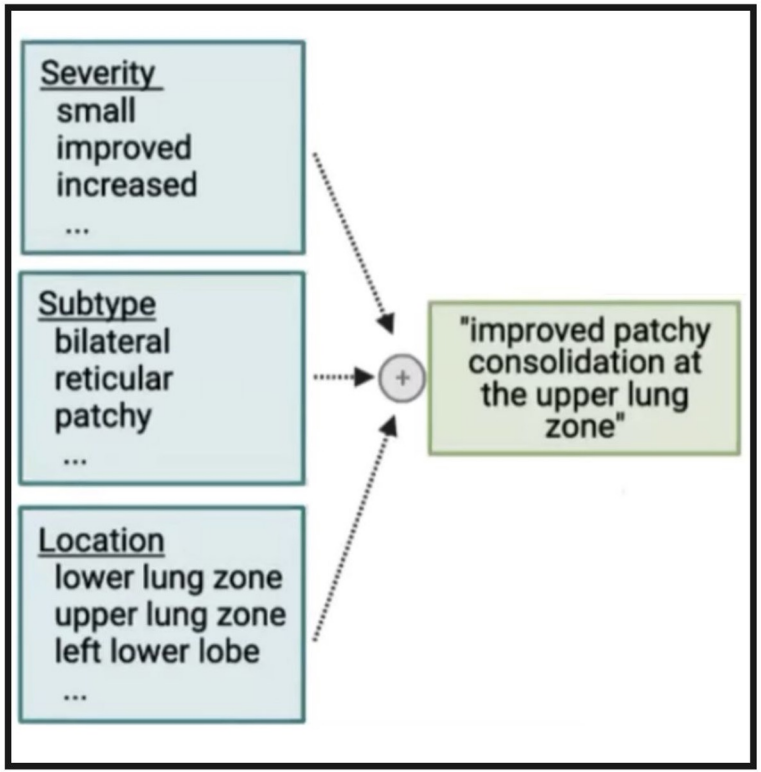
1. Class name as text

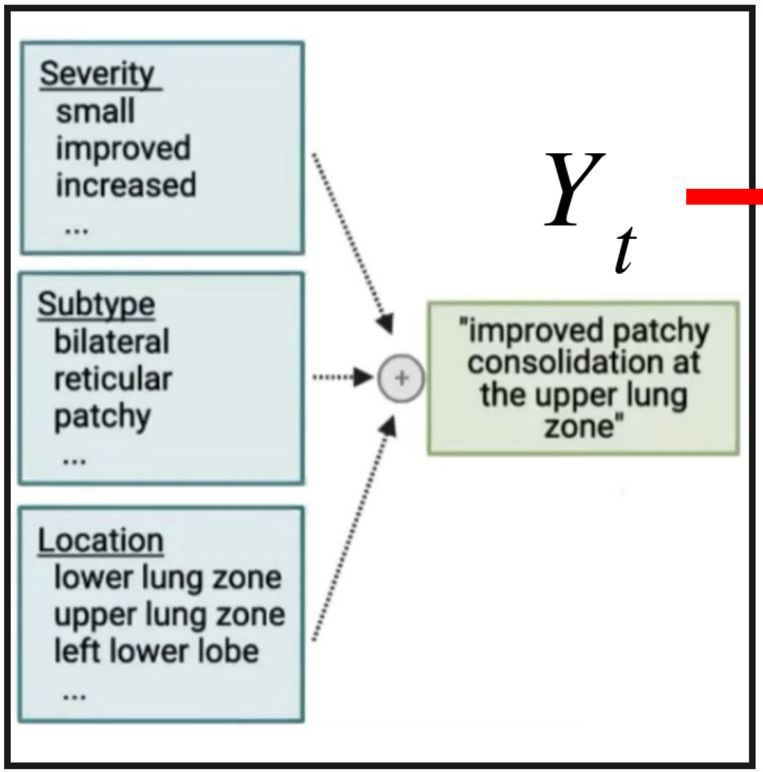


2. Randomly sampled prompt from data



3. Generated prompt





Y_t

Text coder

Global representation

Local representation

\+ Image representations

$$\text{argmax}_i\left[\frac{1}{2}\left(S(t_{gi}, v_g) + Z(t_{li}, v_l)\right)\right]$$

Zero-shot classification result

CheXpert	Acc.	Sens.	Spec.	PPV	NPV	F1
100%	0.57	**0.83**	0.80	0.51	**0.95**	0.63
10%	0.55	0.76	0.82	0.51	0.92	0.61
1%	0.47	0.68	0.85	0.53	0.91	0.59
Zero-shot	**0.61**	0.70	**0.91**	**0.65**	0.92	**0.67**
RSNA	Acc	Sen	Spe	PPV	NPV	F1
100%	**0.79**	0.87	0.76	**0.52**	0.95	**0.65**
10%	0.78	0.78	**0.79**	0.52	0.92	0.63
1%	0.72	0.82	0.69	0.44	0.93	0.57
Zero-shot	0.70	**0.89**	0.65	0.43	**0.95**	0.58

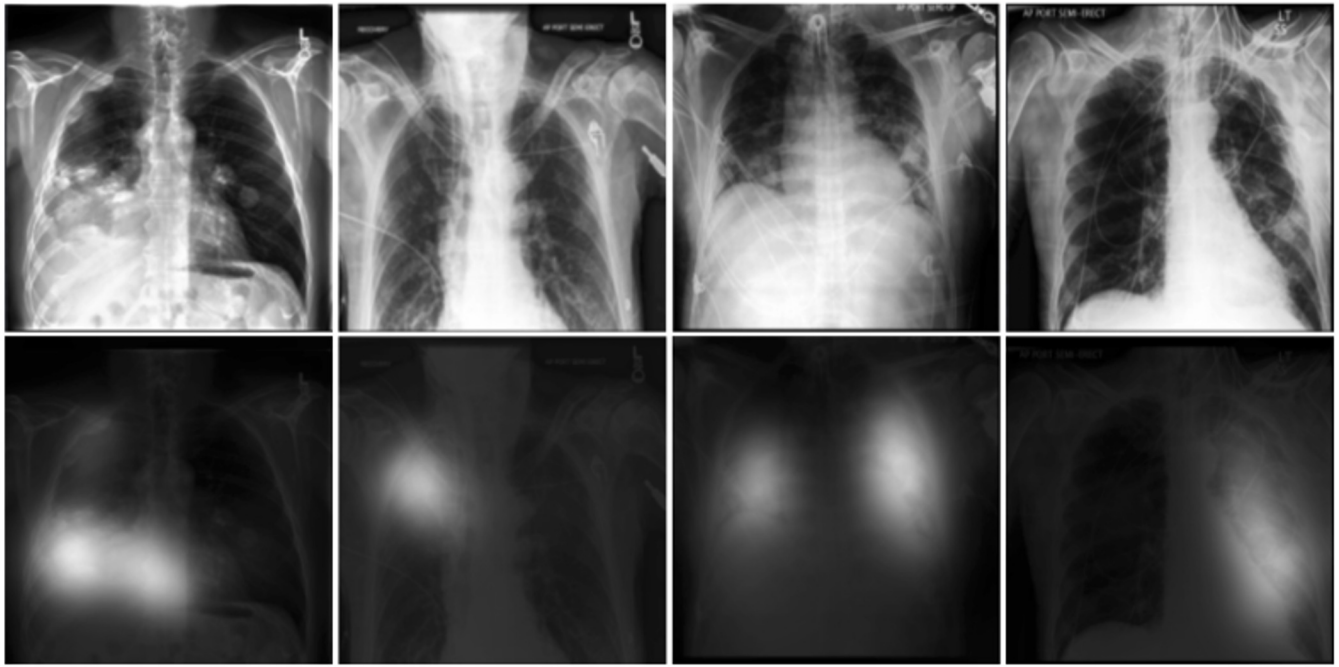
Fine-tune Classification

	CheXpert			RSNA		
	1%	10%	100%	1%	10%	100%
Random	56.1	62.6	65.7	58.9	69.4	74.1
ImageNet	74.4	79.1	81.4	74.9	74.5	76.3
DSVE [8]	50.1	51.0	51.5	49.7	52.1	57.8
VSE++ [9]	50.3	51.2	524	49.4	57.2	67.9
ConVIRT [40]	85.9	86.8	87.3	77.4	80.1	81.3
GLoRIA (Ours)	**86.6**	**87.8**	**88.1**	**86.1**	**88.0**	**88.6**

Segmentation

Initialization Method	Pneumothorax Segmentation		
	1%	10%	100%
Random	0.090	0.286	0.543
ImageNet	0.102	0.355	**0.635**
ConVIRT [40]	0.250	0.432	0.599
GLoRIA (Ours)	**0.358**	**0.469**	0.634

Visualizing Attention



Pneumonia (a) | Pneumothorax (b) | Edema (c) | Opacity (d)

Quiz

How does the paper leverage its method to be used on Image-Text Retrieval task?

Quiz

How does the paper achieve context-aware local image representation?