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

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

Stanford University

ICCV 2021

Introduction

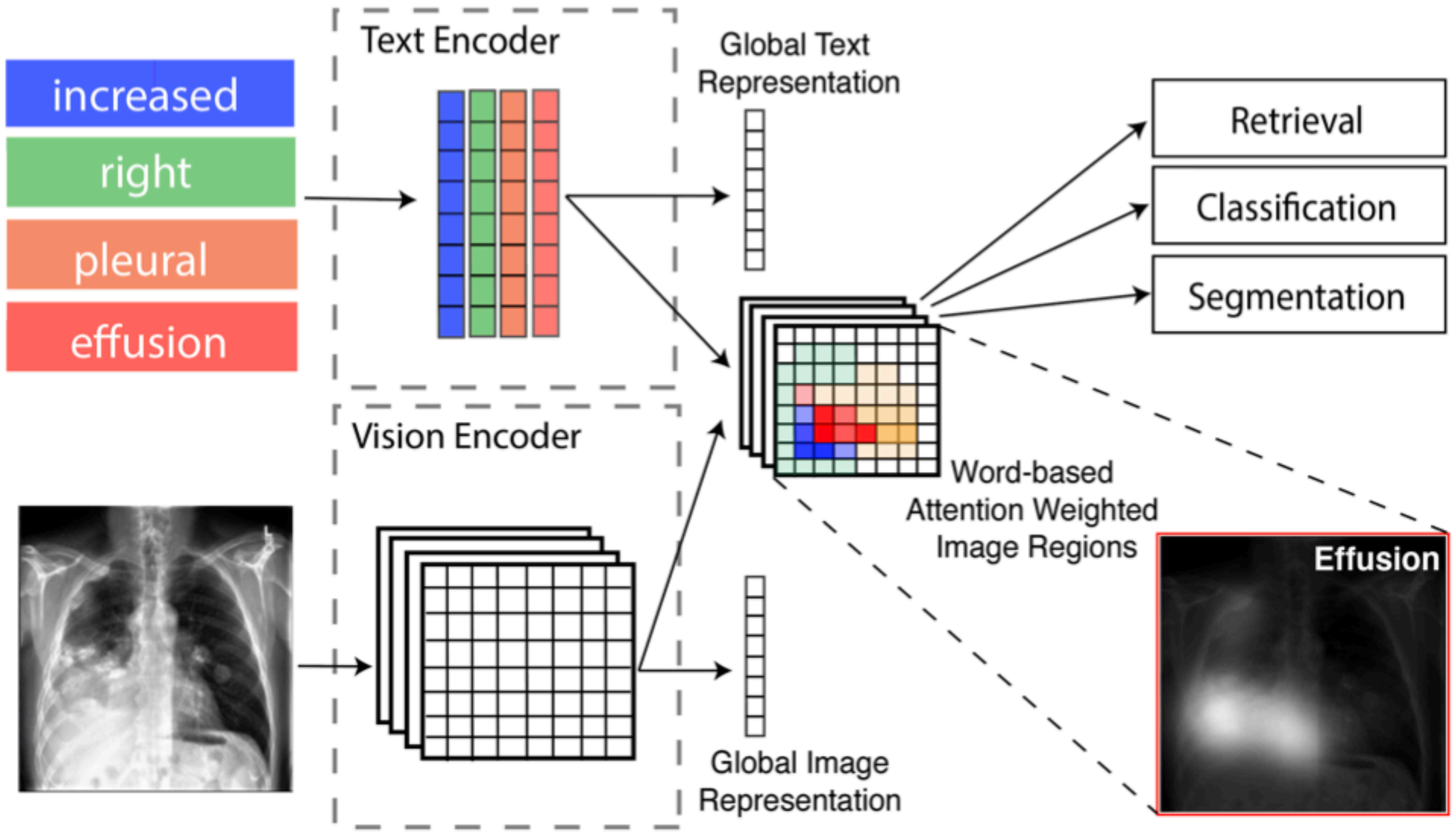
Task:

Multimodal representation learning for medical tasks

< radiology report, radiology image>

Idea:

global-*<u>local</u>* representation learning by contrasting image sub-regions and report words



Transfer:

- image-text retrieval
- classification
- segmentation

Method

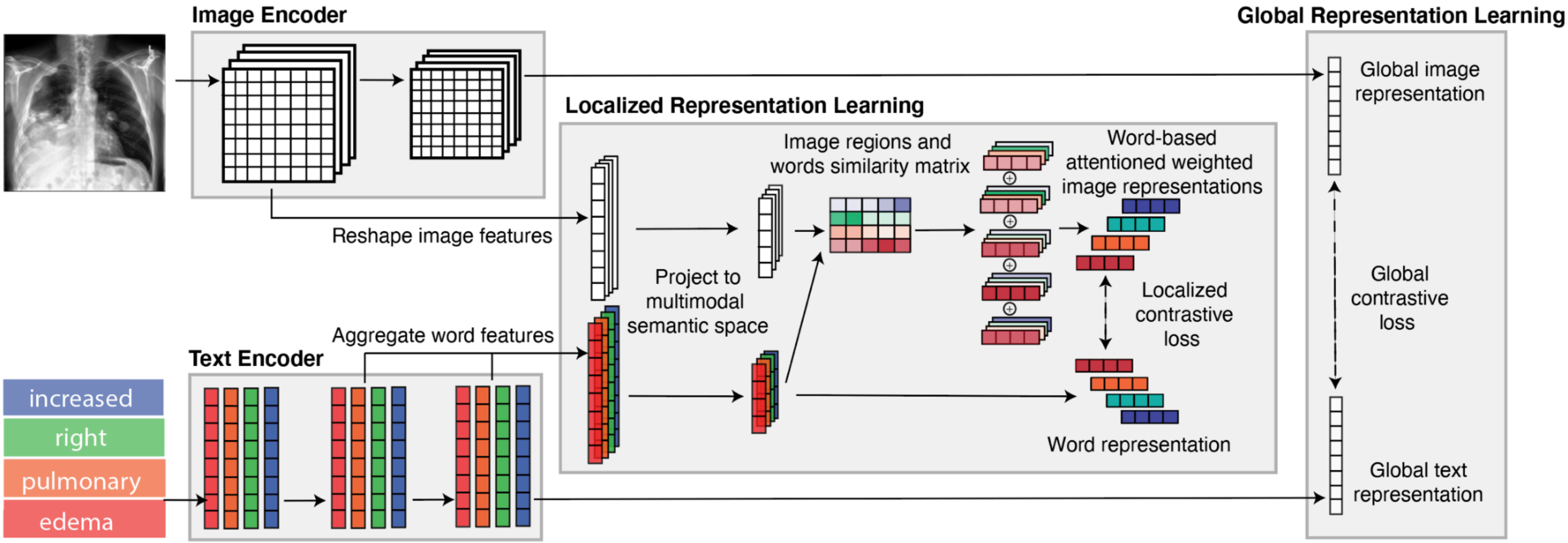


Image Encoder
Global Representation Learning
Global image representation
Localized Representation Learning
Image regions and words similarity matrix
Word-based attentioned weighted image representations
Reshape image features
Project to multimodal semantic space
Localized contrastive loss
Global contrastive loss
Aggregate word features
Text Encoder
increased
right
pulmonary
edema
Word representation
Global text representation

Method

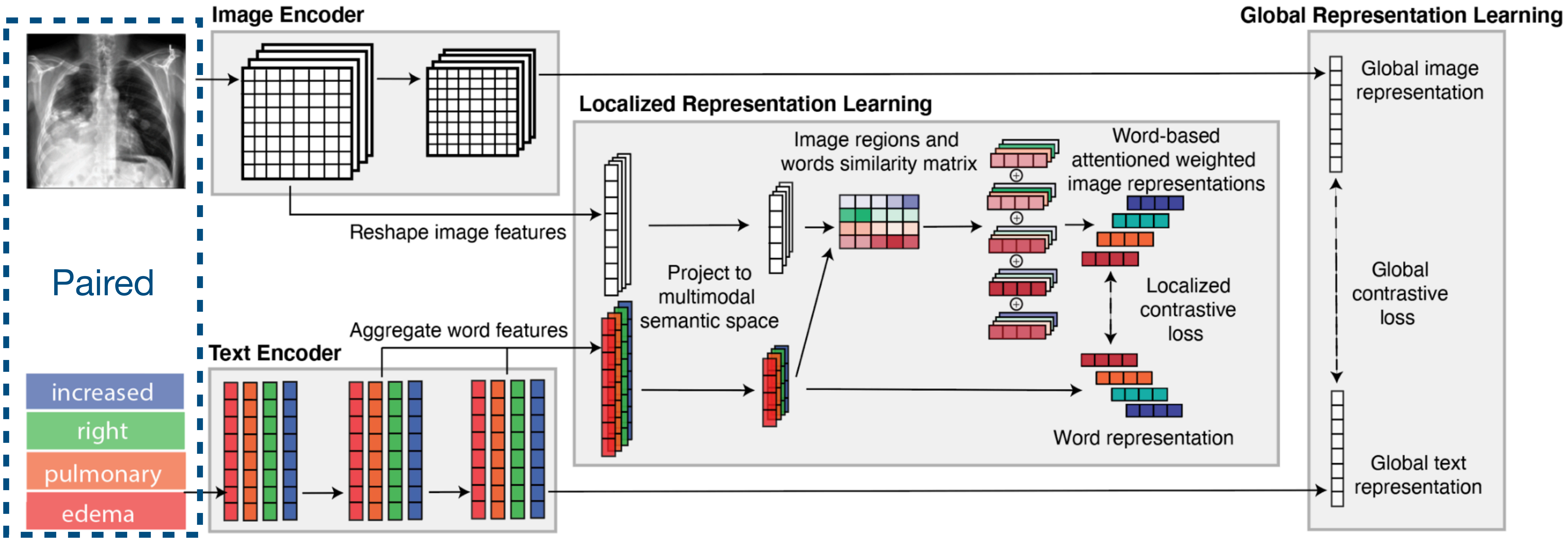
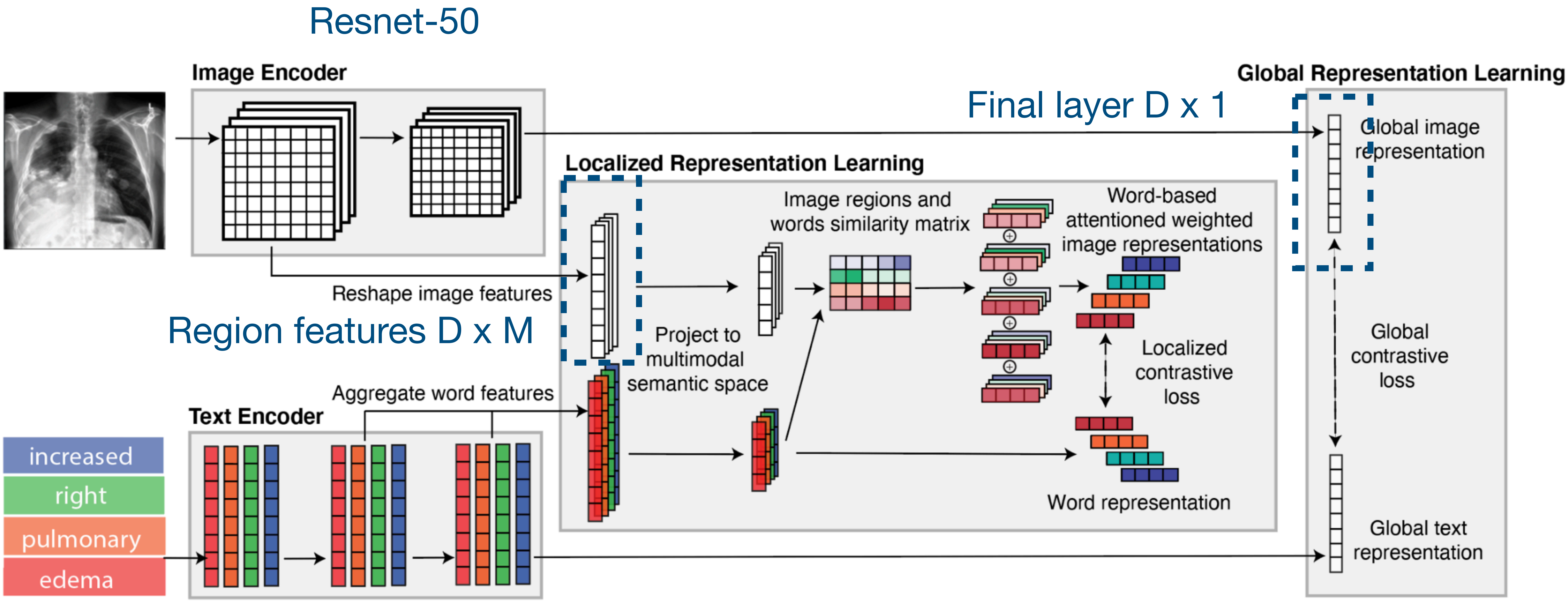


Image Encoder
Global Representation Learning
Global image representation
Localized Representation Learning
Image regions and words similarity matrix
Word-based attentioned weighted image representations
Reshape image features
Paired
Project to multimodal semantic space
Localized contrastive loss
Global contrastive loss
Aggregate word features
Text Encoder
increased
right
pulmonary
edema
Word representation
Global text representation

Method



Resnet-50
Image Encoder
Global Representation Learning
Final layer D x 1
Global image representation
Localized Representation Learning
Image regions and words similarity matrix
Word-based attentioned weighted image representations
Reshape image features
Region features D x M
Project to multimodal semantic space
Localized contrastive loss
Global contrastive loss
Aggregate word features
Text Encoder
increased
right
pulmonary
edema
Word representation
Global text representation

Method

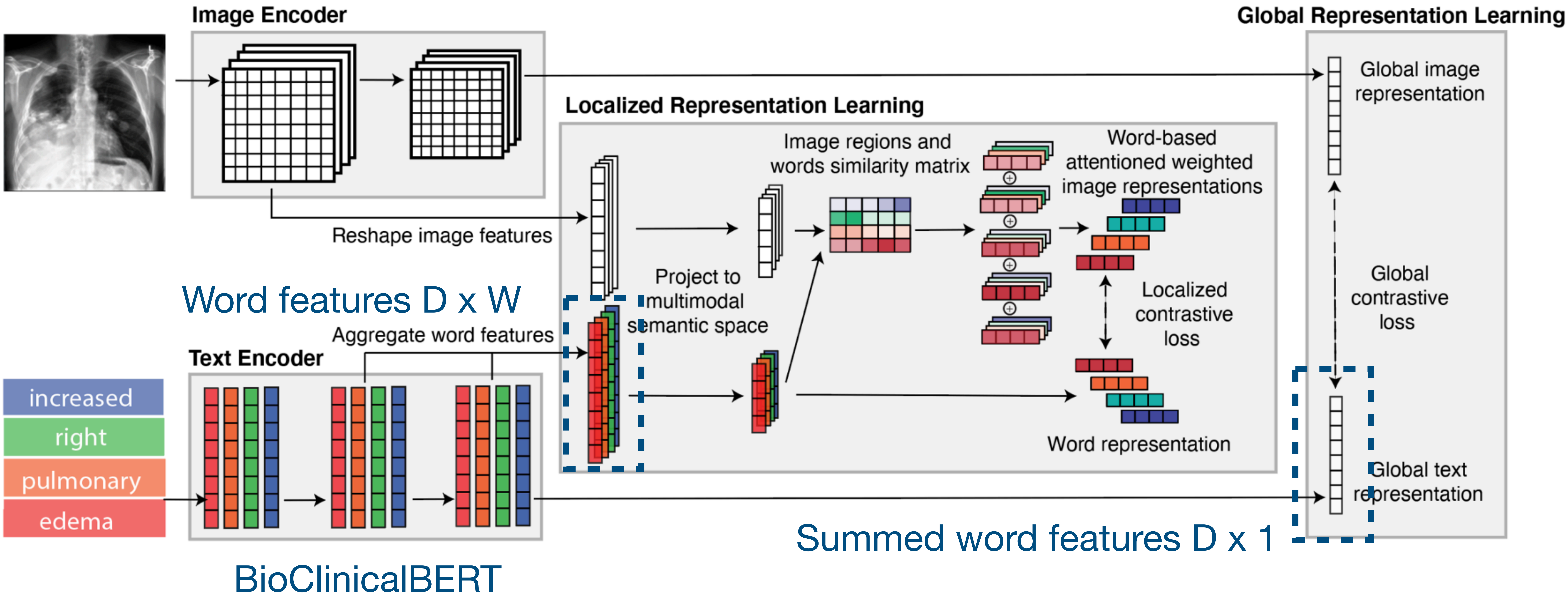
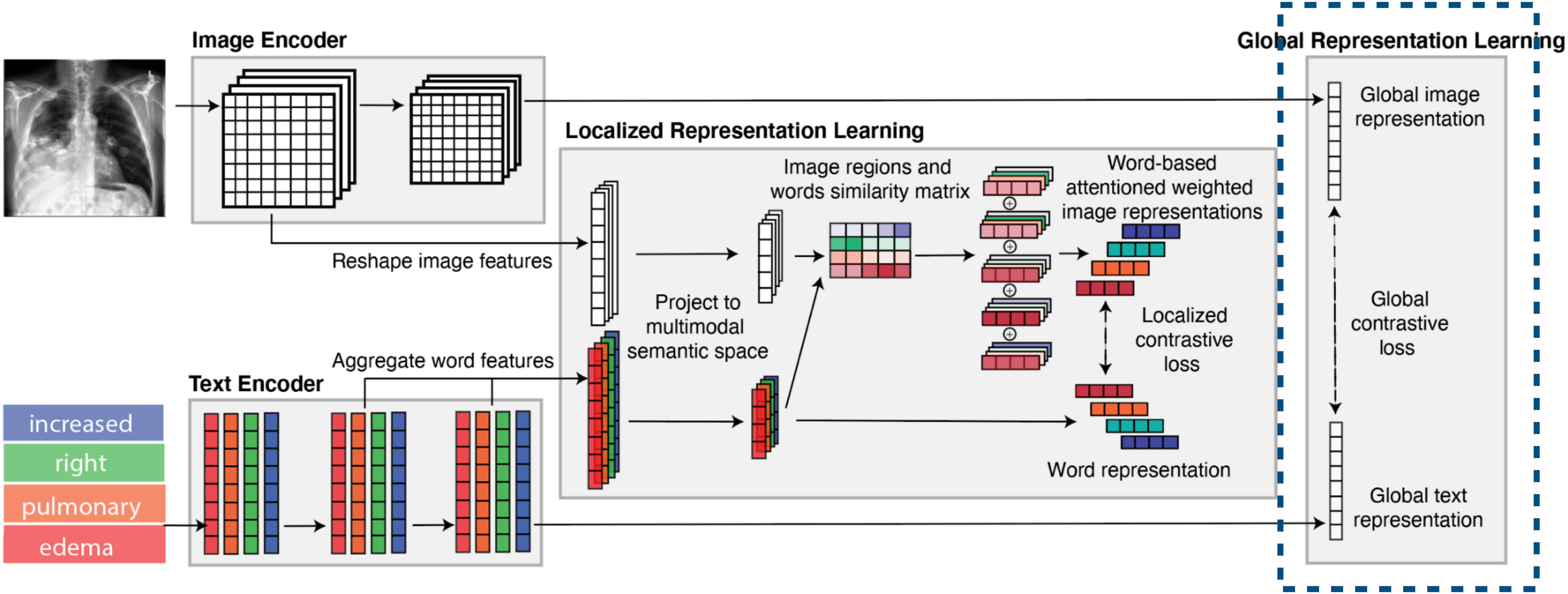


Image Encoder
Global Representation Learning
Global image representation
Localized Representation Learning
Image regions and words similarity matrix
Word-based attentioned weighted image representations
Reshape image features
Project to multimodal semantic space
Word features D x W
Aggregate word features
Text Encoder
Localized contrastive loss
Global contrastive loss
increased
right
pulmonary
edema
Word representation
Global text representation
Summed word features D x 1
BioClinicalBERT

Method



Global contrastive learning

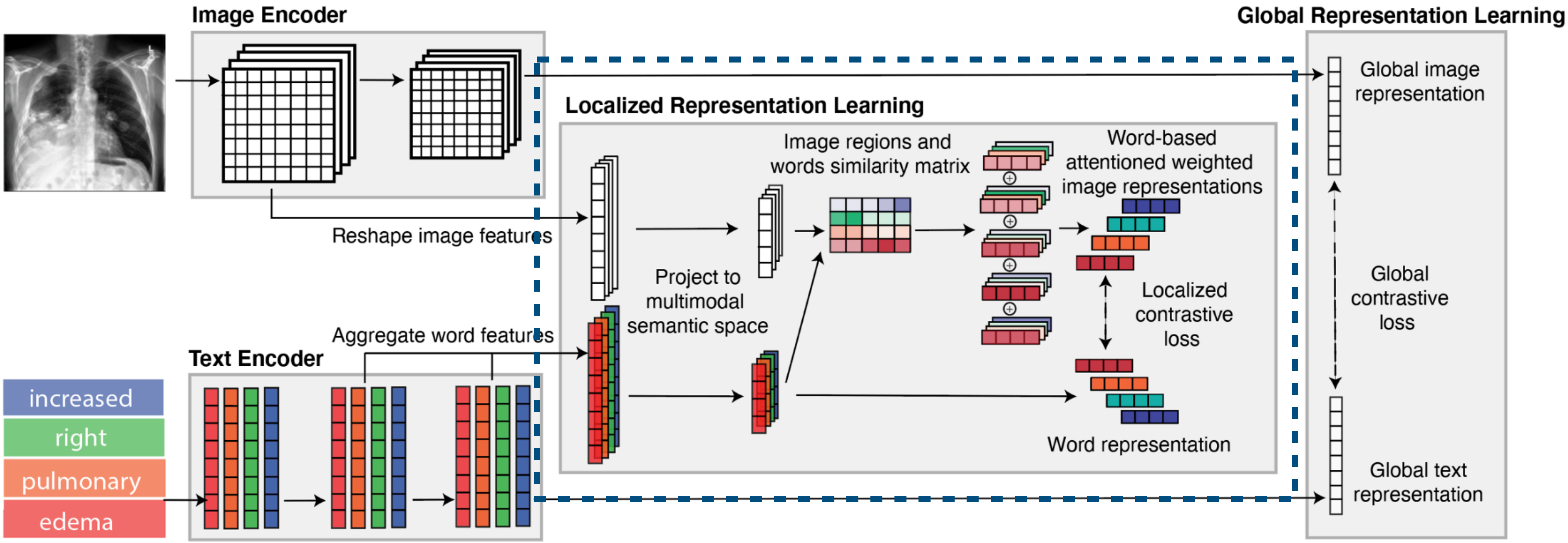
Predict correct text from image

$$L_g^{(v|t)} = \sum_{i=1}^{N} -\log\left(\frac{\exp(\langle v_{gi}, t_{gi}\rangle/\tau_1)}{\sum_{k=1}^{N}\exp(\langle v_{gi}, t_{gk}\rangle/\tau_1)}\right)$$

Predict correct image from text

$$L_g^{(t|v)} = \sum_{i=1}^{N} -\log\left(\frac{\exp(\langle v_{gi}, t_{gi}\rangle/\tau_1)}{\sum_{k=1}^{N}\exp(\langle v_{gk}, t_{gi}\rangle/\tau_1)}\right)$$

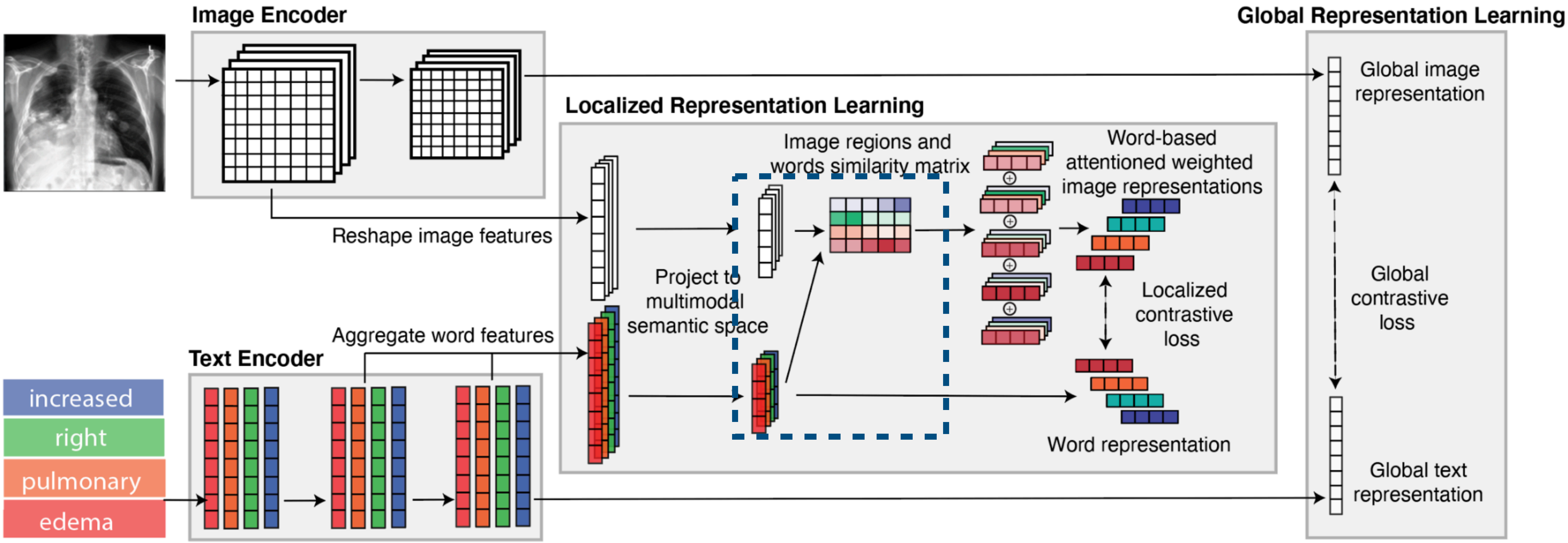
Method



Local contrastive learning:

Learn attentions that weigh different image sub-regions based on their significance for a given word

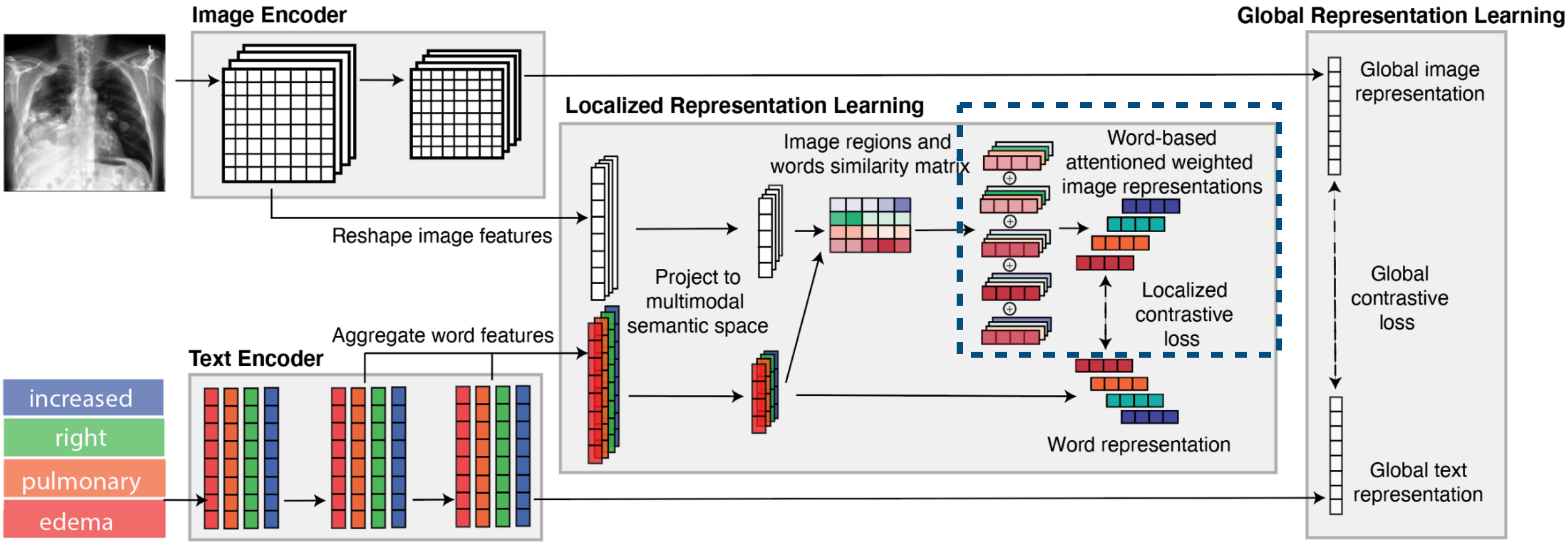
Method



Region-word similarity (M x W)

$$s = v_l^T t_l$$

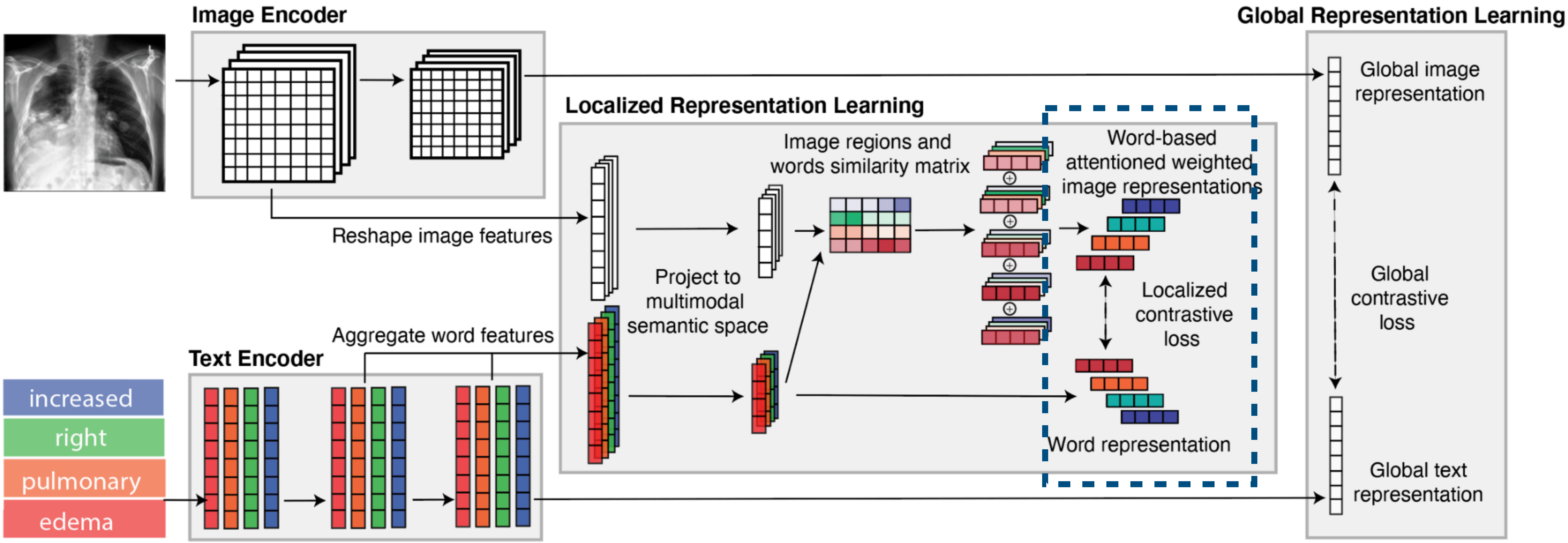
Method



For each word *i*, compute a attention-weighted image region feature

$$a_{ij} = \frac{\exp(s_{ij}/\tau_2)}{\sum_{k=1}^{M} exp(s_{ik}/\tau_2)} \qquad c_i = \sum_{j=0}^{M} a_{ij} v_j$$

Method



Local contrastive learning:

Similar to global contrastive learning, but with a local matching function

$$Z(v\_l, t\_l) = \log\left(\sum_{i=1}^{W} \exp(\langle c_i, t\_li / \tau_3 \rangle)\right)^{\tau_3}$$ i: word

Method

Global contrastive learning

Predict correct text from image

$$L_g^{(v|t)} = \sum_{i=1}^{N} -\log\left(\frac{\exp(\langle v_{gi}, t_{gi}\rangle/\tau_1)}{\sum_{k=1}^{N}\exp(\langle v_{gi}, t_{gk}\rangle/\tau_1)}\right)$$

Predict correct image from text

$$L_g^{(t|v)} = \sum_{i=1}^{N} -\log\left(\frac{\exp(\langle v_{gi}, t_{gi}\rangle/\tau_1)}{\sum_{k=1}^{N}\exp(\langle v_{gk}, t_{gi}\rangle/\tau_1)}\right)$$

Method

Global contrastive learning

Predict correct text from image

$$L_g^{(v|t)} = \sum_{i=1}^{N} -\log\left(\frac{\exp(\langle v_{gi}, t_{gi}\rangle/\tau_1)}{\sum_{k=1}^{N}\exp(\langle v_{gi}, t_{gk}\rangle/\tau_1)}\right)$$

Predict correct image from text

$$L_g^{(t|v)} = \sum_{i=1}^{N} -\log\left(\frac{\exp(\langle v_{gi}, t_{gi}\rangle/\tau_1)}{\sum_{k=1}^{N}\exp(\langle v_{gk}, t_{gi}\rangle/\tau_1)}\right)$$

$$L_l^{(v|t)} = \sum_{i=1}^{N} -\log\left(\frac{\exp(Z(v_{li}, t_{li})/\tau_2)}{\sum_{k=1}^{N}\exp(Z(v_{li}, t_{lk})/\tau_2)}\right)$$

Predict correct text from image

$$L_l^{(t|v)} = \sum_{i=1}^{N} -\log\left(\frac{\exp(Z(v_{li}, t_{li})/\tau_2)}{\sum_{k=1}^{N}\exp(Z(v_{lk}, t_{li})/\tau_2)}\right)$$

Predict correct image from text

Local contrastive learning

Method

Global contrastive learning

Predict correct text from image

$$L_g^{(v|t)} = \sum_{i=1}^{N} -\log\left(\frac{\exp(\boxed{\langle v_{gi}, t_{gi}\rangle}/\tau_1)}{\sum_{k=1}^{N}\exp(\langle v_{gi}, t_{gk}\rangle/\tau_1)}\right)$$

Predict correct image from text

$$L_g^{(t|v)} = \sum_{i=1}^{N} -\log\left(\frac{\exp(\langle v_{gi}, t_{gi}\rangle/\tau_1)}{\sum_{k=1}^{N}\exp(\langle v_{gk}, t_{gi}\rangle/\tau_1)}\right)$$

$$L_l^{(v|t)} = \sum_{i=1}^{N} -\log\left(\frac{\exp(\boxed{Z(v\_li, t\_li)}/\tau_2)}{\sum_{k=1}^{N}\exp(Z(v\_li, t\_lk)/\tau_2)}\right)$$

Predict correct text from image

$$L_l^{(t|v)} = \sum_{i=1}^{N} -\log\left(\frac{\exp(Z(v\_li, t\_li)/\tau_2)}{\sum_{k=1}^{N}\exp(Z(v\_lk, t\_li)/\tau_2)}\right)$$

Predict correct image from text

Local contrastive learning

Regions and words in paired image and text should be better matched

Transfer

- **image-text retrieval:**
  global similarity and local matching score
- **classification**:
  zero-shot classification by image-text similarity
  generate text for each class in terms of sub-types, severities and locations
  find the class with highest average similarity
- **segmentation**:
  fine-tune

Experiment

Image-text retrieval

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] global only	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**

Table 1: Results of image-text retrieval on the CheXpert 5x200 dataset. The top K Precision metrics are reported for $K = 5, 10, 100$. Ours method achieves the best performance by incorporating both global and local representations.

Zero-shot classification

Overfit

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

Table 3: Results of zero-shot image classification on the CheXpert 5x200 and RSNA datasets. Note that representation learning framework is trained using CheXpert. We compare classification results with different amounts of training data for comparison.

Pre-trained on CheXpert Full

Experiment

Supervised classification: Linear classifier

	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**

Table 2: Results of fine-tuned image classification (AUROC score) on CheXpert and RSNA test sets based on different portion of training data: 1%, 10%, 100%.

Experiment

Segmentation: U-Net

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

Table 4: Results of image segmentation (Dice score) on SIIM dataset with different portion of training data: 1%, 10%, 100%.

Attention weights

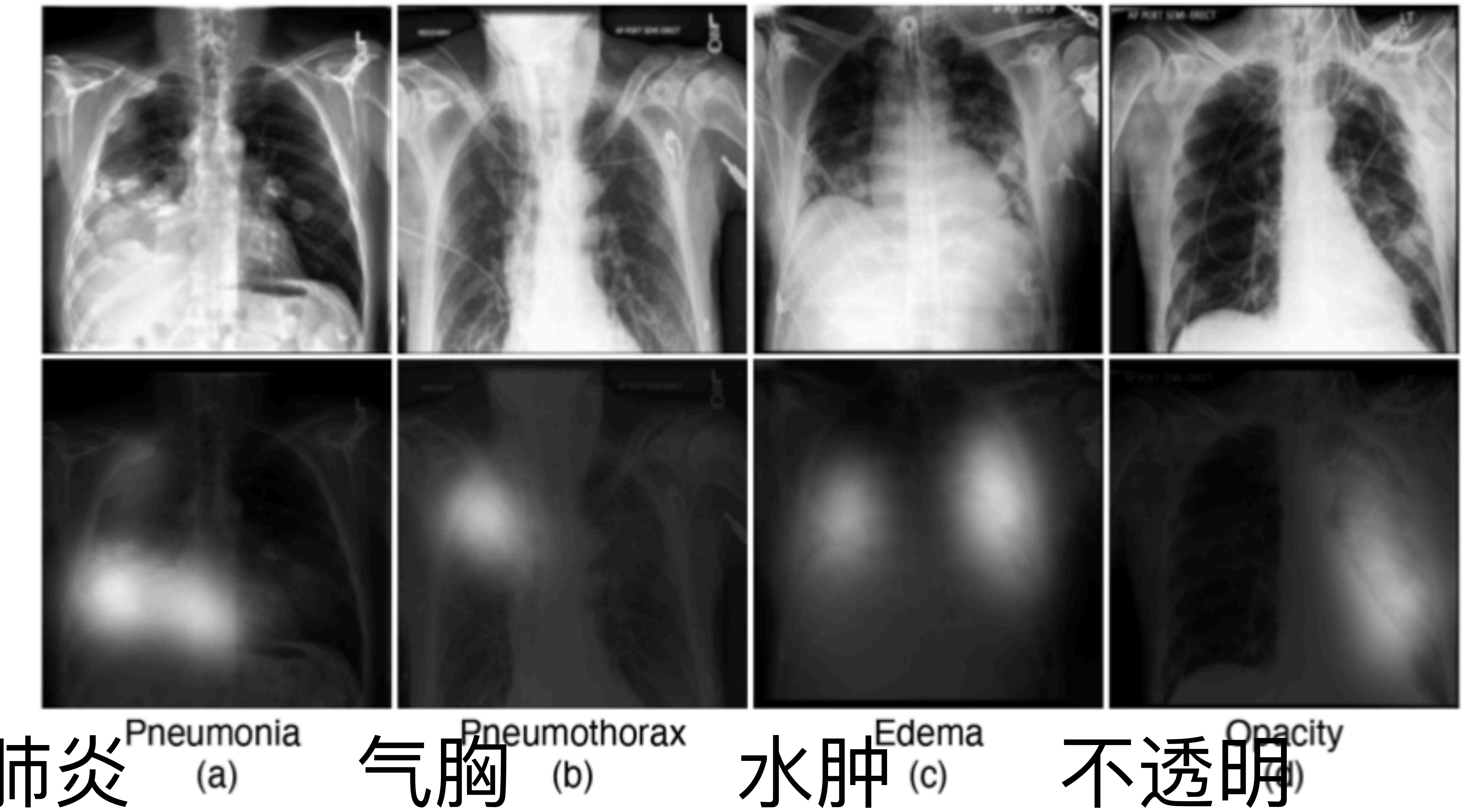


Figure 4: Examples of frontal radiographs of the chest (top) with corresponding attention weights for the given word (below).

Conclusion

- global + local representation learning for medical tasks
- local representation learning with region-word matching
- tested on chest X-ray

[*"Car", "dio", "mega", "ly"*], it is important to understand the direct correspon- dence of the term *"Cardiomegaly"*

Most state-art-the art methods require pre- trained object detection model for local feature extraction, which is not applicable for medical images.