Paper List

MICCAI2021

Improving Pneumonia Localization via Cross-Attention on Medical Images and Reports

Information

Improving Pneumonia Localization via Cross-Attention on Medical Images and Reports

Riddhish Bhalodia[1], Ali Hatamizadeh[2], Leo Tam[2], Ziyue Xu[2], Xiaosong Wang[2], Evrim Turkbey[3], and Daguang Xu[2(✉)]

[1] School of Computing, University of Utah, Salt Lake City, UT, USA
riddhishb@sci.utah.edu
[2] NVIDIA Corporation, Santa Clara, USA
{ahatamizadeh,leot,ziyuex,xiaosongw,daguangx}@nvidia.com
[3] Department of Radiology and Imaging Sciences, National Institutes of Health Clinical Center, Bethesda, USA
evrim.turkbey@nih.gov

Introduction

- Pneumonia localization in chest X-rays and its subsequent characterization is of vital importance
- Medical reports, on the other hand, are highly descriptive and provide a plethora of information
- Distill information from the Medical Report and inform the localization of corresponding images without added supervision

Difficulties

- Textual information from medical reports is hard to distill:
    1. lot of specific terminologies
    2. no clear sentence construction
    3. many redundant/extraneous information

Dataset

MIMIC-CXR dataset:

473,064 chest X-ray images with 206,754 paired radiology reports for 63,478 patients

https://arxiv.org/abs/1901.07042

Only utilize the images corresponding to pneumonia and having at least one of the attributes in the attribute set, which results in 11,308 training samples.

90% for training, 5% for validation, 5% for testing

Dataset

Evaluation:

1. Chest-X-ray-8 dataset: utilizing the 120 annotations given for pneumonia

https://arxiv.org/abs/1705.02315

2. COVID-19 X-Ray dataset: contains 951 X-ray images acquired from different centers across the world

https://github.com/ieee8023/covid-chestxray-dataset

Methodology

1. Attribute Extraction
    - Extract a dictionary of important text-attributes indicative of pneumonia location and characteristics from reports
    - Construct a constant attribute set of 22 keywords
    - Pre-train a Word2Vec[1] model on the entire set to extract the text-features T as $\{\mathbf{m}_i\}_{i=1}^{M}$

Methodology

2. Box Detector and Image Features

- Pretarin a Retinanet with Resnet-50 backbone on ChestX-Ray-8 images, with 2560 pneumonia annotations (bounding-boxes)
- Produces ROIs and pneumonia classification score.
- N ROIs: ROI-feature $\{\mathbf{r}_i\}_{i=1}^N$

  Classification scores $\{s_i\}_{i=1}^N$

  Geometric information $\{\mathbf{g}_i\}_{i=1}^N$

Methodology

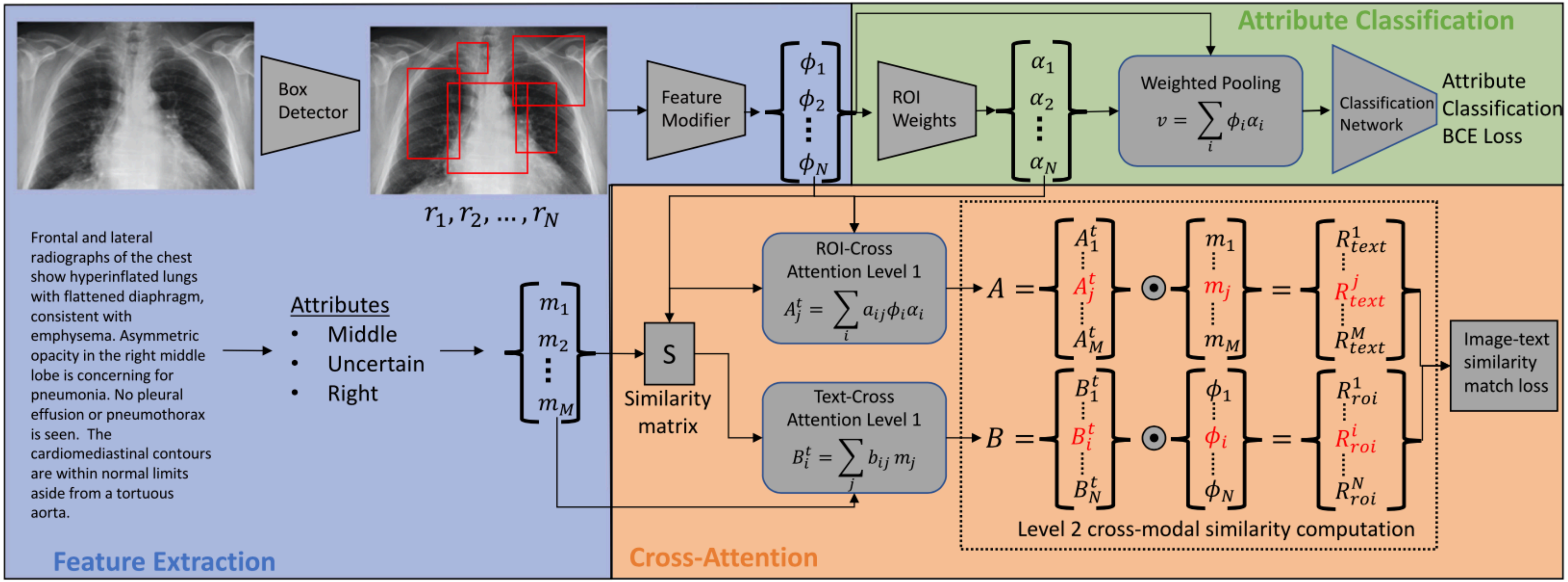

Fig. 1. Network architecture for training the attention based image-text matching for localization. (Color figure online)

Methodology

3. Network Architecture and Attention Model

A. Feature Extractor:

ROI features: $\boldsymbol{\phi}_i = W_1\mathbf{r}_i + W_2[\mathtt{LN}(W_g\mathbf{g}_i)|\mathtt{LN}(W_s s_i)]$

text features: $\{\mathbf{m}_i\}_{i=1}^{M}$

B. Attribute Classification:

Discover the appropriate ROIs that can successfully classify the attribute string

$$\mathbf{v} = \sum_{i=1}^{N} \alpha_i \boldsymbol{\phi}_i$$

Methodology

3. Network Architecture and Attention Model

B. Attribute Classification:

Discover the appropriate ROIs that can successfully classify the attribute string

With computed weights , get an aggregate ROI-feature $\mathbf{v} = \sum_{i=1}^{N} \alpha_i \boldsymbol{\phi}_i$

v, input into a multi-label attribute classification, to produce an attribute probability vector

Loss: BCE loss

Methodology

3. Network Architecture and Attention Model

C. Cross-Attention:

1. construct weighted contribution vectors

s_{ij}: the cosine-similarity between ϕ_i and m_j

$$a_{ij} = \frac{\exp(\lambda_a s_{ij})}{\sum_j \exp(\lambda_a s_{ij})}, \quad b_{ij} = \frac{\exp(\lambda_b s_{ij})}{\sum_i \exp(\lambda_b s_{ij})}$$

A_j represents the aggregate ROI feature based on its contribution to the text attribute m_j

$$\mathbf{A}_j = \sum_{i=1}^{N} \alpha_i \boldsymbol{\phi}_i a_{ij}$$

B_i represents the aggregate attribute feature based on its contribution to the ROI feature ϕ_i.

$$\mathbf{B}_i = \sum_{j=1}^{M} \mathbf{m}_j b_{ij}$$

Methodology

3. Network Architecture and Attention Model

C. Cross-Attention:

2. Calculate mean similarity: reflect how well a given image I matches with the report T

$$R^{j}_{text} = \frac{\mathbf{A}_j^T \mathbf{m}_j}{||\mathbf{A}_j||||\mathbf{m}_j||}, \quad R^{i}_{roi} = \frac{\mathbf{B}_i^T \boldsymbol{\phi}_i}{||\mathbf{B}_i||||\boldsymbol{\phi}_i||}$$

$$S_{roi}(I,T) = \frac{1}{N}\sum_{i=1}^{N} R^{i}_{roi} \text{ and } S_{text}(I,T) = \frac{1}{M}\sum_{j=1}^{M} R^{j}_{text}$$

Methodology

3. Loss Construction and Inference

   A. Loss Construction:

   Negative ROIs, I_n: taking the set of lowest ranking ROIs coming from the Retinanet box detector

   Negative attributes, T_n: finding the nearest word to the given attribute

Loss

$$\mathcal{L}_{trip} = \max(\beta - S_{roi}(I,T) + S_{roi}(I_n,T), 0) \\ + \max(\beta - S_{text}(I,T) + S_{text}(I,T_n), 0) \quad (1)$$

Final Loss:

$$\mathcal{L} = \mathcal{L}_{trip} + \mathcal{L}_{BCE}$$

Methodology

3. Loss Construction and Inference

   B. Inference:

      Only utilize the weights for ROI selection, and following by non-maximal suppression to remove redundant ROI

      No text input for testing

Results

The attribute classification on the test set for MIMIC-CXR 95.6%

Acc with an AUC of 0.84 :95.6%

Table 1. Pneumonia localization performance on different dataset using different methods, the Retinanet [10] refers to the supervised baseline.

Method	Dataset	IoU@0.25	IoU@0.5	IoU@0.75
CAM [22]	MIMIC-CXR	0.521	0.212	0.015
GradCAM [16]	MIMIC-CXR	**0.545**	0.178	0.029
Retinanet [10]	MIMIC-CXR	0.493	0.369	0.071
Proposed w/o classification	MIMIC-CXR	0.510	0.408	0.097
Proposed	MIMIC-CXR	0.529	**0.428**	**0.123**
Retinanet [10]	Chest X-ray-8	0.492	0.430	**0.115**
Proposed w/o classification	Chest X-ray-8	0.484	0.422	0.099
Proposed	Chest X-ray-8	**0.507**	**0.439**	**0.114**

Results

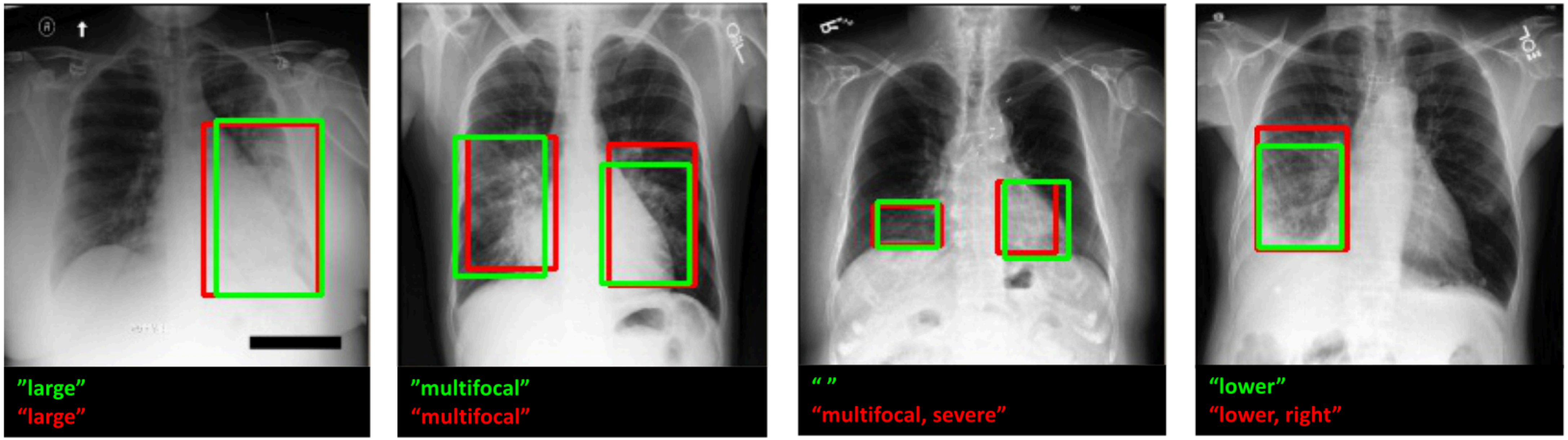

Fig. 2. Examples of localization and attribute classification from MIMIC-CXR test data. Green: expert annotated boxes and extracted attributes, Red: predicted boxes and attributes. (Color figure online)

Results

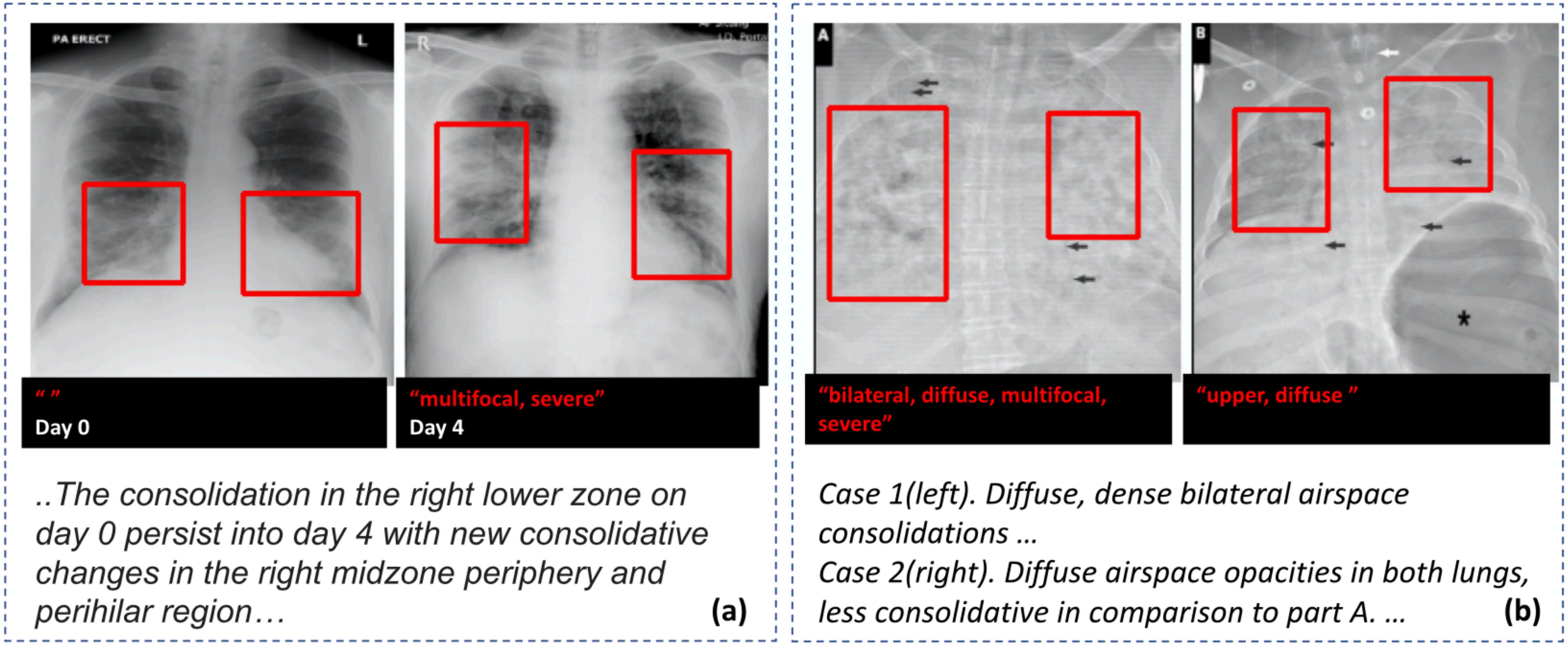

Fig. 3. Example case studies for pneumonia characterization from the COVID-19 Chest X-Ray dataset. The images, predicted attributes and localization, report snippet are shown here.

Reference

[1] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)