Contrastive Attention for Automatic Chest X-ray Report Generation

Fenglin Liu[1], Changchang Yin[2], Xian Wu[3], Shen Ge[3], Ping Zhang[2], Xu Sun[1]

[1] Peking University, China [2] The Ohio State University, USA [3] Tencent Medical AI Lab, China

fenglinliu98@pku.edu.cn; yin.731@osu.edu; zhang.10631@osu.edu; kevinxwu @osu.edu; shenge@tencent.com; xusun@pku.edu.cn

PEKING UNIVERSITY

Introduction

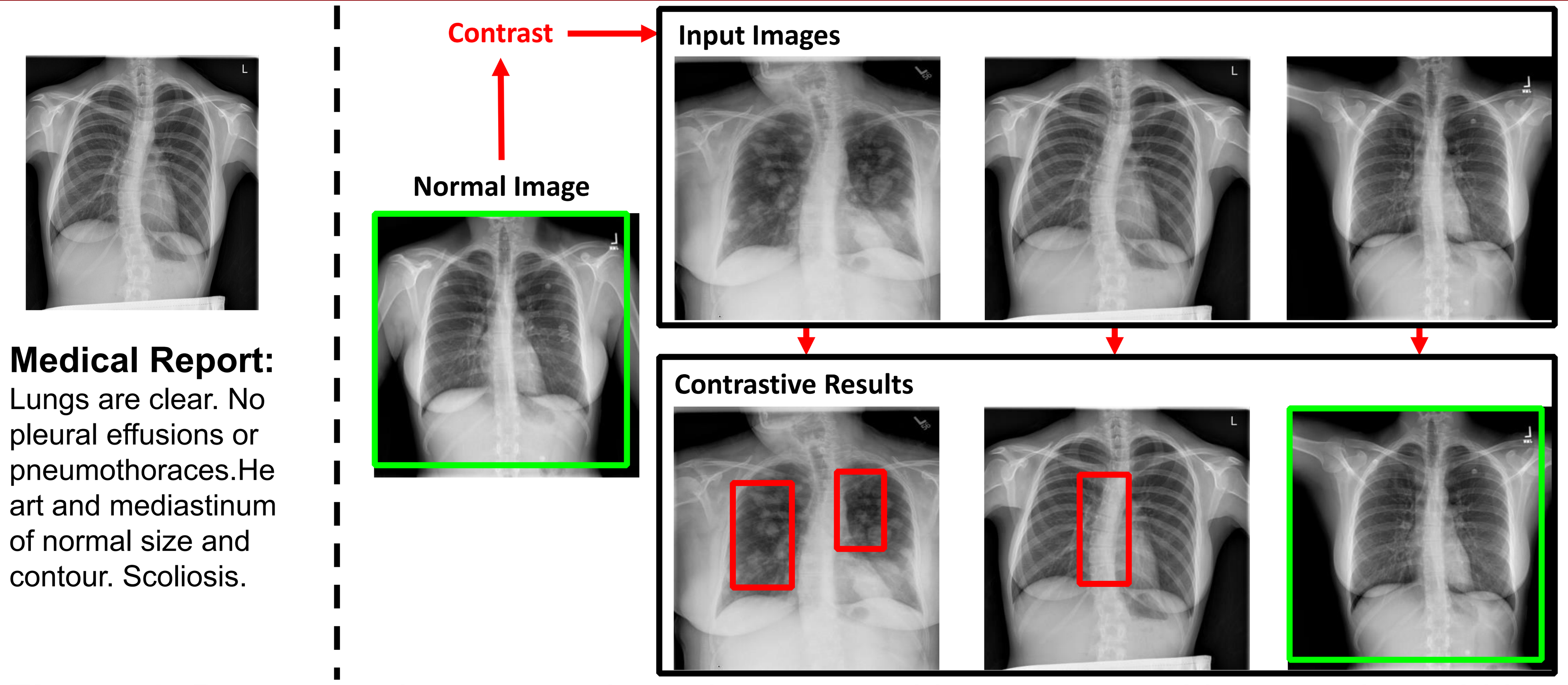

Figure 1. By contrasting current input images and known normal images, it could be easier to capture the suspicious abnormal regions (Red bounding boxes). The images with Green boxes are normal.

Task Objectives:

It aims to generate a long paragraph describing both the normal and abnormal regions, which can assist radiologists in clinical decision-making.

- cover key medical findings: e.g., heart size and lung opacity.
- correctly describe any abnormalities and its details: e.g., the location and shape of the abnormality.
- correctly describe potential diseases: e.g., effusion and consolidation.

Urgent goal/core value: to correctly **capture** and **describe** abnormalities.

Task Challenges:

There are serious data deviation problems in the medical report corpus.

- the normal images dominate the dataset over the abnormal ones [1].
- given an input image, the normal regions usually dominate the image and their descriptions dominate the medical report [2,3].

Solution:

- To capture the abnormal regions of given chest X-ray image, a natural intuition is to compare it with normal images and identify the differences. Therefore, we propose the Contrastive Attention to enable existing methods to better capture and describe the abnormalities.

Approach

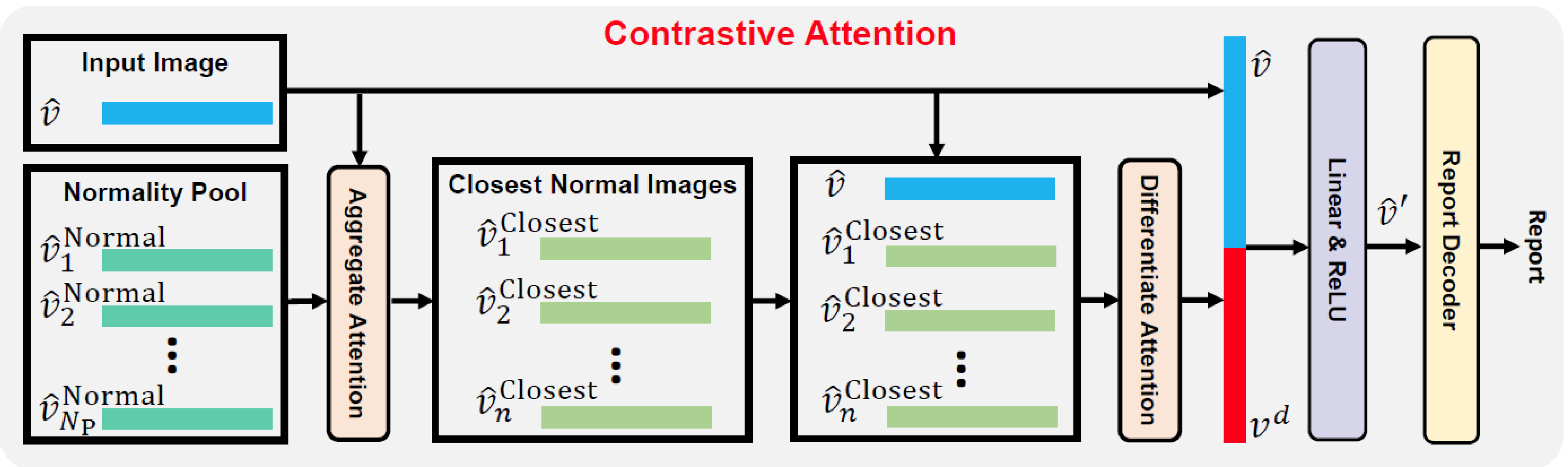

Figure 2. Illustration of our proposed **Contrastive Attention**, which consists of the **Aggregate Attention** and **Differentiate Attention**. In particular, the Aggregate Attention devotes to finding the normal images that are closest to the current input image in the normality pool. The Differentiate Attention devotes to summarizing the common information between the input image and the closest normal images and subtract it from the input image to capture the differentiating properties between the input image and the normal images.

Our approach includes of the **Aggregate Attention** and **Differentiate Attention**.

Aggregate Attention

Given the input image $\hat{v} \in \mathbb{R}^{1\times d}$ and the normality pool P, the aggregate attention is defined as:

$$\hat{v}^{\text{Closest}} = \text{Att}(\hat{v}, P) \qquad \text{Att}(x, y) = \text{softmax}(M)y \quad \text{where} \ \ M = \frac{x\text{W}^x(y\text{W}^y)^T}{\sqrt{d}}$$

$$P' = \text{Aggregate-Attention}(\hat{v}, P) = [\text{Att}_1(\hat{v}, P); \text{Att}_2(\hat{v}, P); \dots; \text{Att}_n(\hat{v}, P)]$$
$$= \{\hat{v}_1^{\text{Closest}}, \hat{v}_2^{\text{Closest}}, \dots, \hat{v}_n^{\text{Closest}}\} \in \mathbb{R}^{n\times d}$$

Differentiate Attention

The first step is learning to summarize the common information $v^c \in \mathbb{R}^{1\times d}$ between the current input image $\hat{v} \in \mathbb{R}^{1\times d}$ and the closest normal images $P' \in \mathbb{R}^{n\times d}$:

$$v^c = \text{AveragePooling}\left(\text{Att}\left([\hat{v}; P'], [\hat{v}; P']\right)\right)$$

Next, we remove (i.e., subtract) the common information $v^c \in \mathbb{R}^{1\times d}$ from the input image $\hat{v} \in \mathbb{R}^{1\times d}$ to obtain the contrastive information $v^d \in \mathbb{R}^{1\times d}$:

$$v^d = \hat{v} - v^c$$

Experiments

- We evaluate our contrastive attention on **ten** baselines.

Settings	Methods	Dataset: MIMIC-CXR (Johnson et al., 2019)						Dataset: IU-X-ray (Demner-Fushman et al., 2016)					
		B-1	B-2	B-3	B-4	M	R-L	B-1	B-2	B-3	B-4	M	R-L
(a)	NIC (Vinyals et al., 2015)†	0.290	0.182	0.119	0.081	0.112	0.249	0.352	0.227	0.154	0.109	0.133	0.313
	w/ Contrastive Attention	0.317	0.200	0.127	0.089	0.120	0.262	0.368	0.232	0.166	0.118	0.144	0.323
(b)	Visual-Attention (Xu et al., 2015)†	0.318	0.186	0.122	0.085	0.119	0.267	0.371	0.233	0.159	0.118	0.147	0.320
	w/ Contrastive Attention	0.309	0.202	0.129	0.093	0.122	0.265	0.384	0.245	0.172	0.125	0.141	0.315
(c)	Spatial-Attention (Lu et al., 2017)†	0.302	0.189	0.122	0.082	0.120	0.259	0.374	0.235	0.158	0.120	0.146	0.322
	w/ Contrastive Attention	0.320	0.204	0.129	0.091	0.122	0.266	0.378	0.236	0.161	0.116	0.146	0.335
(d)	Att2in (Rennie et al., 2017)†	0.314	0.199	0.126	0.087	0.125	0.265	0.410	0.257	0.173	0.131	0.149	0.325
	w/ Contrastive Attention	0.327	0.205	0.132	0.095	0.124	0.271	0.442	0.281	0.200	0.150	0.171	0.344
(e)	Adaptive-Attention (Lu et al., 2017)†	0.307	0.192	0.124	0.084	0.119	0.262	0.433	0.285	0.194	0.137	0.166	0.349
	w/ Contrastive Attention	0.330	0.208	0.134	0.095	0.126	0.270	0.425	0.279	0.198	0.142	0.167	0.347
(f)	Up-Down (Anderson et al., 2018)†	0.318	0.203	0.128	0.089	0.123	0.266	0.389	0.251	0.170	0.126	0.154	0.317
	w/ Contrastive Attention	0.336	0.209	0.134	0.097	0.128	0.273	0.378	0.246	0.169	0.129	0.152	0.330
(g)	HLSTM (Krause et al., 2017)†	0.321	0.203	0.129	0.092	0.125	0.270	0.435	0.280	0.187	0.131	0.173	0.346
	w/ Contrastive Attention	0.352	0.216	0.145	0.105	0.139	0.276	0.453	0.290	0.203	0.153	0.178	0.361
(h)	HLSTM+att+Dual (Harzig et al., 2019)†	0.328	0.204	0.127	0.090	0.122	0.267	0.447	0.289	0.192	0.144	0.175	0.358
	w/ Contrastive Attention	0.323	0.202	0.130	0.102	0.138	0.277	0.464	0.292	0.205	0.149	0.176	0.364
(i)	Co-Attention (Jing et al., 2018)†	0.329	0.206	0.133	0.095	0.129	0.273	0.463	0.293	0.207	0.155	0.178	0.365
	w/ Contrastive Attention	0.351	0.213	0.148	0.106	0.147	0.270	0.486	0.311	0.223	0.178	0.187	0.372
(j)	Multi-Attention (Huang et al., 2019)†	0.337	0.211	0.136	0.097	0.130	0.274	0.468	0.299	0.211	0.155	0.180	0.366
	w/ Contrastive Attention	0.350	0.219	0.152	0.109	0.151	0.283	0.492	0.314	0.222	0.169	0.193	0.381

Table 1. Results on the MIMIC-CXR [4] and IU-X-ray [5]. B-n, M and R-L are short for BLEU-n, METEOR and ROUGE-L, respectively. Higher is better in all columns. Existing methods equipped with our Contrastive Attention consistently outperform baselines.

- As we can see, these baseline models enjoy a comfortable improvement with our method on most metrics.

Methods	Dataset: MIMIC-CXR (Johnson et al., 2019)						Dataset: IU-X-ray (Demner-Fushman et al., 2016)					
	B-1	B-2	B-3	B-4	M	R-L	B-1	B-2	B-3	B-4	M	R-L
HRGR-Agent (Li et al., 2018)	-	-	-	-	-	-	0.438	0.298	0.208	0.151	-	0.322
CMAS-RL (Jing et al., 2019)	-	-	-	-	-	-	0.464	0.301	0.210	0.154	-	0.362
SentSAT + KG (Zhang et al., 2020a)	-	-	-	-	-	-	0.441	0.291	0.203	0.147	-	0.367
Transformer (Chen et al., 2020c)	0.314	0.192	0.127	0.090	0.125	0.265	0.396	0.254	0.179	0.135	0.164	0.342
R2Gen (Chen et al., 2020c)	0.353	0.218	0.145	0.103	0.142	0.277	0.470	0.304	0.219	0.165	0.187	0.371
Contrastive Attention (Ours)	0.350	0.219	0.152	0.109	0.151	0.283	0.492	0.314	0.222	0.169	0.193	0.381

Table 2. Comparison with existing state-of-the-art methods on the test set of the MIMIC-CXR dataset [4] and the IU-X-ray dataset [5].

- We achieve the state-of-the-art results on the two datasets.

References

[1] Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. *In CVPR, 2016.*

[2] Show, describe and conclude: On exploiting the structure information of chest x-ray reports. *In ACL, 2019.*

[3] Exploring and distilling posterior and prior knowledge for radiology report generation. *In CVPR, 2021.*

[4] MIMIC-CXR: A large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042.*

[5] Preparing a collection of radiology examinations for distribution and retrieval. *J. Am. Medical Informatics Assoc., 23(2):304–310.*