LIMITR: Leveraging Local Information for Medical Image-Text Representation

Gefen Dawidowicz 1, Elad Hirsch 1, Ayellet Tal 1,2
1 Technion – Israel Institute of Technology 2 Cornell Tech
{gefen, eladhirsch}@campus.technion.ac.il, ayellet@ee.technion.ac.il

Abstract

Medical imaging analysis plays a critical role in the diagnosis and treatment of various medical conditions. This paper focuses on chest X-ray images and their corresponding radiological reports. It presents a new model that learns a joint X-ray image & report representation. The model is based on a novel alignment scheme between the visual data and the text, which takes into account both local and global information. Furthermore, the model integrates domain-specific information of two types—lateral images and the consistent visual structure of chest images. Our representation is shown to benefit three types of retrieval tasks: text-image retrieval, class-based retrieval, and phrase-grounding. Our code is publicly available 1.

1. Introduction

X-ray imaging is one of the most common medical examinations performed, where in the U.S. alone 70 million scans are performed every year [20]. During a post-scan routine, a medical report is written by a radiologist. This is a challenging task even for trained personnel, since the pathologies, which typically occupy a small portion of the image, might be characterized by subtle (sometimes distributed) anatomical changes. Automating the analysis has the potential to aid experts and to speed-up the process.

The task of producing a joint representation of medical image-report data inherently differs from that of natural image-text, which has been extensively explored recently [6]. The available data is orders of magnitude smaller than that of natural images. Furthermore, the data is highly unbalanced, since most of the examples are normal. Even in abnormal examples, most of the image (report) is normal.

Due to the importance of the task, the available medical data has been used for a variety of applications, including class-based retrieval (retrieving data of the same diagnosis) [8, 27, 17, 25], pathology classification [23, 8, 3, 27, 17, 25], detection [8, 3, 23, 18], and phrase-grounding [3]. Often, these tasks do not require subtle details, such as the severity of the pathology, its exact location, or findings that are beyond the pre-defined pathologies.

We propose a novel model, which learns a joint X-ray image-report representation that is attentive to subtle details, additional descriptions of the pathologies, uncertainties etc., as illustrated in Figure 1. It is based on three key ideas. First, our method learns to utilize both global align-
ment (image-report pairs) that captures high-level information, such as the sheer existence of a disease, and local alignments (region-word pairs) that capture subtle details and abnormalities. Second, it benefits from lateral images, which are usually ignored. Third, it utilizes domain-specific localization information. We elaborate below.

Local alignment in medical imaging is challenging since the data is not annotated locally by bounding boxes and their labels. This is in contrast to natural image datasets, which provide localized ground-truth information [4, 9, 13, 14, 15, 16]. Furthermore, localization ambiguity is inherent in medical imaging, as report findings may correspond to multiple image regions. To this end, we propose a new aggregation method and a new loss, which synthesize both region-word alignments within a single pair and global and local alignments across pairs.

Second, we propose to use lateral views, if they exist, just like radiologists do. These views may provide additional information, yet they are largely ignored by other representation learning works. We introduce an attention model that learns for each portion of the report when to consider both images, when only one, and when none.

Lastly, our model utilizes some basic domain-specific localization information—the structure of the human body; for instance, the heart is always in between the lungs and is approximately in the same image position. We show that we can add global positional encoding for the whole dataset. This encoding also allows the network to better learn local connections between the frontal and the lateral views.

To demonstrate the benefit of our model, we evaluate it for three different retrieval tasks: (1) **Text-image retrieval**: Given a report, the goal is to find the most suitable image and vice versa. In this task, we expect the corresponding report (image) to be retrieved, as illustrated in Figure 1, where all the information that appears in the text and in the image, are taken into account. This task demonstrates the ability of our method to accurately capture subtleties.

(2) **Phrase-grounding**: Given an image and a corresponding phrase, the goal is to produce an attention map of the similarity between the phrase and each region. This task demonstrates the quality of the local alignment. (3) **Classification retrieval**: Given a textual description, the goal is to retrieve images that belong to the same class of the description. This is the most common retrieval task. We present SoTA results for all these tasks.

Hence, our work makes the following contributions:

1. We propose a novel model for learning a joint X-ray image & report representation. It is based on a new local and global alignment scheme. This alignment is shown to boost performance.

2. Our model integrates additional domain-specific knowledge—lateral images and visual structure. This information further improves performance.

3. The benefit of our model is demonstrated on three retrieval tasks, showing its ability to capture fine features in both images and reports. We demonstrate SoTA results on commonly-used datasets.

2. Related work

Recent works on text-image multi-modal representations for chest X-ray (CXR) datasets are used for a variety of tasks. These tasks are either uni-modal or multi-modal, as briefly reviewed hereafter.

**Uni-modal tasks.** The common uni-modal tasks are visual, where the focus is on classification and localization. Specifically, given an image, the goal of [3, 8, 23, 25, 27] is to classify it into pre-defined diagnoses. This can be done for multi-label classes [8, 23, 25, 27] or for binary classification [3, 8, 23, 27].

In localization, we are given images and aim to learn localized information for detection and segmentation tasks [3, 8, 18, 23]. These tasks are designed for single-pathology studies and rely on relatively small datasets. Examples include RSNA pneumonia detection [19], foreign objects detection [26] and SIIM pneumothorax segmentation [1].

Textual analysis in this domain is less prevalent. In [3], the focus is on CXR domain-specific language understanding tasks, such as text classification and masked-token prediction. The joint training with the images provides a superior language model for these tasks.

**Multi-modal tasks.** Our focus is on multi-modal tasks, for which there exist less works. In [8, 27, 25] the task is class-based retrieval, i.e. given an example in one modality (image/text), retrieve examples from the other modality. The requirement is that the retrieved examples should belong to the same class of diagnoses as the query.

In phrase-grounding, introduced by [3], we are given an image and a corresponding phrase describing a pathology in the image, and aim to localize the image regions that match the phrase. This is a challenging task as specific relations between phrases and certain image features need to be learned. However, this work ignores studies with multiple pathologies, as well as more elaborated phrases that express uncertainties and descriptions of normalities.

We introduce an additional retrieval task, text-image (or image-text) retrieval, where the accuracy of the representation can be better evaluated. We expect to retrieve the exact match from one modality, given the other.

3. Model

Our goal is to learn an informative joint representation of X-ray images and their reports. In this joint representation space, an image and its corresponding report should be mapped to close points, whereas mismatched pairs should reside farther apart. Recall that our model realizes three key
Figure 2: Model. The model consists of three blocks. (1) For feature extraction, two CNNs, $E_{\text{img}}^l$ & $E_{\text{img}}^f$ for the lateral & the frontal images, are used as the image encoders and one pre-trained network, $E_{\text{txt}}$, is used as the text encoder. The local visual representations are concatenated and a global representation is created using self-attention. A global representation is similarly created for the report. The distance between these representations forms our global alignment $A_g$, which is used in our global loss, $L_g$. (2) During cross-modal alignment, alignments between the local representations of the two modalities are calculated (Figure 3). These alignments, $\{a_j\}$, are used for our local internal loss $L_{\text{int}}$ (Figure 5). Thanks to our attention mechanism, the model works well with or without lateral images. (3) The local alignments are aggregated using learned significance scores, to create the final image-report similarity score, $A_{\text{agg}}$. These scores are used in our local external loss $L_{\text{ext}}$ (Figure 4).

3.1. Feature extraction

Given an image $x_v$ and its corresponding report $x_t$, the visual and the textual features, are independently extracted. Recall that in medical imaging, we lack local annotated data (bounding boxes and their labels); furthermore, the images suffer from localization ambiguity, as findings may correspond to multiple image regions. Thus, pre-trained object detectors are not as useful as in the natural domain.

Visual feature extraction. We benefit from the localized nature of the intermediate layers of CNNs ($E_{\text{img}}$ in Fig. 2), to obtain $N_r$ region-level visual features $\{v_1, ..., v_{N_r}\}$. In particular, we use the output of the last convolution layer.

Next, we enrich the extracted features with our knowledge of the layout of the human body. Specifically, in chest X-ray the organs are approximately in the same positions (which guide radiologists in the diagnosis process). We leverage this structure by encoding and integrating it within learning. We realize it through adding positional encoding. Since our positional encoding is inherent in the input, it conceptually differs from that being used in transformers, which encode the relative positions. We sum each visual local feature vector $v_i$ with a corresponding vector that encodes its spatial 2D position. In our implementation we use the 2D sinusoidal encoding of [5], as follows. Let $v_i$ be the $i^{th}$-patch features in the image and $(x, y)_i$ be the coordi-
nates of this patch. Our visual feature vector is then defined as the sum of the visual encoder output with the positional encoding corresponding to the patch location in the image \( v_i \leftarrow v_i + PE((x, y)_i) \), where \( PE((x, y)_i) \) is that of \([5]\).

We observe that not all local regions are as important. Specifically, some small regions—those that contain abnormalities—should count more than others. To this end, we apply the self-attention operator of \([22]\) on the local features, in order to extract our global image feature representation, \( \bar{v} \). This \( \bar{v} \) captures the relationships between all the local representations of a given image \( x_v \).

In our implementation we use ResNet-50 \([7]\) as the image encoder \( E_{img} \), as it has been previously shown to benefit medical images \([3, 8, 27]\).

**Frontal and lateral information.** Lateral images, which exist in 50% of the studies, contain information that may improve diagnosis, yet they are largely ignored. Our approach utilizes the information from the lateral images, by learning to weigh information from both views. In addition, as studies do not necessarily contain the lateral view, our model learns from studies with a single view, as well as from studies with both views. Specifically, for each view type, we train a separate image encoder, \( E^i_{img} \) and \( E^f_{img} \). The features from the two encoders are concatenated to form the set \( \{ v^i_1, ..., v^i_{N_r}, v^f_1, ..., v^f_{N_r} \} \) and are inserted to the cross-modal alignment module. When a lateral image is not available, we set \( \{ v^f_{i,j} \}_{i=1}^{N_r} = \emptyset \).

**Textual features extraction.** Given a report \( x_t \), it is first tokenized, using the vocabulary of \([2]\). The sequence of tokens is then fed into a text encoder, \( E_{txt} \). This yields local textual features, at the token level, \( \{ t_1, ..., t_{N_w} \} \). Lastly, in analogy to image regions, certain words are more important than others, we apply self-attention to extract the global textual features, \( \bar{t} \).

In our implementation, we use BioClinicalBERT \([2]\) as our text encoder, due to its high performance.

### 3.2. Cross-modal alignment

Given a visual and textual local representations, \( \{v_1, ..., v_{N_r}\} \) and \( \{t_1, ..., t_{N_w}\} \) respectively, our goal is to learn the alignments between the two modalities, including the alignments between report words and image regions. To do that, our model should address two challenges: (1) the lack of local annotations connecting the two modalities, (2) the pathologies may occupy only a small area of the image, as well as a small part of the report. To this end, a fine-grained approach is sought-after.

Our approach, which is illustrated in Figure 3, takes into account the global representations of images and reports, as well as the local representations of regions and words. This is done both by weighing the image region features with respect to each textual feature and vice versa—by weighing the textual features with respect to each region feature. For clarity, in the following we will describe only how this is done in the first case (weighing image regions with respect to text). However, in the loss, both are summed.

First, the cosine similarity \( c_{ij} \) between \( v_i \) and \( t_j \) is computed, to create \( c_j = [c_{1j}, c_{2j}, ..., c_{N_r,j}] \). It is further normalized using softmax, in order to get an attention weight:

\[
    w_j = \text{softmax}(\lambda c_j), \quad (2)
\]

The attended visual feature \( u_j \) with respect to the \( j \)th word is the weighted sum of all the visual local representations:

\[
    u_j = \sum_{i=1}^{N_r} w_j[i] \cdot v_i. \quad (3)
\]

Next, we calculate the alignment between \( t_j \) and its corresponding \( u_j \). The local alignment \( a_j \) is calculated as:

\[
    a_j = A(u_j, t_j) = \frac{u_j \circ t_j}{||u_j \circ t_j||_2}, \quad (4)
\]

where \( \circ \) is an element-wise multiplication and \( ||\cdot||_2 \) is the \( L_2 \)-norm. Note that unlike \([8, 18, 24]\), our alignment \( a_j \) is a vector rather than a scalar (that averages the entries). This is important for our aggregation method, as will be discussed in Section 3.3, where we learn the importance of the different \( a_j \)’s. Our alignments \( \{a_j\} \) will be later used for the local alignment loss \( L_{int} \).

Similarly to the local alignments, we compute the global alignment, given the global image feature vector, \( \bar{v} \) and the global report feature vector, \( \bar{t} \). It is computed as:

\[
    A_g = A(\bar{v}, \bar{t}) = \frac{\bar{v} \circ \bar{t}}{||\bar{v} \circ \bar{t}||_2}. \quad (5)
\]

This \( A_g \) will be later used for the global loss \( L_g \).
The cross-modal alignment is computed between each region in both the frontal and the lateral images to each token of the report. Hence, if significant information appears in the two images, regions from both receive high weights. But, when important information appears only in one of the images, only its regions will get high weights. In a sense, our model mimics the radiologists actions. When a pathology is available in the two views, they examine both; otherwise, they focus on the relevant view.

3.3. Aggregation

Our goal is to aggregate all the local alignments of a given image-report pair, in order to create the final representation of the pair, \( a_f \). We do not wish to simply sum the alignments [8], since not all regions should be treated equally. In particular, most image regions and their corresponding report descriptions are normal observations, shared by all examples in the dataset, whereas the pathologies mostly occupy small regions. The distinct information, e.g. the pathology, should be given more weight.

To account for this variance, we suggest to weigh the local alignments in accordance with their informativeness. Let the local alignments, computed at the alignment module, be \( \mathcal{A}_T = \{a_1, a_2, a_3, \ldots, a_{N_u}\} \) (Equation 4). These alignments are aggregated into a single alignment vector using a weighted sum. The weight of each vector is computed by self-attention, where (through the interaction between the local regions) it is determined which should be more attended, in order to better perform the task. Thus, for each \( a_t \) we compute its weight relative to the set \( \mathcal{A}_T \). Let \( \bar{a} \) be the mean of \( \mathcal{A}_T \). The weight of \( a_t \) is defined as:

\[
q_t = \left( \text{softmax} \left( \frac{W_q \bar{a} \cdot (W_k \mathcal{A}_T)^T}{\sqrt{d}} \right) \right)_t,
\]

where \( 1 \leq t \leq N_u, W_q \) and \( W_k \) are linear transformations of the self-attention, and \( d \) is the feature dimension.

The final alignment vector between an image and a report is defined as:

\[
a_f = \sum_{t=1}^{N_u} q_t (W_v \cdot a_t).
\]

This aggregated alignment vector represents the image-report pair. It is passed through an FC layer to produce the final scalar alignment score \( \mathcal{A}_{agg} \), which will be later used for the external loss \( L_{ext} \).

3.4. Loss

Recall that our goal is to maximize the similarity between positive image-report pairs and minimize the similarity between negative pairs. To achieve this, we use three instances of the contrastive loss function [21], each expressing a different concept: one global and two local, as seen in Equation 1. The latter two enable different regions or words to be considered as having different significance. We elaborate on these losses hereafter.

The global loss, \( L_g \), attempts to maximize the global alignment \( \mathcal{A}_g \) (Equation 5) of positive image-report pairs and minimize the global alignment of negative pairs. Let \( (x^k_v, x^k_t) \) be a corresponding pair and \( \tau \) be a temperature parameter. The loss is defined as:

\[
L_g(x^k_v, x^k_t) = l^{x^k_v|x^k_t} + l^{x^k_t|x^k_v},
\]

\[
l^{x^k_v|x^k_t} = - \log \left( \frac{\exp(\mathcal{A}_g(x^k_v, x^k_t)/\tau)}{\sum_{j=1}^{N} \exp(\mathcal{A}_g(x^k_v, x^j_t)/\tau)} \right),
\]

\[
l^{x^k_t|x^k_v} = - \log \left( \frac{\exp(\mathcal{A}_g(x^k_v, x^k_t)/\tau)}{\sum_{j=1}^{N} \exp(\mathcal{A}_g(x^j_t, x^k_v)/\tau)} \right).
\]

Here, for a given \( x^k_v, l^{x^k_v|x^k} \) aims to increase its similarity to its corresponding report and decrease its similarity to other reports \( (j \neq k) \). Similarly, for a given report \( x^k_t, l^{x^k_t|x^k} \) aims to increase its similarity to its corresponding image and decrease its similarities to other images in the batch.

The local external loss, \( L_{ext} \), aims to increase the similarity of positive pairs and decrease the similarity of negative ones, this time through the use of the local alignments. Thus, we use the aggregated local similarity score, \( \mathcal{A}_{agg} \) (Section 3.3), instead of \( \mathcal{A}_g \), as the objective for maximization and minimization, as illustrated in Figure 4.

The local internal loss, \( L_{int} \), is given the local textual representation, \( t_j \), as well as its corresponding attention weighted visual representation, \( u_j \), from Equation 3. It aims to improve the local representations by maximizing the similarity between corresponding pairs and minimizing the similarity between non-corresponding pairs of the same example. The loss is defined as follows:

\[
L_{int}(x^k_v, x^k_t) = \sum_{j=1}^{N_u} (l^{t_j|u}\cdot l^{u_j|t}),
\]

\[
l^{t_j|u} = - \log \left( \frac{\exp(a_j(t_j, u_j)/\tau)}{\sum_{i=1}^{N_u} \exp(a_j(t_i, u_j)/\tau)} \right),
\]

\[
l^{u_j|t} = - \log \left( \frac{\exp(a_j(t_j, u_j)/\tau)}{\sum_{i=1}^{N_u} \exp(a_j(t_j, u_i)/\tau)} \right).
\]

Here, for a local textual representation \( t_j \), \( l^{t_j|u} \) aims to increase the similarity between \( t_j \) and its corresponding visual weighted representation, \( u_j \), and to decrease the similarity to other \( u_m \neq j \) from the same study. The same rationale applies to \( l^{u_j|t} \). We calculate \( l^{t_j|u} \) and \( l^{u_j|t} \) separately, for each image-report pair \( (x^k_v, x^k_t) \), as illustrated in Figure 5. Finally, we sum across all local pairs for a given study and then across all studies in the batch. Recall that the procedure described here is performed also for the visual representation in regards to the weighted textual representations.
Figure 4: **Local external loss.** Top: Given a local textual representation $t_j$ (clavicula) and its corresponding weighted visual representation $u_j$, we generate an aggregated alignment score $A_{agg}$, based on the local alignments $\{a_j\}$. Bottom: $A_{agg}$ is used to bring closer (green) the corresponding image-report pair $(x_{t1}, x_{v1})$ and farther away (red) non-corresponding pairs e.g., $(x_{t1}, x_{v2})$.

Figure 5: **Local internal loss.** Given local textual representations, $t_j$, and visual representations, $u_j$, this loss brings closer (green) corresponding representations, e.g. $t_j$ (clavicula) and $u_j$ (clavicula) and further apart (red) non-corresponding representations, e.g. $t_j$ (cardiac) and $u_j$ (cardiac), from the same example.

4. Experimental results

We examine our method on three retrieval applications: text-image retrieval, phrase-grounding, and class-based retrieval. For each application we compare against previous works that evaluate the specific application, as well as against additional methods that we trained. The supplementary material shows that our representations are also useful for classification.

<table>
<thead>
<tr>
<th>Method</th>
<th>Image-to-Text</th>
<th></th>
<th>Text-to-Image</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@1</td>
<td>R@5</td>
<td>R@10</td>
<td>R@1</td>
</tr>
<tr>
<td>MGCA [23]</td>
<td>25.8</td>
<td>51.9</td>
<td>62.1</td>
<td>27.9</td>
</tr>
<tr>
<td>ConVIRT [27]</td>
<td>30.1</td>
<td>53.9</td>
<td>63.8</td>
<td>29.2</td>
</tr>
<tr>
<td>GLoRIA [8]</td>
<td>30.3</td>
<td>57.5</td>
<td>66.5</td>
<td>24.0</td>
</tr>
<tr>
<td>Ours w/o LT&amp;PE</td>
<td>36.1</td>
<td>59.1</td>
<td>69.1</td>
<td>36.4</td>
</tr>
<tr>
<td>Ours</td>
<td>39.7</td>
<td>63.2</td>
<td>71.7</td>
<td>37.7</td>
</tr>
</tbody>
</table>

Table 1: **Text-Image retrieval results.** Given a report, our goal is to retrieve the matching image and vice versa. Our results outperform those of other methods on MIMIC-CXR, even without lateral (LT) images and structural information (PE). Our full method further improves the results.

Datasets. We ran our experiments on three datasets: MIMIC-CXR, CheXpert 5X200, and MS-CXR.

(1) **MIMIC Chest X-ray (MIMIC-CXR)** is a large, publicly available, dataset of chest radiographs, with free-text radiology reports [12]. The dataset contains 377,110 images, corresponding to 227,835 radiographic studies performed at the Beth Israel Deaconess Medical Center. Each study in the dataset contains a report and one or more images of frontal or lateral, where each study contains at least one frontal image. Each radiograph is associated with one or more classes, out of 14 optional diagnostic labels. The two main sections of interest of the report are findings and impression. Studies that do not contain these sections are filtered out. The data is randomly sampled to obtain validation and test sets, containing 1,000 studies each. The remaining training set contains 205,000 studies, of which 100,000 have both frontal and lateral images.

(2) **CheXpert 5X200** contains 200 image-report pairs for 5 abnormality categories [8], sampled from the CheXpert dataset [10]. Each example in the dataset belongs to a single abnormality category.

(3) **MS-CXR** is a subset of MIMIC-CXR, which is extended for phrase-grounding [3]. It contains labeled (text descriptions) bounding boxes for each image. In total, there are 1,153 pairs of region-text pairs.

Text-Image retrieval. Given a report, our goal is to find the most suitable image and vice versa. This task evaluates how close matching pairs are in the feature space. Hence, for this task, only the ground-truth image-report pairs are considered as positive. When both the frontal and the lateral images are available we use them both; when only one exists we fill the missing image with zeros. The accuracy of our model is measured by the Recall@K metric, which returns the percentage of queries whose true match is successfully ranked within the top K matches.

Table 1 compares our performance on MIMIC-CXR to SoTA methods for representation learning, ConVIRT [27], GLORIA [8] and MGCA [23], which we have trained.
Table 2: Phrase-grounding results. Given a phrase and an image, the goal is to produce a similarity map between the phrase and the image. Our results outperform those of other methods, as reported in [3], on MS-CXR. The results are measured using CNR; higher values indicate good localization of the phrase in the image. Qualitative results can be found in the supplementary materials.

Our method outperforms other methods in R@1, R@5 and R@10, both for the image-to-text task and for the text-to-image task, even without utilizing structural knowledge and lateral images. Encoding domain structure and using the additional lateral images further improve the results.

**Phrase-grounding.** Given an image and a corresponding phrase, the goal is to produce an attention map of the similarity between the phrase and each image region [3]. The ground-truth is given as bounding boxes. Hence, this task evaluates the local alignments. Following [3], we measure the performance using contrast to noise ratio (CNR), which measures the attention density inside the ground-truth bounding box. High values indicate that the network accurately detects the regions of interest for the given phrase. It is defined as:

$$\text{CNR} = \frac{|\mu_A - \mu_{\bar{A}}|}{(\sigma_A^2 + \sigma_{\bar{A}}^2)^{\frac{1}{2}}}$$

where $A$ and $\bar{A}$ are the interior and exterior of the bounding box, and $\mu$ and $\sigma$ are the mean and variance of the attention maps in each region.

Table 2 presents the results on the MS-CXR dataset. The results are presented for each of the 8 diagnostic categories, as well as for the average over the whole dataset. Our performance is compared to that of BioViL [3], which introduced this task, and to the results of GLoRIA and ConVIRT that are reported in [3]. Note that BioViL uses a different text encoder (CXR-bert) from all other methods in the table (Bioclinical-Bert). Our model achieves SoTA results, demonstrating the strength of our local representations and the localization ability of our model.

**Class-based retrieval.** Given an image, our goal is to retrieve reports that belong to the same class of the image. For this task, a positive pair is defined as image-report that have the same abnormality label. We follow [8]’s settings, performing image-to-text retrieval and using the Precision@K metric to measure the accuracy of the retrieval. Precision@K is defined as the fraction of retrieved items (images) within the top $K$, which belong to the same class of abnormality as that of the query (textual descriptions).

Table 3 shows that our results outperform those of other methods in almost all metrics. We note that [8] was originally trained and evaluated on the CheXpert dataset [10]. Since the reports of CheXpert are unavailable to the public, all models were trained on MIMIC-CXR [11] and evaluated in a zero-shot manner on the CheXpert 5X200 dataset.

**5. Ablation study**

This section evaluates the contribution of the different components of our method.

**Losses.** Our model optimizes a combination of three losses: global, local-internal and local-external (Equation 1). Table 4 demonstrates the contribution of each loss for both image-to-text and text-to-image retrieval tasks. The best performance is achieved by the combination of the three losses. Thus, attempting to achieve the three goals of these losses indeed improves the learning process.

**Positional encoding.** Recall that we utilize domain-specific
Figure 6: Domain-knowledge importance. (A) The positional information enables the model to consider the signs of right pleural effusion (blue) and small left pleural effusion (orange). Thanks to our positional encoding, our model succeeds in the text-image retrieval task, but would fail otherwise. (B) The visual clue, the flattened diaphragm, which is a sign of over-inflation appears only in the lateral image (highlighted in blue). Thanks to the lateral view, our model succeeds in the text-image retrieval task; it would fail if it used only the frontal image.

<table>
<thead>
<tr>
<th>LT PE</th>
<th>Image-to-Text</th>
<th>Text-to-Image</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@1 R@5 R@10</td>
<td>R@1 R@5 R@10</td>
</tr>
<tr>
<td>✓</td>
<td>36.1 59.1 69.1</td>
<td>36.4 60.7 70.5</td>
</tr>
<tr>
<td>✓</td>
<td>38.1 61.6 72.1</td>
<td>37.3 61.2 71.5</td>
</tr>
<tr>
<td>✓ ✓</td>
<td>39.7 63.2 71.7</td>
<td>37.7 62.1 71.3</td>
</tr>
</tbody>
</table>

Table 5: Domain knowledge. Adding either lateral (LT) images or visual structure information (PE) to the baseline frontal images improves the results in both tasks.

6. Conclusions

This paper presented a new model for learning a joint X-ray image & report representation. The model is based on a new alignment scheme that considers both local and global information. In addition, we propose to enrich the model with domain-specific information.

The benefits of our representation is demonstrated on three types of retrieval tasks, two of which require large precision: text-image retrieval, phrase-grounding, and class-based retrieval. Our model is shown to outperform SoTA models, even when the additional knowledge is unavailable.

The domain-specific knowledge adds to performance. In the future, we would like to study loss functions that allow an image (text) to be paired to multiple mates from the other domain. For instance, the contrastive loss should not push away normal images from normal text of different pairs. This has the potential to improve results across tasks and datasets, especially those of low diversity.

Acknowledgements. We gratefully acknowledge the support of the Israel Science Foundation 2329/22 & the VATAT fellowship.
References


