Knowledge-Augmented Contrastive Learning for Abnormality Classification and Localization in Chest X-rays with Radiomics using a Feedback Loop

Yan Han¹, Chongyan Chen¹, Ahmed Tewfik¹, Benjamin Glicksberg², Ying Ding¹, Yifan Peng³†, Zhangyang Wang¹‡

¹The University of Texas at Austin, ²Icahn School of Medicine at Mount Sinai, ³Weill Cornell Medicine
†yip4002@med.cornell.edu, ‡atlaswang@utexas.edu

Abstract

Accurate classification and localization of abnormalities in chest X-rays play an important role in clinical diagnosis and treatment planning. Building a highly accurate predictive model for these tasks usually requires a large number of manually annotated labels and pixel regions (bounding boxes) of abnormalities. However, it is expensive to acquire such annotations, especially the bounding boxes. Recently, contrastive learning has shown strong promise in leveraging unlabeled natural images to produce highly generalizable and discriminative features. However, extending its power to the medical image domain is under-explored and highly non-trivial, since medical images are much less amendable to data augmentations. In contrast, their prior knowledge, as well as radiomic features, is often crucial. To bridge this gap, we propose an end-to-end semi-supervised knowledge-augmented contrastive learning framework, that simultaneously performs disease classification and localization tasks. The key knob of our framework is a unique positive sampling approach tailored for the medical images, by seamlessly integrating radiomic features as a knowledge augmentation. Specifically, we first apply an image encoder to classify the chest X-rays and to generate the image features. We next leverage Grad-CAM to highlight the crucial (abnormal) regions for chest X-rays (even when unannotated), from which we extract radiomic features. The radiomic features are then passed through another dedicated encoder to act as the positive sample for the image features generated from the same chest X-ray. In this way, our framework constitutes a feedback loop for image and radiomic features to mutually reinforce each other. Their contrasting yields knowledge-augmented representations that are both robust and interpretable. Extensive experiments on the NIH Chest X-ray dataset demonstrate that our approach outperforms existing baselines in both classification and localization tasks.

1. Introduction

The chest X-ray is one of the most common radiological examinations for detecting cardiothoracic and pulmonary abnormalities. Due to the demand for accelerating chest X-ray analysis and interpretation along with the overall shortage of radiologists, there has been a surging interest in building automated systems of chest X-ray abnormality classification and localization [28]. While the class (i.e., outcomes) labels are important, the localization annotations, or the tightly-bound local regions of images that are most indicative of the pathology, often provide richer information for clinical decision making (either automated or human-based).

Automatic robust image analysis of chest X-rays currently faces many challenges. First, recognizing abnormalities in chest X-rays often requires expert radiologists. This process is therefore time-consuming and expensive to generate annotations for chest X-ray data, in particular the localized bounding box region labeling. Second, unlike natural images, chest X-rays have very subtle and similar image features. The most indicative features are also very localized. Therefore, chest X-rays are sensitive to distortion and not amendable to typical image data augmentations such as random cropping or color jittering. Moreover, in addition to high inter-class variance of abnormalities seen in chest X-rays (i.e., feature differences between different diseases), chest X-rays also have large intra-class variance (i.e., differences in presentation among individuals of the same diseases). The appearance of certain diseases in X-rays are often vague, can overlap with other diagnoses, and can mimic many other benign abnormalities. Last but not least, the class distribution of chest X-rays is also highly imbalanced for available datasets.

Recently, contrastive learning has emerged as the frontrunner for self-supervised learning, demonstrating superior ability to handle unlabelled data. Popular frameworks in-
As the trastive knowledge-augmentation for the chest X-ray image. seamless integrates radiomic features as the other con-
tical barrier aging it for this task is not straightforward. One most im-
learning from the mostly unlabeled chest X-rays, but lever-
contrastive learning to be robust to data imbalance [37, 17].
Contrastive learning may offer a promising avenue for
from the mostly unlabeled chest X-rays, but lever-
aging it for this task is not straightforward. One most im-
important technical barrier is that most contrastive learning
frameworks [14, 6, 4, 5, 12] critically depend on maxi-
mizing the similarity between two “views”, i.e., an anchor
and its positive sample, often being generated by applying
random data augmentations to the same image. This
data augmentation strategy, however, does not easily trans-
late to chest X-rays. In addition, the simultaneous demand
for both classification and localization-aware features fur-
ther complicates the issue. Fortunately, classical chest X-
ray analysis has introduced radiomic features [7] as an
auxiliary knowledge augmentation. The radiomic features
can be considered as a strong prior, and therefore can po-
tentially be utilized to guide learning of deep feature extrac-
tors. However, the extraction of reliable radiomic features
via Pyradiomic1 tool [34] heavily depends on the pathology
localization – hence we will run into an intriguing “chicken-
and-egg” problem, when trying to incorporate radiomic fea-
tures into contrastive learning, whose goal includes learning
the localization from unlabeled data.

This paper presents an innovative holistic framework
of Knowledge-Augmented Contrastive Learning, which
seamlessly integrates radiomic features as the other con-
trastive knowledge-augmentation for the chest X-ray image.
As the main difference from existing frameworks, the two
“views” that we contrast now are from two different domain
knowledge characterizing the same patient: the chest X-ray
image and the radiomic features. Notably, the radiomic fea-
tures have to be extracted from the learned pathology lo-
calizations, which are not readily available. As these fea-
tures will be dynamically updated, forming a “feedback
loop” during training in which both modalities’ learning
mutually reinforce each other. The key enabling technique
to link this feedback loop is a novel module we designed,
called Bootstrap Your Own Positive Samples (BYOP). For
an unannotated X-ray image, we utilize Grad-CAM [30] to
generate the input heatmap from the image modality back-
bone, which yields the estimated bounding box after thresh-
olding; and we then extract the radiomic features within
this estimated bounding box, which becomes the alterna-
tive view to contrast with the image view. The usage of ra-
diomic features also adds to the model interpretability. Our
contributions are outlined as follows:

1https://pyradiomic.readthedocs.io/

- A brand-new framework dedicated to improving abnor-
mality identification and localization in (mostly unan-
notated) chest X-rays by knowledge-augmented contrastive
learning, which highlights exploiting radiomic features
as the auxiliary knowledge augmentation to contrast with
the images, given the inability to perform classical image
data augmentation.
- An innovative technique called BYOP to enable the ef-
efective generation of radiomic features, which is neces-
sary as the true bounding boxes are often absent. BYOP
leverages an interpretable learning technique to supply
estimated bounding boxes dynamically during training.
- Excellent experimental results achieved on the NIH
Chest X-ray benchmark [35], using very few annotations.
Besides improving the disease classification AUC from
82.8% to 83.8%, our framework significantly boosts the
localization results, by an average of 2% over different
IoU thresholds, compared to reported baselines. Figure
provides a visualization example showing our localization
results to be more robust and accurate than the pre-
vious results from CheXNet [28].

2. Related Work

Self-supervised Learning and Contrastive Learning:
Self-supervision uses pre-formulated (or “pretext”) tasks to
train with unlabeled data. Popular handcrafted pretext tasks
include solving jigsaw puzzles [24], relative patch predic-
tion [8] and colorization [41]. However, many of these tasks
rely on ad-hoc heuristics that could limit the generaliza-
tion and transferability of learned representations. Conse-
quently, contrastive learning of visual representations has
emerged as the front-runner for self-supervision and has
demonstrated superior performance on downstream tasks
[14, 6, 4, 5, 22, 12]. Most of those successes take place

Figure 1. Visualization of heatmaps of chest X-rays with ground-
truth bounding box annotations (yellow) and its prediction (red)
for localize Cardiomegaly in one test chest X-ray image. The
visualization is generated by rendering the final output tensor as
heatmaps and overlaying it on the original images. The left image
is the original chest X-ray image, the middle is the visualization
result by CheXNet [28] and the right is our model’s attempt. Best
viewed in color.
in the natural image domain due to the ease of creating contrastive views by applying data augmentations.

There have been a few recent attempts towards contrastive learning in multi-domain knowledge tasks. [33] studied knowledge transfer via contrastive learning, including from one sensory domain knowledge to another (e.g., RGB to Depth). [27] contrasted bright-field and second-harmonic generation microscopy images for their registration: two image domain knowledge captured for biomedical applications, with large appearance discrepancy between them. [40] aimed to learn text-to-image synthesis by maximizing the mutual information between image and text, through multiple contrastive losses which capture inter-domain and intra-domain correspondences. The commodity among those existing works is that their two domains (e.g., image and text, RGB and depth, or two types of microscopy images) are both readily available for training. In contrast to our task, the other radiomic features domain knowledge is not available ahead of training and needs to be adaptively bootstrapped from the image domain during training.

Contrastive Learning for Medical Image Analysis: Self-supervised learning using pre-text tasks has been recently popular in medical image analysis [32, 1, 45, 46, 15]. When it comes to contrastive learning, [3] proposed a domain-specific pretraining strategy, by extracting contrastive pairs from MRI and CT datasets using a combination of localized and global loss functions, which relies on the availability of both MRI and CT scans. [43] leveraged contrastive learning to infer transferable medical visual representations from paired images and text modalities.

Prior work adopting contrastive learning on chest X-rays remains scarce, due to the roadblock of creating two contrastive views. [20] explicitly contrasted X-rays with pathologies against healthy ones using attention networks, in a fully supervised setting. The closest recent work to our knowledge [13] also applied radiomic features to guide contrastive learning for detecting pneumonia in chest X-rays. However, their method needs to apply a pre-trained ResNet on the images to generate attention for radiomic features extraction and therefore relied on a multi-stage training heuristic. Their method hence produced no joint optimization of two domain knowledge, and lacked the localiziation capability.

Radiomics in Medical Diagnosis: Radiomics studies have demonstrated their power in image-based biomarkers for cancer staging and prognostication [23]. Radiomics extracts quantitative data from medical images to represent tumor phenotypes, such as spatial heterogeneity of a tumor and spatial response variations. [9] demonstrated that radiomic of CT texture features are associated with the overall survival rate of pancreatic cancer. [7] revealed that the first-order radiomic features (e.g., mean, skewness, and kurtosis) are correlated with pathological responses to cancer treatment. [16] showed that radiomics could increase the positive predictive value and reduce the false-positive rate in lung cancer screening for small nodules compared with human reading by thoracic radiologists. [42] found that multi-parametric MRI-based radiomics nomograms provided improved prognostic ability in advanced nasopharyngeal carcinoma (NPC). In comparison, deep learning algorithms are often criticized for being “black box” and lack interpretability despite high predictive accuracy. That limitation has motivated many interpretable learning techniques including activation maximization [10], network inversion [21], GradCAM [30], and network dissection [2]. We believe that the joint utilization of radiomics and interpretable learning techniques in our framework can further advance accurate yet interpretable learning in the medical image domain.

3. Method

The Framework. Our goal is to learn an image representation \( y_i \) which can then be used for disease classification and localization. Our framework uses two neural networks to learn: the image and radiomics networks. The image network consists of an encoder \( f_i \) (ResNet-18) and a projector \( g_i \) (two-layer MLPs with ReLU). The radiomics network has a similar architecture as the image network, but uses another three-layer MLPs for radiomic encoder \( f_r \) and a different set of weights for the projector \( g_r \). The proposed architecture is summarized in Figure 2.

The primary innovation of our method lies in how we select positive and negative examples, which will be expanded below in Section 3.1 and Section 3.2. We also formulate the semi-supervised loss for our problem when a small amount of annotated data is available in Section 3.3. The entire framework can be trained from end to end, and the representation \( y_i \) will be used for downstream disease classification and localization tasks.

3.1. Finding Positive and Negative Samples: Data-Driven Learning Meets Domain Expertise

The reasons to use contrastive learning as our framework are three-fold. First, contrastive learning leverages unlabeled data and we have few disease localization (bounding boxes) annotations available. Second, empirical findings [37, 17] prove that contrastive learning is robust in classification tasks with class-imbalanced datasets. In clinical settings, most medical image datasets suffer an extreme class imbalance problem [11]. Third, contrastive learning naturally fits “multi-view” concepts. In our case, we are still comparing two different views of the same subject, but unlike classic contrastive learning where two views are from the same domain space, our views for positive sampling are from different domain knowledge ([39] proved that views from multi-domain knowledge should also align), while our
negative sampling is from the same domain knowledge. In the subsequent section, we will describe our unique positive and negative sampling methodologies in more detail.

**Positive Sampling.** To obtain a positive pair of views, we randomly select an image labeled with a given disease and generate two views for it. The first view will be its image features and another view will be its radiomic features. We decided to leverage radiomic features for the second view as traditional image augmentation strategies cannot be leveraged here. Furthermore, radiomic features have labels, are naturally more interpretable than the image features extracted from deep learning-based image encoders.

Obtaining the radiomic features for our dataset is a “chicken-and-egg” problem. **Radiomic features are highly sensitive and dependent on local regions** for which we do not have local bounding box annotations. Meanwhile, we need to make the image features similar to the radiomic features to learn from radiomic features to better learn localization of the abnormalities. This process means that **bounding boxes generation is dependent on radiomic features** which forms a loop cycle. To address this issue, we design the **Bootstrap Your Own Positive Samples (BYOP)** method using such a feedback module. For more details, see Section 3.2.

**Negative Sampling.** The original images are used for views of the negative samples because the same domain is supposed to be more similar and thus harder for the model to distinguish between the positive and negative samples, leading to a more robust model [29]. Besides, the image features focuses on local regions highlighted by the attention map rather than the whole image. To identify harder negative samples, we go one step further, by not only selecting any random image, but “hard similar” images. Here, we first get prior knowledge from the pre-constructed disease hierarchy relationship for image negative sampling, shown in Figure 4, defined by [44]. The pre-constructed disease hierarchy relationship is initialed with 21 nodes. In this hierarchy, each disease (green) belongs to a body part (grey). We therefore only treat normal chest X-rays or images within the same body part but with a different disease as negative examples. We call these negative examples “hard similar” images in this study. As an example, if our “anchor” image is labeled as “Pneumonia/Lung”, our “hard similar” images should include “Atelectasis/Lung”, “Edema/Lung”, or “Normal” but not “Bone Fractures”.

### 3.2. Bootstrap Your Own Positive Samples (BYOP) with Radiomics in the Feedback Loop

The core component of our cross-modal contrastive learning is the **Bootstrap Your Own Positive Samples (BYOP)** module. BYOP leverages a feedback loop to learn region localization from generated radiomic features as the positive sample for the image features. The architecture of BYOP is shown in Figure 3. The BYOP contains two components, bounding box generation and radiomic features extraction.
Bounding Boxes Generation. We feed the fourth layer of the image encoder $f_i$ (i.e., ResNet-18) to the Gradient-weighted Class Activate Mapping (Grad-CAM) [30] to extract attention maps and apply an ad-hoc threshold to generate bounding boxes from the attention maps.

Radiomic Features Extraction. The radiomic features are composed of the following categories:

- First-Order statistics measure the distribution of voxel intensities within the bounding boxes. The features include energy (the measurement of the magnitude of voxel values), entropy (the measurement of uncertainty in the image values), and max/mean/median gray level intensity within the region of interest (ROI), etc.

- Shape-based features include features like Mesh Surface, Pixel Surface, Perimeter, and etc.

- Gray-level features include a gray-level features include a Gray Level Co-occurrence Matrix (GLCM) features, a Gray Level Size Zone (GLSZM) features, a Gray Level Run Length Matrix (GLRLM) features, a Neighboring Gray Tone Difference Matrix (NGTDM) features, and a Gray Level Dependence Matrix (GLDM) features.

Given the original images and generated bounding boxes, we used the Pyradiomic tool to extract radiomic features [34].

3.3. Semi-Supervised Loss Function

Our framework is mixed with supervised classification and unsupervised contrastive learning. For the localization task, we use the knowledge-augmented contrastive loss for unsupervised contrastive learning. For the classification task, we could have used standard cross-entropy loss, but considering that the chest X-ray dataset is highly imbalanced, we instead find focal loss more helpful [25]. We briefly review the two loss functions below.

Unsupervised Knowledge-Augmented Contrastive Loss. Our cross-modal contrastive loss function extends the normalized temperature-scaled cross-entropy loss (NT-Xent). We randomly sample a minibatch of $N$ examples and define the contrastive prediction task on pairs of augmented examples derived from the minibatch. Let $v_{bd}$ be the image in the minibatch with disease $d$ and body part $b$, and $\sin(u, v)$ be the cosine similarity. The loss function $\ell_{v_{bd}}$ for a positive pair of example $(v_{bd}, v'_{bd})$ is defined as

$$\ell_{v_{bd}} = - \log \frac{\exp (\sin (z_i(v_{bd}), z_i(v'_{bd}))/\tau)}{\sum_{k=b, l \neq d} \exp (\sin (z_i(v_{bd}), z_i(v_k))/\tau)}$$

where $1_{[k=b, l \neq d]} \in \{0, 1\}$ is an indicator function evaluating to 1 iff $k = b$ and $l \neq d$. $\tau$ is the temperature parameter. The final unsupervised contrastive loss $L_{ct}$ is computed across all disease-positive images in the minibatch.

Supervised Focal Loss. We feed the output of the image encoder $f_i$ to a simple linear classifier. The supervised classification focal loss is defined as

$$L_{fl} = \begin{cases} -\alpha (1 - y')^\gamma \log y', & y = 1 \\ -(1 - \alpha) y'^\gamma \log (1 - y'), & y = 0 \end{cases}$$

$\alpha$ allows us to give different importance to positive and negative examples. $\gamma$ is used to distinguish easy and hard samples and force the model to learn more from difficult examples.

Eventually, we treat it as multi-task learning (one task is supervised disease classification and one is unsupervised contrastive learning) and the total loss is defined as

$$L = \lambda \times L_{ct} + (1 - \lambda) \times L_{fl}$$

4. Experiments

Dataset and Protocol Setting. We evaluated our framework using the NIH Chest X-ray dataset [35]. It contains 112,120 X-ray images collected from 30,805 patients. As other large chest X-ray datasets, this dataset is also extremely class imbalanced: the healthy cases (84,321 front-view images) are far more than cases with diseases (24,624...
to shuffle the unannotated dataset into three subsets: 70% (111,240 images), called “annotated” (880 images) and “unannotated” boxes from the entire dataset. Hence, we have two sets of pulmonologists. We separate the images with provided bounding disease labels are expected to have above 90% accuracy.

There are 9 classes, specifically one for “No findings” and 8 for diseases (Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, and Pneumothorax). The disease labels were extracted from radiology reports with a rule-based tool [26].

Evaluation Metrics. For the disease classification task, we use Area under the Receiver Operating Characteristic curve (AUC) to measure the performance of our model. For the disease localization task, we evaluate the detected regions against annotated ground truth bounding boxes, using intersection over union ratio (IoU). The localization results are only calculated on the test set of the annotated dataset. The localization is defined as correct only if IoU > T(IoU), where T(*) is the threshold.

Implementation Details. We use the ResNet-18 model as the image encoder. We initialize the image encoder with the weights from the pre-trained ImageNet model except for the last fully-connected layer. We set the batch size as 64 and train the model for 30 epochs. We optimize the model by the Adam method and decay the learning rate by 0.1 from 0.001 every 5 epochs. Furthermore, we use linear warmup for the first 10 epochs only for the disease classification task, which helps the model converge faster to generate stable heatmaps. We train our model on AWS with one Nvidia Tesla V100 GPU. The model is implemented in PyTorch.

4.1. Disease Classification

Table 1 shows the AUC for each class and a mean AUC across the 8 chest diseases. Compared to a series of relevant baseline models, our proposed model achieves better AUC scores for the majority of diseases. The overall improvement in performance is remarkable when compared to other models except CheXNet [28]. One possible reason for our lack of improvement can be that [28]'s backbone is DenseNet-121, which is much deeper than the ResNet-18 in our model. It thus, able to capture much more discriminative features than our ResNet-18. Despite the fact, our model still achieves better or comparable results than CheXNet, which demonstrates that the cross-modal contrastive learning branch boosts the robustness of the image features without the need to increase the complexity of the backbone. Specifically, the performance of our model demonstrates significant improvements for disease abnormalities with larger associated regions on the image, such as “Atelectasis”, “Cardiomegaly”, ” and “Pneumothorax”. In addition, small objects features like “Mass” and “Nodule”, are recognized as well as in CheXNet. In summary, these experimental results show the superiority of our proposed model over relevant other methodologies.

4.2. Disease Localization

We compare our disease localization accuracy to other state-of-the-art models under different IoU thresholds (Table 2). Since disease localization is not an easy task in chest X-ray images, we did not find as many other methods as for disease classification task. To our knowledge, we only have two baseline methods from [35] and [19]. From these comparisons, we find our model significantly outperforms baselines by an average of 2% over different IoU thresholds. Importantly, our model is able to perform well not only on the easier tasks, but also for more difficult ones like localizing “Mass” and “Nodule”, where the disease localization is within a small area. When the IoU threshold is set to 0.1, our model outperforms others on all diseases except for “Cardiomegaly”. As the IoU threshold increases, our framework is superior to other models in terms of better accuracy and maintains this superior performance. For

<table>
<thead>
<tr>
<th>Method</th>
<th>Atelectasis</th>
<th>Cardiomegaly</th>
<th>Effusion</th>
<th>Infiltration</th>
<th>Mass</th>
<th>Nodule</th>
<th>Pneumonia</th>
<th>Pneumothorax</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wang et. al.[35]</td>
<td>0.72</td>
<td>0.81</td>
<td>0.78</td>
<td>0.61</td>
<td>0.71</td>
<td>0.67</td>
<td>0.63</td>
<td>0.81</td>
<td>0.718</td>
</tr>
<tr>
<td>Wang et. al.[36]</td>
<td>0.73</td>
<td>0.84</td>
<td>0.79</td>
<td>0.67</td>
<td>0.73</td>
<td>0.69</td>
<td>0.72</td>
<td>0.85</td>
<td>0.753</td>
</tr>
<tr>
<td>Yao et. al.[38]</td>
<td>0.77</td>
<td>0.90</td>
<td>0.86</td>
<td>0.70</td>
<td>0.79</td>
<td>0.72</td>
<td>0.71</td>
<td>0.84</td>
<td>0.786</td>
</tr>
<tr>
<td>Rajpurkar et. al.[28]</td>
<td>0.82</td>
<td>0.91</td>
<td>0.88</td>
<td>0.72</td>
<td>0.86</td>
<td>0.78</td>
<td>0.76</td>
<td>0.89</td>
<td>0.828</td>
</tr>
<tr>
<td>Kumar et. al.[18]</td>
<td>0.76</td>
<td>0.91</td>
<td>0.86</td>
<td>0.69</td>
<td>0.75</td>
<td>0.67</td>
<td>0.72</td>
<td>0.86</td>
<td>0.778</td>
</tr>
<tr>
<td>Liu et. al.[20]</td>
<td>0.79</td>
<td>0.87</td>
<td>0.88</td>
<td>0.69</td>
<td>0.81</td>
<td>0.73</td>
<td>0.75</td>
<td>0.89</td>
<td>0.801</td>
</tr>
<tr>
<td>Seyyed et. al.[31]</td>
<td>0.81</td>
<td>0.92</td>
<td>0.87</td>
<td>0.72</td>
<td>0.83</td>
<td>0.78</td>
<td>0.76</td>
<td>0.88</td>
<td>0.821</td>
</tr>
<tr>
<td>Our model</td>
<td>0.84</td>
<td>0.93</td>
<td>0.88</td>
<td>0.72</td>
<td>0.87</td>
<td>0.79</td>
<td>0.77</td>
<td>0.90</td>
<td>0.838</td>
</tr>
</tbody>
</table>

Table 1. Comparison with the baseline models for AUC of each class and average AUC. For each column, red values denote the best results.
instance, when the threshold increases, the IoUs of “Cardiomegaly” decrease less than the baselines and even outperform the baselines when IoU threshold is above 0.5.

We prefer a higher IoU threshold, specifically, IoU = 0.7, for disease localization because we expect high-accuracy disease localization application is necessary for clinical applications. To this end, the method we propose is superior to the baseline by a slight margin.

It is also worth noting that, for some diseases, such as Pneumonia and Infiltration, the localization of disease can appear in multiple places while only one bounding box is provided for each image. Hence, it is reasonable that our model does not align well with the ground truth when the threshold is as small as 0.1, especially for Pneumonia and Infiltration. Overall, our model outperforms the reference models for almost all IoU thresholds.

### 4.3. Ablation Discussion

In this section, we study the contribution of our BYOP module on both disease classification and localization tasks.

#### Disease Classification

For this task, note that the use of focal loss should also boost the model with the class-imbalanced chest X-ray dataset. Thus, we compare the performance of our base model with only focal loss (labeled “w. FL”) or with only the BYOP module (labeled “w. BYOP”), respectively. As shown in Table 3, although both focal loss and BYOP improve the model performance, BYOP contributed more strongly. This stronger contribution is expected since BYOP tends to generate more robust radiomic features, which further reinforces the image encoder to focus on the image region that contains the targeted disease.

#### Disease Localization

Note that our base model is a ResNet-18 image encoder, which is not as powerful as CheXNet [28] with DenseNet-121. Thus we compare the performance of our model with CheXNet. As shown in Figure 5, our localization result is superior to the CheXNet. For the example of ‘Atelectasis’, ‘Cardiomegaly’, ‘Effusion’, ‘Nodule’, ‘Pneumonia’ and ‘Pneumothorax’, while the baseline model tends to focus on a large area of the image, our model precisely captures the correct disease location. For harder localization cases like ‘Mass’ and ‘Nodule’, we prefer a higher IoU threshold, specifically, IoU = 0.7, for disease localization because we expect high-accuracy disease localization application is necessary for clinical applications. To this end, the method we propose is superior to the baseline by a slight margin.

It is also worth noting that, for some diseases, such as Pneumonia and Infiltration, the localization of disease can appear in multiple places while only one bounding box is provided for each image. Hence, it is reasonable that our model does not align well with the ground truth when the threshold is as small as 0.1, especially for Pneumonia and Infiltration. Overall, our model outperforms the reference models for almost all IoU thresholds.
the baseline model’s focus is incorrect and does not have any overlap with the ground-truth areas while our model still predicts perfectly. The results demonstrate that the BYOP module significantly boosts the model performance.

5. Conclusion

In this work, we propose a semi-supervised, end-to-end knowledge-augmented contrastive learning model that can jointly model disease classification and localization with limited localization annotation data. Our approach differs from previous studies in the choice of data augmentation, the use of radiomic features as prior knowledge, and a feedback loop for image and radiomic features to mutually reinforce each other. Additionally, the project aims to address current gaps in radiology by making prior knowledge more accessible to image data analytic and diagnostic assisting tools, with the hope that this will increase the model’s interpretability. Experimental results demonstrate that our method outperforms the state-of-the-art algorithms, especially for the disease localization task, where our method can generate more accurate bounding boxes. Importantly, we hope the method developed here is inspiring for the future research on incorporating different kinds of prior knowledge of medical images with contrastive learning.

Acknowledgement

This project is funded by Amazon Machine Learning Grant and NSF AI Center at UT Austin. It also was supported by the National Library of Medicine under Award No. 4R00LM013001.
References


