Holistic and Comprehensive Annotation of Clinically Significant Findings on Diverse CT Images: Learning from Radiology Reports and Label Ontology

Ke Yan[1], Yifan Peng[2], Veit Sandfort[1], Mohammadhadi Bagheri[1], Zhiyong Lu[2], Ronald M. Summers[1]
[1] Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, Clinical Center
[2] National Center for Biotechnology Information, National Library of Medicine
[1,2] National Institutes of Health, Bethesda, MD 20892
{ke.yan, yifan.peng, veit.sandfort, mohammad.bagheri, zhiyong.lu, rms}@nih.gov

Abstract

In radiologists' routine work, one major task is to read a medical image, e.g., a CT scan, find significant lesions, and describe them in the radiology report. In this paper, we study the lesion description or annotation problem. Given a lesion image, our aim is to predict a comprehensive set of relevant labels, such as the lesion's body part, type, and attributes, which may assist downstream fine-grained diagnosis. To address this task, we first design a deep learning module to extract relevant semantic labels from the radiology reports associated with the lesion images. With the images and text-mined labels, we propose a lesion annotation network (LesaNet) based on a multilabel convolutional neural network (CNN) to learn all labels holistically. Hierarchical relations and mutually exclusive relations between the labels are leveraged to improve the label prediction accuracy. The relations are utilized in a label expansion strategy and a relational hard example mining algorithm. We also attach a simple score propagation layer on LesaNet to enhance recall and explore implicit relation between labels. Multilabel metric learning is combined with classification to enable interpretable prediction. We evaluated LesaNet on the public DeepLesion dataset, which contains over 32K diverse lesion images. Experiments show that LesaNet can precisely annotate the lesions using an ontology of 171 fine-grained labels with an average AUC of 0.9344.

1. Introduction

In recent years, there has been remarkable progress on computer-aided diagnosis (CAD) based on medical images, especially with the help of deep learning technologies [23, 36]. Lesion classification is one of the most important topics in CAD. Typical applications include using medical images to classify the type of liver lesions and lung tissues [10, 15, 36], to describe the fine-grained attributes of pulmonary nodules and breast masses [5, 30], and to predict their malignancy [6, 12]. However, existing studies on this topic usually focus on certain body parts (lung, breast, liver, etc.) and attempt to distinguish between a limited set of labels. Hence, many clinically meaningful lesion labels covering different body parts were not explored yet. Besides, in practice, multiple labels can be assigned per lesion and are often correlated.

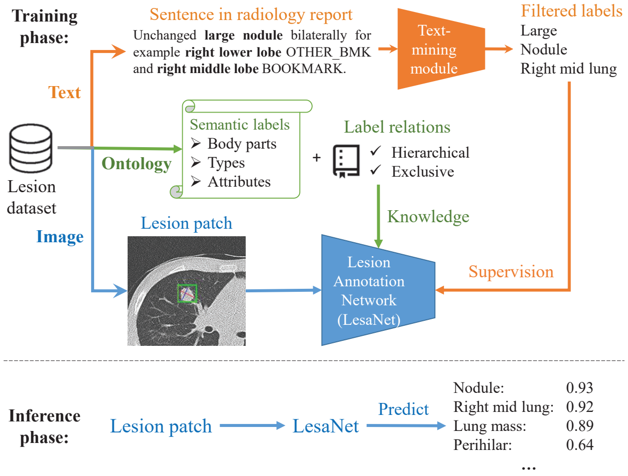


Figure 1. The overall framework. We propose lesion annotation network (LesaNet) to predict fine-grained labels to describe diverse lesions on CT images. The training labels are text-mined from radiology reports. Label relations are utilized in learning.

In this paper, we tackle a more general and clinically useful problem to mimic radiologists. When an experienced radiologist reads a medical image such as a computed tomography (CT) scan, he or she can detect all kinds of lesions in various body parts, identify the lesions' detailed information including its associated body part, type, and attributes, and finally link these labels to the predefined ontology. We aim to develop a new framework to predict these semantic labels holistically (jointly learning all labels), so as to go one step closer to the goal of "learning to read CT images". In brief, we wish the computer to recognize where, what, and how the lesion is, helping the user comprehensively understand it. We call this task lesion annotation due to its analogy to the multilabel image annotation/tagging problem in general computer vision literature [49].

To learn to annotate lesions, a large-scale and diverse dataset of lesion images is needed. Existing lesion datasets [8, 34] are typically either too small or less diverse. Fortunately, the recently-released DeepLesion dataset [47, 48] has largely mitigated this limitation. It contains bounding-boxes of over 32K lesions from a variety of body parts on CT images. However, there are no fine-grained semantic labels given for each lesion in DeepLesion. Manual annotation is tedious, expensive, and not scalable, not to mention it requires experts with considerable domain knowledge. Inspired by recent studies [43, 41], we take an automatic data mining approach to extract labels from radiology reports. Reports contain rich but complex information about multiple findings in the medical image. In the course of interpreting a CT scan, a radiologist may manually annotate a lesion in the image, and place a hyperlink to the annotation (a "bookmark") in the report. We first locate the sentence with bookmark in the report that refers to a lesion, then extract labels from the sentence. We defined a fine-grained ontology based on the RadLex lexicon [19]. This process is entirely data-driven and requires minimal manual effort, thus can be easily employed to build large datasets with rich vocabularies. Sample lesion image, sentence, and labels can be found in Fig. 1.

We propose a LESion Annotation NETwork (LesaNet) to predict semantic labels given a lesion image of interest. This lesion annotation task is treated as a multilabel image classification problem [49]. Despite extensive previous studies [11, 14, 16, 42, 21], our problem is particularly challenging due to several reasons: **1)** Radiology reports are often in the format of free-text, so extracted labels can be noisy and incomplete [43]. **2)** Some labels are difficult to distinguish or learn, e.g. adjacent body parts, similar types, and subtle attributes. **3)** The labels are highly imbalanced and long-tailed. To tackle these challenges, we present the framework shown in Fig. 1. First, we reduce the noise in training labels by a text-mining module. The module analyzes the report to find the labels relevant to the lesion-of-interest. Second, we build an ontology which includes the hierarchical hyponymy and mutually exclusive relations between the labels. With the hierarchical relations, we apply a label expansion strategy to infer the missing parent labels. The exclusive relations are used in a relational hard example mining (RHEM) algorithm to help LesaNet learn hard cases and improve precision. Third, we also attach a simple score propagation layer to enhance recall, especially for rare labels. Finally, metric learning is incorporated in LesaNet to not only improve classification accuracy but also enable prediction interpretability.

The main contributions of this work includes the following: **1)** We study the holistic lesion annotation problem and propose an automatic learning framework with minimum manual annotation effort; **2)** An algorithm is proposed to text-mine relevant labels from radiology reports; **3)** We present LesaNet, an effective lesion annotation algorithm that can also be adopted in other multilabel image classification problems; and **4)** To leverage the ontology-based medical knowledge, we incorporate label relations in LesaNet.

2. Related Work

Medical image analysis with reports: Annotating medical images is tedious and requires considerable medical knowledge. To reduce manual annotation burden, some researchers leveraged the rich information contained in associated radiology reports. Disease-related labels have been mined from reports for classification and weakly-supervised localization on X-ray [43, 41] and CT images [35, 15]. This approach boosts the size of datasets and label sets. However, current studies can only extract image-level labels, which cannot be accurately mapped to specific lesions on the image. The DeepLesion dataset[1] consists of lesions from a variety of body parts on CT images. It has been adopted to train algorithms for universal lesion detection [46], retrieval [48], segmentation and measurement [3, 40]. This paper will explore its usage on lesion-level semantic annotation. Another line of study directly generates reports according to the whole image [44, 50]. Although the generated reports may learn to focus on certain lesions on the image, it is difficult to assess the usability of generated reports. The key information in reports are the labels. If we can accurately predict the labels for each lesion on the image, the creation of high-quality (structured) reports would be straightforward.

Multilabel image classification: Multilabel image classification [49] is a long-standing topic that has been tackled from multiple angles. A direct idea is to treat each label independently and use a binary cross-entropy loss for each [43]. The pairwise ranking loss is applied in [45, 14, 21] to make the scores of positive labels larger than those of the negative ones for each sample. The CNN-RNN framework [42] uses a recurrent model to predict multiple labels one-by-one. It can implicitly model label dependency and avoid the score thresholding issue. In [11], deep metric learning and hard example mining are combined to deal with imbalanced labels.

Noisy and incomplete training labels often exist in datasets mined from the web [7], which is similar to our la-

[1] https://nihcc.box.com/v/DeepLesion

bels mined from reports. Strategies to handle them include data filtering [18], noise-robust losses [31], noise modeling [25], finding reliable negative samples [20], and so on. We use a text-mining module to filter noisy positive labels and leverage label relation to find reliable negative labels. Label relations have been exploited by researchers to improve image classification. Novel loss functions were proposed in [9] for labels with hierarchical tree-like structures. In [24, 16], prediction scores of different labels are propagated between network layers whose structure is designed to capture label relations. We apply label expansion and RHEM strategies to use label relations explicitly, and at the same time employ a score propagation layer to learn them implicitly.

3. Label Mining and Ontology

3.1. Ontology Construction

We constructed our lesion ontology based on RadLex [19], a comprehensive lexicon for standardized indexing and retrieval of radiology information resources [26, 1]. The labels in our lesion ontology can be categorized into three classes: **1. Body parts**, which include coarse-level body parts (e.g., chest, abdomen), organs (lung, lymph node), fine-grained organ parts (right lower lobe, pretracheal lymph node), and other body regions (porta hepatis, paraspinal); **2. Types**, which include general terms (nodule, mass) and more specific ones (adenoma, liver mass); and **3. Attributes**, which describe the intensity, shape, size, etc., of the lesions (hypodense, spiculated, large).

The labels in the lesion ontology are organized in a hierarchical structure (Fig. 2). For example, a fine-grained body part (left lung) can be a part of a coarse-scale one (lung); a type (hemangioma) can be a sub-type of another one (neoplasm); and a type (lung nodule) can locate in a body part (lung). These relations form a directed graph instead of a tree, because one child (lung nodule) may have multiple parents (lung, nodule). Some labels are also mutually exclusive, meaning that the presence of one label signifies the absence of others (e.g., left and right lungs). However, in Fig. 2, chest and lymph node are not exclusive because they may physically overlap; lung nodule and ground-glass opacity are not exclusive either since they may coexist in one lesion. We hypothesize that if labels a and b are exclusive, any child of a and any child of b are also exclusive. This rule can help us in annotation of exclusive labels.

3.2. Relevant Label Extraction

After constructing the lesion ontology, we extracted labels from the associated radiology reports of DeepLesion [47]. In the reports, radiologists describe the lesions and sometimes insert hyperlinks, size measurements, or slice numbers (known as bookmarks) in the sentence to refer to the image of interest. In this work, we only used the sentences with bookmarks to text-mine labels associated with the lesions. First, we tokenized the sentence and lemmatized the words in the sentence using NLTK [2] to obtain their base forms. Then, we matched the named entity mentions in the preprocessed sentences and normalized them to labels based on their synonyms.

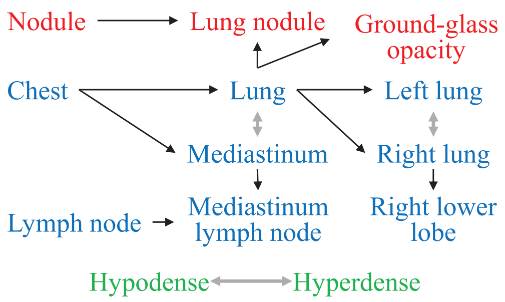


Figure 2. Sample labels with relations. Blue, red, and green labels correspond to body parts, types, and attributes, respectively. Single-headed arrows point from the parent to the child. Double-headed arrows indicate exclusive labels.

The bookmarked sentences often contain a complex mixture of information describing not only the bookmarked lesion but also other related lesions and unrelated things. A sample sentence is shown in Fig. 1, where the word "BOOKMARK" is the hyperlink of interest, while "OTHER_BMK" is the hyperlink for another lesion. There are 4 labels matched based on the ontology, namely large, nodule, right lower lobe, and right middle lobe. Among them, "right lower lobe" is irrelevant since it describes another lesion. In other examples, there are also uncertain labels such as "adenopathy or mass". Since both the irrelevant and uncertain labels may bring noise to downstream training, we developed a text-mining module to distinguish them from relevant labels. Specifically, we reformulate it as a relation classification problem. Given a sentence with multiple labels and bookmarks, we aim to assign relevant labels to each bookmark from all label-bookmark pairs.

To achieve that, we propose to use a CNN model based on Peng et al. [28, 29]. The input of our model consists of two parts: the word sequence with mentioned labels and bookmarks, and the sentence embedding [4]. The model outputs a probability vector corresponding to the type of the relation between the label and the bookmark (irrelevant, uncertain, and relevant). Due to space limit, we refer readers to [29] for details about this algorithm.

4. Lesion Annotation Network (LesaNet)

Fig. 3 displays the framework of the proposed lesion annotation network (LesaNet). In this section, we introduce each component in detail.

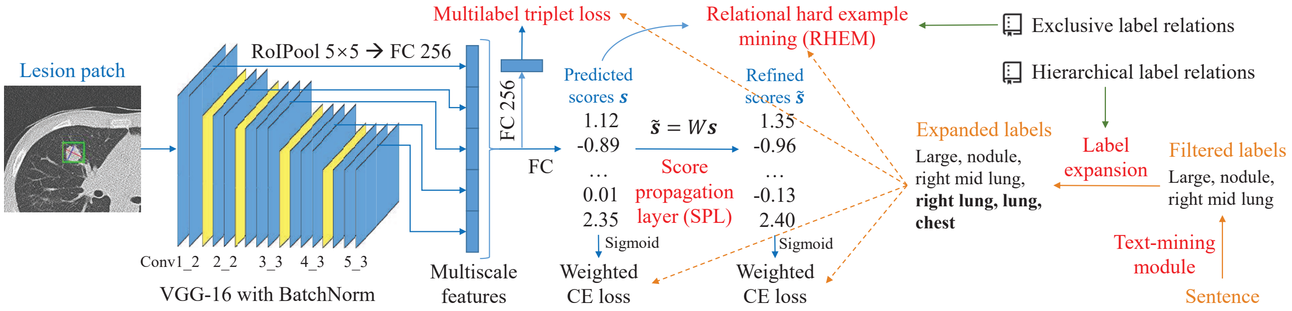

Multilabel triplet loss
Relational hard example mining (RHEM)
Exclusive label relations
Hierarchical label relations
Lesion patch
RoIPool 5×5 → FC 256
FC 256
FC
Predicted scores $\boldsymbol{s}$
1.12
-0.89
…
0.01
2.35
$\tilde{\boldsymbol{s}} = W\boldsymbol{s}$
Score propagation layer (SPL)
Refined scores $\tilde{\boldsymbol{s}}$
1.35
-0.96
…
-0.13
2.40
Sigmoid
Sigmoid
Weighted CE loss
Weighted CE loss
Expanded labels
Large, nodule, right mid lung, right lung, lung, chest
Label expansion
Filtered labels
Large, nodule, right mid lung
Text-mining module
Sentence
Conv1_2 2_2 3_3 4_3 5_3
VGG-16 with BatchNorm
Multiscale features

Figure 3. The framework of LesaNet. The input is the lesion image patch and the final output are the refined scores $\tilde{\boldsymbol{s}}$. The expanded labels are used to train LesaNet and optimize the four losses. Modules in red are our main contributions.

4.1. Multiscale Multilabel CNN

The backbone of the network is VGG-16 [38] with batch normalization [17]. In our task, different labels may be best modeled by features at different levels. For instance, body parts require high-level contextual features while many attributes depict low-level details. Therefore, we use a multiscale feature representation similar to [48]. Region of interest pooling layers (RoIPool) [13] are used to pool the feature maps to 5×5 in each convolutional block. For conv1_2, conv2_2, and conv3_3, the RoI is the bounding-box of the lesion in the patch to focus on its details. For conv4_3 and conv5_3, the RoI is the entire patch to capture the context. Each pooled feature map is then projected to a 256D vector by a fully-connected layer (FC) and concatenated together. After another FC layer, the network outputs a score vector $\boldsymbol{s} \in \mathbb{R}^C$, where C is the number of labels. Because positive cases are sparse for most labels, we adopt a weighted cross-entropy (CE) loss [43] for each label as in Eq. 1, where B is the number of lesion images in a minibatch; $\sigma_{i,c} = \text{sigmoid}(s_{i,c})$ is the confidence of lesion i having label c, whose ground-truth is $y_{i,c} \in \{0, 1\}$; the loss weights are $\beta_c^{\text{p}} = |P_c + N_c|/|2P_c|, \beta_c^{\text{n}} = |P_c + N_c|/|2N_c|$, where P_c, N_c are the numbers of positive and negative cases of label c in the training set, respectively.

$$L_{\text{WCE}} = \sum_{i=1}^{B} \sum_{c=1}^{C} (\beta_c^{\text{p}} y_{i,c} \log \sigma_{i,c} + \beta_c^{\text{n}} (1 - y_{i,c}) \log(1 - \sigma_{i,c})) \,. \tag{1}$$

4.2. Leveraging Label Relations

Label expansion: Labels extracted from reports are not complete. The hierarchical label relations can help us infer the missing parent labels. If a child label is true, all its parents should also be true. In this way, we can find the labels "right lung", "lung", and "chest" in Fig. 3 based on the existing label "right mid lung" in both training and inference.

Relational hard example mining (RHEM): Label expansion cannot complete other missing labels if their children labels are not mentioned in the report. This problem occurs when radiologists did not describe every attribute of a lesion or omitted the fine-grained body part. Although it is hard to retrieve these missing positive labels, we can utilize the exclusive relations to find reliable negative labels. In other words, if the expanded labels of a lesion are reliably 1, then their exclusive labels should be reliably 0.

One challenge of our task is that some labels are difficult to learn. We hope the loss function to emphasize them automatically. Inspired by online hard example mining (OHEM) [37], we define the online difficulty of label c of lesion i as:

$$\delta_{i,c} = |\sigma_{i,c} - y_{i,c}|^\gamma, \tag{2}$$

$\gamma > 0$ is a focusing hyper-parameter similar to the focal loss [22]. Higher γ puts more focus on hard examples. Then, we sample S lesion-label pairs according to δ in the minibatch, and compute their average CE loss. The higher $\delta_{i,c}$ is, the more times it will be sampled. Hence, the loss will automatically focus on hard lesion-label pairs. This stochastic sampling strategy works better in our experiments than the selection strategy in OHEM [37] and the reweighting one in focal loss [22]. An important note is that the sampling is only performed on reliable lesion-label pairs, so as to avoid treating missing positive labels as hard negatives. RHEM also works as a dynamic weighting mechanism for imbalanced labels, thus there is no need to impose weights [37] as the β in Eq. 1. We combine the CE loss of RHEM with Eq. 1 instead of replacing it, since some labels have no exclusive counterparts and have to be learned from Eq. 1.

4.3. Score Propagation Layer

A score propagation layer (SPL) is attached at the end of LesaNet (Fig. 3). It is a simple FC layer that refines the predicted scores with a linear transformation matrix W, followed by a weighted CE loss (Eq. 1). W is initialized with an identity matrix and can learn to capture the first-order correlation between labels. Although the hierarchical and exclusive label relations have been explicitly expressed by label expansion and RHEM, it is still useful to have SPL as it can enhance the scores of positively related labels and suppress the scores of labels with negative correlation and clear separation. On the other side, some exclusive labels can be very similar in location and appearance, for instance, hemangioma and metastasis in liver. When SPL sees a high score of hemangioma, it will know that it may also be a metastasis since in some cases they are hard to distinguish. Therefore, SPL will actually increase the score for metastasis slightly instead of suppressing it. This mechanism is particularly beneficial to improve the recall of rare labels whose prediction scores are often low. This rationale distinguishes SPL from previous knowledge propagation methods [16] that enforce negative weights on exclusive labels, which led to a lower performance in our task. By observing the learned W, we can also discover more label correlation and compare them with our prior knowledge.

4.4. Multilabel Triplet Loss

Interpretability is important for CAD tasks [32]. We expect the algorithm to provide evidence for its predictions. After classifying a lesion, it is desirable if LesaNet can show lesions in the database that have similar labels, which will help the user better understand its prediction as well as the lesion itself. This is a joint lesion annotation and retrieval problem. Lesion retrieval was studied in [48], but only 8 coarse-scale body part labels were used. In this paper, we use the comprehensive labels mined from reports to learn a feature embedding to model the similarity between lesions. As shown in Fig. 3, an FC layer is applied to project the multiscale features to a 256D vector, followed by a triplet loss [33]. To measure the similarity between two images with multiple labels, Zhao et al. [51] used the number of common positive labels as a criterion. However, we argue that each lesion may have a different number of labels, so the number of disjoint positive labels also matters. Suppose X and Y are the set of positive labels of lesions A and B, we use the following similarity criterion:

$$\text{sim}(A, B) = |X \cap Y|^2 / |X \cup Y|. \tag{3}$$

When training, we first randomly sample an anchor lesion A from the minibatch, and then find a similar lesion B from the minibatch so that $\text{sim}(A, B) \geq \theta$, finally find a dissimilar lesion C so that $\text{sim}(A, C) < \text{sim}(A, B)$. θ is the similarity threshold. We sample T such triplets from the minibatch and calculate the triplet loss:

$$L_{\text{triplet}} = \frac{1}{T} \sum_{t=1}^{T} \max(0,\, d(A, B) - d(A, C) + \mu), \tag{4}$$

where $d(A, B)$ is the L2 distance of the embeddings of A and B, μ is the margin. L_{triplet} makes lesions with similar label sets closer in the embedding space.

The final loss of LesaNet combines the four components:

$$L = L_{\text{WCE}} + L_{\text{CE, RHEM}} + L_{\text{WCE, SPL}} + \lambda L_{\text{triplet}}. \tag{5}$$

5. Experiments

5.1. Dataset

From DeepLesion and its associated reports, we gathered 19,213 lesions with sentences as the training set, 1,852 as the validation set, and 1,759 as the test set. Each patient was assigned to one of the subsets only. The total number is smaller than DeepLesion because not all lesions have bookmarks in the reports. We extracted labels as per Sec. 3.2, then kept the labels occurring at least 10 times in the training set and 2 times in both the validation (val) and the test sets, resulting in a list of 171 unique labels. Among them, there are 115 body parts, 27 types, and 29 attributes. We extracted hierarchical label relations from RadLex followed by manual review and obtained 137 parent-child pairs. We further invited a radiologist to annotate mutually exclusive labels and obtained 4,461 exclusive pairs.

We manually annotated the label relevance (relevant / uncertain / irrelevant, Sec. 3.2) in the val and test sets with two expert radiologists' verification. As a result, there are 4,648 relevant, 443 uncertain, and 1,167 irrelevant labels in the test set. The text-mining module was trained on the val set and applied on the training set. Then, labels predicted as relevant or uncertain in the training set were used to train LesaNet. LesaNet was evaluated on the relevant labels in the test set. Because the bookmarked sentences may not include all information about a lesion, there may be missing annotations in the test set when relying on sentences only. Hence, two radiologists further manually annotated 500 random lesions in the test set in a more comprehensive fashion. On average, there are 4.2 labels per lesion in the original test set, and 5.4 in the hand-labeled test set. An average of 1.2 labels are missing in each bookmarked sentence. We call the original test set the "text-mined test set" because the labels were mined from reports. The second hand-labeled test set is also used to evaluate LesaNet.

5.2. Implementation Details

For each lesion, we cropped a 120mm^2 patch around it as the input of LesaNet. To encode 3D information, we used 3 neighboring slices to compose a 3-channel image. Other image preprocessing details are the same as [48]. For the weighted CE loss, we clamped the weights β to be at most 300 to ensure training stability. For RHEM, we set $\gamma = 2$ and $S = 10^4$. For the triplet loss, we empirically set $\theta = 1, \mu = 0.1$, and $T = 5000$. The triplet loss weight was $\lambda = 5$ since this loss is generally smaller than other loss terms. LesaNet was implemented using PyTorch [27] and trained from scratch. Lesions with at least one positive label were used in training. The batch size was 128. LesaNet was trained using stochastic gradient descent (SGD) with a learning rate of 0.01 for 10 epochs, then with 0.001 for 5 more epochs.

Label	AUC	F1	Label	AUC	F1
Chest	96.2	90.2	Nodule	89.1	66.9
Lung	98.6	92.0	Cyst	96.0	40.7
Liver	98.6	78.8	Adenoma	99.9	30.8
Lymph node	93.7	76.2	Metastasis	74.0	10.7
Adrenal gland	99.5	76.2	Hypodense	87.7	50.9
Right mid lung	98.7	56.6	Sclerotic	99.7	75.4
Pancreatic tail	97.5	35.3	Cavitary	94.9	25.0
Paraspinal	97.5	9.8	Large	80.6	17.5

Table 2. Accuracies (%) of typical body parts, types, and attributes.

5.3. Evaluation Metric

The AUC, i.e. the area under the receiver operating characteristic (ROC) curve, is a popular metric in CAD tasks [43, 6]. However, AUC is a rank-based metric and does not involve label decision, thus cannot evaluate the quality of the final predicted label set in the multilabel setting. Thus, we also computed the precision, recall, and F1 score for each label, which are often used in multilabel image classification tasks [49]. Each metric was averaged across labels with equal weights (per-class-averaging). Overall-averaging [49] was not adopted because it biases towards the frequent labels (chest, abdomen, etc.) which are less informative. To turn confidence scores into label decisions, we calibrated a threshold for each label that yielded the best F1 on the validation set, and then apply it on the test set.

5.4. Lesion Annotation Results

A comparison of different methods and an ablation study of our method are shown in Table 1. The baseline method is the multiscale multilabel CNN described in Sec. 4.1. The weighted approximate ranking pairwise loss (WARP) [14] is a widely-used multilabel loss that aims to rank positive labels higher than the negative ones. We applied it to the multiscale multilabel CNN. We defined that fine-grained labels should rank higher than coarse-scale ones if they are all positive. Lesion embedding [48] was trained on DeepLesion based on labels of coarse-scale body parts, lesion location, and size. Among these four methods, LesaNet achieved the best AUC and F1 scores on the two test sets.

The AUCs in Table 1 are relatively high. The algorithms have correctly ranked most positive cases higher than negative ones, proving the effectiveness of the algorithms. However, the F1 scores are relatively low. There are mainly two reasons: 1) The dataset is highly imbalanced with many rare labels. A total of 78 labels have fewer than 10 positive cases in the text-mined test set. These labels may have many more false positives (FPs) than true positives (TPs) when testing, resulting in a low F1. 2) There are missing annotations in the test sets, which is why the accuracies (especially precisions) on the hand-labeled set are significantly higher than the text-mined test set.

Accuracies of some typical labels on the text-mined test set are displayed in Table 2. The average AUC of body parts, types, and attributes are 0.9656, 0.9044, and 0.8384, respectively. Body parts are easier to predict since they typically have more regular appearances. The visual feature of some labels (e.g., paraspinal, nodule) is variable, thus harder to learn. The high AUC and low F1 of "paraspinal" can be explained by the lack of positive test cases (see the explanation in the last paragraph). Some types (e.g., metastasis) can be better predicted by incorporating additional prior knowledge and reasoning. Attributes have lower AUCs partially because some attributes are subjective ("large") or can be subtle ("sclerotic"). Besides, radiologists typically do not describe every attribute of a lesion in the report, thus there are missing annotations in the test set.

Fig. 4 demonstrates examples of our predictions. LesaNet accurately predicted the labels of many lesions. For example, in subplots (a) and (b), two fine-grained body parts (right hilum and pretracheal lymph nodes) were identified; In (c) and (d), a ground-glass opacity and a cavitary lung lesion; In (g) and (h), a hemangioma and a metastasis in liver. Some attributes were also predicted correctly, such as "calcified" in (e), "lobular" in (h), and "tiny" in (i). Errors can occur on some similar body parts and types. In (c), although "left lower lobe" has a high score, "left upper lung" was also predicted, since the two body parts are close. In (g), "metastasis" is a wrong prediction as it may be hard to be distinguished from hemangioma in certain cases. Some rare and / or variable labels were not learned very well, such as "conglomerate" and "necrosis" in (b). Please see the supplementary material for more results.

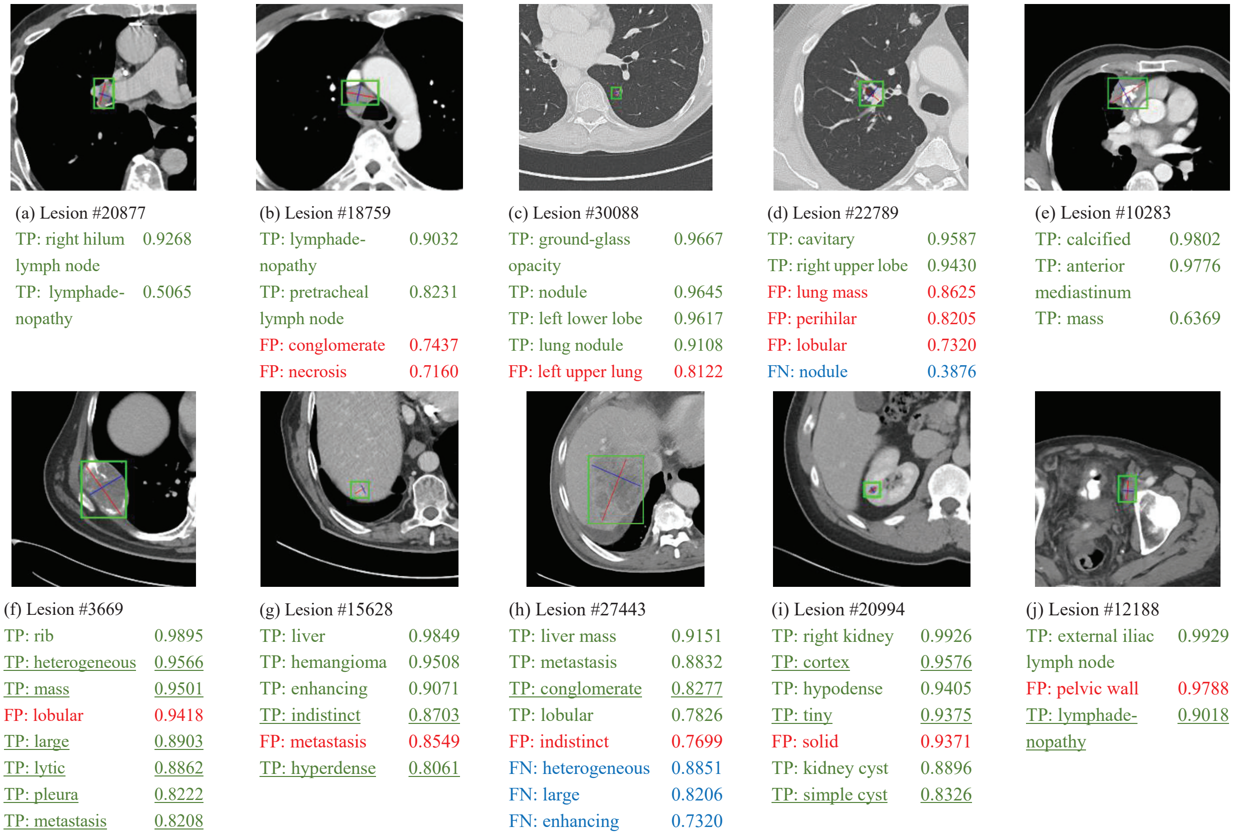
It is efficient to jointly learn all labels holistically. Furthermore, our experiments showed that it does not affect the accuracy of single labels. We conducted an experiment to train and test LesaNet on subsets of labels. For example, subset 1 consists of labels with more than 1000 occurrences in the training set ($n_{\text{tr}} > 1000$); Subset 2 contains labels with $n_{\text{tr}} > 500$. When trained on subset 2, we can test on both subsets 1 and 2 to see if the accuracy on subset 1 has degraded. The results are exhibited in Fig. 5. We can see that for the same test set, the F1 score did not change significantly as the number of training labels increased. Thus, with more data harvested, we may safely add more clinically meaningful labels into training. On the other hand, as more rare labels were added to the test set, the F1 became lower. Fine-grained body parts, types, and many attributes are rare. They are harder to learn due to the lack of training cases. Possible solutions include harvesting more data automatically [47] and using few-shot learning [39].

5.5. Ablation Study and Analysis

Score propagation layer: From the ablation study in Table 1, we find that removing SPL decreased the aver-

Method	Text-mined test set				Hand-labeled test set			
	AUC	Precision	Recall	F1	AUC	Precision	Recall	F1
Multiscale multilabel CNN	0.9048	0.2738	0.5224	0.2823	0.9151	0.3823	0.5340	0.3894
WARP [14]	0.9250	0.2441	**0.6202**	0.3017	0.9316	**0.6677**	0.3273	0.3325
Lesion embedding [48]	0.8933	0.2290	0.5767	0.2610	0.9017	0.3496	**0.5776**	0.3615
LesaNet	**0.9344**	0.3593	0.5327	**0.3423**	**0.9398**	0.4737	0.5274	**0.4344**
w/o score propagation layer	0.9275	**0.3680**	0.4733	0.3233	0.9326	0.4833	0.4965	0.4092
w/o RHEM	0.9338	0.2983	0.5550	0.3178	0.9374	0.4341	0.5327	0.4303
w/o label expansion	0.9148	0.3523	0.5104	0.3270	0.9236	0.4503	0.5420	0.4205
w/o text-mining module	0.9334	0.3365	0.5350	0.3324	0.9392	0.4869	0.5361	0.4250
w/o triplet loss	0.9312	0.3201	0.5394	0.3274	0.9335	0.4645	0.5624	0.4337

Table 1. Multilabel classification accuracy averaged across labels on two test sets. Bold results are the best ones. Red underlined results in the ablation studies are the worst ones, indicating the ablated strategy is the most important for the criterion.



(a) Lesion #20877
TP: right hilum lymph node 0.9268
TP: lymphadenopathy 0.5065

(b) Lesion #18759
TP: lymphadenopathy 0.9032
TP: pretracheal lymph node 0.8231
FP: conglomerate 0.7437
FP: necrosis 0.7160

(c) Lesion #30088
TP: ground-glass opacity 0.9667
TP: nodule 0.9645
TP: left lower lobe 0.9617
TP: lung nodule 0.9108
FP: left upper lung 0.8122

(d) Lesion #22789
TP: cavitary 0.9587
TP: right upper lobe 0.9430
FP: lung mass 0.8625
FP: perihilar 0.8205
FP: lobular 0.7320
FN: nodule 0.3876

(e) Lesion #10283
TP: calcified 0.9802
TP: anterior mediastinum 0.9776
TP: mass 0.6369

(f) Lesion #3669
TP: rib 0.9895
TP: heterogeneous 0.9566
TP: mass 0.9501
FP: lobular 0.9418
TP: large 0.8903
TP: lytic 0.8862
TP: pleura 0.8222
TP: metastasis 0.8208

(g) Lesion #15628
TP: liver 0.9849
TP: hemangioma 0.9508
TP: enhancing 0.9071
TP: indistinct 0.8703
FP: metastasis 0.8549
TP: hyperdense 0.8061

(h) Lesion #27443
TP: liver mass 0.9151
TP: metastasis 0.8832
TP: conglomerate 0.8277
TP: lobular 0.7826
FP: indistinct 0.7699
FN: heterogeneous 0.8851
FN: large 0.8206
FN: enhancing 0.7320

(i) Lesion #20994
TP: right kidney 0.9926
TP: cortex 0.9576
TP: hypodense 0.9405
TP: tiny 0.9375
FP: solid 0.9371
TP: kidney cyst 0.8896
TP: simple cyst 0.8326

(j) Lesion #12188
TP: external iliac lymph node 0.9929
FP: pelvic wall 0.9788
TP: lymphadenopathy 0.9018

Figure 4. Sample predicted labels with confidence scores on the text-mined test set. Green, red, and blue results correspond to TPs, FPs, and FNs (false negatives), respectively. Underlined labels are TPs with missing annotations, thus were treated as FPs during evaluation. Only the most fine-grained predictions are shown with their parents omitted for clarity.

age per-class recall by 3%. Among it, the recall of frequent labels ($n_{\mathrm{tr}} > 1000$) only decreased 0.4%, showing that SPL is important for the recall of rare labels, at the cost of small precision loss. We further examined the learned transformation matrix W in SPL, see Fig. 6 for an example. We can find that $W(\text{liver, hemangioma})$ and $W(\text{enhancing, hemangioma})$ are high. It means SPL discovered the fact that if a lesion is a hemangioma in DeepLesion, it is highly likely in the liver and enhancing, so SPL increased the scores for "liver" and "enhancing". In turn, the scores of liver and enhancing also contributed positively to the final score of hemangioma (see Fig. 4 (g) for an example of hemangioma). Note that these relations were not explicitly defined in the ontology. The label "chest" is exclusive with "abdomen" and "liver", so the learned weights between them are negative. As explained in Sec. 4.3, he-

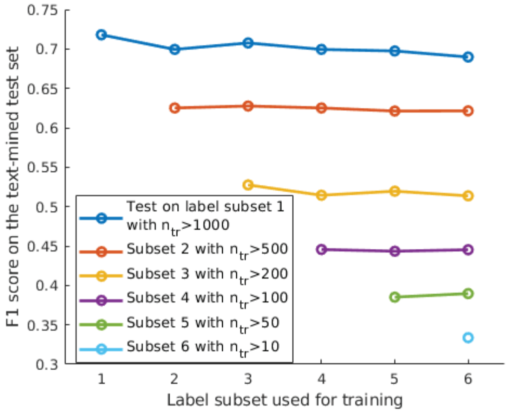


Figure 5. Accuracy of training and testing LesaNet on different subsets of labels. Each curve corresponds to a test subset.

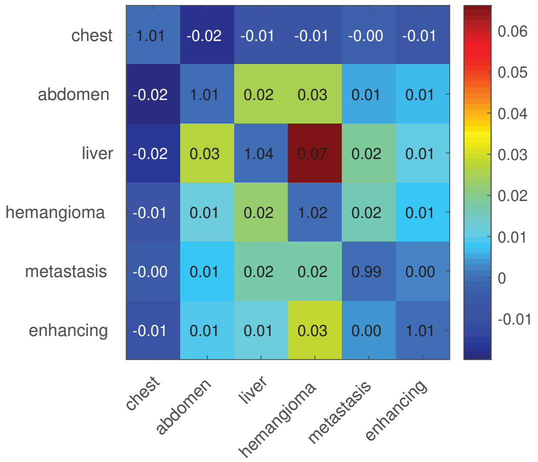


Figure 6. A part of the learned score propagation weights W. The weight in row i, column j is w_{ij}, i.e., the refinement of label i's score received from label j's score. The final scores $\tilde{\boldsymbol{s}} = W\boldsymbol{s}$.

mangioma and metastasis in the liver are hard for the algorithm to distinguish, so SPL also learned positive weights between them. In the future, using our holistic and comprehensive prediction framework, we may try to incorporate more human knowledge into the model, such as "type a locates in body part b and has attribute c", "type d is similar to type e except for attribute f".

Relational hard example mining: RHEM, on the contrary to SPL, is crucial for improving the precision (Table 1), probably because it suppressed the scores of the reliable hard negative labels at the cost of mildly decreased recall. In RHEM, the hard negative labels of a lesion were selected from the exclusive labels of existing positive ground-truths. If we discard this "reliable" requirement and select negative labels from all labels that are not positive, the precision will increase 1.5% because we suppressed more negative labels, but the recall will decrease 4% since many suppressed negative labels are actually positive due to missing annotations.

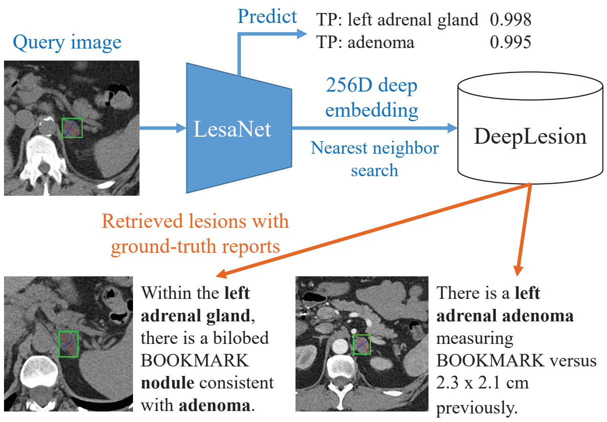

Query image
Predict
TP: left adrenal gland 0.998
TP: adenoma 0.995
LesaNet
256D deep embedding
Nearest neighbor search
DeepLesion
Retrieved lesions with ground-truth reports
Within the **left adrenal gland**, there is a bilobed BOOKMARK **nodule** consistent with **adenoma**.
There is a **left adrenal adenoma** measuring BOOKMARK versus 2.3 x 2.1 cm previously.

Figure 7. Sample lesion retrieval results.

Label expansion: Without it, the training set will lose 40% (parent) labels, thus the accuracy was not good.

Text-mining module: When this was not used, the overall accuracy dropped as the irrelevant training labels brought noises. However, the performance did not degrade substantially, showing that our model is able to tolerate noisy labels to a certain degree [18]. We also found training with the relevant + uncertain labels was better than using relevant labels only, which is because most uncertain labels are radiologists' inferences that are very likely to be true, especially if we only consider the lesion's appearance.

Triplet loss: The triplet loss also contributed to the classification accuracy slightly. The 256D embedding learned from the triplet loss can be used to retrieve similar lesions from the database given a query one. In Fig. 7, LesaNet not only predicted the labels of the query lesion correctly, but also retrieved lesions with the same labels, although their appearances are not identical. The retrieved lesions and reports can provide evidences to the predicted labels as well as help the user understand the query lesion.

More qualitative and quantitative results are presented in the supplementary material.

6. Conclusion and Future Work

In this paper, we studied the holistic lesion annotation problem, and proposed a framework to automatically learn clinically meaningful labels from radiology reports and label ontology. A lesion annotation network was proposed with effective strategies that can both improve the accuracy and bring insights and interpretations. Our future work may include harvesting more data to better learn rare and hard labels and trying to incorporate more human knowledge.

Acknowledgment: This research was supported by the Intramural Research Program of the NIH Clinical Center and National Library of Medicine. We thank NVIDIA for the donation of GPUs and Dr. Le Lu and Dr. Jing Xiao for their valuable comments.

References

[1] BioPortal. Radiology Lexicon, 2018. https://bioportal.bioontology.org/ontologies/RADLEX.

[2] Steven Bird, Steven Bird, and Edward Loper. NLTK: The natural language toolkit. In *Annual Meeting of the Association for Computational Linguistics*, pages 63–70, 2016.

[3] Jinzheng Cai, Youbao Tang, Le Lu, Adam P. Harrison, Ke Yan, Jing Xiao, Lin Yang, and Ronald M. Summers. Accurate Weakly-Supervised Deep Lesion Segmentation using Large-Scale Clinical Annotations: Slice-Propagated 3D Mask Generation from 2D RECIST. In *MICCAI*, pages 396–404, 2018.

[4] Qingyu Chen, Yifan Peng, and Zhiyong Lu. BioSentVec: creating sentence embeddings for biomedical texts. *arXiv preprint arXiv:1810.09302*, 2018.

[5] Sihong Chen, Jing Qin, Xing Ji, Baiying Lei, Tianfu Wang, Dong Ni, and Jie Zhi Cheng. Automatic Scoring of Multiple Semantic Attributes with Multi-Task Feature Leverage: A Study on Pulmonary Nodules in CT Images. *IEEE Transactions on Medical Imaging*, 36(3):802–814, Mar. 2017.

[6] Jie-Zhi Cheng, Dong Ni, Yi-Hong Chou, Jing Qin, Chui-Mei Tiu, Yeun-Chung Chang, Chiun-Sheng Huang, Dinggang Shen, and Chung-Ming Chen. Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans. *Sci. Rep.*, 6(1):24454, 2016.

[7] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. NUS-WIDE: A Real-World Web Image Database from National University of Singapore. In *Proceeding of the ACM International Conference on Image and Video Retrieval - CIVR '09*, page 48, 2009.

[8] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. The cancer imaging archive (TCIA): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6):1045–1057, 2013.

[9] Sergey Demyanov, Rajib Chakravorty, Zongyuan Ge, Seyed-Behzad Bozorgtabar, Michelle Pablo, Adrian Bowling, and Rahil Garnavi. Tree-loss function for training neural networks on weakly-labelled datasets. In *ISBI*, pages 287–291. IEEE, apr 2017.

[10] Idit Diamant, Assaf Hoogi, Christopher F. Beaulieu, Mustafa Safdari, Eyal Klang, Michal Amitai, Hayit Greenspan, and Daniel L. Rubin. Improved Patch-Based Automated Liver Lesion Classification by Separate Analysis of the Interior and Boundary Regions. *IEEE J. Biomed. Heal. Informatics*, 20(6):1585–1594, 2016.

[11] Qi Dong, Shaogang Gong, and Xiatian Zhu. Imbalanced Deep Learning by Minority Class Incremental Rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–14, 2018.

[12] José Raniery Ferreira, Marcelo Costa Oliveira, and Paulo Mazzoncini de Azevedo-Marques. Characterization of Pulmonary Nodules Based on Features of Margin Sharpness and Texture. *Journal of Digital Imaging*, pages 1–13, 2017.

[13] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[14] Yunchao Gong, Yangqing Jia, Thomas Leung, Alexander Toshev, and Sergey Ioffe. Deep Convolutional Ranking for Multilabel Image Annotation. *arXiv preprint arXiv:1312.4894*, 2013.

[15] Johannes Hofmanninger and Georg Langs. Mapping visual features to semantic profiles for retrieval in medical imaging. In *CVPR*, volume 07-12-June, pages 457–465, 2015.

[16] Hexiang Hu, Guang-Tong Zhou, Zhiwei Deng, Zicheng Liao, and Greg Mori. Learning Structured Inference Neural Networks with Label Relations. In *CVPR*, pages 2960–2968, 2016.

[17] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *ICML*, pages 448–456, 2015.

[18] Jonathan Krause, Benjamin Sapp, Andrew Howard, Howard Zhou, Alexander Toshev, Tom Duerig, James Philbin, and Li Fei-Fei. The unreasonable effectiveness of noisy data for fine-grained recognition. In *ECCV*, pages 301–320, 2016.

[19] Curtis P. Langlotz. RadLex: a new method for indexing online educational materials. *Radiographics*, 26(6):1595–1597, Nov 2006.

[20] Xiaoli Li and Bing Liu. Learning to classify texts using positive and unlabeled data. In *IJCAI*, pages 587–592, 2003.

[21] Yuncheng Li, Yale Song, and Jiebo Luo. Improving pairwise ranking for multi-label image classification. In *CVPR*, volume 2017-Janua, pages 1837–1845, 2017.

[22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. In *ICCV*, pages 2980–2988, 2017.

[23] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, dec 2017.

[24] Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. The More You Know: Using Knowledge Graphs for Image Classification. In *CVPR*, pages 20–28, 2017.

[25] Ishan Misra, C. Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. In *CVPR*, pages 2930–2939, 2016.

[26] The National Institutes of Health. RadLex, 2016. https://healthdata.gov/dataset/radlex.

[27] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

[28] Yifan Peng, Anthony Rios, Ramakanth Kavuluru, and Zhiyong Lu. Extracting chemical-protein relations with ensembles of svm and deep learning models. *Database : the journal of biological databases and curation*, 2018, Jan. 2018.

[29] Yifan Peng, Ke Yan, Veit Sandfort, Ronald M. Summers, and Zhiyong Lu. A self-attention based deep learning method for

lesion attribute detection from CT reports. In *IEEE International Conference on Healthcare Informatics*, 2019.

[30] Hariharan Ravishankar, Prasad Sudhakar, Rahul Venkataramani, Sheshadri Thiruvenkadam, Pavan Annangi, Narayanan Babu, and Vivek Vaidya. Medical Image Description Using Multi-task-loss CNN. In *LABELS 2016, DLMIA 2016*, volume 1, pages 121–129, 2016.

[31] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training Deep Neural Networks on Noisy Labels with Bootstrapping. In *ICLR Workshop*, 2015.

[32] Berkman Sahiner, Aria Pezeshk, Lubomir M. Hadjiiski, Xiaosong Wang, Karen Drukker, Kenny H. Cha, Ronald M. Summers, and Maryellen L. Giger. Deep learning in medical imaging and radiation therapy. *Med. Phys.*, oct 2018.

[33] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015.

[34] Arnaud Arindra Adiyoso et al. Setio. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. *Medical Image Analysis*, 42:1–13, 2017.

[35] Hoo Chang Shin, Le Lu, Lauren Kim, Ari Seff, Jianhua Yao, and Ronald Summers. Interleaved text/image deep mining on a large-scale radiology image database for Automated Image Interpretation. *Journal of Machine Learning Research*, 17(9783319429984):305–321, 2016.

[36] Hoo-Chang Shin, Holger R. Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M. Summers. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298, may 2016.

[37] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training Region-based Object Detectors with Online Hard Example Mining. In *CVPR*, pages 761–769, 2016.

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR 2015*, 2015.

[39] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H S Torr, and Timothy M Hospedales. Learning to Compare: Relation Network for Few-Shot Learning. In *CVPR*, pages 1199–1208, 2018.

[40] Youbao Tang, Adam P. Harrison, Mohammadhadi Bagheri, Jing Xiao, and Ronald M. Summers. Semi-Automatic RECIST Labeling on CT Scans with Cascaded Convolutional Neural Networks. In *MICCAI*, pages 405–413, jun 2018.

[41] Yuxing Tang, Xiaosong Wang, Adam P. Harrison, Le Lu, Jing Xiao, and Ronald M. Summers. Attention-Guided Curriculum Learning for Weakly Supervised Classification and Localization of Thoracic Diseases on Chest Radiographs. In *International Workshop on Machine Learning in Medical Imaging*, pages 249–258. Springer, Cham, sep 2018.

[42] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. CNN-RNN: A Unified Framework for Multi-label Image Classification. In *CVPR*, pages 2285–2294, 2016.

[43] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *CVPR*, pages 2097–2106, may 2017.

[44] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-rays. In *CVPR*, pages 9049–9058, 2018.

[45] Jason Weston, Samy Bengio, and Nicolas Usunier. WSABIE: Scaling Up To Large Vocabulary Image Annotation. In *IJCAI*, pages 2764–2770, 2011.

[46] Ke Yan, Mohammadhadi Bagheri, and Ronald M. Summers. 3D Context Enhanced Region-based Convolutional Neural Network for End-to-End Lesion Detection. In *MICCAI*, pages 511–519, 2018.

[47] Ke Yan, Xiaosong Wang, Le Lu, and Ronald M. Summers. DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of Medical Imaging*, 5(3), 2018.

[48] Ke Yan, Xiaosong Wang, Le Lu, Ling Zhang, Adam Harrison, Mohammadhadi Bagheri, and Ronald Summers. Deep Lesion Graphs in the Wild: Relationship Learning and Organization of Significant Radiology Image Findings in a Diverse Large-scale Lesion Database. In *CVPR*, pages 9261–9270, 2018.

[49] Min Ling Zhang and Zhi Hua Zhou. A review on multi-label learning algorithms. *IEEE Trans. Knowl. Data Eng.*, 26(8):1819–1837, aug 2014.

[50] Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. MDNet: A Semantically and Visually Interpretable Medical Image Diagnosis Network. In *CVPR*, pages 6428–6436, 2017.

[51] Fang Zhao, Yongzhen Huang, Liang Wang, and Tieniu Tan. Deep semantic ranking based hashing for multi-label image retrieval. In *CVPR*, volume 07-12-June, pages 1556–1564, 2015.