
TOWARDS RADIOLOGIST-LEVEL THORACIC DISEASE DIAGNOSE ON CHEST X-RAYS USING DEEP NEURAL NETWORK

Xuanyue Yang

xuanyuey@andrew.cmu.edu

Zihua Liu

zihual@andrew.cmu.edu

Ningqian Zhang

ningqianz@andrew.cmu.edu

Wenting Ye

wye2@andrew.cmu.edu

ABSTRACT

X-ray images are one of the most affordable and widely used medical diagnosis method for various diseases, especially thoracic diseases. Significant amount of X-ray imaging data exist in health care organizations across the world. However, analyzing these X-ray images and generating pathology reports may take large amount of effort and require high-level expertise of a radiologist. Deep neural networks, permitting higher levels of abstraction and improving predictions from raw data, can be used to help with this diagnosis process. In this paper, we build a thoracic disease diagnose system using deep neural network, which could identify the disease and generate pathology report simultaneously. The network is composed of a 132-layer DenseNet as encoder to classify the disease and LSTM with self-attention as decoder to generate the report. Through extensive experiments and comparisons with existing work, we show the effectiveness of our proposed model.

1 INTRODUCTION

With the increase of annotated visual data in clinical procedures, deep neural networks have rapidly become a methodology of choice for analyzing medical images. Applying artificial intelligence techniques to automatically detect and localize different types of thoracic diseases, and generate radiological reports from X-rays become a critical need. We use two chest X-ray datasets for this work, the National Institutes of Health (NIH) ChestXray-14 dataset and the Indiana U. Chest X-rays dataset. Our model is composed of two parts: an image-encoder for multi-label classification and a text-decoder for report generation. For the image-encoder, we apply various deep neural networks, including VGGNet-16, ResNet-101, ResNet152, SE-ResNet152, and DenseNet-121. We utilize a weighted binary cross entropy loss to mitigate the problem of unbalanced disease class distribution. To account for the issue of corrupted labels in the ChestXray-14 dataset, we apply mixup, a regularization technique to our DenseNet model. The main focus of our work is three-fold: (1) We propose a joint model for multi-label disease classification and report generation; (2) We handle the problem of unbalanced data in the ChestXray-14 dataset; (3) We address the problem of corrupted data in the ChestXray-14 dataset by training our image encoder on the relabeled ChestXray-14 dataset provided by Rajpurkar et al. (2018).

2 RELATED WORK

Many recent works have been done on both the NIH ChestXray-8 and the NIH ChestXray-14 datasets. Wang et al. (2017) applied a pre-trained Deep Convolutional Neural Network (DCNN) for multi-label classification and achieve an average area under curve (AUC) of 0.738. Yao et al. (2017) proposed a model with an RNN decoder that exploited the conditional dependencies among disease labels and obtained an average ROC AUC of 0.798 on the 14 labels. The CheXNet proposed by Rajpurkar et al. (2017) achieved an F1 score of 0.435. However, this result is not comparable to our results, because their model was tested based on human-reabeled test data. Rajpurkar et al. (2018) later proposed another convolutional neural network CheXNetXt for concurrent disease detection and a method to relabeled the corrupted ChestXray-14 dataset, and released their relabeled training and validation data. Works have also been done on the Indiana U. chest X-rays dataset, mainly for the report generation task. Wang et al. (2018) utilized the ChestXray-14 image data together with publicly-unavailable text reports to train the text-image embedding network for both disease classification and reporting and evaluate the model on both the ChestXray-14 and the Indiana U. chest X-ray datasets. Another report generation model which utilized a co-attention mechanism and a hierarchical LSTM was developed by Jing et al. (2017).

3 METHODS

3.1 DISEASE CLSSIFICATION

The disease classification task could be formulated as a multi-label classification task. Given an image $x \in \mathbb{R}^{3 \times p \times p}$, the system predicts $y \in \mathbb{R}^k$ indicating whether the patient has a certain disease, where p is the scale of the image and k is the number of diseases. We formulate this question as a k -way binary classification question and minimize the following loss function:

$$\ell(X, Y) = \sum_{i=1}^n \sum_{j=1}^k Y_j^{(i)} \log(f_j(X^{(i)})) + (1 - Y_j^{(i)}) \log(1 - f_j(X^{(i)})) \quad (1)$$

where the upperscript and lowerscript denote the index of sample and disease respectively and the n is the number of samples. We use the DenseNet Huang et al. (2017b) for this task, which is currently most widely used image classification model. DenseNets improve flow of information and gradients through the network, making the optimization of very deep networks tractable.

3.1.1 HANDLING UNBALANCED DATA

To address unbalance of the data, we implement the weighted loss and oversampling. In weighted loss, instead of the regular loss, we optimize the following weighted one:

$$\ell(X, Y) = \sum_{i=1}^n \sum_{j=1}^k (1 - w_k) Y_j^{(i)} \log(f_j(X^{(i)})) + w_k (1 - Y_j^{(i)}) \log(1 - f_j(X^{(i)})) \quad (2)$$

Intuitively, we punish the model more if we misclassify a dominating class into another class.

Another approach is to sample the data according to their distribution, i.e. the data with minority label will have higher probability to be sampled. The probability of sampling for i -th sample is:

$$sample_weight^{(i)} = \frac{1}{Z} \sum_{j=1}^k y_j^{(i)} w_j \quad (3)$$

where Z is a normalized coefficient, and w_j is the proportion of j -th class. Since the sampling weight will be zero for data with no findings, we assign a constant to them.

3.1.2 HANDLING CORRUPTED DATA

To handle the inaccurate labels which may influence the performance of our model, we apply the relabeled Chestx-ray14 dataset proposed by Rajpurkar et al. (2018) instead of the original one to train our network. Multiple networks are trained on the original training set to predict the probability of each pathology presenting in the input image. A subset of those networks, which are chosen based on the average error on the tuning set, constitute an ensemble that produces predictions by computing the mean over the predictions of each individual network. The ensemble is used to relabel the original Chestx-ray14 dataset and the relabeled dataset is proved to have a higher labeling accuracy by human experts.

3.2 PATHOLOGY LOCALIZATION

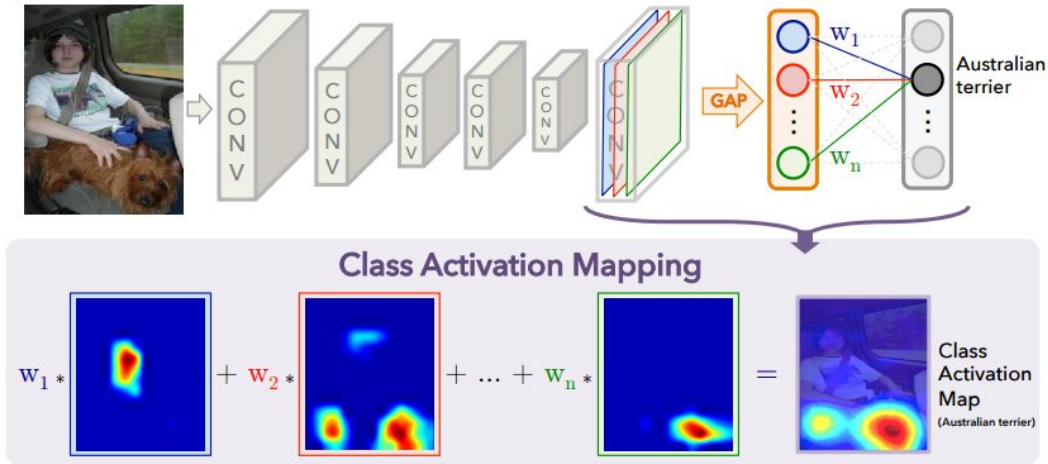


Figure 1: Learning Deep Features for Discriminative Localization, Bolei et al., CVPR 2016

Heatmap Generation: Our multi-label classification network is designed not only for classification, we also use the global and the local features extracted by our network to generate the likelihood map of pathologies, namely a heatmap. We use a general technique called Class Activation Mapping (CAM) proposed by Zhou et al. (2016) for CNNs with global average pooling, which enables our classification-trained CNN to learn to perform pathology localization, without using any bounding box annotations. The generated disease heatmap allows us to visualize and highlight the discriminative disease regions of the input chest X-ray image detected by our CNN.

Our heatmaps are generated using CAM as illustrated in figure 1. For a given image, let $f_k(x, y)$ represent the activation of channel k in the last convolutional layer of our CNN at location (x, y) . If the last convolutional layer has K channels, then the global average pooling layer which follows the last convolutional layer would have K neurals and the k^{th} neural would be the global average pooling of the k^{th} channel of the last convolutional layer, $F^k = \sum_{x,y} f_k(x, y)$. For a given disease class c , the input to the final softmax layer S_c , equals to $\sum_k w_k^c F_k$, where w_k^c is the weight of the class c for the k^{th} neural in the pooling layer. Since w_k^c directly indicates the significance of the feature extracted by the k^{th} channel of the last convolutional layer, the class activation map for disease class c is calculated with the weighted sum of each channel in the last convolutional layer.

$$M_c(x, y) = \sum_k w_k^c f_k(x, y)$$

We re-scale the class activation map to the original size of the input image and obtain the heatmap of the given image.

Bounding Box Generation: The bounding box for each input image is generated based on the heatmap corresponding to that image. We pick up the local maximum points of the heatmap and

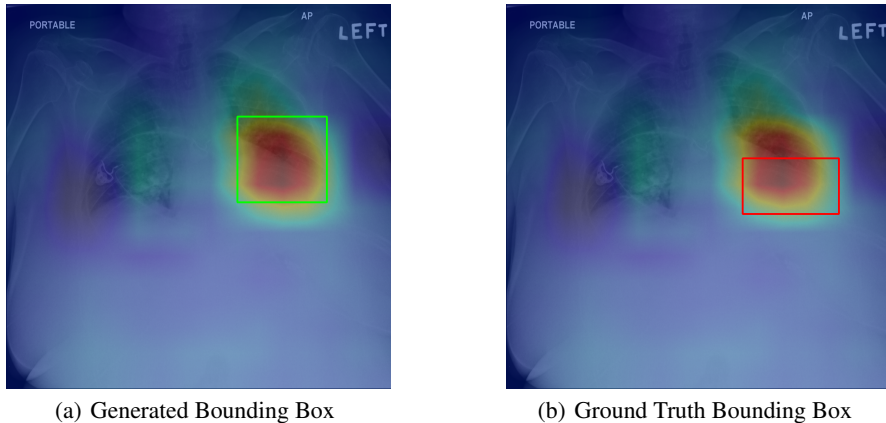


Figure 2: Generated heatmap and bounding box

extend the bounding box horizontally and vertically from the local maximum point until the heatmap value is lower than a threshold. We use the bounding box with the largest size as the final bounding box for the given image. Figure 2 shows an example of our generated bounding box with comparison to the ground truth bounding box.

3.3 REPORT GENERATION

We experiment on the task of report generation using the Indiana U. Chest X-rays dataset, where an amount of 7K images with the corresponding reports are available. As DenseNet Huang et al. (2017b) gives the best performance on our image classification task, we use it as the image encoder and directly adapt the parameters trained on the larger ChestXray-14 dataset. We baseline our report generation model on a image-attended 2-layer RNN decoder, which in each time step, takes as input the word embedding and the attention context over the image. The output of the image encoder is

$$\{\mathbf{f}_{i,j} \in \mathbb{R}^F\}_{i=0,j=0}^{W_o-1,H_o-1}, \quad (4)$$

where F is the number of feature (output channel) and (W_o, H_o) is the output shape. We flatten it as a sequence of feature vectors

$$\mathbf{F} = [\dots, \mathbf{f}_{i \times W_o + j}, \dots]. \quad (5)$$

We adopt Scaled Dot-Product Attention in Vaswani et al. (2017):

$$\mathbf{c} = \mathbf{V}(\text{Softmax}(\mathbf{q}^\top \mathbf{K}))^\top, \quad (6)$$

where the query \mathbf{q} , keys \mathbf{K} and values \mathbf{V} are MLP transformations of the current hidden state h and the flatten image feature vectors:

$$\mathbf{q} = \text{MLP}_q(\mathbf{h}), \quad (7)$$

$$\mathbf{K} = \text{MLP}_k(\mathbf{F}), \quad (8)$$

$$\mathbf{V} = \text{MLP}_v(\mathbf{F}). \quad (9)$$

To improve the performance of decoding a long sequence, we use self-attention with second layer RNN attending over all previous hidden states of the first layer:

$$\mathbf{q} = \text{MLP}_q(\mathbf{h}_{t-1}^l), \quad (10)$$

$$\mathbf{K} = \text{MLP}_k([\mathbf{h}_0^{l-1}, \dots, \mathbf{h}_{t-1}^{l-1}]), \quad (11)$$

$$\mathbf{V} = \text{MLP}_v([\mathbf{h}_0^{l-1}, \dots, \mathbf{h}_{t-1}^{l-1}]). \quad (12)$$

Jing et al. (2017) proposed a hierarchical decoder with a sentence-level RNN generating sentence topic vectors and a word-level RNN unwrap each sentence based on the topic vector, which achieved a much better result in terms of BLEU scores and other metrics. We also implement their architecture but are not yet able to reproduce their performance.

4 EXPERIMENTS

4.1 DATASET

The experiments are conducted on two chest X-ray datasets. The first one is the ChestXray-14 dataset released by NIH. It is an extended version of the ChestXray-8 dataset by Wang et al. (2017). It contains 111,753 frontal view X-ray images of 30,805 unique patients with 14 disease labels which were text-mined from publicly-unavailable radiological reports. The disease class distribution is unbalanced with over half of the data points being labeled as 'No Finding'. The text-mined labels are claimed to have an accuracy of greater than 90%. However, this accuracy is the text-mining accuracy. Based on our experiments, we believe that the label accuracy according to the images is much lower than this for both the training set and the test set. The split of this dataset is based on the split performed by Rajpurkar et al. (2017). They ensured that no patient overlap exists between the splits. The other dataset we use is the Indiana U. Chest X-rays dataset with 7,470 X-ray images and 3,955 radiology reports.

4.2 EXPERIMENTAL SETTINGS

To train the classification model (encoder), we first rescale the image to 640×640 resolution, randomly crop into 512×512 size, and normalize based on the mean and standard deviation of images in the ImageNet training set. Horizontal flipping is performed in 50% possibility. The weights of the network are initialized with weights from a model pretrained on ImageNet. We train the network in end-to-end manner using Adam with default parameters. The batch size is set to 8. The learning rate is 0.001 and decay by a factor of 10 when the AUROC in validation set plateaus after an epoch. No matter training on original or relabeled data, we only evaluate on the original test split.

4.3 RESULT

4.3.1 DISEASE CLASSIFICATION

Table 1: Disease classification Area under Receiver Operating Characteristic (AURO)

Pathology	Wang et al.	Yao et al.	Ours	Rajpurkar et al.
Atelectasis	0.716	0.772	0.838	0.809
Cardiomegaly	0.807	0.904	0.911	0.925
Effusion	0.784	0.859	0.883	0.864
Infiltration	0.609	0.695	0.752	0.735
Mass	0.706	0.792	0.849	0.868
Nodule	0.671	0.717	0.803	0.780
Pneumonia	0.633	0.713	0.837	0.768
Pneumothorax	0.806	0.841	0.868	0.889
Consolidation	0.708	0.788	0.821	0.790
Edema	0.835	0.882	0.901	0.888
Emphysema	0.815	0.829	0.938	0.937
Fibrosis	0.769	0.767	0.794	0.805
Pleural Thickening	0.708	0.765	0.805	0.806
Hernia	0.767	0.914	0.709	0.916
Average AUROC	0.738	0.798	0.826	0.841

Table 1 shows our classification results as well as other state-of-the-art models. Our model trained on relabeled data achieves averages AUROC of 0.826 and outperforms Wang et al. (2017) Yao et al. (2017) by a distinct margin. Although Rajpurkar et al. (2017) has better classification score than us, but the results are not comparable. They use a private test dataset with 420 images to evaluate their model, which is also not consistent with our settings.

Table 2: The AuROC of our model with different training techniques

Model	AUROC	F1 Score	Best Epoch
Baseline	0.826	0.41	12
+ 224×224 scale	0.783	0.32	10
+ weight decay (1e-4)	0.792	0.33	11
+ dropout	0.806	0.382	3
+ oversampling	0.818	0.40	5
+ weighted loss	0.810	0.39	9
+ mixup + oversampling	0.787	0.25	4

We also try different techniques shown in Table 2. Although none of these methods improve model’s performance, it shed the light of this dataset. First, since the DenseNet is pretrained on 224×224 resolution, we try to feed the network with the same image size. The reason of failure might be the image is compressed from 2048 × 2048, such high degree of compression will lose important information. Secondly, we observe that the loss on train split is much lower than in development set, so we add some common regularizers as well as mixup Hongyi Zhang (2018). But it doesn’t work as expected because they might eliminate the noise during the training process, which also exists in the test set. Oversampling can dramatically speed up the training process and we expect the performance gap is due to randomness during the training process.

4.3.2 PATHOLOGY LOCALIZATION

Table 3: Pathology Localization Accuracy

	T(IoU) = 0.2		T(IoU) = 0.6	
	ChestX-ray14	ChestX-ray8	ChestX-ray14	ChestX-ray8
Atelectasis	0.3611	0.4722	0.0167	0.0222
Cardiomegaly	0.9247	0.6849	0.2397	0.0753
Effusion	0.4444	0.4509	0.0196	0.0457
Infiltration	0.4878	0.4796	0.0650	0.0243
Mass	0.4824	0.2588	0.0471	0.0000
Nodule	0.1266	0.0506	0.0000	0.0126
Pneumonia	0.5083	0.3500	0.0500	0.0166
Pneumothorax	0.3878	0.2346	0.0714	0.0306
Average AUROC	0.4654	0.3727	0.0637	0.0284

We evaluate the performance of our generated heatmaps and compute bounding boxes against the hand annotated ground truth bounding boxes included in ChestX-ray14. It enables us to get a reasonable estimate on how well our models perform on feature extraction and disease localization. We use the standard Intersection over Union ratio (IoU) to measure the performance of our localization. We define a correct localization requiring $IoU > T(IoU)$. Table 3 illustrates the accuracy of our localization result on Chestx-ray14 and we found that the dataset only contains annotated bounding boxes of about 900 images for 8 class of diseases. Wang et al. (2017) also explored disease localization on dataset ChestX-ray8 and this dataset contains annotated bounding boxes for the same 8 classes of diseases on 1600 images. Despite of the fact that our model is working on more number of disease classes, we still get a fair result compared with their performance and we perform even better on some classes of disease like Cardimogaly and Pneumothorax.

4.3.3 REPORT GENERATION

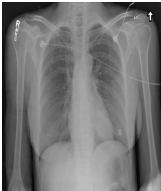

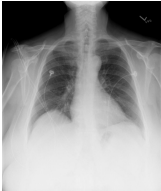
We give in Table 4 the performance of our report generation models in terms of BLEU- $\{1, 2, 3, 4\}$ and ROUGE scores (using <https://github.com/harpriobot/nlp-metrics>). Due to constraint of time and GPU resources, we complete the experiments on just our baseline decoder and self-attentive decoder, and the results show a better capability of self-attentive decoder to generate sequence. We also attempted to reproduce the hierarchical decoder proposed by Jing et al. (2017) which has a BLEU-4 score of 0.247 on their test set but have not yet worked it out. Nevertheless, our self-attentive decoder has achieved a relatively better or comparable performance compared to other report generation models mentioned in Jing et al. (2017) (not listed here due to different dataset partitioning).

Table 4: Report Generation Results

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE
Baseline	0.3615	0.1910	0.1145	0.0746	0.2803
+ Self-attention	0.4195	0.2578	0.1740	0.1231	0.3259

Table 5 below shows several examples of our generated reports, which could to some extent reasonably describe the X-ray image, as compared to the ground truth. Limitations are that the dataset is too small for the model to learn a better understanding of the image, and that a majority of reports in the dataset follow a similar pattern (typically for normal images), leading to the bias of generated reports. We believe that with more data available, e.g., if NIH would release the corresponding reports with the ChestXray-14 dataset, the task could be better solved. Wang et al. (2018) from NIH actually train the report generation model on the ChestXray-14 dataset with the reports available, but kept privately within NIH.

Table 5: Gound Truth and Generated Report Comparisons

Image	Ground Truth	Generated Report
	heart size normal. mediastinal silhouettes and pulmonary vascularity are within normal limits. calcified lingular granuloma. No focal consolidations or pleural effusions. No pneumothorax. Breast implants there is a moderate wedge xxxx deformity of the midthoracic vertebrae, xxxx t6, age-indeterminate.	1. no acute cardiopulmonary disease. the heart and mediastinum are unremarkable. the lungs are clear without infiltrate. there is no effusion or pneumothorax. there is a mild wedge xxxx deformity of the midthoracic vertebral body.
	very low lung volumes, bronchovascular crowding and bibasilar areas of atelectasis. no lobar consolidataion. no appreciable pleural effusion or pneumothorax. heart size within normal limits.	the cardiomedial silhouette is within normal limits. there are low lung volumes with bronchovascular crowding. there is no large pleural effusion. there is no pneumothorax.
	negative for acute cardiopulmonary process. negative for cardiac enlargement or vascular congestion. minimal subsegmental atelectasis at the left base otherwise negative for focal confluent airspace disease. the visualized bony structures are intact. there are minimal degenerative disc changes of the mid/lower thoracic spine. no pneumothorax.	1. no acute cardiopulmonary disease. the heart and mediastinum are unremarkable. the lungs are clear without infiltrate. there is no effusion or pneumothorax. there is mild degenerative changes of the thoracic spine.

5 DISCUSSION

5.1 LIMITATION OF DATASET

5.1.1 CORRUPTED TEST LABELS

One major limitation of this dataset is that both the training and test labels are corrupted to some extent, since they are both text-mined from text reports. We applied the mixup regularization method which to help increase the robustness of a neural model when the training labels are corrupted. However, it fails to improve the performance of a DenseNet model trained on data without relabeling. This result supports our previous hypothesis that both the training and test labels are not exactly accurate. It is likely that without mixup the model can predict the similarly corrupted test labels with more success.

5.1.2 SPLIT OF DATASET

The original split of the ChestXray-14 dataset available on Kaggle does not guarantee no patient overlap between splits, which may lead to the problem of data leakage. As a result, we adopt the split performed by Rajpurkar et al. (2017) which provides this guarantee.

5.2 FUTURE WORK

For future work, one thing we could do is to refine the ChestXray-14 dataset, especially the labels. Even though Wang et al. (2017) mentioned the text-mined labels are at least 90% accurate, they did not take any possible mismatch between text reports and X-ray images into consideration and did not mention how accurate a label is to an image. The meaning of some labels is also obscure. For example, there is no information about whether 'No Finding' equals to 'No Disease' in this dataset. What's more, we could continue working on reproducing the hierarchical LSTM model with co-attention mechanism proposed by Jing et al. (2017) which demonstrates better performance in the report generation task. We could introduce other state-of-the-art vision object detection neural models like R-CNN for better disease localization and utilize both frontal and lateral view of the X-ray images to further improve classification performance.

REFERENCES

- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016. doi: 10.1109/cvpr.2016.90. URL <http://dx.doi.org/10.1109/CVPR.2016.90>.
- Yann N. Dauphin David Lopez-Paz Hongyi Zhang, Moustapha Cisse. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations*, 2018. URL <https://openreview.net/forum?id=r1Ddpl-Rb>.
- Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks, 2017.
- Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017a. doi: 10.1109/cvpr.2017.243. URL <http://dx.doi.org/10.1109/CVPR.2017.243>.
- Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, volume 1, pp. 3, 2017b.
- Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports, 2017.
- Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning, 2017.

-
- Pranav Rajpurkar, Jeremy Irvin, Robyn L. Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P. Langlotz, and et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLOS Medicine*, 15(11), 2018. doi: 10.1371/journal.pmed.1002686.
- Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.
- 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. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. doi: 10.1109/cvpr.2017.369. URL <http://dx.doi.org/10.1109/CVPR.2017.369>.
- 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9049–9058, 2018.
- Li Yao, Eric Poblenz, Dmitry Dagunts, Ben Covington, Devon Bernard, and Kevin Lyman. Learning to diagnose from scratch by exploiting dependencies among labels, 2017.
- Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929, 2016.