# Lung Nodule Detection and Classification

Isabel Bush
Stanford Computer Science
353 Serra Mall, Stanford, CA 94305
ibush@stanford.edu

## Abstract

*Detection of malignant lung nodules in chest radiographs is currently performed by pulmonary radiologists, potentially with the aid of CAD systems. Recent advancements in convolutional neural network (CNN) models have improved image classification and detection for many tasks, but there has been little exploration of their use for nodule detection in chest radiographs. In this paper we explore using a ResNet CNN model with transfer learning to classify complete chest radiographs as non-nodule, benign nodule, or malignant nodule, and to localize the nodule, if present. The model is able to classify radiographs as nodule or non-nodule with 92% sensitivity and 86% specificity, but is less able to distinguish between benign and malignant nodules. The model is also able to determine the general nodule regions but is unable to determine exact nodule locations.*

## 1. Introduction

Lung cancer is the second-most common type of cancer in both men and women and is the leading cause of cancer death in the United States [1]. The best chances of survival come from early detection and treatment, which could potentially be aided by improved automated malignant nodule detection methods.

A lung nodule is a small, round growth of tissue within the chest cavity. Nodules are generally considered to be less than 30mm in size, as larger growths are called masses and are presumed to be malignant. Nodules between 5-30mm may be benign or malignant, with the likelihood of malignancy increasing with size. Smooth nodules with signs of calcification are more likely benign while lobulated or spiculated nodule edges may indicate malignancy [3].

There are two main chest imaging techniques, basic X-ray imaging and computed tomography (CT). Chest X-ray images, or radiographs, provide a single view of the chest cavity. Posteroanterior views, in which the X-ray beam travels through the patient's chest from back to front, are most common. CT scans are 3-dimensional images produced using X-ray images taken from many orientations using a rotational scanner. CT scans can provide a more complete view of the chest internals and can thus be used to more easily detect shape, size, location, and density of lung nodules. However, CT scan technology is expensive and is often not available in smaller hospitals or rural areas. By contrast, basic chest radiographs are relatively cheap and fast, and expose the patient to little radiation, so they are usually the first diagnostic step for detecting any chest abnormalities.

Pulmonary radiologists typically detect lung nodules in radiographs by considering the shape and brightness of circular objects within the ribcage [10]. Studies have found that only 68% of retrospectively detected lung cancer nodules in radiographs were originally detected by a single reader and only 82% with two readers [5] [15]. Computer-aided detection (CAD) techniques have been explored to make the identification of lung nodules quicker and more accurate. Nodule detection algorithms have been designed using traditional image processing techniques to identify regions of the chest radiograph that potentially contain a bright object of the expected size, shape, and texture of a lung nodule.

With recent improvements in convolutional neural networks (CNNs), some researches have looked at using these models to classify lung nodules. Unfortunately, as is often the case with medical imaging, the available datasets are relatively small. Since deep networks depend on lots of data to learn, training a complex neural network from scratch on lung nodule images may not prove very successful. However, transfer learning, or training a network on a large dataset and then using these trained weights for new tasks on new datasets, has been shown to work well for a wide range of image datasets and tasks [11].

In this paper, we examine using CNNs with transfer learning for nodule classification and localization. The inputs to the model are full posteroanterior chest radiographs that may or may not contain nodules. The outputs from the classification task are probability scores for each radiograph containing a benign nodule, a malignant nodule, or no nod-

ule. For images that contain a nodule, the CNN is also used in a localization task, where the outputs are four box coordinates indicating the predicted nodule location and expanse.

## 2. Related Work

Lung nodule detection is still primarily done by hand by trained pulmonary radiologists. Existing CAD systems are designed to aid radiologists in this task by performing an initial highly-sensitive nodule detection pass and alerting radiologists to potential nodules. The high sensitivity of the CAD systems means that they also detect many false-positives, which the radiologists are expected to then weed out. However studies have indicated that radiologists have a difficult time effectively differentiating true nodules from false positives, so CAD systems that can reduce the number of false positives would be desirable. Current CAD systems also do not find all nodules (100% sensitivity), so radiologists are still expected to scan the entire radiograph to look for other nodules [17].

A CAD system works by following a sequence of defined steps. The system initially segments a radiograph into anatomic regions (left and right lung, heart, clavicles) using segmentation algorithms such as shape models and pixel classification. It then detects potential candidate nodule regions using filtering and thresholding. Features are extracted from these candidate regions and trained on a classifier, which returns a degree of suspicion for the region. Highly-suspicious regions are then shown to the radiologist. In these CAD systems, the detection of potential candidate nodule regions and the feature extraction steps are hand-designed for this particular task. Thus the development of these CAD systems is time-consuming and does not translate to new medical diagnostic tasks [17].

Determining the effectiveness of these CAD systems is difficult as they are used in conjunction with human readers and thus results are affected by human error. Results from studies of reader nodule detection with and without a CAD system vary from 49% to 65% sensitivity without a CAD system to 68% to 93% sensitivity with the CAD aid, but all show some increase with the use of the CAD system [16] [8] [14].

The use of convolutional neural networks for nodule detection or classification in radiographs is much less prevalent. Two studies from 1995 and 1996 used small CNNs (with two or three layers) to classify candidate regions as nodule or non-nodule [10] [9]. Both studies used pre-scan CAD systems to initially crop out nodule candidate regions from the radiographs.

A couple more recent studies have looked at using deeper CNNs and transfer learning for similar tasks, although not for detecting lung nodules in radiographs. Bar *et al*. explore using a AlexNet trained on the large image dataset ImageNet to detect Right Pleural Effusion and Enlarged heart conditions in radiographs [2]. The researchers relied on this transfer learning technique since their dataset was very small (only 93 images). Features were extracted from many different layers within AlexNet and trained with an SVM classifier.

A recent paper from Bram van Ginneken *et al*. discussed using the pre-trained convolutional neural network OverFeat to classify regions from chest CT scan images as nodule or non-nodule [18]. As with the early radiograph nodule studies, their process involved using a commercially-available CAD system to identify possible nodule locations within the scan. Crops at each potential location were then fed through the neural network to be classified as true nodules or not.

Since convolutional neural networks are able to examine similar graphics in differing locations within an image using sliding filters, in this present study we examine skipping the initial step of using an existing CAD system to crop out potential nodule regions. We explore whether a convolutional neural network model is able to detect nodules within a larger chest radiograph without using specialized CAD systems which are time-consuming to develop and highly task-dependent.

## 3. Methods

For both localization and classification, we examine using transfer learning with a 50-layer residual network (ResNet) model. The ResNet models are deep CNNs that are designed to ease backwards propagation of gradients through the network and thus improve training. The building block of a ResNet is a small stack of convolutional layers in which the input is summed with the output of the layers to create skip connections. In the 50-layer ResNet, the small layer stack between skip connections consists of three convolutional layers (1x1 filter, 3x3 filter, 1x1 filter) and is called a "bottleneck" as the 1x1 filters decrease and then restore dimensions to speed computation [4]. The ResNet used in this study was implemented in Caffe [7] and pre-trained on a subset of the ImageNet dataset, with about 1.3 million labeled color images of common animals and objects in 1000 classes.

Batches of 64 radiographs were fed forward through the ResNet with fixed pre-trained weights, and features were extracted from five different layers. Earlier convolutional network layers generally pick out edges and generic visual features that may be more transferable to a new classification task, while later layers may be more specific to the initial ImageNet classes. On the other hand, deeper networks generally perform better for image classification as they allow for more non-linear relationships between pixels and output classes. So in this paper, we explore extracting features from multiple layers spread throughout the network to determine which features are best for the nodule classifica-

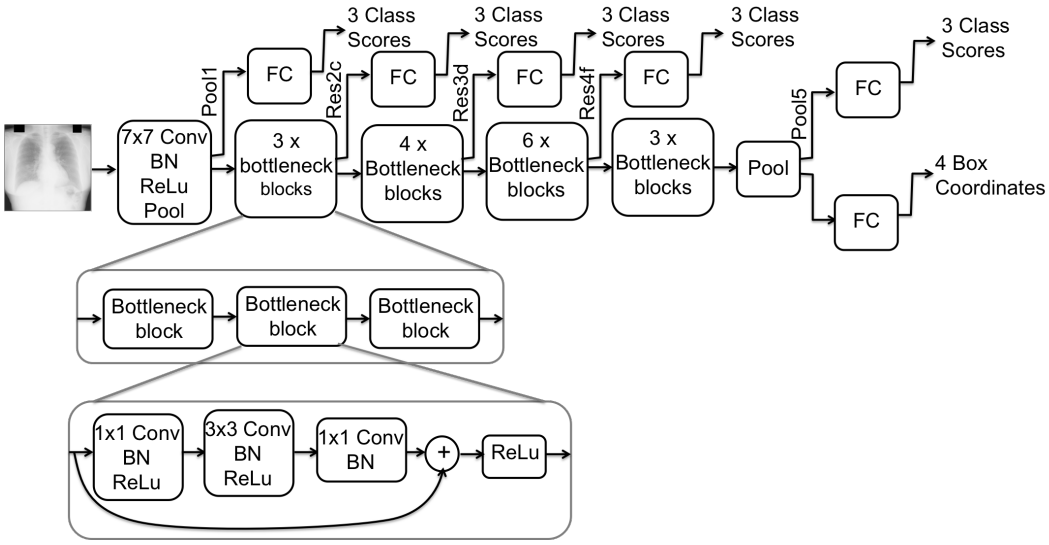


Figure 1. A 50-layer residual network used for both classification and localization of nodules. Features were extracted from 5 layers within the ResNet and passed through a final fully-connected layer to output the three class scores. Features from the final pool5 layer were also passed through a fully-connected layer to output predicted box coordinates of the nodule location.

tion task.

Features were extracted from ResNet layers pool1, res2c, res3d, res4f, and pool5, as shown in figure 1. The pool1 layer is after a single convolutional layer (followed by batchnorm and ReLu layers) and a max pooling layer. The res2c layer is after three ResNet bottleneck blocks, the res3d layer is after seven bottleneck blocks, and the res4f layer is after thirteen bottleneck blocks. Finally, the pool5 layer is after sixteen bottleneck blocks and a final average pooling layer. The pool5 layer is the final layer of the 50-layer ResNet before the fully-connected layer that produces predicted class scores. The extracted features were saved to disk to speed computation. Since the ResNet layers are frozen, there is no need to back-propagate through them during training.

The extracted features were input into a final fully-connected layer with three outputs. This final layer was trained as a Softmax classifier by interpreting the outputs as unnormalized log probabilities of the three classes and minimizing the cross-entropy loss between these labels and the correct image labels (non-nodule, benign nodule, or malignant nodule). The cross-entropy loss for a single image $i$ is given by equation (1), where $s_j$ is the $j^{th}$ component of the output vector and $y_i$ is the correct image label.

$$L_i = -log\left(\frac{e^{s_{y_i}}}{\sum_j e^{s_j}}\right) \tag{1}$$

The total loss for the batch is the sum of the mean of the image losses and an L2 norm regularization term to favor small weights and avoid overfitting to the training data. This full loss is shown in equation (2), where $\lambda$ is a regularization constant and $W$ is the weight matrix.

$$L = \frac{1}{N}\sum_i L_i + \lambda||W||_2^2 \tag{2}$$

The algorithm attempts to find the weights that minimize this loss function. This minimization is accomplished through mini-batch gradient descent with momentum. At each iteration, weights are updated according to the momentum update equation (3), where $\alpha$ is the learning rate, $\mu$ is a momentum parameter, and the gradient of the loss $\nabla_W L$ is calculated over the batch of training data.

$$\begin{aligned} v &\leftarrow \mu v - \alpha\nabla_W L \\ W &\leftarrow W + v \end{aligned} \tag{3}$$

Features extracted from the final pool5 layer for images with a benign or malignant nodule were also fed into a separate regression head for nodule localization. This fully-connected layer had four outputs, which represented the network's prediction for box coordinates around the nodule. The weights for this layer were trained by minimizing the Euclidean distance between the predicted coordinates and the true nodule box coordinates. As before, the total loss function is a sum of the mean of the Euclidean distances for the batch of images and a regularization term, as shown in (4), where $s_j$ is the $j^{th}$ component of the output vector and $b_j$ is the $j^{th}$ component of the true box coordinate vector.

$$L_i = \sqrt{\sum_j (s_j - b_j)^2}$$

$$L = \frac{1}{N} \sum_i L_i + \lambda ||W||_2^2 \quad (4)$$

After all weights were trained, a final end-to-end model was created using weights from both the original ResNet and the final fully-connected layer. This model was used to create saliency maps, as was done in [13]. Saliency maps were computed by passing an image forward through the model, setting the output gradient for the predicted class to one (zeroing all others), and then back-propagating the gradient to the image. Then each pixel in the saliency map is the max among the three color channels of the normalized absolute value of the image gradients. Brighter pixels in the saliency map indicate regions that have a larger impact on the final classification.

## 4. Dataset

The dataset is from the Japanese Society of Radiological Technology (JSRT) [6]. This dataset of posteroanterior chest radiographs includes 93 non-nodule images, 54 benign nodule images, and 100 malignant nodule images. Each radiograph is a grayscale image of 2048 x 2048 pixels at a resolution of 145 ppi. Chest images are at slightly varying scales, are not always centered in the frame, and many images have black blocks at the top from the X-ray process as in Figure 2. In the images, a nodule may be found in varying locations within the ribcage, including upper, middle, or lower lobe on either left or right side. The nodules range from about 30 to 170 pixels in diameter.

All radiograph classifications as nodule or non-nodule in the dataset were confirmed using CT scans. Nodule classifications of benign or malignant were made not only through their appearance in this dataset and the CT scans, but also through testing of tissue samples and monitoring for nodule changes over time. Nodules that shrunk or disappeared with antibiotics or did not change over a period of two years were considered benign.

In addition to classification labels for each image, the JSRT dataset also includes the size and (x, y)-coordinate of the center of the lung nodule, if present. These values were used to approximate a square bounding box for the nodule.

For this study, a set of 32 radiographs (13% of the data) was reserved as a test set. The remaining images were split into four folds for cross-validation. Hyper-parameters were tuned by training with three folds and testing with the forth, and repeating this process with each fold as the validation set. Once hyper-parameters were chosen, the network was retrained using images from all four folds.

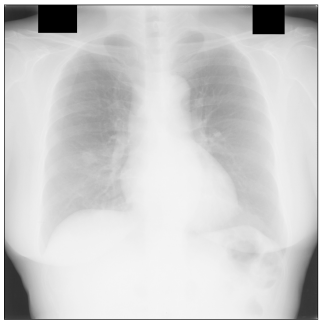

Figure 2. Chest radiograph of a patient with a malignant nodule in the right middle lobe.

Since the dataset is small, data augmentation techniques were used to artificially increase the number of training images as well as to even the class distribution. Since a nodule may be of different sizes and appear at various locations within the radiographs, data was augmented through varying image scales and translational crop locations, as well as horizontal mirroring.

To use the 50-layer ResNet on this radiograph dataset, a few modifications were required. Since the ResNet expects color images of size 224 x 224, the radiographs were scaled randomly to between 256 x 256 and 384 x 384 and then cropped to 224 x 224 pixels. To handle the grayscale radiographs, the pixels were duplicated to the three color channels, and the mean image was subtracted.

## 5. Results and Discussion

We can quantify the model's ability to classify radiographs by observing the classification accuracy on the test set. Accuracy is calculated as the number of correctly classified images (predicted label and true label are the same) divided by the total number of images. With three classes, random guessing would yield an accuracy of 33%, so this is the baseline accuracy to which we may compare our model accuracies.

A plot of the classification test accuracies over training epochs can be seen in figure 3. These plots were produced by running the test-set through the network periodically while doing the final model training using training data from all four folds. The learning rates found during cross-validation and used for this final training ranged from 0.001 to 0.01 for features extracted from different layers, with the regularization parameter ranging from 0.05 to 0.2. For all layers, the learning-rate was halved every 12 epochs, and the momentum parameter was maintained at 0.9.

In figure 3, we see that after 30 epochs, models trained using features extracted from all five layers perform better than the baseline 33% accuracy. We can also observe that higher accuracies were achieved using features extracted

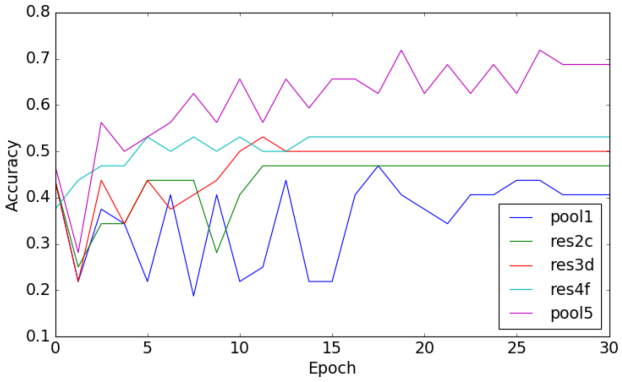


Figure 3. Classification test accuracies over training epochs using features extracted from five different ResNet layers: pool1, res2c, res3d, res4f, and pool5. Higher accuracies are seen for features extracted later in the network, with pool5-feature accuracies reaching 68%.

later in the network. Using features extracted from the pool5 layer, which is equivalent to using the full 50-layer ResNet with the last fully-connected layer modified to output only three class scores, we can observe test accuracies as high as 68%.

Higher accuracies from features extracted later in the ResNet is consistent with recent findings that deeper CNNs perform better than shallower ones. Deeper networks are generally more difficult to train, but by using transfer learning, we do not need to train the many ResNet layers. The better performance using layers later in the network indicates that even higher-level visual features learned on ImageNet transfer well to this new radiograph classification task.

To better understand which image classes the network is accurately predicting, we can look at the precision and recall for each class. Precision values for each class indicate the fraction of images that were predicted to be of that class that were indeed in the class. Recall values indicate the fraction of images from each class that were correctly predicted to be in the class. Using features from pool5, precision and recall values for the three classes are shown in table 1. The model appears to be best at identifying non-nodule radiographs, but has more difficulty classifying benign and malignant nodule radiographs.

In the confusion matrix in figure 4, we can visualize the fraction of radiographs from each class in the test set that were assigned each of the predicted labels. Here we again see that the model is able to classify non-nodule images quite well, but has more trouble distinguishing between benign and malignant nodules, as we saw with the relatively low precision and recall values for these classes.

It is likely that the network can often detect the presence or absence of the bright round nodule-like objects, but has

Table 1. Three class precision and recall

| Class | Precision | Recall |
|---|---|---|
| Non-nodule | 0.75 | 0.86 |
| Benign | 0.6 | 0.55 |
| Malignant | 0.57 | 0.57 |

Table 2. Two class precision and recall

| Class | Precision | Recall |
|---|---|---|
| Non-nodule | 0.75 | 0.86 |
| Nodule | 0.96 | 0.92 |

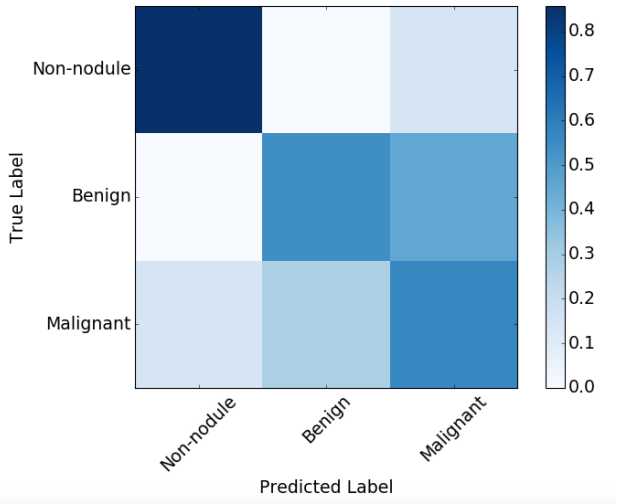


Figure 4. Confusion matrix for radiograph classification.

more difficulty detecting the nuanced differences between benign and malignant nodules. This is not very surprising since the nodules in these images may be as small as three pixels, so detecting difference in the shape and texture of the nodule will be a very difficult task for the model.

This task of distinguishing benign and malignant nodules in radiographs is also difficult for pulmonary radiologists, and if a person is suspected to have any type of nodule in their radiograph, they are generally sent for further testing using a CT scan and/or examination of tissue samples. Since the true class labels of benign and malignant for these radiographs were made based on external information such as further imaging results and prolonged monitoring, it is not clear whether there is even enough visual information within the radiograph to make the distinction.

If we focus on simply detecting nodules, we can combine benign and malignant nodule images into a single class. Doing so, we get precision and recall above 92% for nodule images, as shown in table 2. For medical diagnostic, it is often desirable to have higher recall or sensitivity for the positive or diseased class (nodule) to the potential detriment of specificity, which measures the recall of the negative class (non-nodule), so as not to miss giving further testing and

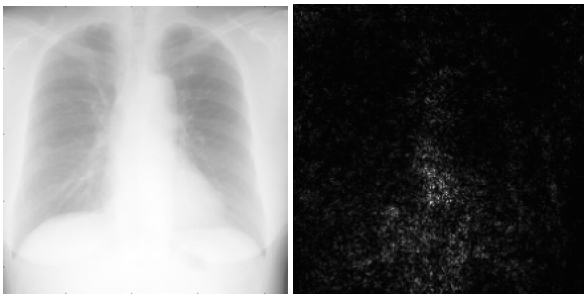

Figure 5. Saliency map for an image with a nodule.

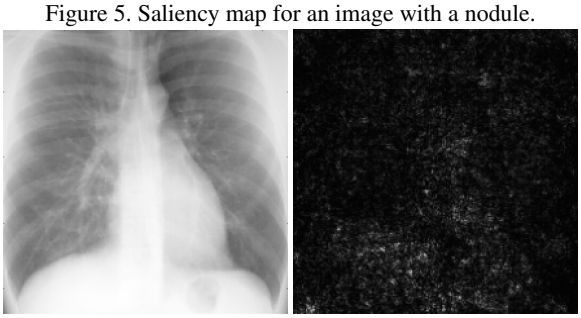

Figure 6. Saliency map for a non-nodule image.

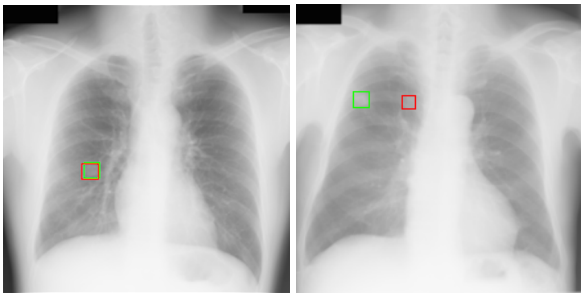

Figure 7. True and predicted nodule bounding boxes on a training set image with an overfit model (left) and a test set image for a model with high regularization (right). The green box is the true bounding box, red is the predicted bounding box.

treatment to a patient who needs it. In this case, we have nodule sensitivity or true positive rate of 92% with a false positive rate (1 - specificity) of 14%.

We can also examine the saliency maps to get a better sense for which parts of the radiographs contribute most to the model's predicted classes. Nodule and non-nodule radiographs, as well as their associated saliency maps, may be seen in figures 5 and 6, respectively. Saliency maps for non-nodule radiographs tend to have a brightness spread throughout the maps, which makes sense since there is no unique visual attribute that is common to all non-nodule images (the class is rather defined by the absence of an attribute). Saliency maps for nodule images have more bright patches of pixels towards the center of the image. It appears that the model has learned to ignore the pixels towards the outer edge of the radiograph and focus closer to the spine. However, the saliency maps are not entirely interpretable and it does not appear that the model can learn to focus exactly on the nodule locations.

From the localization task, we can analyze the model's ability to learn nodule locations when explicitly provided the true locations during training. When tuning hyperparameters for the final fully-connected layer in the regression head, it was very easy for the model to overfit to the training set. The Euclidean loss for both test and train data drops greatly at first, but then the test loss begins to increase while the training loss drops to near zero. Looking at training images from this overfit model such as in figure 7 (left), we see that the predicted bounding boxes have a near-perfect overlap with the true bounding boxes, however for test images the boxes were relatively far apart. By increasing the regularization parameter to 0.5, we can observe both train and test images such as that in figure 7 (right). Although there is rarely intersection between boxes, they are often in the correct general region of the chest.

To quantify the error between true and predicted bounding boxes, we can observe the mean distance between box centers and their mean size difference, as in figures 8 and 9. By the end of training, the sizes differ by less than four pixels, which is about 6mm or 30% of the size of the average nodule in the dataset. For the test set, mean distances differ by about 60 pixels. Although this is certainly far from overlapping boxes, it is also much better than random. Choosing two box centers uniformly at random from the 224 x 224 image, we would expect a mean distance of 117 pixels, or nearly twice as great as that seen in the test set [12]. This again indicates that the model is able to determine the general region of the nodule but is unable to ascertain its exact location.

The difficulty in localizing the nodules and tendency for the model to overfit is likely due to the small nodules sizes and the limited number of training images. It is a fairly large request to provide images and four associated numbers to a model and expect it to determine that these numbers localize a portion of the image that is approximately 10 x 10 pixels. It is an even greater request to do so using augmentations from about 130 nodule training images. With so few images, it is easy for the model to fit noise in order to decrease the distances between bounding boxes. With increased regularization, there is less overfitting, but the model is unable to determine generalized aspects of the nodule images to decrease error enough for overlapping bounding boxes. The bright centers in the saliency maps and bounding boxes in the correct general region indicate the network has begun to generalize to nodule locations in unseen data, but more training images would be needed for better performance.

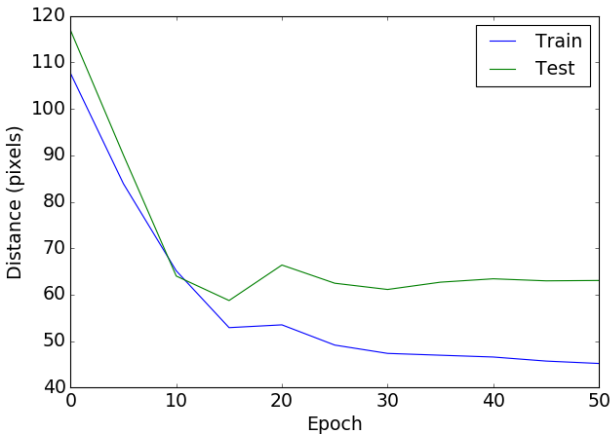


Figure 8. Mean pixel distance between true and predicted nodule boxes for both train and test sets over training epochs.

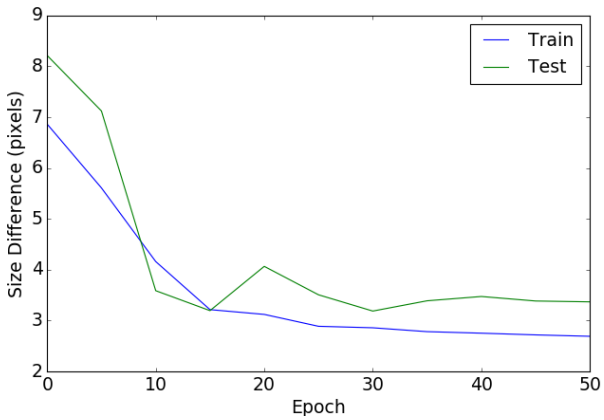


Figure 9. Mean pixel difference between true and predicted nodule sizes for both train and test sets over training epochs.

## 6. Conclusion

We have seen that a CNN model with transfer learning does fairly well at classifying chest radiographs as having a nodule or not, but is less adept at determining malignancy or localizing nodules.

With only modifications to the final network layer, a 50-layer ResNet model trained to classify color images of animals and everyday objects was able to also classify black-and-white medical images with 68% accuracy. Models constructed using features from earlier network layers had accuracies above random, but did not perform as well as features from the final pooling layer.

Although the best model was able distinguish nodule radiographs with 92% sensitivity and 86% specificity, the model had more difficulty classifying the difference between benign and malignant nodules. If this model were used in practice, further imaging and testing should be recommended for any patient that receives a nodule classification, whether benign or malignant. This is indeed generally the practice with nodules detected by radiologists today, and a 92% sensitivity is on par with the best results seen in studies of sensitivities of radiologists using CAD systems. Although it should be noted that it is difficult to compare results given the small test set size in this study and varied results from radiologist studies.

We have also seen that the ResNet model is able to localize the general region of a nodule but is unable to determine its precise location. This result is not surprising given the small nodule size and limited training images.

The greatest future model improvements would likely come from training with more chest radiographs. More training images would help prevent overfitting and allow the model to generalize better to unseen radiographs. With a larger dataset, we could also try fine-tuning the weights for a few of the ResNet convolutional layers before the final fully-connected layer. This would allow the model to adapt more to the new dataset, which may have some different high-level visual features than the original ImageNet dataset.

Future work could also involve cropping 224 x 224 sections of the original 2048 x 2048 radiographs around the nodules and attempting to localize the nodules within these cropped images and classify them as benign or malignant. This would improve the resolution of the nodules as seen by the network, increasing their mean pixel size from 10 to 90 pixels in diameter. Although this would not allow for classification and localization in full radiographs as the nodule locations would need to be known in advance, it would help determine whether there are identifiable visual features in the radiographs to distinguish benign and malignant nodules and provide an upper bound for how well the model can perform.

Additionally, since the current model is able to localize the general nodule regions, it is possible that the original 2048 x 2048 images could be cropped based on the results of a first localization pass and then the nodules in the cropped regions could be classified and localized using a second pass through the network with the higher resolution images.

## References


[1] American Cancer Society. Key statistics for lung cancer, 2016.

[2] Y. Bar, I. Diamant, L. Wolf, and H. Greenspan. Deep learning with non-medical training used for chest pathology identification. *SPIE*, 9414, 2015.

[3] W. Brant and C. Helms. *Fundamentals of Diagnostic Radiology*. Lippincott Williams and Wilkins, 4 edition, 2012.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv:1512.03385*, 2015.

[5] R. T. Heelan, B. J. Flechinger, and M. R. Melamed. Non small cell lung cancer: Results of the new york screening program lung cancer: Results of the new york screening program. *Radiology*, 151:289–293, 1984.

[6] S. J, K. S, I. J, M. T, K. T, K. K, M. M, F. H, K. Y, and D. K. Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *AJR*, 174:71–74, 2000.

[7] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv:1408.5093*, 2014.

[8] S. Kasai, F. Li, J. Shiraishi, and K. Dois. Usefulness of computer-aided diagnosis schemes for vertebral fractures and lung nodules on chest radiographs. *Academic Radiology*, 15:571–575, 2008.

[9] J.-S. Lin, S.-C. Lo, A. Hasegawa, M. Freedman, and S. Mun. Reduction of false positives in lung nodule detection using a two-level neural classification. *IEEE Transactions on Medical Imaging*, 15(2):206–217, 1996.

[10] S.-C. Lo, S.-L. Lou, J.-S. Lin, M. Freedman, M. Chien, and S. Mun. Artificial convolution neural network techniques and applications for lung nodule detection. *IEEE Transactions on Medical Imaging*, 14(4), 1995.

[11] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. *arXiv:1403.6382*, 2014.

[12] L. Santalo. *Integral Geometry and Geometric Probability*. Addison-Wesley Publishing Co., 1976.

[13] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv:1312.6034*, 2014.

[14] W. Song, L. Fan, Y. Xie, J. Qian, and Z. Jin. A study of interobserver variations of pulmonary nodule marking and characterizing on dr images. *Proc SPIE*, 5749:272–280, 2005.

[15] F. P. Stitik, M. S. Tockman, and N. F. Khouri. Screening for cancer. *Chest Radiology*, pages 163–191, 1985.

[16] E. van Beek, B. Mullan, and B. Thompson. Evaluation of a real-time interactive pulmonary nodule analysis system on chest digital radiographic studies: a prospective study. *Academic Radiology*, 15(571-575), 2008.

[17] B. van Genneken, C. Schaefer-Prokop, and M. Prokop. Compter-aided diagnosis: How to move from the laboratory to the clinic. *Radiology*, 261(3), 2011.

[18] B. van Ginneken, A. Setio, C. Jacobs, and F. Ciompi. Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. *Biomedical Imaging (ISBI)*, pages 286–289, 2015.