Weakly Supervised Pneumonia Localization

Shih-Cheng (Mars) Huang [1] **Medi Monam** [2] **Emanuel Cortes** [3]

.

Abstract

We have developed a weakly supervised method for localizing Pneumonia on chest X-rays. Our model includes two parts 1) a 10-layer Convolutional Neural Network (CNN) that predicts the presence of Pneumonia and 2) a Class Activation Map (CAM) that localizes the Pneumonia manifestation without requiring bounding box labels. By having our weakly supervised approach achieve slightly better performance than a supervised method (R-CNN), we believe that this brings tremendous value in labeling diseases in images that are often unannotated in medical records. Thus, our method has the potential to provide care to populations with inadequate access to imaging diagnostic specialists, while automate other medical image data-sets.

1. Introduction

Pneumonia is an inflammatory condition of the lung that is responsible for more than 1 million hospitalizations and 50,000 deaths every year in the United States. Globally, Pneumonia is responsible for over 15% of all deaths of children under the age of 5(4). Currently, chest X-rays are the best available method for diagnosing pneumonia (8), playing a crucial role in clinical care (3) and epidemiological studies (2).

However, Chest X-Ray images generated from hospitals do not specify the precise location of the Pneumonia, which is a significant challenge when training a machine learning model for this purpose. At most, doctors will keep a brief description, such as ”aggregation of lung opacity on the patient’s lower right lung”. This is because the precise pixel location of lung opacity on the X-ray image is only part of the equation for diagnosing, and only the final conclusion is recorded in the Electrical Health Record (EHR). To developed a machine learning algorithm that predicts Pneumonia location using traditional supervised methods requires the precise x,y coordinate-labels that datasets lack. Hence, this deficiency in labelled datasets commonly observed in the medical imaging field is the motivation behind our work.

In this work, we tackle this challenge in a novel approach: we use a weakly supervised approach to automate localizing Pneumonia in chest X-rays. Our model is considered ”weakly” supervised because it only needs the binary labels (Pneumonia vs. No Pneumonia) during training to estimate a bounding box around the region of the lung opacity. At a high-level, our ”weakly” supervised algorithm works as follows: 1) input an X-ray image in U-Net architecture for data augmentation, 2) input augmented image and original image in a 10-layer CNN architecture to classify if given image is Pneumonia positive, and 3) if image is Pneumonia positive, apply CAM to precisely localize the Pneumonia aggregation.

2. Related Work

There have been recent efforts in detecting Pneumonia using X-ray images. For instance, the CheXNet(10) team modified ChestX-ray14’s (13) algorithm to increase the accuracy in detecting 14 diseases, including Pneumonia. However, neither effort localizes using bounding boxes, and both use ImageNet to pretrain their models. Despite the fact that both works achieve high accuracy, neither solves the problem of clearly annotating the Pneumonia manifestation using bounding boxes in the X-ray images. As such, we leverage the work of four algorithms in our approach: 1) R-CNN(6), 2) CAM(14), 3) VGG architecture(12), and 4) U-Net . Each of these algorithms and approaches was implemented in a different part of the project. For instance, we used a VGG model for the supervised portion to make accurate classification of Pneumonia images. However, we had to modify the VGG architecture to optimize it for our data-set, while making the model compatible with the Class Acticatiom Map. The CAM paper gave us the inspiration to extract regions of the input image contributed to the prediction of Pneumonia without training labels. Furthermore, we bench-marked our CAM output results by comparing it to that of a supervised R-CNN model. Lastly, the U-Net architecture allowed us to segment the lung portion of our input image and provide extra features to our model.

[0]

3. Dataset and Feature Engineering

3.0.1. Dataset

We acquired our dataset from Kaggles RSNA Pneumonia Detection Competition (13). The dataset consists of 28,989 X-ray images (8964 with pneumonia, 8525 Healthy, 11,500 not healthy/ no Pneumonia). A diseased/no Pneumonia label is for any diseased lung that has no Pneumonia but the manifestation of any of other disease, such as fluid overload (pulmonary edema), bleeding, volume loss (atelectasis or collapse), lung cancer, or post-radiation/surgical changes. For the purpose of demonstrating the feasibility of our model without complicating the training and evaluation, we removed the diseased/no Pneumonia labelled images from our dataset. This was valuable in balancing our dataset: 51.25% Pneumonia and 48.74% Healthy. Further, we segmented our data into a 70/20/10 train, validation, and test split.

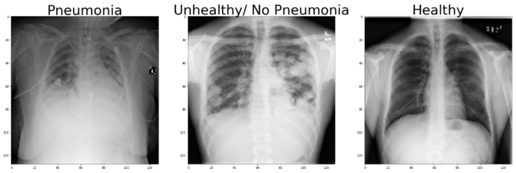

Figure 1. (left) Lung with Pneumonia, (mid) Unhealthy/Diseased Lung (no Pneumonia), (right) Healthy lung

Figure 1 contains examples of the three types of labels that are found in our dataset (Pneumonia, Diseased/No Pneumonia, Healthy). Both the Pneumonia and Diseased/No Pneumonia images have a opaque areas in the lung that make the lung more cloudy than the healthy lung. The main difference between Pneumonia and a Diseased/No Pneumonia lung is the shape of these opaque areas. Pneumonia's exact region tends to be hard to define, while diseased lungs, generally, have clearly defined opacities.

Images with Pneumonia labels are also associated with ground truth bounding box of Pneumonia regions. On average, the bounding box area is approximately 50,000 pixels, with an average dimension of 300w by 400h pixels. These bounding boxes were recorded as the X,Y coordinates of the lower left corner of ground truth bounding boxes of regions with Pneumonia, along with the widths and heights of these boxes.

3.0.2. Feature Engineering

Each pixel of the images was normalized by subtracting the mean and dividing by the standard deviation at that location. We also compressed the original image from 1024*1024 pixels down to 128*128 pixels, allowing us to expedite the training of our neural network. A U-Net neural network was used to predict the confidence of each pixel belonging to the lung. Then we segmented the lung by multiplying the original image matrix with the localization matrix (figure 2). Both the original and the segmented images were fed into our model as inputs to provide our model with a hypothesis of lung location.

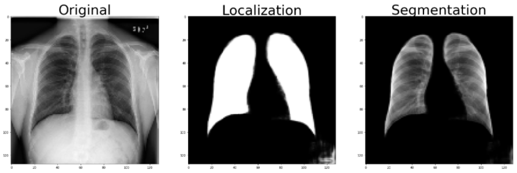

Figure 2. (left) Original input image, (mid) predicted lung location, (right) segmented lung

4. Methods

4.0.1. Classification

The first part of our work is to build a CNN model that can accurately classify whether a given image is labeled as Pneumonia or not. The images that were predicted as Pneumonia positive are then fed into our localization model.

Baseline
Since the first part of this project is a supervised classification task, SVM, Random Forest and Logistic Regression were used to baseline our classification model. These models could not take a matrix of pixels as an input, so we flatten the images into one dimensional vectors.

CNN architecture
Our best model contains 10 convolutional layers, each with zero padding to keep the size of the original image, and we used ReLU as the activation function. The convolution filters are matrix of weights that slides through the original image to pick up patterns for prediction. The CAM requires our model to only have one Fully Connected (FC) layer and a Global Average Pool (GAP) layer before that. The GAP layer takes the average of the output for each of the convolution filters. The FC layer connects the flattened averages to the two classes. The model was trained with an Adam optimizer with 0.0001 learning rate on 20 epochs.

4.0.2. Localization

The second part of our project is to build a weakly supervised model that can predict the localization of Pneumonia on the positively classified images, without the training labels of the locations.

Supervised R-CNN Approach (Benchmark)
To generate region proposals, we slide a small network over

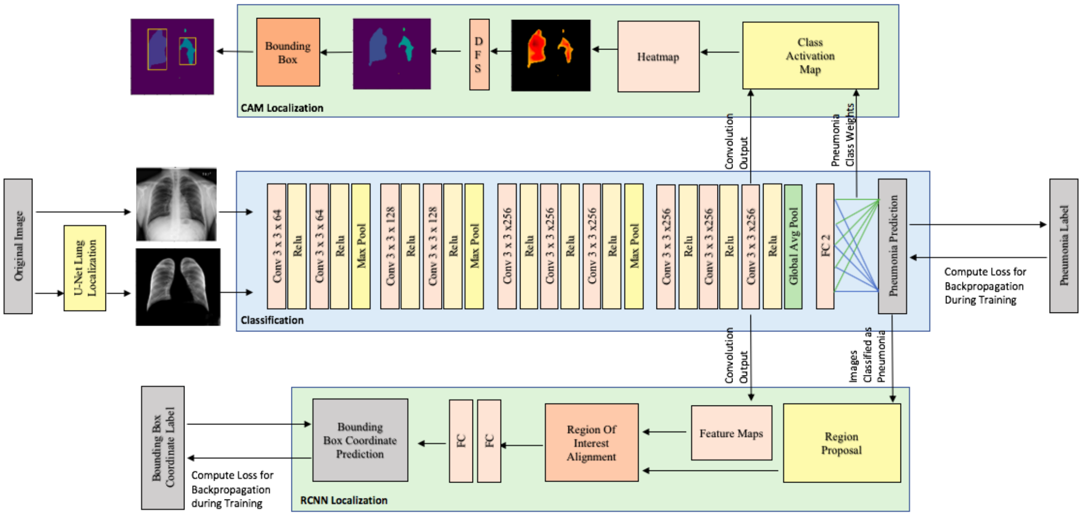

Figure 3. a) classification model b) R-CNN localization baseline c) CAM localization model

the convolutional feature map output by the last shared convolutional layer. This layer comes from the classifier that is trained on predicting Pneumonia. This small network takes in a small spatial window from the CNN feature map and predicts whether or not these windows contain pneumonia or not Pneumonia. A window is defined as having four coordinates: x1, y1, x2, y2. We only keep the windows that are classified as having pneumonia and by how much these spatial windows overlap with the ground truth labels. Then for each spatial window, the features from the original CNN feature map is mapped back to the CNN feature map from the classifier and these windows are pooled to the same size and are feed to two networks, one network to predict class (background or pneumonia) and another network to predict the coordinates (figure 3 b).

Weakly Supervised Approach

Our weakly supervised portion of the model consists of the following components (figure 3C):

1. CAM
   A CAM that takes in the output of the final CNN model and the FC layer weights for the Pneumonia class neuron and sums up the weighted outputs using the following formula:

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

   Where x is the input image features, f_k give the output from the last convolution layer given x, and w_k^c is the fully connect weight for the k^{th} filter output to class c. In our case, class c is the Pneumonia class.

2. Heatmap
   The output from CAM was then scaled into a 3-channel RGB heatmap.

3. DFS Clustering
   To find individual clusters of predictions on the heatmap, we applied a Depth First Search Clustering algorithm (Algorithm 1) on a random non-zero pixel on the heatmap, and repeated until all non-zero pixels are clustered.

Algorithm 1 DFS Cluster Algorithm

```
class index = 0
while still exist non-zero pixel without class label do
    pick a random non-zero pixel without class label, assign
    pixel to class index
    for each neighbor pixel do
        if if neighbor is also a non-zero pixel without class
        then
            recursively apply DFS    end
        end
    end
    class index += 1
```

4. Bounding box
   Lastly, we drew a bounding box around each clusters by finding the minimum and maximum X,Y coordinates of the clusters, and only kept boxes that are within 2 standard deviations of all predictions.

5. Experimentation

Input features

Running any localizing model on the original X-rays yielded low IoU scores. We discovered that often time the model will localize matter denser than the lung, such as muscle tissue and equipment, as Pneumonia positive (figure 4 right). And since a high IoU score correspond to a more accurate and tightly fitted localization, we experimented with a number of approaches to increase our IoU score.

We started by only feeding in the segmented lungs from U-Net as the input for our model. Though initially the classification results were promising, the IoU score did not improve significantly. We soon discovered that in instances of sever Pneumonia infection, where the density of that part of the lung and surrounding tissue were almost identical, the U-Net algorithm segmented out that part of the lung. This, in turn, yielded inaccurate localization results, where the algorithm localizes on the healthy part of the lung (figure 4 left, mid)

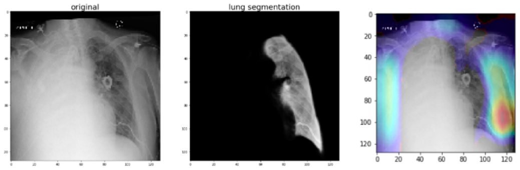

Figure 4. (Left) Original image, (mid) segmented lung (right) Class activation heatmap when we feed in only original image

After testing out different combination, we found the best results can be achieved when we use both the original and segmented image simultaneously by running them through two channels of the network. We hypothesize that including the segmented healthy part of the lung provides the model with extra information on where the likely locations of the lung opacity.

Model architecture and parameter tuning
We started our classification model using the VGG16 model, which includes 26 layers and 138,357,544 weights. It was clear that the large number of weights caused our model to quickly overfit, with training accuracy as high as 96% but 72% for validation. We then modified the architecture by removing layer by layer until bias of the model dropped but still were able to make above 90% accuracy on the validation set.

Then, we tested different filter sizes of the convolutional layers to improve the classification. Our highest validation accuracy was achieved starting with small 3x3 filters and gradually increase the filter size to 16x16 at the last layer. However, a big filter size at the last convolution layer gave us imprecise prediction of the the bounding boxes, and caused a sharp decrease in the IoU score. Therefore, we decided to sacrifice some classification accuracy for an increased IoU score, by keeping all filters to 3x3.

Finally, different optimizers were tested to train our model. The standard Stochastic Gradient Descent Algorithm trains very slowly and does not converge to above 85% accuracy. Adam and Adagrad converge faster, but Adam achieved a higher validation accuracy. We also tested out different learning rates (0.001, 0001, 0.0001) for each optimizer. Learning rate of 0.001 caused the weights to blow up and achieve lower than 50% accuracy. However, a learning rate of 0.0001 with Adam optimizer gave us our higher accuracy in the shortest period of time. (7)

Localization For the clustering portion of our work, we experimented using K-mean and EM with mixture of Gaussian to cluster the pixels. For K-means, we ran it with the possible number of bounding boxes observed in the dataset as the number of clusters (k), and choose the k with the highest silhouette score. However, we realized that silhouette score cannot be calculated with one cluster. Also, we seem to get higher silhouette score as the number of clusters increase. We did consider using EM to cluster the heatmap pixels, however that requires us to know the number of clusters beforehand.

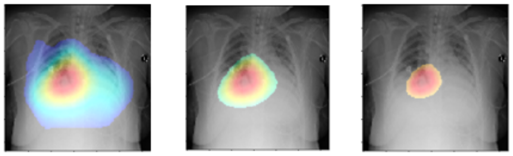

Figure 5. Heatmap with threshold cutoff value of (left) 0.2 (mid) 0.45 (right) 0.7

6. Results

6.0.1. Classification

Classifier	LR	SVM	RF	Our Model
Training	75.86%	74.17%	86.39%	93.07%
Test	73.02%	58.18%	83.00%	92.47%

Table 1: Classification accuracy for our model & baselines

Since we have a balanced dataset, we used accuracy as a metric to evaluate the performance of our classifier as compared to the baselines (Table 1). As it is important to correctly label as many Pneumonia positive images as possible to draw a the bounding box on, a confusion matrix was also generated for our model to evaluate the true positive rate and sensitivity of our model (Table 2).

Confusion Matrix	Predicted True	Predicted False
Actual True	2230	210
Actual False	110	1823

Table 1: Confusion matrix for our model

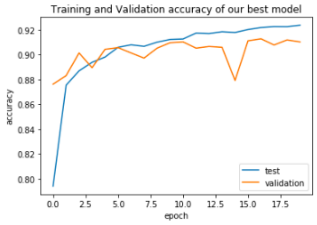

Figure 6. Train and validation accuracy

6.0.2. LOCALIZATION

Localization	R-CNN	CNN+AM
Training	0.1859	N/A
Test	0.1266	0.1508

To evaluate the localization task, we used the IoU metric (Formula 2) by calculating the intersection over union of the prediction and ground truth bounding boxes. Our best weakly supervised model achieved an IoU of 0.1508.

$$IoU(A,B) = \frac{A \cap B}{A \cup B}$$

Formula 2: The IoU formula

7. Discussion

Our CNN significantly outperformed traditional classifiers without over-fitting (Table 1), with almost a 10% increase of accuracy as compared to the best baseline model. We acknowledge that CNNs are designed for images, and the flattened images that the baseline models took in as an input might lose some of the local pattern information. Since the flattened images has more features than training data (128*128), it is hard for the traditional supervised learning algorithm to learn.

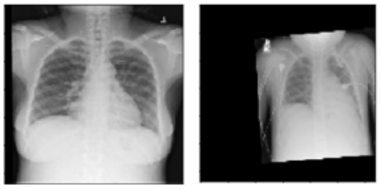

Figure 7. Examples of False Negative Predictions

Analyzing false-negatives images gave us an insight on how we can improve on our classifier. For instance, even though the left image in figure 7 is clear, it is labeled as Pneumonia positive. The right image did not fit into the frame and is slightly rotated, also causing misclassification. Going forward, our model can be improved by introducing random image augmentations such as rotation and zoom. Full resolution images should also be experimented if computing power permits.

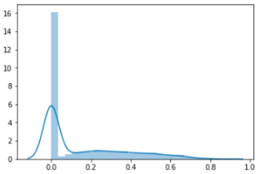

Figure 8. IoU distribution (in 100s)

With regards to localization, our model localizes Pneumonia with higher IoU than the supervised approach, with an increase of 0.0242 (Table 1B). This is significant as we do not need to train a localization model nor location labelled training data. Though this is still far from Human level labelling we see a great potential in our approach.

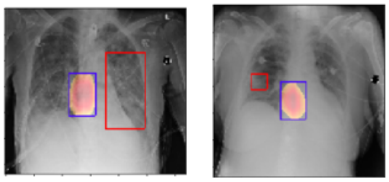

Figure 9. Examples of prediction with no overlapping

Finally, by analyzing the predictions, we deduced a few ways to improve our models. First, when our CNN classified a lung to be Pneumonia negative, our algorithm does not draw a bounding box. Each example with no bounding boxes will receive an IoU score of 0 (figure 8), which significantly lowers the average of IoU score. If all the images were correctly labeled and fed into CAM, our IoU score increases to 0.379. Figure 9 shows us that our network also tends to classify the human spine as part of the lung opacity. Second, our model also struggles to localize very small bounding boxes. We can possibly improve this by using an even smaller filter size for our CNN. Lastly, the heatmap cutoff should be dynamic, as different images might have different severity of lung opacity or pixel contrast.

8. Conclusion

Based on our result, we have shown that our weakly supervised method is able to localize Pneumonia slightly better than a supervised method. We predict that our model can perform even better if we have the computing power to train on the full images, as a lot of information are lost during compression. We also expect improvements by including more training data or transferring learned models from similar works, such as ChestXNet. If improved to human level performance, our weakly supervised model not only can automate pneumonia location annotation and classification tasks, but also can be used to automate other medical image datasets.

References

[1] K Berbaum, Jr EA Franken, and WL Smith. The effect of comparison films upon resident interpretation of pediatric chest radiographs. *Investigative radiology*, 20(2):124–128, 1985.

[2] Thomas Cherian, E Kim Mulholland, John B Carlin, Harald Ostensen, Ruhul Amin, Margaret de Campo, David Greenberg, Rosanna Lagos, Marilla Lucero, Shabir A Madhi, et al. Standardized interpretation of paediatric chest radiographs for the diagnosis of pneumonia in epidemiological studies. *Bulletin of the World Health Organization*, 83:353–359, 2005.

[3] T Franquet. Imaging of pneumonia: trends and algorithms. *European Respiratory Journal*, 18(1):196–208, 2001.

[4] Richard A Garibaldi. Epidemiology of community-acquired respiratory tract infections in adults: incidence, etiology, and impact. *The American journal of medicine*, 78(6):32–37, 1985.

[5] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):142–158, 2016.

[6] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017.

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[8] World Health Organization et al. Standardization of interpretation of chest radiographs for the diagnosis of pneumonia in children. 2001.

[9] EJ Potchen, JW Gard, P Lazar, P Lahaie, and M Andary. Effect of clinical history data on chest film interpretation-direction or distraction. In *Investigative Radiology*, volume 14, pages 404–404. LIPPINCOTT-RAVEN PUBL 227 EAST WASHINGTON SQ, PHILADELPHIA, PA 19106, 1979.

[10] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[11] Suhail Raoof, David Feigin, Arthur Sung, Sabiha Raoof, Lavanya Irugulpati, and Edward C Rosenow III. Interpretation of plain chest roentgenogram. *Chest*, 141(2):545–558, 2012.

[12] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[13] 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 *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 3462–3471. IEEE, 2017.

[14] 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*, pages 2921–2929, 2016.

9. Contribution

Mars Huang - Came up with project idea and methodology. Build the CNN classifier and tested different architectures. Modified the classifier to fit in to Class Actiation Mapping. Implemented CAM, DFS clustering algorithm. Made functions to draw bounding box, calculate IoU and feature engineering. Tried to implement EM and kmeans to cluster heatmap regions. Attempted to reduce dimentions of the data for baseline by using factor analysis. Tested all baselines for classification portion of the project and experimentation in the classification and weakly supervised localization. Created mltoolkit for baselines. Generated all figures, major contributed to the paper and poster. Set up google cloud.

Medi Monam - Lead in reading literature to gather knowledge in the field. UNet segmentation feature engineering. Experimented with methods to cluster Heatmap islands. Experimented with implementation of VGG16. Contributed to poster and paper. Printed poster.

Emanuel Cortes - Built the supervised model for classification and localization. Implemented a resnet backbone classifier, custom region proposal layer, ROI pooling, and a bounding box regressor that is pretrained on the COCO dataset and finetuned on Kaggles RSNA Pneumonia Detection Competition dataset. Experimented with feeding other CNN-based backbone classifier architectures, whose output feature mapes were used as inputs to the custom region proposal layer. Contributed to the poster and paper.