Reports of
The Machine Learning and Inference Laboratory

Towards
Active Learning for User-Defined Chest X-Ray Diagnosis System

Ying Wang
Janusz Wojtusiak

MLI 20-2
October 2020

RESEARCH AND EDUCATION IN MACHINE LEARNING

Abstract:

The current supervised training of deep neural networks on medical image diagnosis requires typically large pools of labeled data, which is scarce, expensive, and limited by types of annotations. There are enormous medical demands that require the detection of subtle differences in the medical images, such as recognizing the severity of pulmonary edema or diagnosing complications of multiple diseases, etc. To enable and minimize the cost of medical images auto diagnosis based on user-defined criteria, we proposed an innovative system of active learning to maximize the usage of annotations effect, precisely connecting the medical images to corresponding patients' history records and provide a variety of resolutions of medical imaged similarities and differences. The system supports distributed training across multiple locations and periods. This paper proposed a new active learning method taking the cost value for each specific misdiagnosis and annotation into consideration. When selecting records for queuing, this application will effectively reduce the overall cost for annotation and reduce the cost for high-risk misdiagnosis. The system also shows the regression between the percentage of annotated images and the AUC results on different diseases.

Introduction:

Chest x-ray images are commonly used in patients for assessment and diagnosis, including the condition of lungs, heart-related lung problems, the size and outline of heart, blood vessels, calcium deposits, fractures, postoperative changes, a pacemaker, defibrillator or catheter, etc.; and certain combinations among them. Current public available labeled Chest X-ray datasets have driven deep learning methods to achieve expert-level performance on diagnosing common related diseases. CheXpert design a labeler to automatically detect the presence of 14 observations in radiology reports[Irvin 2019], including Enlarged Cardiom, Cardiomegaly, Lung Lesion, Lung Opacity, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other, Fracture, Support Devices, and No Finding. CheXNet detects 14 diseases similar to CheXpert outperforms the state of the art database[[Rajpurkar2017].

The current supervised training of deep neural networks requires typically large pools of labeled data. In medical imaging, however, available datasets tend to be limited in size due to privacy issues. Labeled research image data is scarce since manual annotation requires tedious, a time-consuming effort by professional radiologists and physicians. Although some large labeled datasets are available, they can be severely imbalanced by the over-representation of common problems and under-representation of rare conditions. The resolution of differential diagnosis in these diseases are generally low so that the fine detection and classification for diagnosis among images is not feasible. Two largest Chest x-ray images databases currently available, the NIH Chest x-ray database and MIMIC CXR, both labeled with 14 different diagnoses.

However, there is a gap between the current deep learning-based methods and the medical demands that require the detection of subtle differences in the medical images. For example, pulmonary edema is one of the most direct symptoms of Chronic Heart Failure

(CHF) [Mahdyoon1989]. Heart failure patients have extremely heterogeneous responses to treatment [Francis, 2014]. The assessment of pulmonary edema severity will enable clinicians to make better treatment plans based on prior patient responses. It will facilitate clinical research studies that require quantitative phenotyping of the patient status [Chakko 1991]. Quantifying pulmonary edema is extremely important but more challenging than detecting pathologies in chest x-ray images[Rajpurkar2017][Wang2017]. The grading of pulmonary edema severity relies on much more subtle image findings.

To fill the gap between the current deep learning-based methods and the medical demands of detecting subtle differences in the medical images, we proposed a distributed, scalable system enabling a user-defined target for specific specialty areas through active learning.

There has been an increasing amount of application in semi-supervised learning and active learning in the medical and healthcare sectors. In [Smailagic 2018], a novel sampling method was proposed that queries the unlabeled examples that maximize the average distance to all records in training set in a learned feature space. In [Wu2018], a low-rank modeling-based multi-label active learning method for effective medical image annotation was developed. In a recent work by Kim [2020], an active learning method, confident coreset, which considers both uncertainty and distribution for effectively selecting informative samples were proposed. The development in this area has increased the annotation procedure's efficiency and enabled more applications in medical image analysis. There hasn't been much direct application on using active learning in medical image diagnosis due to the computation complexity in semi-supervised learning and active learning in image recognition to reach certain accuracy. However, if we could extract the features from CNN precisely and apply the features to semi-active learning, it is possible to develop a lightweight medical image diagnosis system.

In our proposed system, we leverage the general medical domain knowledge through the labeling of common 14 different diagnoses and existing CNN based supervised deep learning trained model, introduces specified domain knowledge through label propagation, enables user-defined target, maximizes the value of labeling through active learning by introducing the concept of records resolution, and precisely connects the medical images to corresponding patients history records. The purpose of leveraging an existing CNN model is to extract the necessary information from the medical images, and other CNN based models can replace it. The system supports distributed training across multiple locations and time frames and joints the efforts together. It can be used by specialists in different fields to detect subtle differences in the medical images. It also provides an alternative approach for tracking the progression of certain diseases through medical images.

The rest of the paper is organized as follows. We will first provide an overview of the design, followed by a system architecture description. After that, we will discuss the details of the model and its results. Conclusion and future work are shown at the end of this paper.

Design Overview and System Architecture:

An inference modeling for our proposed system can be described in Figure 1. We first need to construct features representation learned through CNN based CheXNet model and include the initial labeling in the database. In [Selvaraju2019], a generalized Class Activation Mapping(CAM), which is called Grad-CAM, was first introduced for ‘visual explanations' for decisions from a large class of Convolutional Neural Network (CNN)-based models. And we have adopted a similar approach to present the features of each sample.

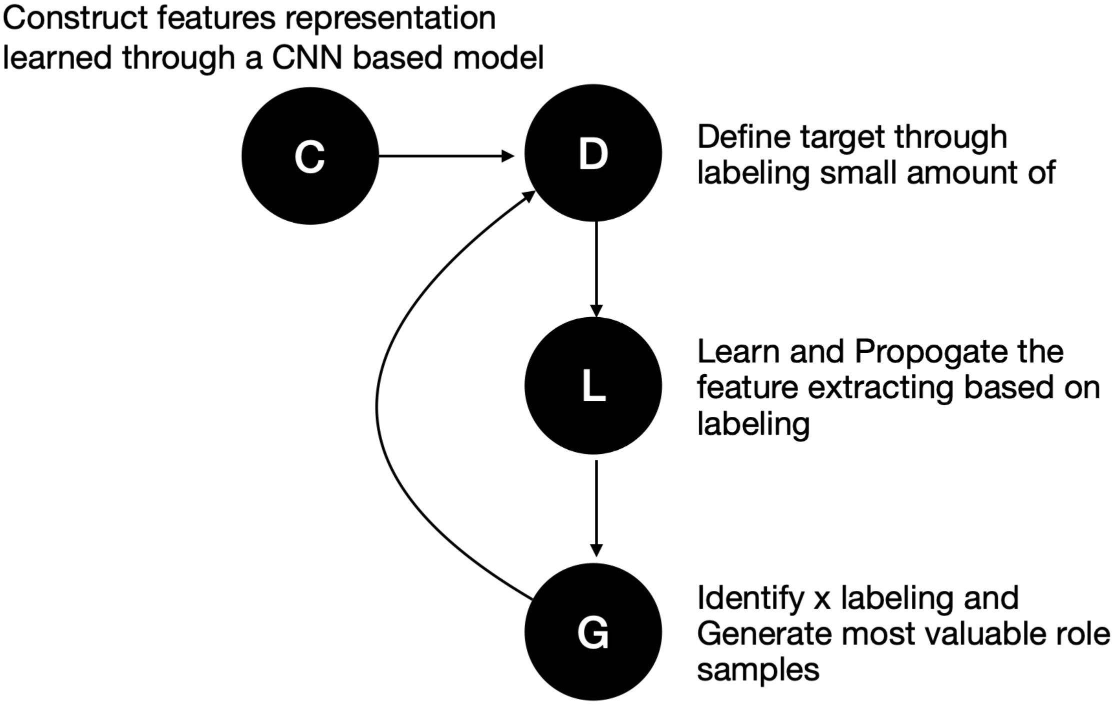

Construct features representation learned through a CNN based model
C
D
Define target through labeling small amount of
L
Learn and Propogate the feature extracting based on labeling
G
Identify x labeling and Generate most valuable role samples

Figure 1: Inference Model

As shown in Figure 1, there are four stages for constructing the proposed system:

1. (C): Preprocessing: extract images information from the cheXnet model or other CNN based models. The purpose of the extraction is to add generic domain knowledge of each image into the proposed model. The information extracted from the existing CNN model, including the last layer of the convolution layer and the previous full connection layer before the softmax.
2. (D): Cross-referencing: as radiologists and physicians labeling the primary seeds of initial labeled data, it is essential to set up a framework that is based on information from the initial labeling to automatically link information $p \in \{p_i\}_{i=1}^{n}$ from patients' history records for further enhancement learning where p_i is the *ith* patient's record and patients are indexed by numbers of 1 to n.
3. (L): Semi-supervised learning: a selected set of $x \in \{x_i\}_{i=1}^{N}$ is labeled to $y = \{y_i\}_{i=1}^{N_l}$ by professional radiologists or physicians, where x_i is the *ith* record and the records are indexed by number from 1 to N. Using Semi-Supervised Learning, the model mix all labeled data $x \in \{x_i\}_{i=1}^{N}$ and unlabeled data $x \in \{x_i\}_{i=N+1}^{n}$, through the process of label propagation to predict the labeling for unlabelled data.
4. (G) Active Learning: the models select the most desirable medical images $x_k, where\ k \in \{k_i\}_{i=1}^{K}$ for professional radiologies to label. The selection of K is essential to the performance of the model.

Preprocessing sets up the ground information of samples and domain knowledge and the initial definition of the targets. The definition of the targets is determined by $\{y_i\}_{i=1}^{N_l}$ initially in this step. The selection of $\{y_i\}_{i=1}^{N_l}$ is from professional radiologists and not necessarily the optimal choice for step two semi-supervised learning algorithms to understand and follow in execution. Semi-supervised learning identifies each sample by following the definition of the targets. Label propagation is used in the experiment session, and other methods can be used as well. This step predicts the label for all images, as well as the confidence level of each prediction. The confidence level is used in step 4 for active learning. With the prediction result, a cross-referencing step can directionally search the patient's historical record related to the target definition, which can enhance the learning. This step also depends on the accessibility of the history data record associated with the specific image. This step is not implemented in the experiment session currently. Step4, active learning is essential to the proposed model. As stated in [Tarvainen 2017]: Mean teachers are better role models; the target definition is the key to the success of the model. The initial selected labeled image may not be the best role models for students (unlabeled images) to learn. The model chose the images with the highest return value or contribution to the following up prediction and present them to the professional radiologists to add labels. The extended labeling group became a better definition of the targets. With a better description of the target, step 2 will be able to provide high accuracy. Figure 2 shows the architecture of implementation for the proposed model.

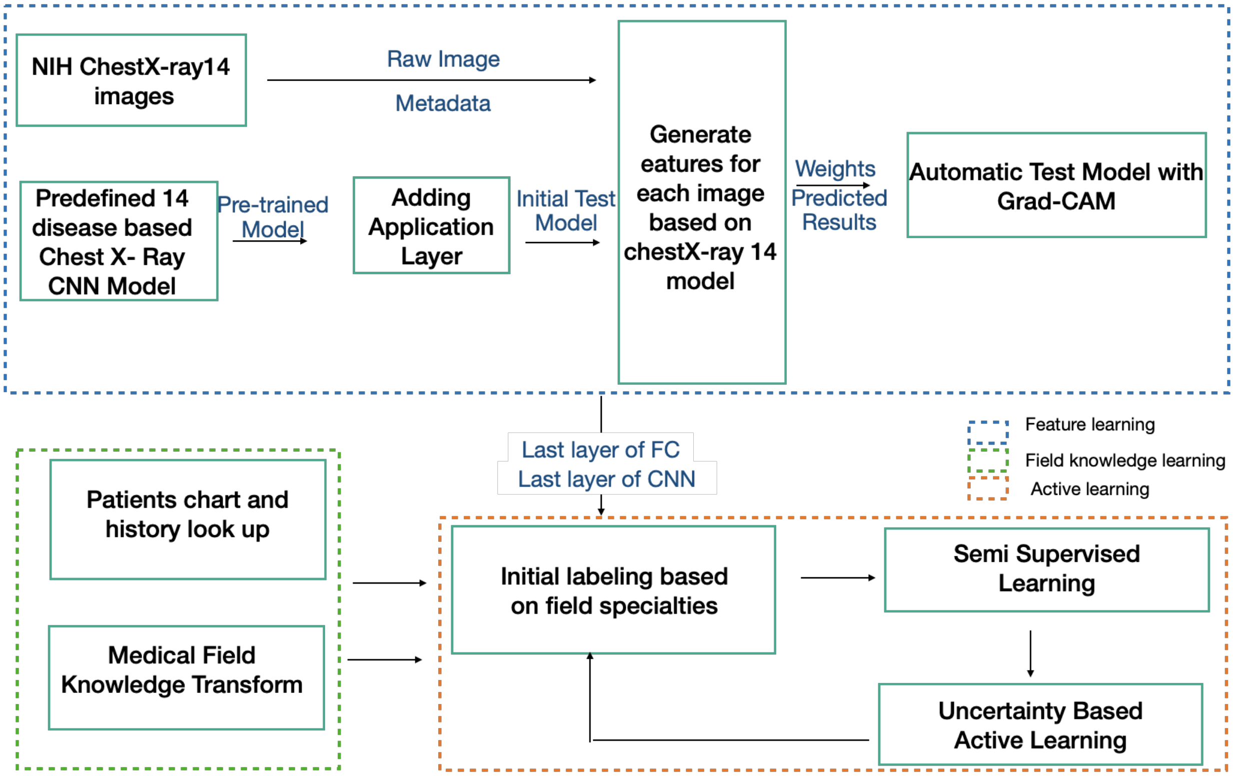

NIH ChestX-ray14 images
Raw Image
Metadata
Predefined 14 disease based Chest X- Ray CNN Model
Pre-trained Model
Adding Application Layer
Initial Test Model
Generate eatures for each image based on chestX-ray 14 model
Weights
Predicted Results
Automatic Test Model with Grad-CAM
Last layer of FC
Last layer of CNN
Feature learning
Field knowledge learning
Active learning
Patients chart and history look up
Medical Field Knowledge Transform
Initial labeling based on field specialties
Semi Supervised Learning
Uncertainty Based Active Learning

Figure 2: System Architecture

The training of discriminative models for a specific specialty area is also hardware expensive and time-consuming. By contrast, generative modeling facilitates semi-supervised training with limited labeled data. A generative model can be described as a conditional probability of the observable X, given a target y, symbolically $P(X|Y = y)$. A discriminative model is a model of the conditional probability of the target Y, given an observation x, symbolically, $P(Y|X = x)$. In the scenario of limited labeled data, a generative model can guide the model towards the interested area and a more efficient way of leveraging labeled data.

We describe and demonstrate semi-supervised learning based multi-iteration generative model to guide medical images diagnosis according to labeled medical images in the targeted direction. It combines domain knowledge through the labeling and machine learning method. It also provides physicians and researchers the flexibility of defining the categories of diagnosis.

The approach leverages the feature extraction from a comprehensive discriminative model cheXNet and utilizes the characteristic features of CheXNet as the input. CheXNet is a 121-layer convolutional neural network trained on ChestX-ray14, the largest publicly available chest X-ray dataset [Rajpurkar2017]. Other than CheXNet, additional CNN based medical image recognizing systems can be used for feature extraction. In Table 1, the 121-layer CheXNet structure was shown.

Table 1: Dense Net-121 based CheXNet

Layers	Dense Net-121
Convolution	$7 \times 7\ conv$, stride 2
Pooling	$3 \times 3\ max\ pool$, stride 2
Dense Block 1	$\left(\frac{1 * 1\ conv}{3 * 3\ conv}\right) * 6$
Transition Layer 1	$1 \times 1\ conv$ $2 \times 2\ average\ pool$
Dense Block 2	$\left(\frac{1 * 1\ conv}{3 * 3\ conv}\right) * 12$
Transition Layer 2	$1 \times 1\ conv$ $2 \times 2\ average\ pool$
Dense Block 3	$\left(\frac{1 * 1\ conv}{3 * 3\ conv}\right) * 24$

Transition Layer 3	$1 \times 1\ conv$ $2 \times 2\ average\ pool$
Dense Block 4	$\left(\frac{1 * 1\ conv}{3 * 3\ conv}\right) * 16$
Transition Layer 4	$1 \times 1\ conv$ $2 \times 2\ average\ pool$
Classification Layer	$1 \times 1\ global\ average\ pool$ $14D\ fully\ connected, element\ wise\ sigmoid$

For calculation convenience, we first resize images in the dataset to be $n \times n$pixels. There are no restrictions on the original images' size as long as the resolutions are high enough to detect the features we need. Let x ∈ $R^{n \times n}$ be a 2D x-ray image and y ∈ $\{c_0, c_1, c_2, c_3, \ldots, c_k\}$ be the corresponding label that contains information of diagnosis, severity, complication, or other categories. A dataset includes a set of N images $x = \{x_i\}_{i=1}^{N}$, in which N_l images annotated with labels $y = \{y_i\}_{i=1}^{N_l}$ Then, a probabilistic feature representation Z to predict the label y information was constructed.

The domain-specific features from CheXNet are leveraged to construct the feature representation. Instead of identifying pre-defined diagnoses, the proposed system enables a user-defined target through the labeling of images. The labeling can have multiple sources: medical experts, cross sourcing patient's history EHR records or referencing other models, etc. Through active learning, the labeling is not a prerequisite but an ongoing process with the utilization of the model. The system evaluates the uncertainty and other aspects of each record and marks the images that, if labeled, will be most valuable to the application of the model. During the utilization of the model, the algorithm can select the images for labeling. The labeled image improves model performance by guiding the direction of the model.

Model Description:

Before getting into the detailed description, there are some variables need to be defined first:

g : number of elements in the last convolution layer

f : number of elements in the last full connection layer

p : the union array of last convolution layer and full connection layer

y : the labeled records

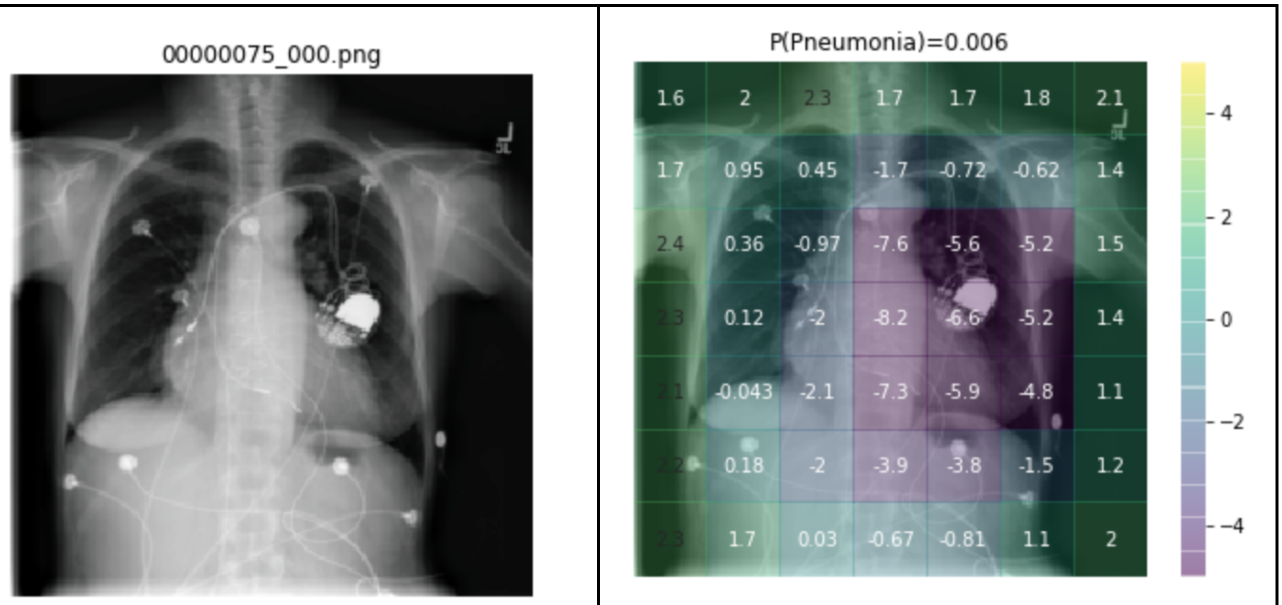
Let p ∈ R^{g+f} be a one-dimensional array that is derived from the CheXNet model.

$$p \ = \{p_i\}_{i=1}^{g} \cup \{p_j\}_{j=g+1}^{f+g}$$

Where $\{p_i\}_{i=1}^{g}$ represent the calculated array of results of the last layer of convolution layer, and $\{p_j\}_{j=g+1}^{f+g}$ represents the calculated array of results of the last full connection layer before the softmax.

Let $y \ = \{y_i\}_{i=1}^{N_l}$ be the labeled samples. The samples are defined based on demand for a particular specialty or defined target. For example, it can be labeled as predicted pulmonary edema severity[Liao2019].

The last convolution layer $\{p_i\}_{i=1}^{g}$ represents how sensitive each part of an image is to the classification result[Selvaraju2019][Zech2018]. Figure 3 shows an example of the last convolution layer. We can see that throughout CNN based CheXNet training model, the focused area is highlighted in the purple color. This area has a higher sensitivity level to the predicted result compared to the rest of the area.



a): Original Chest X-Ray Diagnosed as Pneumonia

(b) Image with Grad-Cam Integrated

Figure 3: Example for Grad Cam based on CheXNet model

The sensitive area is positively correlated to the model itself. The prediction accuracy in a given model will affect the focused, sensitive area. To work across different chest X-ray reading CNN models and depend less on the accuracy of the select chest X-ray reading

CNN models, it is necessary to include the last full connection layer, which represents the decision made from a particular model. In summary, by adding both the last convolution layer and the last layer of full connection, we obtained the information data from both the decision matrix and the reasoning matrix.

The parameters of last full connection layer before softmax is $\{p_j\}_{j=g+1}^{f+g}$. This is the layer that is linearly used to predict the probability of each diagnosis. The relationship between them can be seen as follows, where A is the character matrix of each image:

$$y^c = \{p_j\}_{j=g+1}^{f+g},$$

the gradient backpropagation is:

$$gb = \frac{\partial y^c}{\partial A_{ij}^k} \quad (1)$$

$$\alpha_k^c = \frac{1}{z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{i,j}^k} (2)$$

$$L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right) \ (3)$$

$$p = \{p_i\}_{i=1}^{g} = L_{Grad-CAM}^c \ (4)$$

In (2), $\frac{1}{z} \sum_i \sum_j$ is the global average pooled over the width and height dimensions to obtain the neuron importance weights α_k^c . $\frac{\partial y^c}{\partial A_{ij}^k}$ is the gradients via back propagation. $\frac{1}{Z}$ is a proportionality constant that is used for normalization, and $Z = \sum_{i,j} 1$.

Thus, through the convolution model training, we can transfer the domain knowledge of each image, and we use it here as the input to component D in Figure 2.

The target of an application is reflected by the labeling of a small percentage of the images. The target is defined as y. The labeling came from professional radiologists' input, or it could come from a guided searching of a patient's historical record. It is a continuous effort to enrich the database and make the model more accurate or, more specifically, towards the target of application. $y = \{y_i\}_{i=1}^{N_l}$ is the labeled samples. As N^l increases, the accuracy will improve as well. However, there is a cost function associated with the number of labeling and a cost for low accuracy. The goal here is to minimize the overall cost. The cost model can be summarized as:

$$\min \left(f_0(N_l) + \sum_{i,j=1}^{k} f_{i,j} C_{i,j}\right) \quad (5)$$

In function (5), $f_0(N_l)$ is the cost of labeling $N_l$samples. $f_{i,j}$is the cost for misdiagnosis type i to type $j$where ($f_{i,j}$= 0 when $i = j$) using the current model. $C$is the confusion matrix where $C_{i,j}$is the number of samples for each misdiagnosis. The cost model is integrated as the loss function with Z in figure 2, which is the model of a semi-supervised based learning model to shape the feature extraction towards the fields of the applications. When the benefit of increasing the accuracy of the model brought by the number of annotated records is less than the cost of annotation itself, the model is stabilized and stops generating new records for physicians to diagnose. When setting $f_0(N_l)$= 0, the model will always be in the query model, and keep on increasing the accuracy by generating new records for physicians to diagnose. When a particular misdiagnosis is set to have a high price, this type of misdiagnose records takes a larger percentage over the generated records for querying.

The semi-supervised learning algorithm maximizes the log probability $log(prob())$ of the sample with respect to parameters θ, and the objective is:

$$Max(P), where\ P = \sum_{i=1}^{N_l} prob(x_{i,}y_{i,}\theta) + \sum_{i=N_{l+1}}^{N} prob(x_{i,}\theta) \text{ (6)}$$

A label propagation algorithm(LPA) based algorithm was implemented to achieve the goal. The process includes five steps:

1. Initialize the labels y at all samples. For a given sample x_i, $y_i(0) = C_0$

2. Set $t = 1$.

3. Arrange the samples in the network in random order and set it to C_0.

4. For each $x \in X$ chosen in that specific order, let $C_x(t) = f(C_{xi1}(t), \ldots, C_{xim(t)}, C_{xi(m+1)}, \ldots, C_{xik}(t-1))$. f here returns the label occurring with the highest frequency among neighbors. Select a label at random if there are multiple highest frequency labels.

5. If every node has a label that the maximum number of their neighbors have, then stop the algorithm, else, set $t = t + 1$ and go to (3).

In contrast with other algorithms, label propagation can result in various community structures from the same initial condition. The range of solutions can be narrowed if some nodes are given preliminary labels while others are held unlabelled. Consequently, unlabelled nodes will be more likely to adapt to the labeled ones. The selection of semi-supervised learning also depends on the type of application and available data set. During implementation, the component L is an exchangeable module that allows replacing other semi-supervised learning algorithms without affecting the overall architecture.

Component G in Figure 2 is an essential part of the proposed model. x output the predicted result and point out the weak point of the current target definition. In component D, profession radiologies or physicians cherry-picked the cases that they would like to identify. The message of definition was delivered to the algorithm and component L interpret the definition and make the judgement. In component G, not only it delivers the result, but also tries to ask the right question for more accurate understanding of the target definition. The selecting of samples $q \in \{q_j\}_{j=1}^{e}$ that need to be reassured by the professional radiologists is the process of asking the right question and calibrating itself. Multiple methods have been proposed in the literature for selecting the records for querying. Here, we proposed targeting on lowering the overall cost. In equation (5), the overall cost includes the labeling cost and the cost for misdiagnosis type i to type j. We adopted a simplified and valid assumption that the least uncertain records in the test result has the most information, which has been used in many active learning models. $n_1$is the total number of selected records for querying. We take $n_2$most untain records R, where $n_2 \geq n_1$ .By sorting the misdiagnosis weight matrix, $F = [f_{i,j}]$, $where\ F[0] =$

$max(f_{i,j}) = f_{m,k}$, we can get that type k is the most expensive cost if the ground truth is m for that record. If a record that is predicted to be k is in R, we select a random records from R that is predicted as k. If there is no record in of k, we will look for the second most expensive one where $F[1] = max(F - max(f_{i,j})) = f_{m,k}$. If the ground truth after the physician's diagnosis is m, then, we will continue to look for records predicted as k. If not, we will update m and k to be $F[1] = max(f_{i,j}) = f_{m,k}$, and follow the same steps. The searching stops when the select number of records equals to n_1.

Recall equation (5), if $n2 = n1$, then the expected cost is

$$f_0(n_1) + \sum_{i,j=1}^{k} f_{i,j}$$

The loop among D, L, G forms a process of active learning or interactive learning. The time scale between loops can be distributed through the utilizing of the system. The geographic location can also be distributed among any medical organizations and shared through the cloud.

Experiment and Result:

Approximately 6,000 chest x-ray images and their associated radiology reports were collected as proof of concept in this work. A relatively small sample to evaluate the system performance was included considering the use case when less data is available. We extract the features of the images based on the CheXNet model. The NIH label is used for the labeling. Due to lack of interaction with physician's on evaluating the cost for misdiagnosis, we simplified our model cost equation (5) to be $c_{i,j} = 0 \, when \, i \neq j$. In future research, we will run expert interviews and other format of survey to determine the values of $c_{i,j} \, when \, i \neq j$.

The algorithms were evaluated based on several aspects: model accuracy over various percentages of labeled samples, model accuracy over several categorized diagnoses. The model accuracy is evaluated by the average value of F1 scores in all categories.

Figure 4 shows that the recall value varies by the percentage of labeled samples in the overall samples. In Figure 4, Cardiomegaly Detection is used as an example. As the percentage of the labeled sample increases, the recall value also increases.

Figure 5 shows that the average F1 score varies by the percentage of labeled samples in overall samples. In Figure 5, Cardiomegaly Detection is also used as an example. As the percentage of the labeled sample increases, the F1 score also increases.

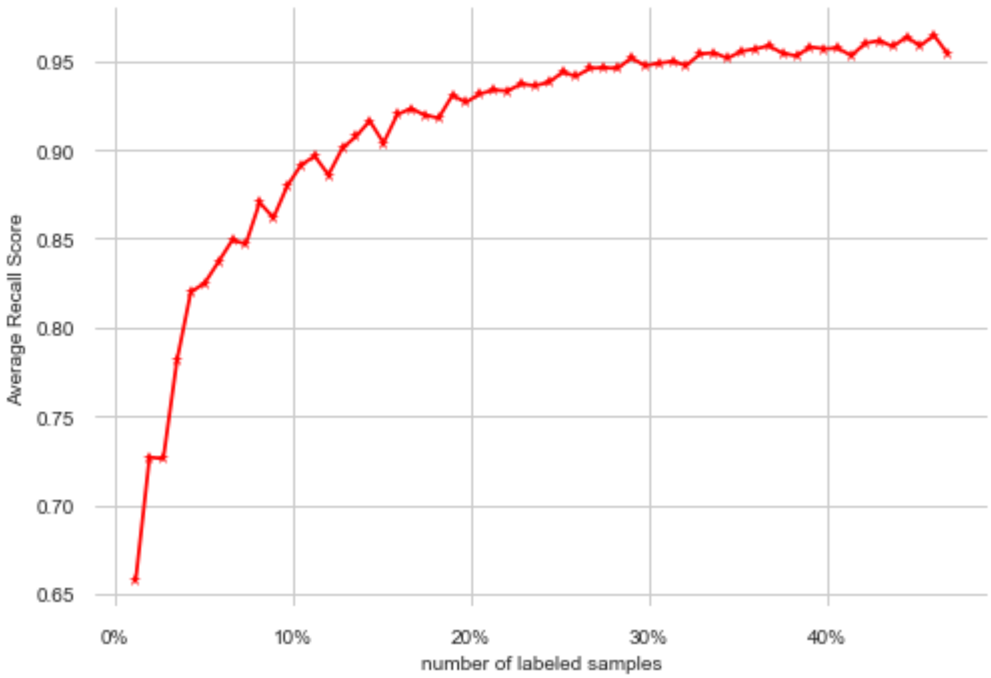


Figure 4: Recall score vs. labeled sample percentage for Cardiomegaly Detection

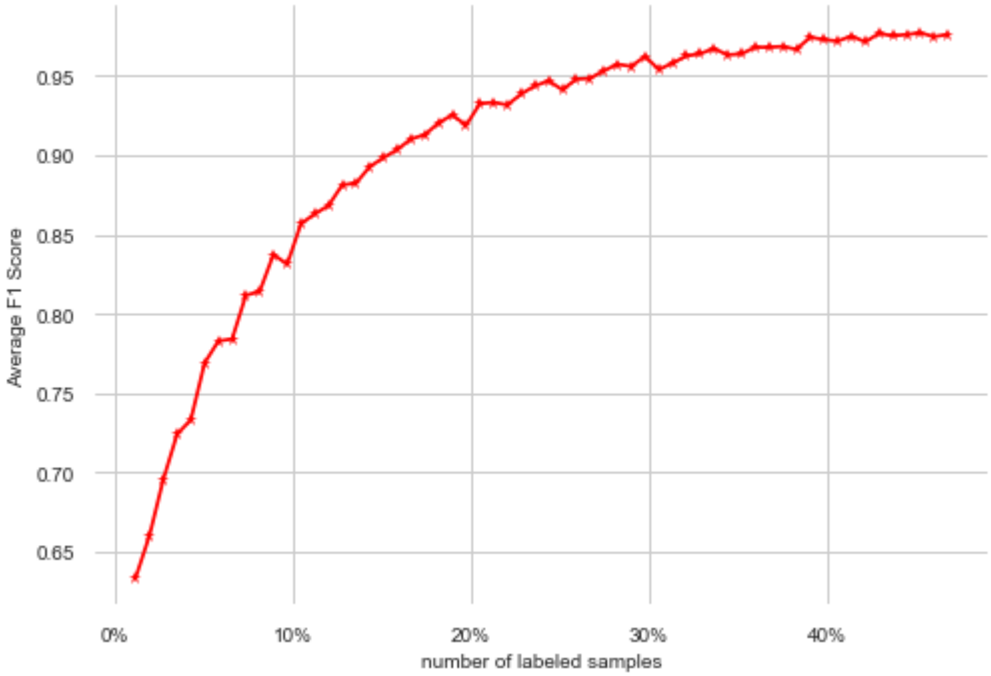


Figure 5: Average F1 score vs. labeled sample percentage for Cardiomegaly Detection

Cardiomegaly Detection in CNN based Chest X-Ray Diagnosis has relatively higher accuracy. For a more comprehensive evaluation of the system, given the current constrained on labeled data, we also compared the curve among different disease detection.

The total number of samples is the same for each disease detection in the experiment shown in Figure 6. We can see that given the total number of samples, two factors are affecting the performance of the system: type of disease, percentage of labeled data. For some diseases, the percentage required is higher than others in order to achieve precise accuracy.

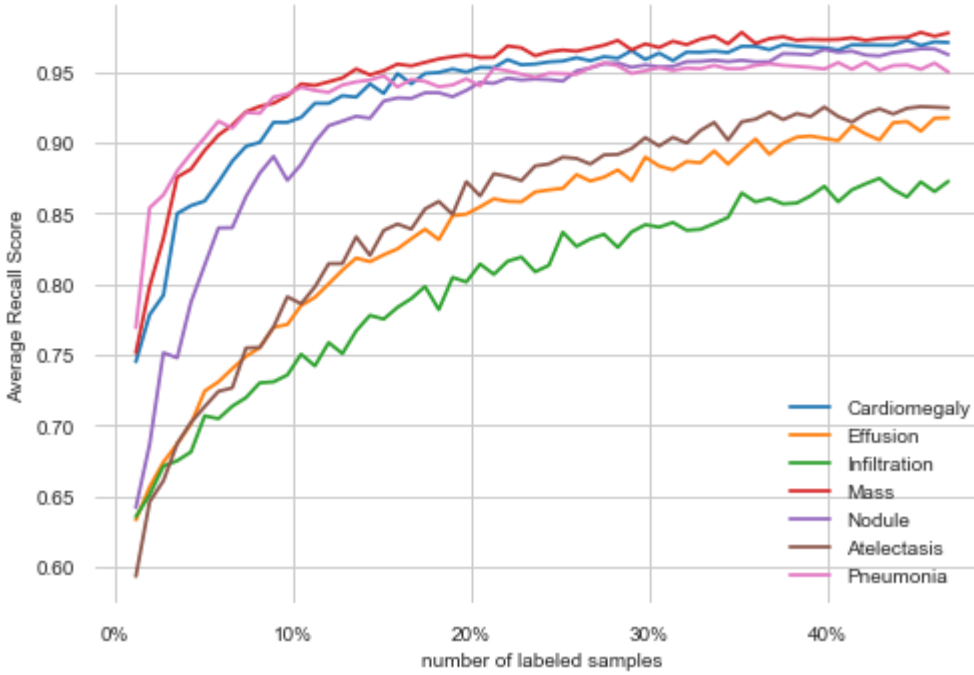


Figure 6: Comparison between different diseases and diagnosis

In the clinical application, the classification is not restricted to a predefined disease. It could be a select type of Pneumonia. With certain labeled data, the model is customized as demand and provides an effective method with low computation complexity. In Figure 7, we make the diagnosis that Effusion and Infiltration simultaneously appeared, and try to identify these images. From the recall rate, it shows a better performance compared to when only Effusion or only Infiltration is diagnosed.

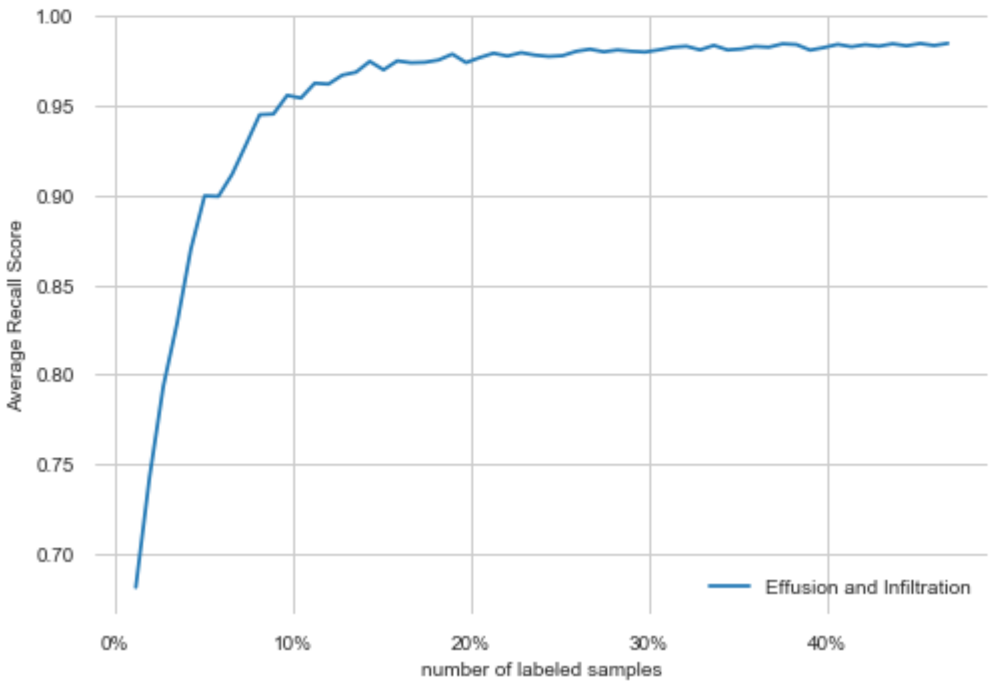


Figure 7: Performance for a complication of Effusion and Infiltration

Conclusion:

With approximately 2 billion procedures per year, chest X-rays are the most common imaging examination tool used in practice, critical for screening, diagnosing, and managing a variety of diseases(Raoof2012). There is a shortage of experts who can interpret X-rays, leading to increased mortality from treatable diseases (Kesselman2016). For example, ChestNet detects Pneumonia from frontal-view chest X-ray images and detects multiple conditions and outperforms the previous state of the art on ChestX-ray14[Rajpurkar2017]. This technology increases access to medical imaging expertise, however, is limited by the pre-defined 14 types of diseases. It is challenging for a generative model to learn distinct data clusters for the labels that rely on subtle image features. This paper demonstrates that Generative Semi-Supervised Active Learning could be an effective way for user-defined chest X-Ray diagnosis identification. We proposed a label propagation approach that relies on features abstracted from existing CNN based models and guided by medical domain knowledge through user-defined labeling. With a limited percentage of labeling among a large image dataset, the system can identify the chest X-ray types defined by the labeled data.

Our results suggest learning compact feature representations jointly from image information and limited labels can help inform prediction by capturing structure shared by the image distribution and the conditional distribution of labels given images. Subtle details on a patient's status captured in medical images and labels can be extracted, enabling clinicians to deliver better care by quantitatively clustering and summarizing large scales of patients' medical quality, even in the appearance of a new or undefined type of disease. Given the consideration of high labeling cost, the model works effectively with limited amounts of labels. Furthermore, without the limitation of a pre-defined category or diagnosis, more valuable information can be abstracted from a patient's medical chart and use the labeling information, which significantly reduces the cost of manually labeling.

Training on CNN based Chest X-ray diagnosis is both hardware and technically intensive. It requires high power GPU stations and expertise in machine learning. A lightweight solution for user-defined diagnosis was provided. It leverages the trained model of ChestXNet or other CNN based models, guided by the expertise labeling, and reshapes the model to fit user needs. The low requirement of training cost and labeling cost enables its application in fast spread disease detection.

Discussion and Future Work:

Uncertainty sampling is the most commonly used in actively learning framework and is used in the data experiment. The shortcoming is the assumption that all samples in the set are roughly equal size [ElRafey 2017]. Other models are proposed and need to be compared with the current result. For an imbalanced dataset, we will explore anomaly detection based methods and compare the performance.

Generative Adversarial Networks (GAN) have gained a lot of attention in the computer vision community due to their data generation capability without explicitly modeling the probability density function. The discriminator's adversarial loss provides a creative way of incorporating unlabeled samples into training and imposing higher-order consistency. This has proven useful in many cases, such as domain adaptation, data augmentation, and image-to-image translation[Yi 2019]. One of the focuses of future work is to integrate GAN into the system.

Cross-referencing can directionally search the patient's historical record related to the target definition, which can enhance learning. The implementation of this step will be implemented in future work.

A thorough investigation of labeling cost optimization is essential. Policy, regulation, the labor cost need to be added into consideration.

The comparison between LPA and other semi-supervised learning models, including "Label Propagation for Deep Semi-supervised Learning" [Iscen2019], is needed in future work. The adaptivity, association, and matching between the semi-supervised learning models and medical applications need further investigation.

We have proposed to associate the uncertainty with the cost of misdiagnosis for active learning. The implementation will be included in future work.

As part of field tests and calibration is an essential part. Combining patients' history charts and collaborating with physicians on testing, tuning, and expanding the model will be performed.

We also plan to compare the constraint-based and metric-learning methods. The two semi-supervision ways individually improve clustering accuracy and build a multi-interaction label propagation approach that integrates their strengths.

Reference:

[Madani 2018] A. Madani, M. Moradi, A. Karargyris and T. Syeda-Mahmood, "Semi-supervised learning with generative adversarial networks for chest X-ray classification with ability of data domain adaptation," *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, Washington, DC, 2018, pp. 1038-1042.

[Chakko 1991] Chakko, S., Woska, D., Martinez, H., De Marchena, E., Futterman, L., Kessler, K.M., Myerburg, R.J.: Clinical, radiographic, and hemodynamic correlations in chronic congestive heart failure: conflicting results may lead to inappropriate care. The American journal of medicine 90(1), 353–359 (1991)

[Francis 2014] Francis, G.S., Cogswell, R., Thenappan, T.: The heterogeneity of heart failure: will enhanced phenotyping be necessary for future clinical trial success? (2014)

[Mahdyoon1989] Mahdyoon, H., Klein, R., Eyler, W., Lakier, J.B., Chakko, S., Gheorghiade, M.: Radiographic pulmonary congestion in end-stage congestive heart failure. Ameri- can Journal of Cardiology 63(9), 625–627 (1989)

[Liao2019] Liao, R., et al.: Semi-supervised learning for quantification of pulmonary edema in chest x-ray images. arXiv preprint arXiv:1902.10785 (2019)

[Rajpurkar2017] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al.: Chexnet: Radiologist-level Pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225 (2017)

[Wang2017] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on CVPR. pp. 3462–3471. IEEE (2017)

[Rajpurkar2017] Rajpurkar, Pranav & Irvin, Jeremy & Zhu, Kaylie & Yang, Brandon & Mehta, Hershel & Duan, Tony & Ding, Daisy & Bagul, Aarti & Langlotz, Curtis & Shpanskaya, Katie & Lungren, Matthew & Ng, Andrew. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning (2017).

[Selvaraju2019] Selvaraju, Ramprasaath R. et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization.” International Journal of Computer Vision 128.2 (2019): 336–359. Crossref. Web.

[Iscen2019] Ahmet Iscen and Giorgos Tolias and Yannis Avrithis and Ondrej Chum, Label Propagation for Deep Semi-supervised Learning arXiv 1904.04717, 2019

[Tarvainen 2017] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In NIPS, 2017.

[ElRafey 2017] ElRafey, A. and Wojtusiak, J. (2017), Recent advances in scaling-down sampling methods in machine learning. WIREs Comput Stat, 9: e1414. doi:10.1002/wics.1414

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

[Irvin 2019] Irvin, Jeremy et al. “CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison.” arXiv.org (2019): n. pag. Web.

[Yi 2019]Yi, Xin, Ekta Walia, and Paul Babyn. “Generative Adversarial Network in Medical Imaging: A Review.” Medical Image Analysis 58 (2019): 101552. Web.

[Smailagic2018]Smailagic, Asim et al. “MedAL: Deep Active Learning Sampling Method for Medical Image Analysis.” *arXiv.org* (2018): n. pag. Web.

[Wu2018]Wu, Jian et al. “Active Learning with Noise Modeling for Medical Image Annotation.” *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. Vol. 2018-. IEEE, 2018. 298–301. Web.

[Kim2020]Kim, Seong, Farrukh Mushtaq, and Nassir Navab. “Confident Coreset for Active Learning in Medical Image Analysis.” *arXiv.org* (2020): n. pag. Web.

[Zech2018]J. Zech, “reproduce-chexnet,”GitHub repository, 2018

A publication of the *Machine Learning and Inference Laboratory*
College of Health and Human Services
George Mason University
Fairfax, VA 22030-4444 U.S.A.
http://www.mli.gmu.edu

Editor: J. Wojtusiak

The *Machine Learning and Inference (MLI) Laboratory Reports* are an official publication of the Machine Learning and Inference Laboratory, which has been published continuously since 1971 by R.S. Michalski's research group (until 1987, while the group was at the University of Illinois, they were called ISG (Intelligent Systems Group) Reports, or were part of the Department of Computer Science Reports).