Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

A Deep Community Based Approach for Large Scale Content Based X-Ray Image Retrieval

Nandinee Fariah Haq[a,*], Mehdi Moradi[b], Z. Jane Wang[a]

[a] *The University of British Columbia, Vancouver, Canada*
[b] *IBM Almaden Research Center, San Jose, USA*

ARTICLE INFO

Article history:

Keywords: content based image retrieval (CBIR), deep learning, graph, community detection

ABSTRACT

A computer assisted system for automatic retrieval of annotated medical images with similar image contents can serve as an efficient management tool for handling and mining large scale data, and can also be used as a tool in clinical decision support systems. In this paper, we propose a deep community based automated medical image retrieval framework for extracting similar images from a large scale X-ray database. The framework integrates a deep learning-based image network generation approach and a network community detection technique to extract similar images. When compared with the state-of-the-art medical image retrieval techniques, the proposed approach demonstrated improved performance. We evaluated the performance of the proposed method on two large scale chest X-ray datasets, where given a query image, the proposed approach was able to extract images with similar disease labels with a precision of 85%. To the best of our knowledge, this is the first deep community based image retrieval application on large scale chest X-ray database.

1. Introduction

Chest X-rays are the most frequently performed radiological examinations in clinical routines to identify different abnormal thoracic and cardiopulmonary conditions (Folio, 2012). With the advances in medical imaging technology and the subsequent hike in the number of radiology examinations ordered, there is a substantial surge in the workload of radiologists (Hosny et al., 2018). This, in turn, results in a longer radiology turnaround time hence reducing the overall quality of patient care (Bastawrous and Carney, 2017; Rimmer, 2017). A computer assisted system to automatically analyze and extract previously diagnosed X-rays with similar image content can be a helpful tool to guide the diagnosis and the process of generating a radiology report (Akgül et al., 2011). This in turn can accelerate the radiology workflow and thereby improve the overall quality of healthcare.

Content based image retrieval (CBIR) has been an active area of research in the field of computer vision for the past 20 years. In content based retrieval systems, the database images are first represented in terms of a set of associated features computed directly on the image content. During retrieval, given a query image, similar images are selected from a database of images based on their feature similarity with the query image. Traditionally, CBIR systems were developed by designing discriminant handcrafted features. However, with handcrafted features, the challenge remains to reduce the "semantic gap", which is the information lost in the process of designing low-dimensional features to represent all information an image contains (Qayyum et al., 2017). This gap can be reduced with the help of machine learning based techniques where an intelligent system is trained to automatically generate a discriminant feature space. Given sufficient data, deep learning networks

[*] Corresponding author
e-mail: nandinee@ece.ubc.ca (Nandinee Fariah Haq)

automatically learn complex features at multiple levels of abstraction without using handcrafted features. With the recent advancement of deep learning based techniques and with the availability of large scale and ever-increasing number of digital image databases, the research community has now moved towards the implementation of deep learning based CBIR systems (Torralba et al., 2008; Lai et al., 2015; Liu et al., 2016a).

Although image retrieval systems have been extensively studied for natural image retrieval tasks, the application of the retrieval framework in medical images, especially in radiology images still remains a challenging task (Zhang and Metaxas, 2016; Akgül et al., 2011). This is partly because medical images are more difficult to analyze when compared to natural images, owing to the complex imaging parameters, interactions between different diseases, and subtle differences between images with different diagnosis decisions (Li et al., 2018). Nevertheless, efforts have been made to develop medical image retrieval systems in recent years. Most of the literature reports retrieval based on handcrafted or shallow learning based features from different medical image modalities (Quellec et al., 2011; Rahman et al., 2011; Zhang et al., 2014; Lan et al., 2018). However, shallow learning based features are not applicable when designing a retrieval system for a large scale database (Li et al., 2018). Deep learning based systems have the potential to become a suitable tool for large scale medical image retrieval. However, content based image retrieval has not seen many successful applications of deep learning yet (Litjens et al., 2017), partly due to the unavailability of large scale radiology datasets. The current deep learning based radiology image retrieval approaches mostly use models pre-trained on other image databases (Anavi et al., 2015; Shah et al., 2016), or a model with fewer layers trained with a smaller dataset (Liu et al., 2016b; Conjeti et al., 2017; Chen et al., 2018). However medical images can be very different from natural images and hence pre-trained models might not be the proper feature extractor for medical settings. The dominance of deep learning is mainly a result of the availability of large training datasets. Therefore, similarly, a domain specific model trained on a well-annotated single modality large image dataset can be a meaningful way to leverage the full potential of deep learning techniques for medical image retrieval systems.

In this work, we present a deep-community based large scale medical image retrieval framework for radiology images. The framework utilizes a deep neural network model trained on chest X-ray images to generate image representative codes. To implement an efficient search engine for the medical image retrieval task, the database is divided offline into communities of most similar images using a network community extraction approach. The extraction of similar images is formulated as a novel region growing based sub-network extraction problem from a graph network of database images. To extract similar images we maximize the community quality metric named weighted modularity that takes into account the strength of the formed image community and the difference between the edges within the image community from that of a randomly distributed network. The framework is evaluated on two recently published large scale chest X-ray image datasets. To the best of our knowledge, this is the first deep community based image retrieval application on large scale chest X-ray datasets.

2. Materials and Methods

In this section, we describe the datasets used for this work and the proposed framework for large scale medical image retrieval. We formulate the medical image retrieval from a large scale dataset as a deep leaning based community extraction problem. Our framework for large scale medical image retrieval consists of three parts: image code generation from a deep neural network model, graph network formulation, and similar image community formation. The framework is shown in Fig. 1 which includes the major components described in Sections 2.2-2.5.

2.1. Datasets

For this work, we used two publicly available large scale datasets of chest X-rays. The first dataset is the ChestX-ray8 dataset from the National Institutes of Health (NIH) (Wang et al., 2017) that contains 112,120 frontal-view X-ray images from 30,805 unique patients, among which 16,630 are male and 14,175 are female. The images were collected from the year 1992 to 2015 and have associated text-mined disease labels. The disease labels were mined from the associated radiological reports using natural language processing algorithms. Thirteen common thoracic disease labels are used in this work which are reported in Table 1, along with the numbers of positive and negative samples in the dataset. The dataset is multi-label, i.e. each chest X-ray image can bear more than one positive disease label. The dataset includes 67,310 PA view images and 44,810 AP view images.

The second dataset used here is the CheXpert dataset released by a team at Stanford University (Irvin et al., 2019). The dataset contains 223,648 chest X-ray images from 64,740 unique patients, with 35,917 male and 28,822 female subjects. The images were obtained between 2002 and 2017. Among the X-ray images, 191,229 images are frontal chest X-rays and the rest 32,419 are lateral X-rays. The disease labels were generated using an automatic rule-based labeler from their associated radiology reports. The labeler classified the labels as positive, negative and uncertain value. The uncertain label is assigned when it has no positive mentions and at least one uncertain mention in the associated radiologist reports. In this work, we used nine disease labels based on their prevalence in the dataset. The disease labels and their number of samples are reported in Table 1. Fig. 2 shows a few frontal posterior-anterior (PA) images from these datasets with different disease labels.

2.2. Image Code Generation

We start with a database of N X-ray images I_n's with associated disease labels y_n's, $\mathcal{D} = \{I_n|y_n\}_{n=1}^{N}$. The dataset is normalized and passed through a data augmentation block that horizontally flips the X-ray images randomly on the fly. A deep CNN based model is then trained on the augmented data to generate the disease likelihood from the X-ray images. In this work we considered a DL model proposed in Huang et al. (2017)

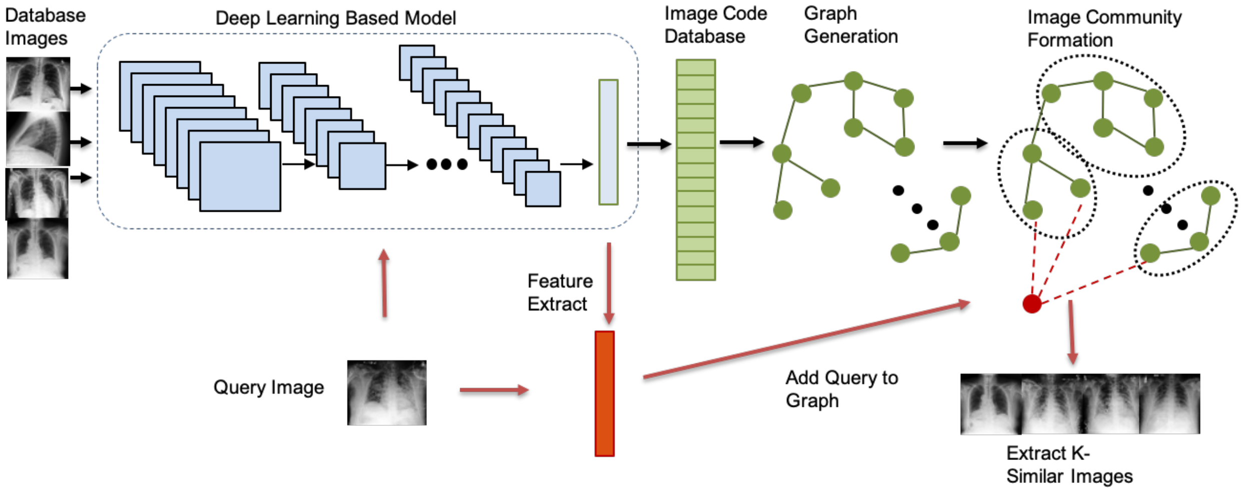

Fig. 1. Illustration of the proposed framework.

Table 1. Description of the datasets. Number of cases from different disease types with positive, negative and uncertain labels.

Disease label	NIH Dataset		Stanford Dataset		
	Positive	Negative	Positive	Negative	Uncertain
Atelectasis	11,559	100,561	33,456	156,453	33,739
Enlarged Cardiomediastinum	-	-	10,907	200,338	12,403
Cardiomegaly	2,776	109,344	27,068	188,493	8,087
Consolidation	4,667	107,453	14,816	181,090	27,742
Edema	2,303	109,817	52,291	158,373	12,984
Pneumonia	1,431	110,689	6,047	198,831	18,770
Pneumothorax	5,302	106,818	19,456	201,047	3,145
Pleural Effusion	13,317	98,803	86,254	125,766	11,628
Lung Opacity	-	-	105,707	112,343	5,598
Infiltration	19,894	92,226	-	-	-
Emphysema	2,516	109,604	-	-	-
Pleural Thickening	3,385	108,735	-	-	-
Fibrosis	1,686	110,434	-	-	-
Nodule	6,331	105,789	-	-	-
Mass	5,782	106,338	-	-	-

which consists of densely connected convolutional layer blocks, known as DenseNet. Within the dense blocks, each convolutional layer has a feed-forward connection to every other layer. Neural networks based on dense blocks have shown superior performances for chest X-ray based applications (Irvin et al., 2019). Moreover, due to the flow of gradients throughout the model, dense block based architectures are easy to train and hence are preferable for training a deep neural network model with smaller training set (Lee et al., 2015; Huang et al., 2017).

In this work, specifically, we have employed a DenseNet model with four dense blocks and four convolutional blocks. The structure of the model is shown in Fig. 3. The output layer consists of a fully-connected sigmoid function-based dense layer to allow for multi-label classification. The weights are initialized from a model pre-trained on ImageNet dataset (Deng et al., 2009) and then the model architecture is trained on the chest X-ray dataset.

The medical image datasets used in this work are highly imbalanced datasets, with different numbers of samples for different diseases, and almost 2 to 14 times more negative samples than the positive ones, as can be seen from Table 1. To handle the class imbalance concern present in the datasets, we propose optimizing the following weighted binary cross-entropy function as the loss function during training :

$$\mathcal{L}_{I|y} = \sum_{y^j \in \{0,1\}}^{j} -w_{j+} \times y^j log(\bar{y}^j) - w_{j-} \times (1-y^j) log(1-\bar{y}^j) \quad (1)$$

$$where, \quad w_{j+} = \frac{|n_{j-}|}{N}; \quad w_{j-} = \frac{|n_{j+}|}{N}$$

where $\mathcal{L}_{I|y}$ is the loss term for image I with the label y, y^j is the ground truth label for disease class-j and $\bar{y}^j$ is the predicted likelihood. $|n_{j+}|$ and $|n_{j-}|$ are the total number of positive and negative samples respectively for class-j.

After training the DenseNet model, the 1024 dimensional feature vector from the second last layer was extracted for each

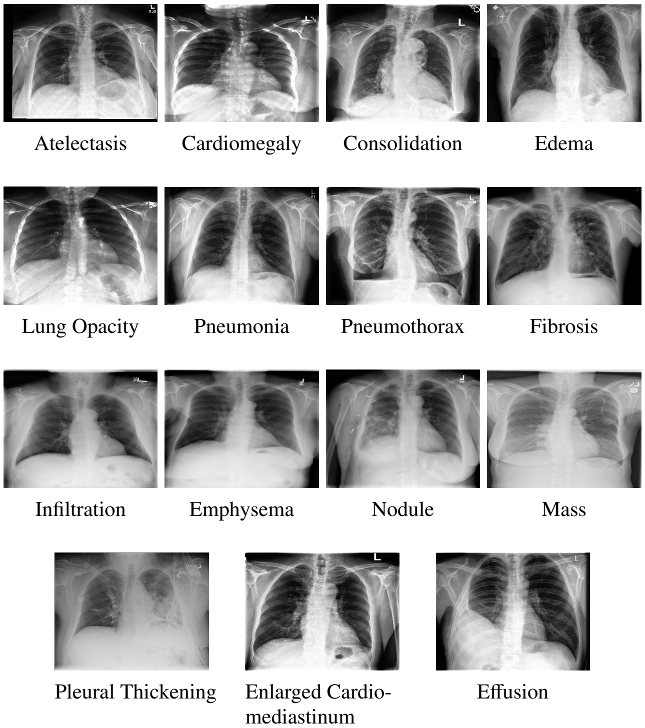
Atelectasis
Cardiomegaly
Consolidation
Edema
Lung Opacity
Pneumonia
Pneumothorax
Fibrosis
Infiltration
Emphysema
Nodule
Mass
Pleural Thickening
Enlarged Cardio-mediastinum
Effusion

Fig. 2. Examples of frontal posterior-anterior (PA) images from the datasets for different disease labels.

image. The feature vector, i_n was then used as the image representative code for image I_n in the dataset and the codes were saved with the associated images in the database, $\mathcal{D} = \{I_n \rightarrow i_n | y_n\}_{n=1}^{N}$. We denote the deep learning process of generating image codes from the associated images as $\mathcal{F}_{code} : I \rightarrow i$.

2.3. Graph Network Formulation

The next step of the framework consists of generating a graph network of similar images. The database can be considered as a network of images where similar images are strongly connected with each other whereas images that are different are either loosely connected or unconnected. From the database, a graph network $\mathcal{G}(I, E)$ was generated where $I = \{I_n | n \in [1, N]\}$ represents the nodes of the network and $E = \{e_{mn} | e_{mn} = \mathcal{F}_{edge}(i_m, i_n), m \in [1, N], n \in [1, N]\}$ denotes the edges. The edge between two image nodes I_m and I_n were calculated from their corresponding code vectors $i_m = \mathcal{F}_{code}(I_m)$ and $i_n = \mathcal{F}_{code}(I_n)$ by an edge generating function $\mathcal{F}_{edge}$ as follows:

$$e_{mn} = \mathcal{F}_{edge}(i_m, i_n) = \delta_T\Big(\frac{i_m \cdot i_n}{\sqrt{i_m \cdot i_m}\sqrt{i_n \cdot i_n}}\Big) \tag{2}$$

where,

$$\delta_T(x) = \begin{cases} x, & \text{if } x \leq T \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

The edge vector $e_{mn} \in \{0, 1\}$ and $e_{mn} = 1$ denotes the presence of an edge between samples I_m and I_n. The threshold T is selected to ensure that the graph network is sparse. In our large scale network problem, for a network with $N > 10^5$ nodes, the threshold is heuristically selected so that the total number of edges is in the order of 10^{10}.

Table 2. Definition of variables.

$\mathcal{G}(I, E)$	Graph with nodes I and edges E
N	Total number of nodes in the network, $N = \lvert I \rvert$
L	Total number of edges in the network, $L = \lvert E \rvert$
n_u	Total number of vertices in community-u
d_u	Sum of degrees of vertices in community-u
l_u	Total number of edges within community-u
l_{uv}^{ext}	Total number of edges between communities u and v
q_u	Modularity term of community-u, defined in Eqn. 4
λ_u	Weight term for community-u, defined in Eq.5
Q	Weighted modularity, defined in Eq.4
$\mathbf{C}^{N\times1}$	Community label vector

Algorithm 1 Image Community Formation

```
Input: Network, $\mathcal{G}(I, E)$
Output: Image community labels, $C$
Initialize: $\bar{\mathcal{G}}(\bar{V}, \bar{W}) \leftarrow \mathcal{G}(I, E)$, $\bar{V} \leftarrow \{i | \forall i \in I\}$
$outer \leftarrow TRUE$
while $outer$ do
    $inner \leftarrow TRUE$
    $C \leftarrow \{\{i\}\}, \forall i \in \bar{V}$
    $l_c \leftarrow \sum w_{ij}, \forall i \in c, \forall j \in c$
    $d_c \leftarrow \sum w_{ij}, \forall i \in c, \forall j \in \bar{V}$
    while $inner$ do
        for $i \in \bar{V}$ do
            $\bar{c} \leftarrow \underset{c*}{\texttt{argmax}}\{\Delta q_{i\rightarrow c*} \Leftrightarrow q_{i\rightarrow c*} > 0\}$ ; $\forall c* \ \ s.t. \ w_{ij} > 0, j \in c*$
            $d_{\bar{c}} \leftarrow d_{\bar{c}} + \sum_{j\in\bar{V}} w_{ij}$;   $l_{\bar{c}} \leftarrow l_{\bar{c}} + \sum_{j\in\bar{c}} w_{ij}$
            $d_c \leftarrow d_c - \sum_{j\in\bar{V}} w_{ij}$;   $l_c \leftarrow l_c - \sum_{j\in c} w_{ij}$
            $\bar{c} \leftarrow \bar{c} \cup \{i\}$;   $c \leftarrow c \setminus \{i\}$
        end for
        if $no\ movement\ possible$ then
            $inner \leftarrow FALSE$
        end if
    end while
    $\bar{V} \leftarrow \{c\}$; $\forall c \in C$
    $\bar{W} \leftarrow \{w_{c\bar{c}} | w_{c\bar{c}} = \sum_{i\in c, j\in\bar{c}} w_{ij}, c \in C, \bar{c} \in C\}$
    if $no\ change\ in\ communities$ then
        $outer \leftarrow FALSE$
    end if
end while
return C
```

2.4. Image Community Formation

The next step of the framework consists of finding similar image clusters from the database of images. From the network $\mathcal{G}$, we can extract similar image clusters by optimizing a community quality metric named the *weighted modularity*, that has the ability to find clusters from networks without requiring any prior knowledge regarding the number and sizes of clusters (Haq et al., 2019). For a graph $\mathcal{G}$ with N nodes which are divided into c communities, the weighted modularity of the partition is defined as:

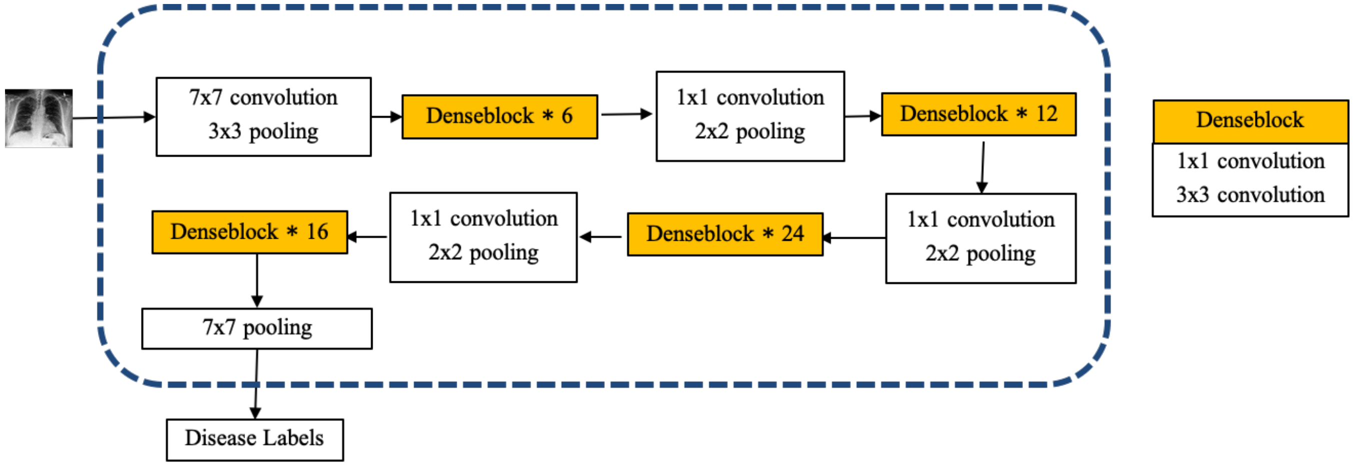

Fig. 3. Architecture of the deep neural network model.

$$Q = \sum_{u=1}^{c} \lambda_u \left[\frac{l_u}{L} - \left(\frac{d_u}{2L}\right)^2 \right] = \sum_{u=1}^{c} \lambda_u q_u \qquad (4)$$

where,

$$\lambda_u = 1 + \zeta_u \qquad (5)$$

$$\zeta_u = \begin{cases} 0, & for\ coarse\ communities; \\ \frac{2l_u}{n_u(n_u-1)}, & for\ finer\ communities. \end{cases} \qquad (6)$$

Here, n_u is the total number of nodes and l_u is the total number of edges within community-u. d_u means the sum of degrees of vertices in community-u and L is the total number of edges in $\mathcal{G}$. The related variables are defined in Table 2. The first term λ_u from Eqn. 4 denotes how strong the community is, and the second term q_u represents the difference between the fraction of edges that exist within the members of the community-u and the expected such fraction if the edges were distributed at random. Modularity maximization based methods target to find such a partition of the network for which the modularity in Eq. 4 is the maximum. $\zeta_u = 0$ is used to find bigger communities, and $\zeta_u = \frac{2l_u}{n_u(n_u-1)}$ is used to extract smaller communities from a large network. Here we used $\lambda_u = 1$ for coarse-level image community generation. Eqn. 4 then reduces to the traditional modularity equation (Newman and Girvan, 2004).

Since our network is large, to extract image communities from the graph network $\mathcal{G}$, we adopt an approach proposed in Blondel et al. (2008), which is a heuristic method to extract approximate communities from very large networks. This method has two phases that are repeated iteratively. With this approach, we start by assigning each node into separate communities. Then for each image node, I_n, we consider the neighbour I_m of I_n and calculate the gain in terms of Q if I_n is placed in the community of I_m. When $\zeta_u = 0$, the change in Q when an isolated node I_n with its own community n is placed in the community c can be computed by:

$$\Delta q_{n\to c} = \left[\frac{l_c + l_{nc}^{ext}}{2L} - \left(\frac{d_c + d_n}{2L}\right)^2\right] - \left[\frac{l_c}{2L} - \left(\frac{d_c}{2L}\right)^2 - \left(\frac{d_n}{2L}\right)^2\right] \qquad (7)$$

Here, l_c is the total number of edges within community-c, d_c is the sum of degrees of nodes within c. d_n denotes the degree of the isolated node I_n and l_{nc}^{ext} is the total number of edges between I_n and the nodes within community-c. The method computes the gain in Q from Eqn. 7 for each node I_n and the node is placed in the community c for which the gain $\Delta q_{n\to c}$ is the maximum. The process is applied sequentially for all nodes until no further increment in Q can be achieved. Suppose we get a total number of $\bar{c}$ communities from the first phase. The second phase of the algorithm builds a new graph, $\bar{\mathcal{G}}(\bar{V}, \bar{W})$ whose nodes, $\bar{V} = \{i | i \in [1, \bar{c}]\}$ represent the communities found from the first phase, and the weights of the edges, $\bar{W} = \{w_{ij} | i \in [1, \bar{c}], j \in [1, \bar{c}]\}$ between these new nodes are denoted by the total number of edges between the nodes in the corresponding communities. The total number of edges between the nodes of a community generate self-loops in the newly formed graph. Then the first phase of the algorithm is applied to the newly formed graph. Since this is a weighted network the variables from Eqn. 7 becomes:

$$\begin{gathered} l_c = \sum_{i\in c, j\in c} w_{ij}; \quad d_c = \sum_{i\in c, j\in \bar{V}} w_{ij} \\ d_n = \sum_{j\in \bar{V}} w_{nj}; \quad l_{nc}^{ext} = \sum_{j\in c} w_{nj}; \\ L = \frac{\sum_{i\in \bar{V}, j\in \bar{V}} w_{ij}}{2} \end{gathered} \qquad (8)$$

These two phases are applied to the network repeatedly. The above method to generate image community is outlined in Algorithm 1. Since the number of communities decreases at each two phase iteration, the network size decreases. Hence this approach is applicable for extracting communities faster from larger networks. The process terminates when no further improvement in modularity is observed, and the resulting partition is returned. The database is then updated with the clustering label associated with each image, $\mathcal{D} = \{I_n \to i_n | y_n, C_n\}_{n=1}^{N}$, where $C = \{C_n | C_n \in [1, c], n \in [1, N]\}$ means the community labels representing the resulting partition that divides the image nodes into c-communities.

Table 3. Disease labels considered for the datasets.

Dataset	Disease Labels Considered
NIH	Atelectasis, Cardiomegaly, Effusion, Pneumothorax, Emphysema, Pleural Thickening, Fibrosis, Consolidation, Edema, Pneumonia, Infiltration, Nodule, Mass
NIH-U (consolidated labels)	Atelectasis, Cardiomegaly, Effusion, Pneumothorax, Emphysema, Pleural Thickening, Fibrosis, Opacities (includes Consolidation, Edema, Pneumonia, Infiltration), Lesion (includes Nodule, Mass)
Stanford	Atelectasis, Enlarged Cardiomediastinum, Cardiomegaly, Pleural Effusion, Pneumothorax, Lung Opacity, Consolidation, Edema, Pneumonia
Stanford-U (consolidated labels)	Atelectasis, Enlarged Heart (includes Enlarged Cardiomediastinum, Cardiomegaly), Pleural Effusion, Pneumothorax, Opacities (includes Lung Opacity, Consolidation, Edema, Pneumonia)

Algorithm 2 Top-$\mathcal{K}$ similar image retrieval for a query image, I_q

```
Input: $\mathcal{G}(I, E)$ ,$C$, $\mathcal{F}_{code}$, $\mathcal{F}_{edge}$, $I_q$, $\mathcal{K}$
Output: Retrieved Images, $\mathcal{R}$
Initialize: $\bar{\mathcal{G}}(\bar{I}, \bar{E}) \leftarrow \mathcal{G}(I, E)$
// Add query image to $\mathcal{G}$ and assign a separate community, $c_q$
$\mathcal{R} \leftarrow \{\{I_q, c_q\}\}$
$i_q \leftarrow \mathcal{F}_{code}(I_q)$
$\bar{I} \leftarrow \bar{I} \cup \mathcal{R}$
$\bar{E} \leftarrow \bar{E} \cup \{e_{qn} | e_{qn} = \mathcal{F}_{edge}(i_q, i_n), \forall n \in I\}$
$\bar{C} \leftarrow C \cup \{c_q\}$
while $|\mathcal{R} \setminus I_q| < \mathcal{K}$ do
    $c = \underset{\tilde{c} \in \bar{C} \setminus \{c_q\}}{\texttt{arg max}} \Delta Q_{\mathcal{R} \cup \tilde{c}}$
    $\mathcal{R}_s \leftarrow \{\{u\}\};\ \forall u \in c$
    if $|\mathcal{R} \cup \mathcal{R}_s| < \mathcal{K}$ then
        $\mathcal{R} \leftarrow \mathcal{R} \cup \mathcal{R}_s$
    else
        $\mathcal{E}_s \leftarrow \{e_{uv} | u \in \mathcal{R} \cup \mathcal{R}_s, v \in \mathcal{R} \cup \mathcal{R}_s\}$
        $C_s \leftarrow \{\{u\}\}, \forall u \in \mathcal{R}_s$
        Generate subgraph: $\mathcal{G}_s(\mathcal{R} \cup \mathcal{R}_s, \mathcal{E}_s)$
        for $u \in \mathcal{R}_s \setminus (\mathcal{R}_s \cap \mathcal{R})$ do
            $I_u \leftarrow \underset{u}{\texttt{arg max}} \Delta Q^{\mathcal{G}_s}_{u \rightarrow \mathcal{R}}$
            $\mathcal{R} \leftarrow \mathcal{R} \cup \{I_u\}$
            if $|\mathcal{R} \setminus \{I_q\}| \geq \mathcal{K}$ then
                break while loop
            end if
        end for
    end if
end while
return $\mathcal{R}$
```

2.5. Similar Image Retrieval

After training the deep neural network model and formation of similar image community, the next step of the framework consists of extracting $\mathcal{K}$ similar images for a query image. Given a query image, I_q, at first its image representative code, i_q, is extracted from the trained model, $i_q = \mathcal{F}_{code}(I_q)$. We then place the query image node in the graph network, $\mathcal{G}$. The edges between the query image node and the database image nodes are generated based on their code similarity as follows:

$$e_{qn} = \mathcal{F}_{edge}(i_q, i_n) = \delta_T\Big(\frac{i_q \cdot i_n}{\sqrt{i_q \cdot i_q}\sqrt{i_n \cdot i_n}}\Big);\ \forall n \in I \tag{9}$$

The updated graph network $\bar{\mathcal{G}} = (I \cup \{i_q\}, E \cup \{\{e_{qn}\}\})$ is then used to extract the $\mathcal{K}$ most similar images. For this purpose, we implement a region growing algorithm based on the maximization of the weighted modularity to extract $\mathcal{K}$ similar images. The proposed algorithm is outlined in Algorithm 2. We start with a region, $\mathcal{R}$, with a community label c_q that only includes the query image node i_q. We then assign database images nodes to $\mathcal{R}$ iteratively until $\mathcal{K}$ images are retrieved. This is a two step process which is repeated until $\mathcal{K}$ images are retrieved. At the first step, we solve for the image community, c, which is closest to the query image node in terms of weighted modularity. Then the second step begins, where we confine our search space to find $\mathcal{K}$ most similar images only to those nodes that belong to the image community c. If the query image node i_q, with a separate community c_q, is placed in the image community c, then from Eqn. 4, the gain in terms of the weighted modularity can be expressed as:

$$\Delta Q_{c_q \cup c} = \lambda_{c_q \cup c} \times q_{c_q \cup c} - [\lambda_{c_q} \times q_{c_q} + \lambda_c \times q_c] \tag{10}$$

Here, q_c and λ_c are the modularity and the weight term of the image community c, and q_{i_q} and λ_{i_q} are the terms for the query image i_q. $q_{c_q \cup c}$ and $\lambda_{c_q \cup c}$ are the modularity and the weight term for the community generated if i_q is placed in the community c. Since here $\mathcal{K} \ll N$, for the retrieval we used the ζ value for finer communities in Eqn. 6. From Eqn. 4, following the notations defined in Table 2, these terms can be expressed as:

$$\begin{aligned} \lambda_c &= 1 + \frac{2l_c}{n_c(n_c - 1)};\ \ q_c = \frac{l_c}{L + d^{i_q}} - \frac{1}{4}\Big(\frac{d_c}{L + d^{i_q}}\Big)^2 \\ \lambda_{c_q} &= 1 + \frac{2l_{c_q}}{n_{c_q}(n_{c_q} - 1)};\ \ q_{c_q} = \frac{l_{c_q}}{L + d^{i_q}} - \frac{1}{4}\Big(\frac{d_{c_q}}{L + d^{i_q}}\Big)^2 \end{aligned} \tag{11}$$

Here d^{i_q} means the degree of the query node, i_q. Similarly the terms $\lambda_{c_q \cup c}$ and $q_{c_q \cup c}$ can be expressed as:

$$\begin{aligned} \lambda_{c_q \cup c} &= 1 + \frac{2l_{c_q \cup c}}{n_{c_q \cup c}(n_{c_q \cup c} - 1)} \\ q_{c_q \cup c} &= \frac{l_{c_q \cup c}}{L + d^{i_q}} - \frac{1}{4}\Big(\frac{d_{c_q \cup c}}{L + d^{i_q}}\Big)^2 \end{aligned} \tag{12}$$

where,

$$
\begin{aligned}
n_{c_q \cup c} &= n_c + n_{c_q} \\
d_{c_q \cup c} &= d_c + d_{c_q} \\
l_{c_q \cup c} &= l_c + l_{c_q} + l^{ext}_{cc_q}
\end{aligned}
\tag{13}
$$

To find the closest community, we solve for c which generates the maximum increase of weighted modularity if the query image node i_q is placed in the community-c, i.e. $c = \underset{\tilde{c}}{\operatorname{arg\,max}} \Delta Q_{\mathcal{R}\cup\tilde{c}}$. We then extract the sub-network forming the community c, $\mathcal{G}_s(V_s, E_s)$ and confine our search space to select similar images to the nodes of the community-c, V_s.

To retrieve $\mathcal{K}$ similar images, we first assign each node of the sub-network into their own separate communities. So at the first step, we start from $|V_s|$ number of communities, where $|V_s|$ is the total number of nodes and $|E_s|$ is the total number of edges in $\mathcal{G}_s$. Then the increase in weighted modularity is calculated if any node is placed in the region, $\mathcal{R}$ with the query node, i_q. Finally, the node that generates the maximum increase in weighted modularity is placed in the region $\mathcal{R}$. Following the notations of Table 2, the gain in weighted modularity if node-j with community j is moved to $\mathcal{R}$ can be expressed as:

$$
\Delta Q^{\mathcal{G}_s}_{j\to\mathcal{R}} = \lambda_{j\to\mathcal{R}} \times q_{j\to\mathcal{R}} - [\lambda_{\mathcal{R}} \times q_{\mathcal{R}} + \lambda_j \times q_j] \tag{14}
$$

where $q_{j\to\mathcal{R}}$ and $\lambda_{j\to\mathcal{R}}$ are the modularity and the weight term for the new community generated if j is merged with $\mathcal{R}$. Let us assume that $l_{\mathcal{R}}$ denotes the number of edges within $\mathcal{R}$, $d_{\mathcal{R}}$ is the sum of degrees of the nodes in $\mathcal{R}$, d_j is the degree of node-j, and $l^{ext}_{j\to\mathcal{R}}$ denotes the total number of edges between node-j and the nodes in the region $\mathcal{R}$. Then the terms of Eqn.14 can be expressed as follows:

$$
\begin{aligned}
\lambda_j = 1;\;\; \lambda_{\mathcal{R}} &= 1 + \frac{2 l_{\mathcal{R}}}{|\mathcal{R}|(|\mathcal{R}| - 1)} \\
q_j &= -(d_j/2|E_s|)^2; \\
q_{\mathcal{R}} &= (l_{\mathcal{R}}/|E_s|) - (d_{\mathcal{R}}/2|E_s|)^2 \\
\lambda_{j\to\mathcal{R}} &= 1 + 2\frac{l_{\mathcal{R}} + l^{ext}_{j\to\mathcal{R}}}{|\mathcal{R}|(|\mathcal{R}| + 1)} \\
q_{j\to\mathcal{R}} &= \frac{l_{\mathcal{R}} + l^{ext}_{j\to\mathcal{R}}}{|E_s|} - \left(\frac{d_{\mathcal{R}} + d_j}{2|E_s|}\right)^2
\end{aligned}
\tag{15}
$$

At each iteration step, we solve for the image node u that generates the maximum weighted modularity gain if merged with $\mathcal{R}$, i.e. $u \leftarrow \underset{j}{\operatorname{arg\,max}} \Delta Q^{\mathcal{G}_s}_{j\to\mathcal{R}}$. This process is repeated and image nodes are merged repeatedly to $\mathcal{R}$. The process terminates when $\mathcal{K}$ image nodes are retrieved. If $|V_s| < \mathcal{K}$, then after retrieving the nodes of community c, the first step of the algorithm is repeated, i.e. the algorithm searches for the closest image community from the rest of the communities. This two step process is repeated until $\mathcal{K}$ image nodes are retrieved or no increase in weighted modularity is possible, i.e. no increment in weighted modularity is observed by merging any remaining nodes to $\mathcal{R}$. The images that belong to the region $\mathcal{R}$ is then returned as the retrieved images.

Table 4. Results of the area under receiver operating characteristics curves (AUC) for the neural network model trained with the NIH dataset.

Disease label	AUC_{LR}	AUC_{HR}
Atelectasis	0.79	0.82
Cardiomegaly	0.88	0.89
Consolidation	0.79	0.81
Edema	0.88	0.90
Pneumonia	0.71	0.73
Pneumothorax	0.76	0.83
Pleural Effusion	0.85	0.84
Infiltration	0.73	0.73
Emphysema	0.98	0.98
Pleural Thickening	0.79	0.82
Fibrosis	0.77	0.78
Nodule	0.80	0.83
Mass	0.84	0.82
Mean AUC	0.81	0.83

Table 5. Results of the area under receiver operating characteristics curves (AUC) for the neural network model trained with the Stanford dataset.

Disease label	AUC_{LR}	AUC_{HR}
Atelectasis	0.81	0.79
Enlarged Cardiomediastinum	0.61	0.68
Cardiomegaly	0.84	0.83
Consolidation	0.93	0.93
Edema	0.93	0.92
Pneumonia	0.76	0.71
Pneumothorax	0.80	0.90
Pleural Effusion	0.93	0.94
Lung Opacity	0.91	0.92
Mean AUC	0.84	0.88

3. Experiments

We report the retrieval performances with two settings. In the High Resolution (HR) setting, the chest X-ray images were of size 512×512, and in the Low Resolution (LR) setting the images were downsampled to the size of 224×224. The weights of the deep neural network model were initialized from a model pretrained on ImageNet dataset. We then trained the model on chest X-ray image datasets. For both datasets, the model was trained with a batch size of 32 for 20 epochs. For the NIH dataset, the initial learning was set as 10^{-3}, and for the Stanford dataset that was 10^{-4}. The initial learning rate was 10^{-3} and the learning rate was reduced by a factor of 10 each time a plateau was reached up to a learning rate of 10^{-8}. Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ was used to train the model. We calculated the area under the receiver operating characteristic (ROC) curves for each of the labels and the model with the highest mean area under ROC curve was picked as the best model for generating image codes. Note that the Stanford dataset has uncertain labels for one or more diseases in a large number of samples. Here we have included all the samples to train the model and ignored the uncertain labels while updating the gradient during training. During the retrieval the uncertain labels were counted as positive labels.

We report the performances of the proposed image retrieval system using all disease labels provided for both datasets. The labels for both datasets are generated from their corresponding radiologic reports. However, certain clinical information is also taken into account while generating the radiologic reports to predict the thoracic diseases, since some diseases have similar appearances on chest X-rays. Hence we also considered relabeling the diseases based on their appearances on a chest X-ray and reported the retrieval performances. Table 3 reports the disease labels considered for each of the datasets. For NIH-U and Stanford-U dataset setup, we considered consolidating all the opacity related labels under one label and the lesion related labels under one label.

We compared the performance of the proposed approach with six state-of-the-art literature-based approaches in terms of their normalized discounted cumulative gain. The Discounted Cumulative Gain (DCG) of a retrieval system retrieving $\mathcal{K}$ similar images is defined as:

$$DCG = \sum_{n=1}^{\mathcal{K}} \frac{2^{r_n} - 1}{\log(n+1)} \qquad (16)$$

where r_n is the graded relevance of the retrieved image at position-n with respect to the query image. The graded relevance is defined as the number of common positive labels shared by the images. The Normalized Discounted Cumulative Gain (nDCG) is the defined as the ratio of the DCG over the DCG that a perfect retrieval system would achieve given the database, as follows:

$$nDCG = \frac{DCG}{DCG_{ideal}} \qquad (17)$$

To have a fair comparison, we would like to compare with the reported results of previous methods directly by testing our approach on the same dataset. It is worth mentioning that, to our knowledge, there has been no reported retrieval result on the two large datasets studied in this work. Existing methods were evaluated at much smaller datasets. We therefore report the performances of three shallow learning based and three deep neural network-based methods using a small subset of the NIH chest X-ray dataset, for which the the performances of the compared methods in terms of their normalized discounted cumulative gain were reported in the literature (Chen et al., 2018). The dataset consists of 12,000 training images and 1,000 test images. For the CNN-based methods, the raw pixels were used as inputs; and for other deep learning methods, the 1024 dimensional GIST features were used as inputs. For hashing based methods, the bit sizes of different length were considered (16,3,48,64 bits), and we report the best performance that was achieved for each compared method.

We further report the performances of the proposed approach on the large scale datasets by investigating two literature based retrieval metrics that are used in medical image retrieval frameworks (Müller et al., 2004; Li et al., 2018). For a system retrieving $\mathcal{K}$ similar images, the Average Cumulative Gain (ACG) is defined as:

$$ACG = \frac{\sum_{n=1}^{\mathcal{K}} s_n}{\mathcal{K}} \qquad (18)$$

Table 6. Comparison of the retrieval performances of the proposed framework with literature-based approaches. The retrieval performances of the literature-based methods were reported in Chen et al. (2018) on a subset of NIH data.

Method	nDCG
Wang et al. (2012)	0.15
Gong et al. (2012)	0.16
Erin Liong et al. (2015)	0.19
Liu et al. (2016a)	0.17
Chen et al. (2018)	0.24
Lan et al. (2018)	0.15
Proposed Framework	0.31

where s_n is the graded similarity of the image retrieved at position n with respect to the query image. Here we defined s_n as the ratio of common positive labels between the retrieved image at n and the query image to the total positive labels in the query. The metric *Precision* is defined as the percentage of relevant images retrieved over the total number of retrieved images. The relevance of each retrieved image is assigned based on the existence of common positive disease labels between the query image and the retrieved images. If $\delta(\cdot) \in \{0, 1\}$ is an indicator function, the precision is defined as:

$$precision = \frac{\sum_{n=1}^{\mathcal{K}} \delta(r_n > 0)}{\mathcal{K}} \qquad (19)$$

4. Results

The performances of the deep neural network models trained on the X-ray datasets are reported here using the area under receiver operating characteristics curves (AUC). Table 4 reports the AUC for the model trained on NIH dataset, and Table 5 reports the AUC for the Stanford dataset. As reported here, we were able to achieve a mean AUC of 0.81 for the NIH dataset with the model trained on LR images, and that of 0.83 was achieved with the HR image trained model. For the Stanford dataset we also observed an increase in mean AUC with HR images (0.88) as opposed to the mean AUC with the LR images (0.84). Higher AUC values were achieved on the Cardiomegaly, Edema and Pleural Effusion labels for both datasets. For both datasets, Pneumonia was one of the low-performing labels in terms of AUC. This is partly because pneumonia is not entirely a radiologic diagnosis, rather it is a clinical diagnosis where other clinical information is incorporated to make a decision.

We then compare the proposed method with existing methods on a specific smaller dataset, citing their performances as reported in the literature. We report the performances of the proposed framework, together with six state-of-the-art literature-based methods, in Table 6 in terms of the normalized discounted cumulative gain. As can be seen from Table 6, the proposed approach performs better than other literature-based approaches and achieves an nDCG value of 0.31, outperforming the other methods reported.

We further report the performance of the proposed approach on the large scale datasets defined in Table 3 using two retrieval

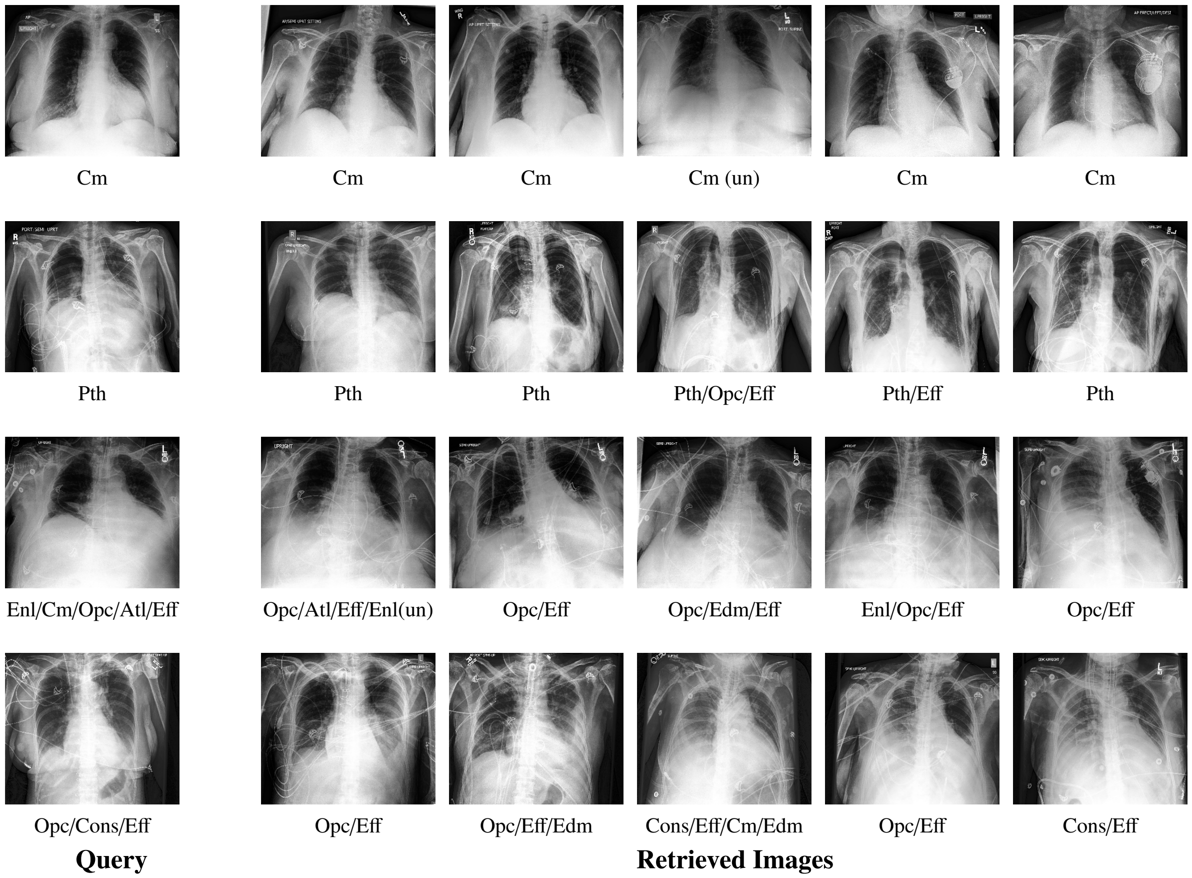

Fig. 4. Examples of the retrieval framework from the Stanford Dataset with nine disease labels. Each row represents the top-five retrieved images for the query image in the first column. The disease labels are enlisted at the bottom of each image using the following abbreviations– Cm:Cardiomegaly, Pth:Pneumothorax, Opc:Lung Opacity, Eff:Pleural Effusion, Enl:Enlarged Cardiomediastinum, Atl:Atelectasis, Edm:Edema, Cons:Consolidation, un: uncertain.

Table 7. Performances of the proposed image retrieval framework on top-10 retrieved images.

Dataset	ACG	Precision
NIH_{LR}	0.36	48%
NIH_{HR}	0.40	52%
NIH-U_{LR}	0.46	58%
NIH-U_{HR}	0.51	63%
$\text{Stanford}_{\text{LR}}$	0.38	76%
$\text{Stanford}_{\text{HR}}$	0.43	81%
$\text{Stanford-U}_{\text{LR}}$	0.53	81%
$\text{Stanford-U}_{\text{HR}}$	0.59	85%

metrics. Table 7 reports the performances in terms of ACG and precision. As it can be seen for the LR setting, for NIH dataset with all thirteen disease labels, we achieved an ACG value of 0.36 with a 48% precision. The retrieval performances are better for the Stanford dataset. When all nine disease labels were used, the framework was able to retrieve images where 76% of the retrieved images' disease labels matched exactly with the query image. With the HR setting, we were able to achieve better retrieval performances. For the NIH dataset the ACG increased to 0.40 and the precision increased to 52%. For the Stanford dataset, the HR setting resulted in an ACG value of 0.43 with a 81% precision. In both LR and HR settings, we observed a better retrieval performance on the Stanford dataset. The superior performance of the proposed framework on the Stanford dataset might be related to the quality of labels in these two datasets. The labeler used for the Stanford dataset to extract the disease labels from the associated radiologic reports is shown to outperform the NIH labeler, and hence the disease labels in Stanford dataset are less noisy. However, in this work, we used the disease labels generated by the NIH labeler for the NIH and NIH-U datasets to report the performance metrics.

When we combined the disease labels based on their similar appearances, the retrieval performances improved for both the NIH and the Stanford datasets. With NIH-U labeling, we observe more than 20% increase in precision and ACG, for both LR and HR settings. For Stanford dataset, combining similar labels results in an ACG of 0.53 with the LR setting, as opposed to 0.38 when the labels generated from an automated labeler are used. The precision also improves to 81%. With the HR setting, ACG of 0.59 was achieved with consolidated labels as opposed to 0.43 with the labeler generated labels. The retrieval also increased to 85% with the consolidated labels. The improvements in performances indicate the potential of the method on the disease labels that can be diagnosed based on image contents only.

Fig.4 shows examples of the retrieval results on the Stanford dataset. The first image of each row is the query image and the rest of the images are the top-five retrieved images. The positive disease labels are shown at the bottom of each image. The first two example shows the retrieved images for a query image with a single positive label, and the last two query images have multiple disease labels. As can be seen, all the positive disease labels can be retrieved from the top few retrieved images.

5. Discussion and Conclusion

In this work, we have presented a large scale medical image retrieval framework and reported its performances on the two largest available chest X-ray image datasets. The proposed framework consists of a deep neural network-based image code generator trained on the medical image dataset. A graph network is then formed based on code-similarity values, and a graph community detection scheme is applied to form similar image communities. Although developed for chest X-ray image retrieval, the proposed approach is generally applicable to other medical image-based retrieval tasks.

The deep community model was able to distinguish the radiologic disease labels, while performance on clinical diagnosis-based disease labels indicates the necessity of incorporating clinical information in the decision making process. Nevertheless, the proposed framework is shown to outperform other state-of-the-art approaches for content-based medical image retrieval.

We observed better performance of the proposed framework on the Stanford dataset. The disease labels used in this work were generated from the associated radiological reports using automatic rule-based labeler. However, the accuracy of labels provided with the NIH data is challenged in the literature (Irvin et al., 2019), and hence the disease labels in the Stanford dataset are potentially less noisy. Therefore we also observed a better retrieval performance on the Stanford dataset.

When both radiologic diagnosis-based and clinical diagnosis-based disease labels were used, the performance across diseases vary in terms of classification AUCs, as seen from Tables 4 and 5, which is consistent with other studies on these datasets (Wang et al., 2017; Allaouzi and Ahmed, 2019). One of the reasons of such variation originates from the fact that a few of the thoracic diseases have similar appearances in X-ray film. For example, the disease labels– Consolidation, Infiltration and Lung Opacity all appear as clouds or opacities in an X-ray film. The labels used in both of these datasets came from radiology reports, which also incorporates clinical information to distinguish between the diseases. However, in this work, we only considered the image content to extract similar images. Since we did not incorporate clinical information in the current implementation, we observed that the performance on some of the clinical labels were lower. One such example is Pneumonia, which is a clinical disease, as opposed to an image finding. We observed that when only image content is considered, the performance is among the lowest among disease labels. However, further improvement in the retrieval performance was observed when diseases with similar appearances in chest X-rays are consolidated under one label. In the absence of clinical information, annotating the X-rays solely based on image appearances can thereby improve the efficiency of the proposed content-based image retrieval framework.

One possible limitation of this work is the exclusion of metadata and other clinical information from the framework. The framework currently only considers image contents to extract similar cases from the database. However, while content-based image descriptors and similarity measures can help bridge the gap between visual content and semantic features, image-only approaches have a limited resolving ability for the clinical decision process (Akgül et al., 2011). Medical diagnosis and prognosis is a complex process that can be benefit from a retrieval framework that also incorporates clinical metadata, patient history and other relevant information along with the medical images. Another limitation is the unbalanced number of positive samples across different diseases in the datasets. The Stanford dataset also has uncertain labels. Although we addressed this unbalance by incorporating a weight term in the loss function, nonetheless a balanced dataset could help in learning disease signatures. Nevertheless, the promising performance of the proposed image content based framework indicates its potential as a large scale image retrieval tool and paves its way for future developments.

The proposed framework targets to extract image codes from a deep neural network model and generates image similarity network from the image codes. The similarity network is then used to divide the database images into similar image communities. This approach is different from the geometric deep learning or graph convolutional network based learning approaches (Kipf and Welling, 2017; Hamilton et al., 2017; Bronstein et al., 2017). Geometric deep learning attempts to generalize deep neural models to non-Euclidean domains, such as graphs, where the goal is to learn a function of features on a graph, where each node is associated with a set of of attributes or features. These attributes or features are modelled as signals on the nodes of the graph. In our current implementation, we extracted similar image communities on a traditional graph, and while forming communities, no node-level attributes were considered. In future we plan to extend the current work to incorporate the clinical information associated with the image nodes by formulating the image retrieval task as a geometric learning problem.

With the availability of the growing number of images in hospitals along with their associated reports, the proposed framework for large scale retrieval can be a powerful tool to guide the diagnosis.

References

Akgül, C.B., Rubin, D.L., Napel, S., Beaulieu, C.F., Greenspan, H., Acar, B., 2011. Content-based image retrieval in radiology: current status and future directions. Journal of digital imaging 24, 208–222.

Allaouzi, I., Ahmed, M.B., 2019. A novel approach for multi-label chest x-ray classification of common thorax diseases. IEEE Access 7, 64279–64288.

Anavi, Y., Kogan, I., Gelbart, E., Geva, O., Greenspan, H., 2015. A comparative study for chest radiograph image retrieval using binary texture and deep learning classification, in: 2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC), IEEE. pp. 2940–2943.

Bastawrous, S., Carney, B., 2017. Improving patient safety: avoiding unread imaging exams in the national va enterprise electronic health record. Journal of digital imaging 30, 309–313.

Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment 2008, P10008.

Bronstein, M.M., Bruna, J., LeCun, Y., Szlam, A., Vandergheynst, P., 2017. Geometric deep learning: going beyond euclidean data. IEEE Signal Processing Magazine 34, 18–42.

Chen, Z., Cai, R., Lu, J., Feng, J., Zhou, J., 2018. Order-sensitive deep hashing for multimorbidity medical image retrieval, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 620–628.

Conjeti, S., Roy, A.G., Katouzian, A., Navab, N., 2017. Hashing with residual networks for image retrieval, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 541–549.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee. pp. 248–255.

Erin Liong, V., Lu, J., Wang, G., Moulin, P., Zhou, J., 2015. Deep hashing for compact binary codes learning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2475–2483.

Folio, L.R., 2012. Chest imaging: an algorithmic approach to learning. Springer Science & Business Media.

Gong, Y., Lazebnik, S., Gordo, A., Perronnin, F., 2012. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence 35, 2916–2929.

Hamilton, W., Ying, Z., Leskovec, J., 2017. Inductive representation learning on large graphs, in: Advances in neural information processing systems, pp. 1024–1034.

Haq, N.F., Moradi, M., Wang, Z.J., 2019. Community structure detection from networks with weighted modularity. Pattern Recognition Letters 122, 14–22.

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L.H., Aerts, H.J., 2018. Artificial intelligence in radiology. Nature Reviews Cancer 18, 500.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al., 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. arXiv preprint arXiv:1901.07031 .

Kipf, T.N., Welling, M., 2017. Semi-supervised classification with graph convolutional networks, in: International Conference on Learning Representations (ICLR).

Lai, H., Pan, Y., Liu, Y., Yan, S., 2015. Simultaneous feature learning and hash coding with deep neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3270–3278.

Lan, R., Zhong, S., Liu, Z., Shi, Z., Luo, X., 2018. A simple texture feature for retrieval of medical images. Multimedia Tools and Applications 77, 10853–10866.

Lee, C.Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z., 2015. Deeply-supervised nets, in: Artificial intelligence and statistics, pp. 562–570.

Li, Z., Zhang, X., Müller, H., Zhang, S., 2018. Large-scale retrieval for medical image analytics: A comprehensive review. Medical image analysis 43, 66–84.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Medical image analysis 42, 60–88.

Liu, H., Wang, R., Shan, S., Chen, X., 2016a. Deep supervised hashing for fast image retrieval, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2064–2072.

Liu, X., Tizhoosh, H.R., Kofman, J., 2016b. Generating binary tags for fast medical image retrieval based on convolutional nets and radon transform, in: 2016 International Joint Conference on Neural Networks (IJCNN), IEEE. pp. 2872–2878.

Müller, H., Michoux, N., Bandon, D., Geissbuhler, A., 2004. A review of content-based image retrieval systems in medical applications—clinical benefits and future directions. International journal of medical informatics 73, 1–23.

Newman, M.E., Girvan, M., 2004. Finding and evaluating community structure in networks. Physical review E 69, 026113.

Qayyum, A., Anwar, S.M., Awais, M., Majid, M., 2017. Medical image retrieval using deep convolutional neural network. Neurocomputing 266, 8–20.

Quellec, G., Lamard, M., Cazuguel, G., Cochener, B., Roux, C., 2011. Fast wavelet-based image characterization for highly adaptive image retrieval. IEEE Transactions on Image Processing 21, 1613–1623.

Rahman, M.M., Antani, S.K., Thoma, G.R., 2011. A learning-based similarity fusion and filtering approach for biomedical image retrieval using svm classification and relevance feedback. IEEE Transactions on Information Technology in Biomedicine 15, 640–646.

Rimmer, A., 2017. Radiologist shortage leaves patient care at risk, warns royal college. Bmj 359, j4683.

Shah, A., Conjeti, S., Navab, N., Katouzian, A., 2016. Deeply learnt hashing forests for content based image retrieval in prostate MR images, in: Medical Imaging, SPIE. p. 978414.

Torralba, A., Fergus, R., Weiss, Y., et al., 2008. Small codes and large image databases for recognition., in: CVPR, Citeseer. p. 2.

Wang, J., Kumar, S., Chang, S.F., 2012. Semi-supervised hashing for large-scale search. IEEE Transactions on Pattern Analysis and Machine Intelligence 34, 2393–2406.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M., 2017. 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 computer vision and pattern recognition, pp. 2097–2106.

Zhang, S., Metaxas, D., 2016. Large-scale medical image analytics: Recent methodologies, applications and future directions. Medical Image Analysis 33, 98 – 101.

Zhang, X., Liu, W., Dundar, M., Badve, S., Zhang, S., 2014. Towards large-scale histopathological image analysis: Hashing-based image retrieval. IEEE Transactions on Medical Imaging 34, 496–506.