CHEST PATHOLOGY DETECTION USING DEEP LEARNING WITH NON-MEDICAL TRAINING

Yaniv Bar[1], Idit Diamant[2], Lior Wolf [1], Hayit Greenspan[2]

[1] The Blavatnik School of Computer Science, Tel-Aviv University, Tel Aviv 69978, Israel
[2] Department of Biomedical Engineering, Tel-Aviv University, Tel Aviv 69978, Israel

ABSTRACT

In this work, we examine the strength of deep learning approaches for pathology detection in chest radiographs. Convolutional neural networks (CNN) deep architecture classification approaches have gained popularity due to their ability to learn mid and high level image representations. We explore the ability of CNN learned from a non-medical dataset to identify different types of pathologies in chest x-rays. We tested our algorithm on a 433 image dataset. The best performance was achieved using CNN and GIST features. We obtained an area under curve (AUC) of 0.87-0.94 for the different pathologies. The results demonstrate the feasibility of detecting pathology in chest x-rays using deep learning approaches based on non-medical learning. This is a first-of-its-kind experiment that shows that Deep learning with ImageNet, a large scale non-medical image database may be a good substitute to domain specific representations, which are yet to be available, for general medical image recognition tasks.

Index Terms *— Chest Radiography, Computer-Aided Diagnosis Disease Categorization, Deep Learning, Deep Networks, CNN.*

1. INTRODUCTION

Chest radiographs are the most common examination in radiology. They are essential for the management of various diseases associated with high mortality and display a wide range of potential information, many of which is subtle. Most of the research in computer-aided detection and diagnosis in chest radiography has focused on lung nodule detection. Although the target of most research attention, lung nodules are a relatively rare finding in the lungs. The most common findings in chest X-rays include lung infiltrates, catheters and abnormalities of the size or contour of the heart [1]. Fig. 1 shows examples of healthy and pathological chest X-rays. Distinguishing the various chest pathologies is a difficult task even to the human observer. Therefore, there is an interest in developing computer system diagnosis to assist radiologists in reading chest images.

Initial studies on chest pathology detection in radiographs can be found in the literature [2, 3, 4]. In [2] the healthy versus pathology detection in chest radiography was explored using Local Binary Patterns (LBP) [5]. Avni et al. [3, 4] used the Bag-of-Visual-Words (BoVW) model [6] to discriminate between healthy and four pathological cases.

Deep neural networks have recently gained considerable interest due to the development of new variants of CNNs and the advent of efficient parallel solvers optimized for modern GPUs. Deep learning refers to machine learning models such as Convolutional Neural Networks (CNNs) that represent mid-level and high-level abstractions obtained from raw data (e.g. images) [7]. Recent results indicate that the generic descriptors extracted from CNNs are extremely effective in object recognition, and are currently the leading technology [8, 9]. Deep learning methods are most effective when applied on large training sets. In the medical field such large datasets are usually not available. Initial studies can be found in the medical field that uses deep architecture methods [10, 11]. However, we are not aware of any works that use generic, non-medical, training sets in order to address a medical imaging task. Moreover, we are not aware of any deep architecture methods for the specific task of pathology detection in chest radiographs.

In our work, we utilize the strength of deep learning approaches in a wide range of chest-related diseases. We also explore categorization of healthy versus pathology which is an important screening task. We empirically explore the use of CNNs for these tasks with a particular focus on pre-trained CNN that is learned from a large scale real-life and non-medical image database. We later show that categorization rates can be slightly improved by combining features extracted from CNN and common low-level visual features that are optimal for the task of object categorization.

2. METHODS

CNNs constitute a feed-forward family of deep networks, where intermediate layers receive as input the features generated by the former layer, and pass their outputs to the next layer. The strength of a deep network is in learning hierarchical layers of concept representation, corresponding to different levels of abstraction. For visual data, the low levels of abstraction might describe the different orientated edges in the image; middle levels might describe parts of an object while high layers refer to larger object parts and even the object itself. In this work, we tested the deep learning network capabilities in chest pathology detection

2.1. Pre-trained CNN model using ImageNet

We focus on the Decaf pre-trained CNN model [12], an adaptation of a CNN which closely follows the CNN which was constructed by Krizhevsky et al. [13], with the exception of small differences

in the input data and the cancelation of the split of the network into two pathways. The CNNs in [12, 13] were learned over a subset of images from ImageNet [14], a comprehensive real-life large scale image database (>20M) that is arranged according to concepts/categories (>10K). Specifically, [12] learned a CNN on more than one million images that are categorized into 1000 categories, which is illustrated in Figure 2.

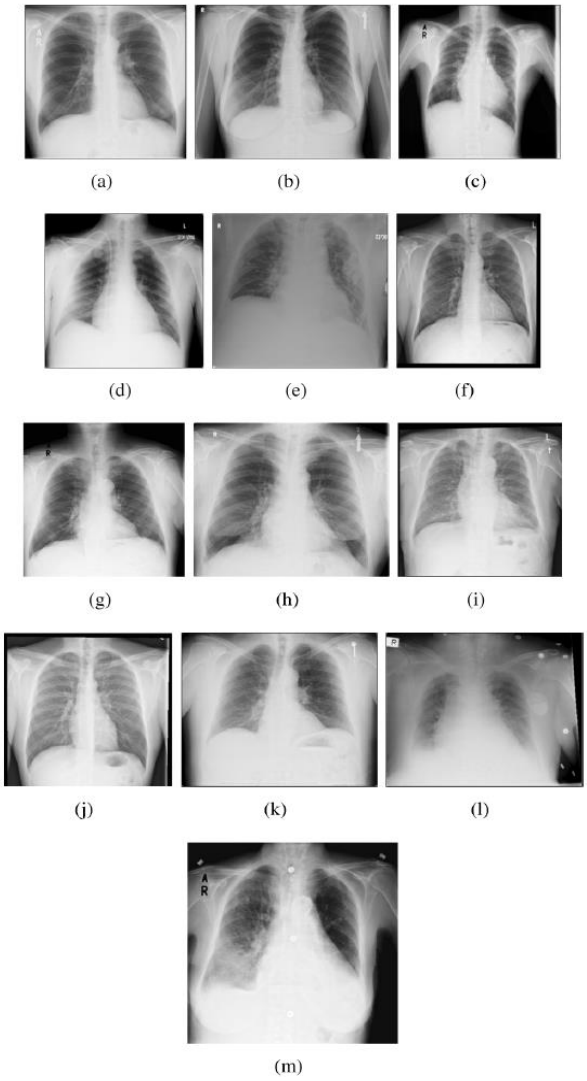


Fig. 1. Chest x-rays categories examples: (a)-(c) healthy; (d)-(f) enlarged heart (cardiomegaly); (g)-(i) enlarged mediastinum; (j)-(l) left or right effusion; (m) multiple pathologies: enlarged heart and mediastinum, left and right effusion.

Using the notation of [12] to denote the activations of the *n*-th hidden layer of the obtained network as *Decaf*$_n$, the 5th layer (*Decaf*$_5$) and 6th layer (*Decaf*$_6$) and 7th layer (*Decaf*$_7$) features were extracted and defined as descriptors. *Decaf*$_5$ denotes the last convolutional layer and is the first set of activations that has been fully propagated through the convolutional layers of the network and *Decaf*$_6$ denotes the first fully-connected layer.

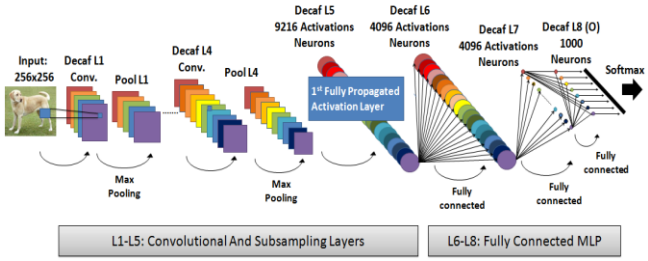


Fig. 2. An illustration of the architecture of the CNN used by [12].

2.2. Selecting and Combining feature sets

We tested several common descriptors that are known in the literature, including GIST [15] and Bag-of-Visual-Words (BoVW). The GIST descriptor [15] is derived by resizing an image to 128 x 128 and iterating over different scales (4 scales in our case) where for each scale the image is divided into 8x8 cells. For each cell, orientation (every 45 degrees), color and intensity histograms are extracted, and the descriptor is a concatenation of all histograms, for all scales and cells.

BoVW is a state-of-the-art method that has been previously tested for this specific task [3, 4]. The BoVW [6] image representation is adapted from the bag-of-words (BoW) representation of text documents. Therefore, to represent an image using BoVW model, it must be treated as a document, which means that the image is treated as a distribution of visual elements. We thus need to find these visual elements and discretize their space in order to create the visual word dictionary. Once we generated the visual word dictionary, an image can be represented as a histogram of visual word occurrences based on the collection of its local descriptors. We implemented the visual words model [3, 4] as follows: (1) extracting patches from each training image, (2) Applying Principal Component Analysis (PCA) for dimensionality reduction to reduce noise level and computational complexity, (3) adding the patch center coordinates to the feature vector. This introduces spatial information into the image representation, without the need to explicitly model the spatial dependency between patches, (4) clustering of all patches using K-means into representative visual words generating a dictionary. K-means is a common clustering method which clusters the input feature vectors into K groups where their centers used as the visual words which build the dictionary. A given (training or testing) image can now be represented by a unique distribution over the generated dictionary of words. Empirically, we found that using 7 PCA components of variance-normalized raw patches and using a dictionary size of 1000 words results in the best classification performance.

All features used in our work were normalized: each feature across all images has its mean subtracted, and is divided by its standard deviation. Following normalization, we applied fusion of different feature groups. Empirically we claim that fusing CNN intermediate layers features with GIST features captures the salient information that allows a more accurate categorization of our problem. We implemented a fusion approach of the different features by averaging the class probabilities obtained for each feature group. The algorithm flowchart is described in Figure 3.

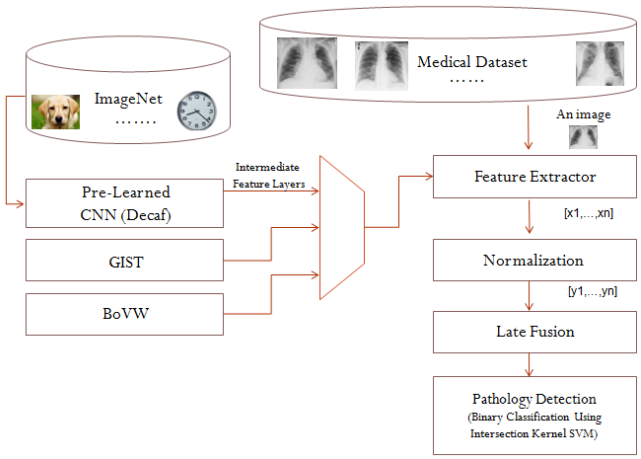


Fig. 3. Algorithm flowchart.

3. EXPERIMENTS AND RESULTS

3.1. Data

Our dataset consists of 443 frontal chest x-ray images (DICOM format). The images were acquired from the Diagnostic Imaging Department of Sheba Medical Center (Tel-Hashomer, Israel). Two radiologists interpreted the X-rays, and this served as the reference gold standard. The radiologists examined all of the images independently. They then discussed and reached a consensus regarding the label of every image. For each image and pathology type, a positive or negative label was assigned. The images depict 3 chest pathology conditions: Right Pleural Effusion (44 images), Cardiomegaly (99 images) and Abnormal Mediastinum (110 images). Overall, the dataset contains 219 images with at least one pathology condition. The digitized images were cropped and centered.

3.2. Experimental Results

We started with a binary categorization task, per pathology. Classification was performed using a Support Vector Machine (SVM) classifier with leave-one-out-cross-validation method. For each binary categorization task, cases diagnosed with the examined pathology were labeled as positive cases, while cases that weren't diagnosed with this pathology were labeled as negative cases. We investigated the different linear and nonlinear kernels (linear, polynomial and RBF) using standard grid-search technique, and empirically selected the efficient non-linear intersection kernel. Three accuracy measures were examined: sensitivity, specificity and the area under the ROC curve (AUC). Sensitivity and Specificity are derived based on the optimal cut point on the ROC – the point on the curve closest to (0,1).

Tables 1-3 present the experimental results. We report *Decaf5 and Decaf6* baseline descriptors (*Decaf7* gave less significant results). We note the boost in performance following the introduction of deep architecture descriptors. For all cases, the deep architecture descriptors outperform the GIST descriptor. Another improvement was gained by applying late fusion on the baseline descriptors: *Decaf5*, *Decaf6* and GIST. In almost all cases the fused descriptor outperforms the deep architecture single-layered descriptors. In several of the cases a clear improvement over the BoVW descriptor is shown.

Figure 4 shows comparative ROC curve analysis performed on our dataset using a leave-one-out-cross-validation method. It can be seen that our fused method matches or outperforms all other tested methods.

			Deep Learning		Late Fusion
Descriptor	**GIST**	**BoVW**	**L5**	**L6**	**GIST+L5+L6**
Right Pleural Effusion	0.87	0.89	0.94	0.90	0.92
Cardiomegaly	0.92	0.94	0.93	0.91	0.94
Mediastinum	0.83	0.86	0.87	0.87	0.88
Healthy vs Pathology	0.84	0.88	0.86	0.86	0.87

Table 1. AUC accuracy metric.

			Deep Learning		Late Fusion
Descriptor	**GIST**	**BoVW**	**L5**	**L6**	**GIST+L5+L6**
Right Pleural Effusion	0.82	0.84	0.89	0.89	0.86
Cardiomegaly	0.84	0.89	0.85	0.87	0.89
Mediastinum	0.79	0.80	0.81	0.74	0.80
Healthy vs Pathology	0.83	0.80	0.80	0.86	0.84

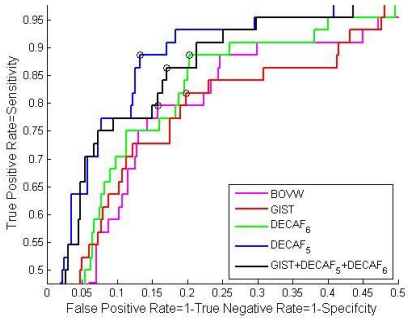
Table 2. Sensitivity accuracy metric.

			Deep Learning		Late Fusion
Descriptor	**GIST**	**BoVW**	**L5**	**L6**	**GIST+L5+L6**
Right Pleural Effusion	0.80	0.85	0.87	0.80	0.83
Cardiomegaly	0.82	0.87	0.89	0.82	0.86
Mediastinum	0.78	0.80	0.82	0.84	0.85
Healthy vs Pathology	0.72	0.79	0.79	0.72	0.78

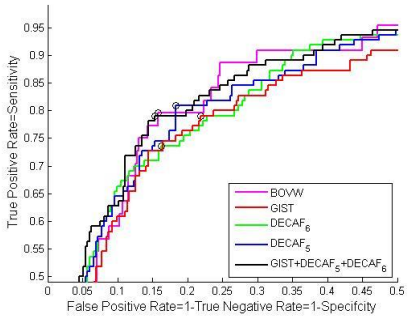
Table 3. Specificity accuracy metric.

4. CONCLUSIONS

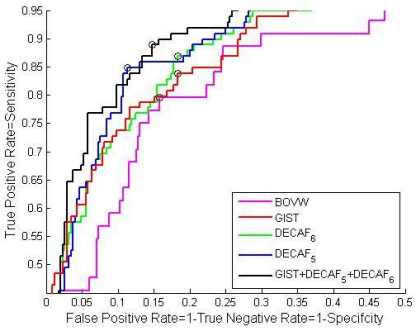
In conclusion, in this work we present a system for the medical application of chest pathology detection in radiograph images which uses CNN that is learned from a non-medical dataset (ImageNet). Unlike previous work on using pre-trained CNNs as a feature extraction method [8], in our case *Decaf5* baseline descriptor is the leading representation. This representation alone is an effective off-the-shelf descriptor for chest x-ray retrieval tasks. We have demonstrated that this result can be improved by fusing the baseline descriptors of *Decaf5*, *Decaf6* and GIST, assuming that the combination captures information that eludes each one of the descriptors alone. Future work entails further tuning of the CNN with actual x-ray data. We believe such tuning may augment the CNN performance even further. Our results demonstrate the feasibility of detecting pathology in chest x-ray using deep learning approaches based on non-medical learning. This is a general approach that is also applicable to other medical classification tasks.



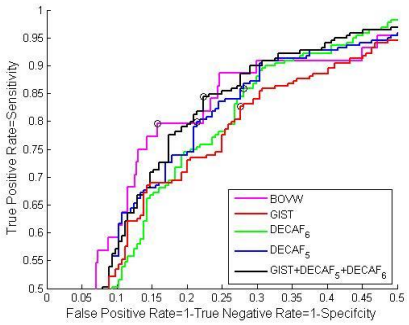
Right Pleural Effusion Detection.



Abnormal Mediastinum Detection.



Cardiomegaly Detection,



Healthy vs. Pathology Detection.

Fig. 4. ROCs of different examined pathologies.

Acknowledgments. We thank our colleagues from the Diagnostic Imaging Department of Sheba Medical Center, Tel Hashomer, Israel - for the data collection used in this work.

5. REFERENCES

[1] B. van Ginneken, L. Hogeweg, and M. Prokop, "Computer-aided diagnosis in chest radiography: Beyond nodules," European Journal of Radiology, vol. 72, no. 2, pp. 226-230, 2009.

[2] J.M. Carrillo-de-Gea, G. García-Mateos, "Detection of Normality / Pathology on Chest Radiographs using LBP," In Bioinformatics, pp. 167-172, 2010

[3] U. Avni, H. Greenspan, E. Konen, M. Sharon, and J. Goldberger, "X-ray categorization and retrieval on the organ and pathology level, using patch-based visual words," IEEE Transactions on Medical Imaging, vol. 30, no. 3, pp. 733-746, 2011.

[4] U. Avni, H. Greenspan, and J. Goldberger, "X-ray categorization and spatial localization of chest pathologies," Medical Image Computing and Computer-Assisted Intervention–MICCAI, Springer Berlin Heidelberg, pp. 199-206, 2011.

[5] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," Pattern recognition, vol. 29, no, 1, pp. 51-59, 1996.

[6] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," Workshop on statistical learning in computer vision, ECCV, vol. 1, no. 1-22, pp. 1-2, 2004.

[7] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS), pp. 253-256, 2010.

[8] A.S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN Features off-the-shelf: an Astounding Baseline for Recognition," IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 512-519, 2014.

[9] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1717-1724, 2014.

[10] A. Prasoon, K. Petersen, C. Igel, F. Lauze, E. Dam, and M. Nielsen, "Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network," Medical Image Computing and Computer-Assisted Intervention–MICCAI, Springer Berlin Heidelberg , pp. 246-253, 2013.

[11] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," Medical Image Computing and Computer-Assisted Intervention–MICCAI, Springer Berlin Heidelberg, pp. 411-418, 2013.

[12] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. "Decaf: A deep convolutional activation feature for generic visual recognition," arXiv preprint arXiv:1310.1531, 2013.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems, pp. 1097-1105, 2012.

[14] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248-255, 2009.

[15] A. Oliva, and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," International journal of computer vision, vol. 42, no. 3, pp. 145-175, 2001.