Deep learning with non-medical training used for chest pathology identification

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 radiograph data. 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 a CNN to identify different types of pathologies in chest x-ray images. Moreover, since very large training sets are generally not available in the medical domain, we explore the feasibility of using a deep learning approach based on *non-medical* learning. We tested our algorithm on a dataset of 93 images. We use a CNN that was trained with *ImageNet*, a well-known large scale non-medical image database. The best performance was achieved using a combination of features extracted from the CNN and a set of low-level features. We obtained an area under curve (AUC) of 0.93 for Right Pleural Effusion detection, 0.89 for Enlarged heart detection and 0.79 for classification between healthy and abnormal chest x-ray, where all pathologies are combined into one large class. This is a first-of-its-kind experiment that shows that deep learning with large scale non-medical image databases may be sufficient for general medical image recognition tasks.

Keywords: Deep learning, classification, chest x-rays, convolutional neural networks.

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]. 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.

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) [2]. Recent results indicate that the generic descriptors extracted from CNNs are extremely effective in object recognition, and are currently the leading technology [3,4]. 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 use deep architecture methods [5,6]. 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 this work, we examine the strength of deep learning approaches for pathology detection in chest radiograph data. We also explore categorization of healthy versus pathology which is an important screening task. In our experiments we explore the possibility to use convolutional neural networks (CNNs) that are learned from ImageNet, a large scale non-medical image database, for the task of medical image analysis.

Medical Imaging 2015: Computer-Aided Diagnosis, edited by Lubomir M. Hadjiiski, Georgia D. Tourassi,
Proc. of SPIE Vol. 9414, 94140V · © 2015 SPIE · CCC code: 1605-7422/15/$18 · doi: 10.1117/12.2083124

2. METHODS

The strength of deep networks is in learning multiple layers of concept representation, corresponding to different levels of abstraction. For visual datasets, the low levels of abstraction might describe edges in the image, while high layers in the network refer to object parts and even the category of the object viewed.

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. Two popular choices are CNNs suggested by [7] and [8] for the Large Scale Visual Recognition Challenge of ImageNet [9]. ImageNet is a comprehensive real-life non-medical large scale image database consisting of approximately 15 Million images with more than 20,000 categories (e.g. musical instrument, tool, fruit). The CNNs described in [7] and [8] are constructed of a few layers that learn convolutions, interleaved with non-linear and pooling operations, followed by locally or fully connected layers. The features extracted from intermediate layers of these networks produce highly discriminative features that achieve state of the art performance in visual classification tasks [3].

In this work we tested the deep learning networks capabilities in chest pathology detection. We extracted several different descriptors and compared among them. Our main descriptor is extracted using the Decaf implementation of a CNN [10] which follows the CNN in [7]. The CNN in [10] was trained over a subset of images from ImageNet of more than one million images that are categorized into 1000 categories. Using the notation of [10] to denote the activations of the *n*-th hidden layer of the obtained network as $Decaf_n$, the 5th layer ($Decaf_5$), 6th layer ($Decaf_6$) and 7th layer ($Decaf_7$) features were extracted. $DeCAF_5$ contains 9216 activations of the last convolutional layer and is the first set of activations that has been fully propagated through the convolutional layers of the network, $DeCAF_6$ contains 4096 activations of the first fully-connected layer and $DeCAF_7$ denotes features taken from the final hidden layer - i.e., just before propagating through the final fully connected layer to produce the class predictions. Figure 1 illustrates a schematic view of the Decaf implementation of the CNN in [10].

A second baseline descriptor covered in this work is the "Picture Codes" (PiCoDes) descriptor [9]. PiCoDes is a compact high level representation of popular low-level features (SIFTs, GIST, PHOG, and SSIM) which is optimized over a subset of the ImageNet dataset containing approximately 70,000 images. For PiCoDes, a preliminary offline stage is performed which constructs a classification basis consisting of kernels that are learned over the low-level visual features obtained from ImageNet subset. PiCodes then uses this classification basis to define a recognition model for object category recognition by transforming an image data using the classification basis such that the entries in the image descriptor are thresholded projections of the low-level visual features extracted from the image. This encoding schema yields a binary image descriptor with high performance rates on object category recognition. As a benchmark for our approach we have tested several common descriptors. These include Local Binary Patterns (LBP) [12] and GIST [13]. The GIST descriptor, initially proposed for scene recognition [13], is derived by the concatenation of orientation, color and intensity histograms over different scales and cells segmentation.

The algorithm flowchart is described in Figure 2. Classification was performed using SVM with linear kernel using leave-one-out-cross-validation. Three accuracy measurements 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 – that is the point on the curve closest to (0,1). For all features except binary ones, values are standardized: each column has its mean subtracted, and is divided by its standard deviation.

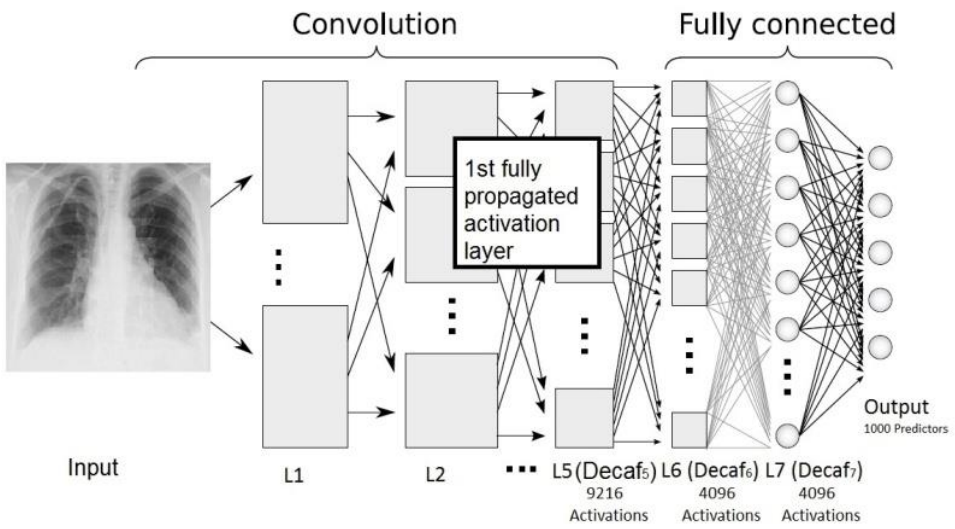


Figure 1: CNN schematic view of the Decaf CNN described in [11].

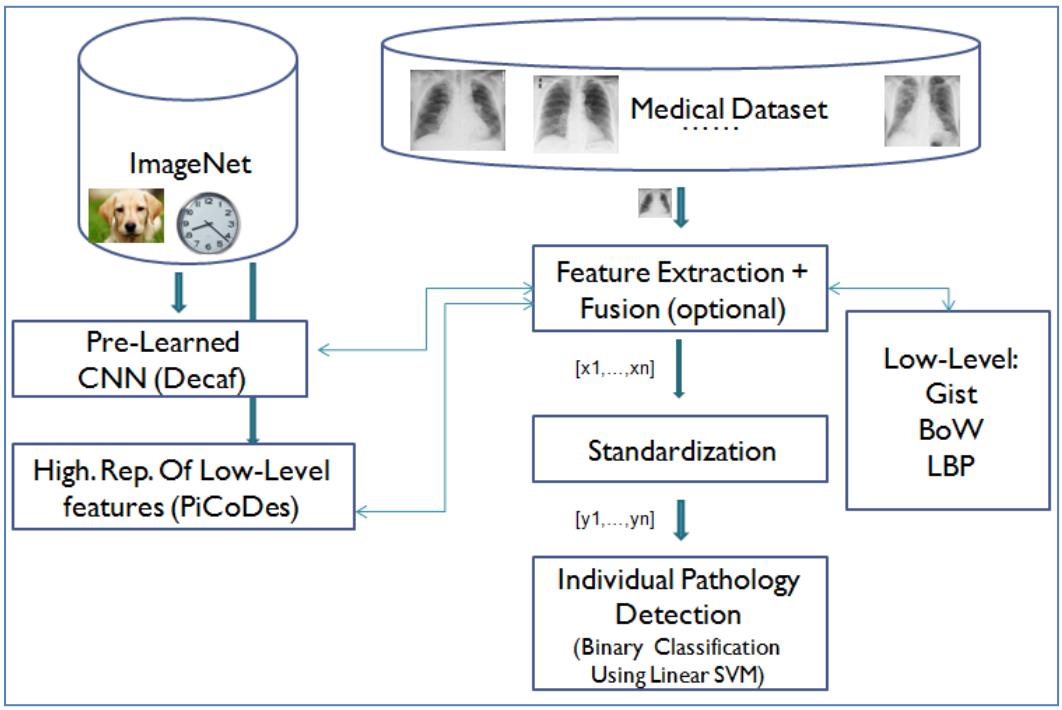


Figure 2: Algorithm flowchart.

3. EXPERIMENTS AND RESULTS

Our Dataset consists of 93 frontal chest x-ray images that were acquired from Sheba Medical Center. The digitized images are of variable size. They are cropped, centered and contain several artifacts such as reading directives (e.g. arrows, left/right indicators) and medical equipment. X-ray interpretations were conducted by two radiologists. First, 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 2 chest pathology conditions including enlarged heart (69 images) and right pleural effusion (79 images) as well as a healthy chest x-rays (56 images). Figure 3 shows examples of healthy and pathological chest X-rays.

Tables 1-3 present the experimental results. We note the boost in performance following the introduction of deep architecture descriptors. For all cases the deep architecture descriptors outperforms all low-level descriptors and matches or outperform the high level PiCoDes representation. Another improvement is gained by applying fusion on both baseline descriptors, $DeCAF_5$ and PiCoDes. This is done by concatenating the baseline descriptors before the classification is taking place. Empirically, we claim that fusing these two baseline descriptors ($Decaf_5$ and PiCodes) captures the salient information that allows a more accurate separation between an individual pathology set and its complement set. It seems that due to the different nature of both baseline descriptors, the combination of the two captures information that eludes each one of the descriptors alone. Figure 4 depicts a comparative ROC curves analysis for all examined conditions. It is evident that our fused method outperforms all other tested methods.

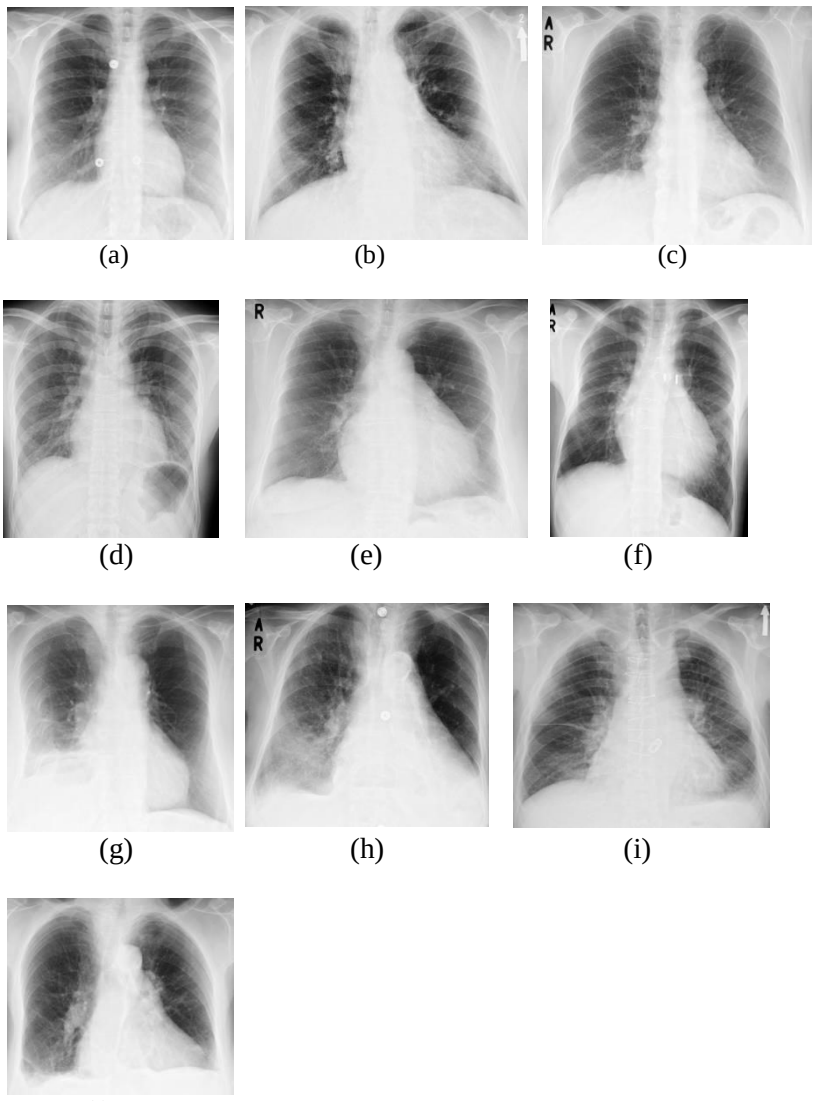


Figure 3: Chest x-rays categories examples: (a)-(c) healthy; (d)-(f) enlarged heart; (g)-(i) right effusion; (j) multiple pathologies: enlarged heart and right effusion.

Table 1. Right Pleural Effusion Condition.

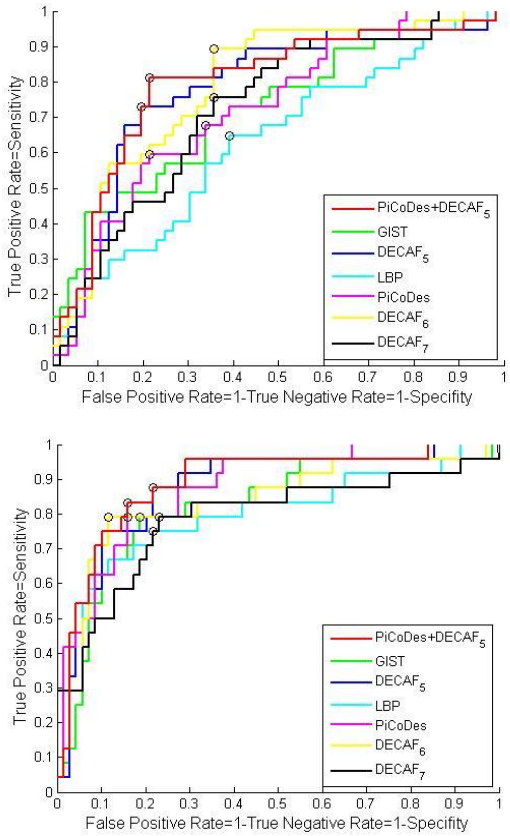
	Low Level		High Level	Deep			Fusion
	LBP	GIST	PiCoDes	Decaf L5	Decaf L6	Decaf L7	PiCoDes+Decaf L5
Sensitivity	0.71	0.79	0.79	0.93	0.86	0.86	**0.93**
Specificity	0.77	0.92	0.91	0.84	0.86	0.80	**0.84**
AUC	0.75	0.93	0.91	0.92	0.91	0.84	**0.93**

Table 2. Healthy vs. Pathology.

	Low Level		High Level	Deep			Fusion
	LBP	GIST	PiCoDes	Decaf L5	Decaf L6	Decaf L7	PiCoDes+Decaf L5
Sensitivity	0.65	0.68	0.59	0.73	0.89	0.76	**0.81**
Specificity	0.61	0.66	0.79	0.80	0.64	0.64	**0.79**
AUC	0.63	0.72	0.72	0.78	0.79	0.72	**0.79**

Table 3. Enlarged Heart Condition.

	Low Level		High Level	Deep			Fusion
	LBP	GIST	PiCoDes	Decaf L5	Decaf L6	Decaf L7	PiCoDes+Decaf L5
Sensitivity	0.75	0.79	0.79	0.88	0.79	0.79	**0.83**
Specificity	0.78	0.81	0.84	0.78	0.88	0.77	**0.84**
AUC	0.80	0.82	0.87	0.87	0.84	0.79	**0.89**



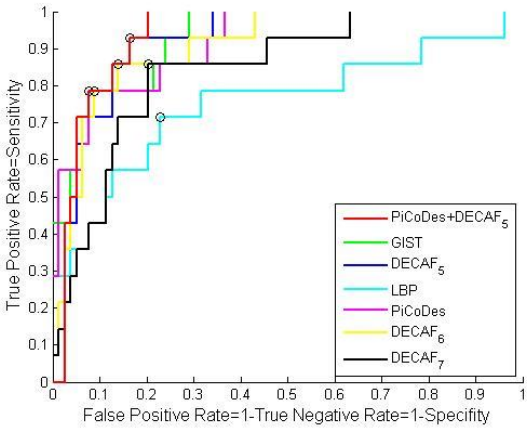


Figure 4: (Top) Healthy vs. Pathology condition ROC; (Middle) Enlarged heart condition ROC; (Bottom) Right pleural effusion condition ROC.

4. DISCUSSION AND CONCLUSIONS

In this work we present a system for medical application of chest pathology detection in x-rays which uses convolutional neural networks that are learned from a non-medical archive (ImageNet). We show that a combination of deep learning (Decaf) and PiCodes features achieves the best performance. Additionally we show that Decaf layer 5 achieves better results compared to layers 6 and 7. Our results demonstrate the feasibility of detecting pathology in chest x-ray using deep learning approaches based on non-medical learning. Deep learning methods have not been tested for chest pathology detection for our knowledge, especially not with non-medical archive learning. This is a first-of-its-kind experiment that shows that Deep learning with ImageNet training may be sufficient for general medical image recognition tasks.

REFERENCES

[1] Ginneken, B., Hogeweg, L., and Prokop, M., "Computer-aided diagnosis in chest radiography: Beyond nodules," Eur. J. Radiol., 72(2), 226–230, (2009).

[2] LeCun, Y., Koray K., and Clément F., "Convolutional networks and applications in vision," In International Symposium on Circuits and Systems (ISCAS), 253-256, (2010).

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

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

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

[6] Ciresan, D. C., Giusti, A., Gambardella, L. M., and Schmidhuber, J., "Mitosis Detection in Breast Cancer Histology Images using Deep Neural Networks," In Medical Image Computing and Computer-Assisted Intervention–MICCAI, Springer Berlin Heidelberg, (2013).

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

[8] Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y., "Overfeat: Integrated recognition, localization and detection using convolutional networks," arXiv preprint arXiv:1312.6229, (2013).

[9] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., and Fei-Fei, L., "Imagenet: A large-scale hierarchical image database," In: IEEE Conference on Computer Vision and Pattern Recognition, 248-255, (2009).
[10] Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T., "Decaf: A deep convolutional activation feature for generic visual recognition," arXiv preprint arXiv:1310.1531, (2013).
[11] Bergamo, A., Torresani, L., and Fitzgibbon, A.W., "Picodes: Learning a compact code for novel-category recognition," In: Advances in Neural Information Processing Systems, 2088-2096, (2011).
[12] Ojala, T., Pietikäinen, and M., Harwood, D., "A comparative study of texture measures with classification based on featured distributions," Pattern recognition, 29(1), 51-59, (1996).
[13] Oliva, A., Torralba, A., "Modeling the shape of the scene: A holistic representation of the spatial envelope," International journal of computer vision, 42(3), 145-175, (2001).