Taylor & Francis
Taylor & Francis Group

Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization

ISSN: 2168-1163 (Print) 2168-1171 (Online) Journal homepage: http://www.tandfonline.com/loi/tciv20

Chest pathology identification using deep feature selection with non-medical training

Yaniv Bar, Idit Diamant, Lior Wolf, Sivan Lieberman, Eli Konen & Hayit Greenspan

To cite this article: Yaniv Bar, Idit Diamant, Lior Wolf, Sivan Lieberman, Eli Konen & Hayit Greenspan (2018) Chest pathology identification using deep feature selection with non-medical training, Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 6:3, 259-263, DOI: 10.1080/21681163.2016.1138324

To link to this article: https://doi.org/10.1080/21681163.2016.1138324

Published online: 16 May 2016.

COMPUTER METHODS IN BIOMECHANICS AND BIOMEDICAL ENGINEERING: IMAGING & VISUALIZATION, 2018
VOL. 6, NO. 3, 259–263
https://doi.org/10.1080/21681163.2016.1138324

Chest pathology identification using deep feature selection with non-medical training

Yaniv Bar[a], Idit Diamant[b], Lior Wolf[a], Sivan Lieberman[c], Eli Konen[c] and Hayit Greenspan[b]

[a]The Blavatnik School of Computer Science, Tel Aviv University, Tel-Aviv, Israel; [b]Department of Biomedical Engineering, Tel-Aviv University, Tel Aviv, Israel; [c]Diagnostic Imaging Department, Sheba Medical Center, Ramat Gan, Israel

ABSTRACT

We demonstrate the feasibility of detecting pathology in chest X-rays using deep learning approaches based on non-medical learning. Convolutional neural networks (CNN) learn higher level image representations. In this work, we explore the features extracted from layers of the CNN along with a set of classical features, including GIST and bag-of-words. We show results of classification using each feature set as well as fusing among the features. Finally, we perform feature selection on the collection of features to show the most informative feature set for the task. Results of 0.78–0.95 AUC for various pathologies are shown on a dataset of more than 600 radiographs. This study shows the strength and robustness of the CNN features. We conclude that deep learning with large-scale non- medical image databases may be a good substitute, or addition to domain-specific representations which are yet to be available for general medical image recognition tasks.

ARTICLE HISTORY
Received 21 November 2015
Accepted 2 January 2016

KEYWORDS
Radiography; chest X-rays; computer-aided diagnosis; deep learning; CNN; feature selection

1. Introduction

Radiographs are the most common examination in radiology, with approximately 2B procedures per year. Chest X-rays are a major part of the procedures. They are essential for the management of various diseases associated with high mortality and display a wide range of potential information, much of which is subtle. Distinguishing the various chest pathologies is a difficult task even to the human observer. Figure 1 shows examples of healthy and pathological chest X-rays where a given chest X-ray might contain multiple pathologies. Examining the pleural effusion case – lungs usually appear very dark on an X-ray image because they contain mostly air which allows the X-rays through very easily. If the lungs start accumulating fluid or pus in its surrounding, fewer of the X-rays will make it through to the film and those areas appear whiter.

Although chest X-ray is the most common imaging examination, in many radiology departments the lack of radiologists and the increasingly growing workload leads to inability to read all the X-rays by radiologists. Therefore, there is an interest in developing computer system diagnosis to assist radiologists in reading chest images. An automated software which will be reliable in determining which X-ray is suspicious would help to improve the quality of diagnostic imaging service. This can be a very attractive application for many markets in the western world – to improve accuracy and efficiency in the radiology departments, and in the third-world countries (e.g. China) – where no substantial radiology service is available.

To date, most of the research in computer-aided detection and diagnosis in chest X-rays is task specific and has focused on lung nodule detection (van Ginneken et al. 2009). 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 (van Ginneken et al. 2009). Works can be found which are task specific and use methods such as lung segmentation, suppression of ribs and location of typical lung textures (Detection of Emphysema, Diagnosis of interstitial lung diseases). Initial studies on chest pathology detection and classification as an image-level labelling can be found in the literature (de Gea et al. 2010; Avni et al. 2011). In de Gea et al., (2010) the healthy versus pathology detection in chest radiography was explored using Local Binary Patterns. Avni et al. (2011) used the Bag-of-Visual-Words (BoVW) model (Csurka et al. 2004) to discriminate between healthy and four pathological cases. BoVW methodology has proven strong in ImageClef competitions (http://www.imageclef.org): categorising X-rays on organ level.

Deep Learning is a class of machine learning techniques, where many layers of information processing stages in hierarchical supervised architectures are exploited for feature learning and for pattern analysis/classification. The essence of deep learning is to compute hierarchical features or representations of the observed data, where the higher level features or factors are defined from lower level ones (Deng et al. 2013). Deep learning, or the use of deep (i.e. many-layered) convolutional neural networks (CNNs) for machine recognition and classification, is advancing the limits of performance in domains as varied as computer vision, speech and text (Dean et al. 2012). Recent results indicate that the generic descriptors extracted from CNNs are extremely effective in object recognition (Oquab et al. 2014), and they are currently the leading recognition technology.

In this work, we utilise the strength of deep learning approaches in a wide range of chest-related diseases and in addition we explore categorisation of healthy versus pathology which is an important screening task. We empirically explore the use of a pre-trained *Decaf* CNN (Donahue et al. 2013) that is

CONTACT Hayit Greenspan hayit@eng.tau.ac.il

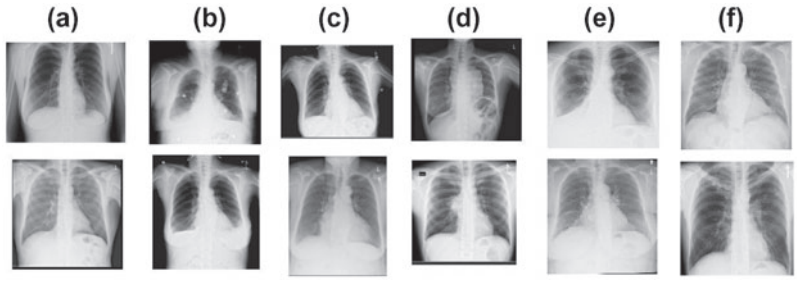

Figure 1. Chest X-rays categories examples: (a) healthy, (b) left or right effusion, (c) enlarged heart (cardiomegaly), (d) enlarged mediastinum, (e) left or right consolidation, (f) multiple pathologies: enlarged heart, mediastinum, left and right effusion and left or right consolidation (source: Diagnostic Imaging Department, Sheba Medical Center, Tel Hashomer, Israel)

learned from ImageNet, a large-scale real-life and non-medical image database (Deng et al. 2009). Deep learning methods are most effective when applied on large training sets. However, since such large data are generally not available in the medical domain, we explore the feasibility of using a deep learning approach based on non-medical learning. We also apply a feature selection technique to the deep learning representations and show that it can improve the performance in our task.

Preliminary studies can be found in the medical field that use deep architecture methods (e.g. Cireşan et al. 2013; Li et al. 2014). In our earlier work (Bar et al. 2015), we showed initial results on three pathologies in a small image data-set. We used a leave-one-out cross-validation method and explored the combination of low-level features (GIST) with features extracted from deep architecture network. In the current work, we have an augmented data-set that includes six pathologies. We explore feature selection on a large set of features – to investigate what features are the more informative for the task. Finally, we provide more clinically relevant results using a separate training and test set scenario.

2. Methods

2.1. Extracting deep features from a pre-trained CNN model

In this work, we tested the *Decaf* pre-trained CNN model (Donahue et al. 2013) which closely follows the popular CNN model which was constructed by Krizhevsky et al. (2012) with the exception of a minor difference in the cancellation of the split of the network into two computational pathways. Both CNNs were learned over a subset of images from ImageNet (Deng et al. 2009), a comprehensive real-life large-scale image database (>20M) that is arranged into more than 20K non-medical concepts/categories (e.g. musical instrument, fruit). More specifically, the CNN of Donahue et al. (2013) was learned on more than 1M images that are categorised into 1K categories and was constructed of a few layers that learn convolutions, interleaved with non-linear and pooling operations, followed by locally or fully connected layers and an output layer. An illustration is provided in Figure 2.

The strength of deep networks is in learning multiple layers of conceptual representations, corresponding to different levels of abstraction, where the low levels of abstraction might describe edges of an image, while high layers in the network refer to object parts or the category of the object in view. 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.

Using the notation of Donahue et al. (2013) to denote the activations of the *n*-th hidden layer of the obtained network, the fifth layer (*Decaf* 5), sixth layer (*Decaf* 6) and the seventh 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. An illustration of feature extraction from a network is provided in Figure 3.

2.2. Combining feature sets and deep feature selection

In addition to network features, we use a set of classical descriptors that is known in the literature to perform well in image classification/categorisation tasks. These include GIST (Oliva et al. 2001) features and BoVW (Csurka et al. 2004). We combine the features using a linear weighted fusion. In this method, the [support vector machine (SVM)] classifier outputs a probability for each class, and the probabilities are combined linearly using an averaging rule.

In addition to exploring possible combinations of features, we explore feature selection as a means for improved representation. Feature selection, a process of selecting a subset of original features according to certain criteria, is an important and frequently used dimensionality reduction technique for data mining. Reducing data dimensionality is achieved by removing the irrelevant, and often, redundant features. In the past few decades, researchers have developed large amount of feature selection algorithms. These algorithms are designed to serve different purposes, are of different models, and all have their own advantages and disadvantages. A repository of feature selection algorithms can be found in Zhao et al. (2010) work.

In the current work, we use the Kruskal–Wallis feature selection method (Hollander et al. 2013), which is a non-parametric analysis method that makes no assumptions about the distribution of the data (e.g. normality). Kruskal–Wallis method ranks all features across all groups together by comparing the feature medians. The method tests, across all data samples, if two features have equal median and produces the *P*-value. If the value of *P*-value is close to '0' it is an evidence against the null hypothesis (feature medians are equal). A strong separation of the medians indicates that the feature under consideration contains discriminative information which means that it has a high clustering

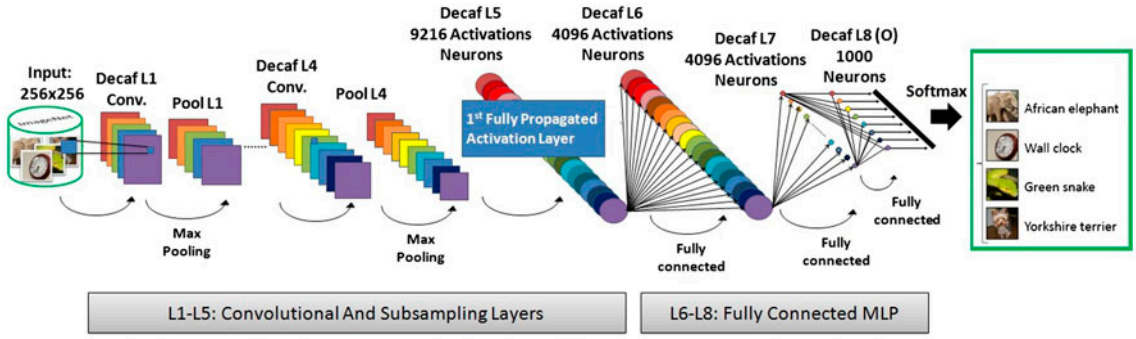

Figure 2. A schematic illustration of Donahue et al. (2013) CNN architecture and training process.

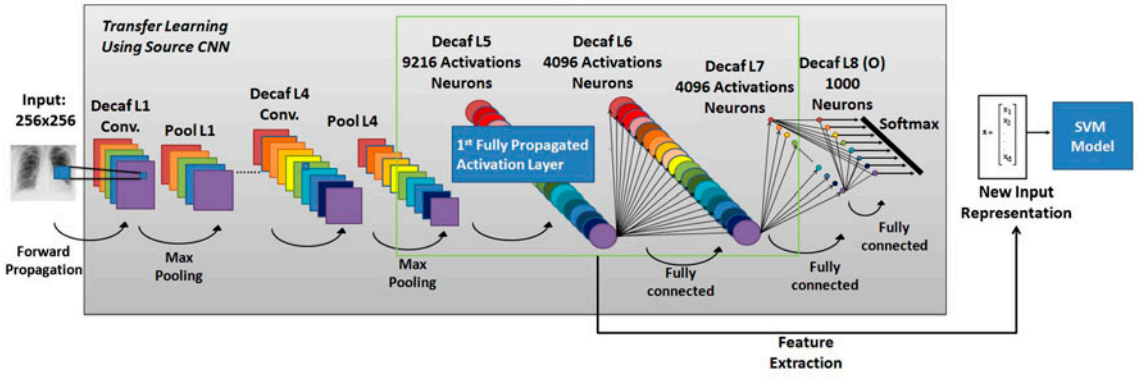

Figure 3. Feature extraction from Donahue et al. (2013) pre-trained CNN.

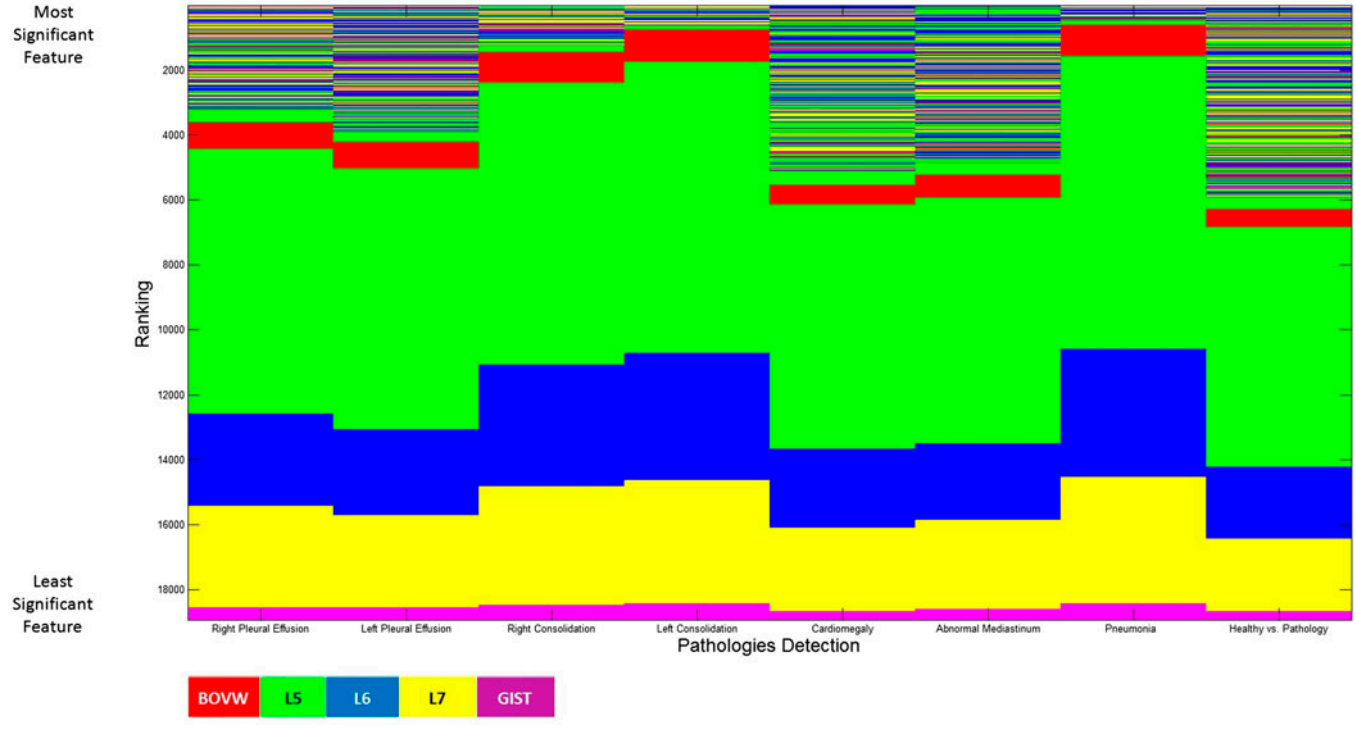

Figure 4. Features ranking, evaluated on the testing set, applied on the combined feature vector [BoVW, Deep features (*Decaf* 5, *Decaf* 6 and *Decaf* 7), Gist]. Ranking is performed on all examined identification cases. We use the following abbreviation: Ln for *Decafn*.

Table 1. AUC accuracy metric classification performance.

Descriptor	Right pleural effusion	Left pleural effusion	Right consolidation	Left consolidation	Cardiomegaly	Abnormal mediastinum	Healthy vs. pathology
GIST	0.85	0.79	0.77	0.41	0.96	0.73	0.88
BoVW	0.89	0.87	0.78	0.65	0.94	0.74	0.85
L5	0.91	0.81	0.80	0.75	0.95	0.79	0.90
L6	0.91	0.82	0.85	0.76	0.94	0.80	0.90
L7	0.90	0.79	0.75	0.79	0.93	0.79	0.89
L5 + L6 + GIST	0.92	0.82	0.83	0.68	0.96	0.79	0.91
L5+L6+L7	0.92	0.82	0.83	0.78	0.94	0.80	0.91
FS (5000)	0.93	0.82	0.84	0.78	0.95	0.80	0.92

Note: We use the following abbreviation: Ln for *Decafn*, + for classifiers fusion, FS for feature selection method.

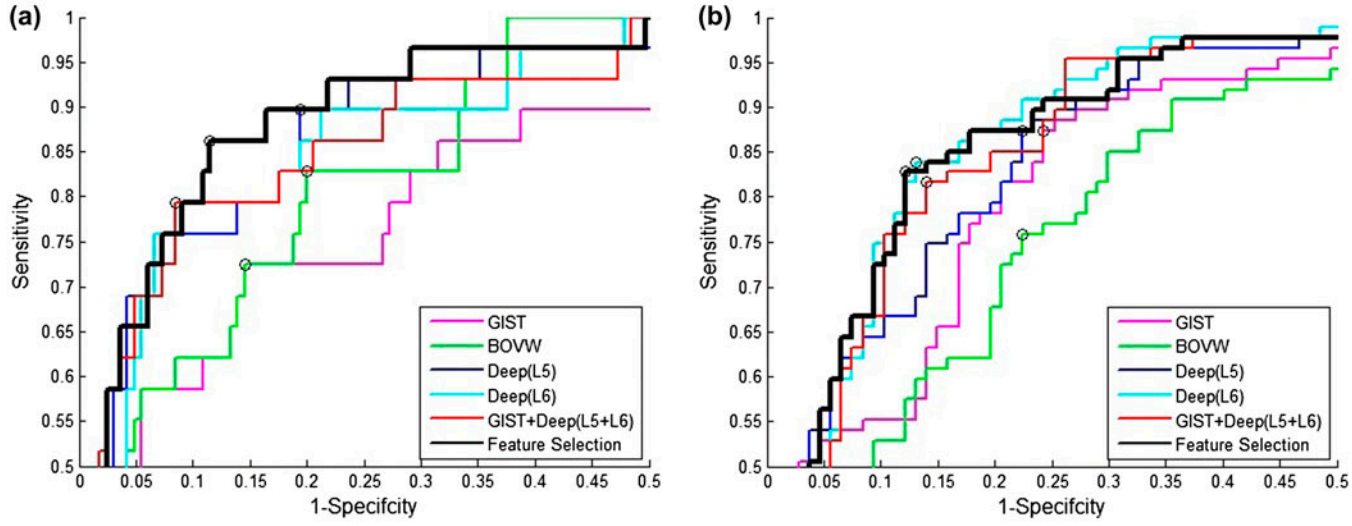

Figure 5. ROCs of different examined pathologies using Deep feature selection. We use the following abbreviation: Ln for *Decafn*, + for classifiers fusion, FS for feature selection method. (a) Right pleural effusion and (b) healthy vs. pathology.

power. Consequently, features with discriminative information are selected.

We have applied this feature selection method to extract the top 5000 most significant components out of the entire 18,920 features that we have: GIST features, BoVW features, *Decaf* 5, *Decaf* 6, *Decaf* 7, representing low-level to high-level features. The number of components was tested empirically until optimised.

3. Experiments and results

3.1. Data

Our data-set consists of 637 frontal chest X-ray images in DICOM format. The images are of variable size. They are cropped, centred and contain several artefacts such as reading directives (e.g. arrows, left/right indicators) and medical equipment but otherwise were not preprocessed (e.g. equalisation). We have replicated the intensity channel to support the CNN three-channel RGB input data expectations. The pertained CNN resizes the images into specific accepted resolution automatically.

The images were collected from the Diagnostic Imaging Department of Sheba Medical Center, Tel Hashomer, Israel. Gold standard was achieved using image interpretation done by two expert radiologists. The radiologists examined all of the images independently and then reached a decision regarding the label of every image. For each image and pathology type, a positive or negative label was assigned.

The images depict six chest pathology conditions: Right Pleural Effusion (73 images), Left Pleural Effusion (74 images), Right Consolidation (58 images), Left Consolidation (45 images), Cardiomegaly (154 images) and Abnormal Mediastinum (145 images). Overall, the data-set contains 325 images with at least one pathology condition. We split the data randomly into a training set of 443 images and an independent testing set of 194 images, reflecting a 70–30 split.

3.2. Experimental results

Kruskal–Wallis ranking of features was performed on the entire augmented 18,920 features, representing low-level to high-level features: GIST features (512 features), BoVW features (1000 features) and *Decaf* 5 (9216 features), *Decaf* 6 (4096 features) and *Decaf* 7 (4096 features) CNN intermediate layer features. The ranking, illustrated in Figure 4, reveals that most of the significant features are obtained from the Deep network and consists of a mix of features from the various intermediate layers. We also note that starting from a certain position within the hierarchy, typically following thousands of mixed features, the features are organised in the hierarchy without meaningful ordering, and ordering reflects the augmented order of features from which selection and ranking is performed: [BoVW, Deep features (*Decaf* 5, *Decaf* 6 and *Decaf* 7), Gist]. Overall, the features selected for the categorisation task include around 60% of the *Decaf* 5

layer features , 20% of the *Decaf* 6 layer features and 7% of the *Decaf* 7 layer features.

We next conduct a performance comparison across the feature sets used, on a set of binary categorisation tasks, task per pathology. For each task, cases diagnosed with the examined pathology were labelled as positive cases, while cases that were not diagnosed with this pathology were labelled as negative cases. Classification was performed using a SVM with a non-linear intersection kernel. We build a model from the training set and evaluate it on the testing set. Accuracy measures that were examined include: 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).

Table 1 presents the experimental results. Using deep architecture descriptors (L5,L6 and L7 layer features) provides results that are either comparable to or superior to the classical GIST and BoVW descriptors, across all pathologies. A closer look at the results emphasises the fluctuate nature of the results on different pathologies. In a few cases, GIST and BoVW stand out while in other pathologies single-layer network features or fused features achieve the higher results. Looking at the Feature selection (FS) results, where the top 5000 most significant features were extracted, we note robustness in performance across all pathologies, with AUC between 0.78 and 0.95. Moreover, although the number of components (5000) was tested empirically until optimised, convergence of the AUC metric to similar results within an accuracy error range of 0–3% started from approximately 1000 features.

Figure 5 shows comparative ROC curve analysis on the test set for two pathologies. In the healthy versus pathology scenario (a), an important screening task, we obtain an AUC of 0.92 with corresponding 0.83 Sensitivity and 0.88 Specificity. For the case of Right pleural effusion (b) we obtain 0.86 Sensitivity and 0.88 Specificity, respectively.

4. Discussion and conclusion

In this work, we show that by adopting the CNN representation to tackle medical image classification of various pathologies we can achieve excellent results which surpass relatively recent state of the art results using BoVW. We explore features for the chest pathology categorisation task. We use both classical as well as CNN-based features. We show results for categorisation based on each feature set independently, as well as the fusion across several classifiers and the result of using a feature selection stage prior to the categorisation task. We conclude that the most informative feature set consists of a selection of features from the CNN layers. Using this selected set of features gives very strong performance across all pathologies as well as for screening (healthy vs. pathology). Intuitively, one could argue that the learned weights which constitute the deep feature layers are optimised to the images of the CNN training dataset and the task it is trained for, thus one could imagine the optimal representation for each problem lies at an intermediate layer of the CNN but without knowing which layer in advance. Feature selection algorithms can assist us in this problem by picking the most significant Deep features from the different layers, in an automated and robust process, while preserving and augmenting the classification performance. This is a general approach that may be applicable to other medical classification tasks.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Dr. Greenspan Lab is funded in part for Deep Learning in Medical Imaging by the INTEL Collaborative Research Institute for Computational Intelligence (ICRI-CI).

ORCID

Yaniv Bar http://orcid.org/0000-0002-9538-0119

References

Avni U, Greenspan H, Konen E, Sharon M, Goldberger J. 2011. X-ray categorization and retrieval on the organ and pathology level, using patch-based visual words. IEEE Trans Med Imaging. 30:733–746.

Bar Y, Diamant I, Wolf L, Lieberman S, Konen E, Greenspan H. 2015. Chest pathology detection using deep learning with non-medical training. In: 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI). Brooklyn (NY): IEEE; p. 294–297.

Carrillo-de-Gea JM, García-Mateos G. 2010. Detection of normality/pathology on chest radiographs using lbp. In: 2010 Proceedings of the First International Conference on Bioinformatics, Valencia, Spain. p. 167–172.

Ciresan DC, Giusti A, Gambardella LM, Schmidhuber J. 2013. Mitosis detection in breast cancer histology images with deep neural networks. In: 2013 Medical Image Computing and Computer-Assisted Intervention (MICCAI)Springer: Berlin; p. 411–418.

Csurka G, Dance C, Fan L, Willamowski J, Bray C. 2004. Visual categorization with bags of keypoints. In: 2004 Workshop on statistical learning in computer vision, ECCV, Prague, Czech Republic. vol. 1; p. 1–2.

Dean J, Corrado G, Monga R, Chen K, Devin M, Mao M, Senior A, Tucker P, Yang K, Le QV, Ng AY. 2012. Large scale distributed deep networks. In: Advances in Neural Information Processing Systems 25 (NIPS 2012). Lake Tahoe, NV.12231231

Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. 2009. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Miami (FL): IEEE; p. 248–255.

Deng L, Yu D. 2014. Deep learning: methods and applications. Foundations Trends Signal Process 7:197–387. http://dx.doi.org/10.1561/2000000039

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

Hollander M, Wolfe DA, Chicken E. 2013. Nonparametric statistical methods. New York: John Wiley & Sons. Print..

Krizhevsky A, Sutskever I, Hinton GE. 2012. Imagenet classification with deep convolutional neural networks. In: 2012 Advances in Neural Information Processing Systems 25 (NIPS), Lake Tahoe, NV, USA. p. 1097–1105.

Li R, Zhang W, Suk HI, Wang L, Li J, Shen D, Ji S. 2014. Deep learning based imaging data completion for improved brain disease diagnosis. In: 2014 Medical Image Computing and Computer-assisted Intervention (MICCAI). Berlin: Springer; p. 305–312.

Oliva A, Torralba A. 2001. Modeling the shape of the scene: a holistic representation of the spatial envelope. IJCV. 42:145–175.

Oquab M, Bottou L, Laptev I, Sivic J. 2014. Learning and transferring mid-level image representations using convolutional neural networks. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus (OH): IEEE; p. 1717–1724.

van Ginneken B, Hogeweg L, Prokop M. 2009. Computer-aided diagnosis in chest radiography: beyond nodules. Eur J Radiol. 72:226–230.

Zhao Z, Morstatter F, Sharma S, Alelyani S, Anand A, Liu H. 2010. Advancing feature selection research. ASU feature selection repository, 2010 Technical Report; p. 1–28.