Using Deep Learning for Pulmonary Nodule Detection & Diagnosis

Full Paper

Ross Gruetzemacher
Auburn University
rossgritz@gmail.com

Ashish Gupta
Auburn University
ashishg4@gmail.com

Abstract

This study uses a revolutionary image recognition method, deep learning, for the classification of potentially malignant pulmonary nodules. Deep learning is based on deep neural networks. We report results of our initial findings and compare performance of deep neural nets using a combination of different network topologies and optimization parameters. Classification accuracy, sensitivity and specificity of the network performance are assessed for each of the four topologies.

Keywords: deep learning, deep neural networks, lung cancer, big data, analytics

Introduction

Lung cancer is one of the most aggressive cancers and is projected by the American Cancer Society to result in mortality of over 70% with approximately 225,000 Americans newly diagnosed in 2016 (Society, A. C. 2016). It is responsible for roughly one quarter of all cancer deaths. The early identification of pulmonary nodules is an important task for the management of lung cancer; if detected early malignant nodules can be treated with a much higher success rate. The National Lung Screening Trial has demonstrated that frequent screening using low-dose Computed Tomography (CT) is effective at reducing mortality from lung cancer.

However, reading of CT scans by radiologists for detecting the presence of pulmonary nodules and their malignancy is a tedious and time-consuming task. Due to the high stakes associated with false negatives, it is desirable to have CT scans read by multiple readers. In order to improve workflow and reduce workload for radiologists charged with detecting and diagnosing lung cancer, several computer-aided detection (CADe) and computer-aided diagnosis (CADx) tools have been developed by researchers in industry and academia.

CADe/x systems are tools intended to help radiologists in managing workflow, thus, high numbers of false positives are undesirable. Several such CADe/x tools rely on using traditional analytics approaches such as segmentation-based techniques for the detection of pulmonary nodules. Some of the more successful of these techniques have used successive k-means classifiers for localization and segmentation of images (Gurcan, et al. 2002; Murphy, et al. 2009), although good results have been documented with alternate methods of segmentation and classification (Messay, et al. 2010) as well. These findings typically report having sensitivity that ranges from 80% to 85% with between 3 to 5 false positives per scan on an average. Most recently, a study reported sensitivity of 94% but with a higher false positive rate of 7 per scan (Firmino, et al. 2016).

We use a state-of-art framework, *Deep Learning*, and apply it to improve the identification of malignant pulmonary nodules. *Deep learning* is based on using '*deep*' neural networks (DNNs) comprised of a large number of hidden layers. This approach has emerged over the past several years as the preferred method for a variety of complex pattern recognition tasks. Research on using DNNs for CADe is in nascent stages. However, initial studies that have explored the efficacy of applying deep learning have demonstrated a very low number of false positives when compared with typical results reported by deploying traditional segmentation techniques (Tajbakhsh, et al. 2015). Furthermore, such studies have indicated that DNNs have great potential for application in a variety of CADe tasks involving volumetric medical data (Tajbakhsh, et al. 2015; Van Ginneken, et al. 2015). Two of these have explored the use convolutional neural networks for classification of pulmonary nodules (Kumar, et al. 2015; Shen, et al. 2015), but the use of extensive '*deep*' neural networks in this scope have only been demonstrated with off-the-shelf, pre-trained networks (Van Ginneken, et al. 2015).

The objective of this preliminary work is to explore the effectiveness of using DNNs to distinguish between large and small pulmonary nodules. This study demonstrates the application of a novel method for classifying large and small pulmonary nodules without using the nodule segmentation techniques associated with large numbers of false positives per scan. The method described here can be easily extended to include more easily distinguished classes such as parenchyma and non-nodules. Furthermore, a novel hierarchical DNN model is proposed as a complete CADe/x system in which the classifier being demonstrated represents a central component. The potential for such a system is of particular significance because it offers an alternative to the state-of-art existing methods that have typically reported high numbers of false positives per scan (Murphy, et al. 2009; Messay, et al. 2010; Firmino, et al. 2016).

Background

Deep learning more broadly describes a variety of computational models composed of multiple processing layers (*i.e.* a deep network of layers) used for learning representations of data with various levels of abstraction. Stemming from the seminal work of Krizhevsky *et al.* (2010), the subsequent half decade has seen remarkable progress in general image classification tasks, More specifically, Krizhevsky *et al.* used a deep convolutional neural network, which has since been the catalyst for fundamental change in the study of computer vision and the foundation of a new branch of machine learning called *deep learning*. Convolutional neural networks have also long been known as highly effective for visual recognition tasks (LeCun, Y., et al. 1998), however, due to computational costs associated with their use, adoption in mainstream science had remained somewhat limited to classification problems focused on grayscale images with very limited resolution. Krizhevsky *et al.* used multiple graphics processing units (GPUs) to apply deep convolutional neural networks to the 1000 classes identified in the over 1.2 million images comprising the ImageNet dataset. GPUs, designed for highly-parallel vectorized operations and most commonly used for graphics rendering in gaming applications, have since become a requirement for computer vision researchers (Sermanet, et al. 2013; Szegedy, et al. 2015) and all researchers exploring large, complex datasets with deep learning methods.

The primary benefit that deep learning has brought to computer vision is in the domain of feature learning. With the powerful, feature extraction properties of DNNs, hand-tuned features painstakingly defined by experts are no longer required (LeCun, et al. 2015). However, feature

extraction and representation properties of DNNs only improve as the size of the training dataset increases. The benefits of deep learning are not, therefore, limited strictly to computer vision.

The application of deep learning methods to medical images is still in nascent stages. Early studies using deep neural networks for applications in medical images successfully demonstrated improvements in segmentation tasks (LeCun, et al. 2015; Ciresan, D., et al. 2012).

Dataset

The dataset used for training was obtained from the public Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI) (Clark, et al. 2013; Armato, et al. 2012). This reference database is comprised entirely of CT scans containing pulmonary nodules. Datasets from the LIDC-IDRI have been widely used for studying nodule detection methods, including various studies of relevance to this work (Firmino, et al. 2016; Kumar, et al 2015; Shen, et al. 2015).

The LIDC-IDRI database comprises of over 1000 CT scans, each of which is annotated by 4 different radiologists. Three types of objects are identified in the annotations of reading radiologists; nodules equal to or greater than 3mm (large nodules), nodules less than 3mm in diameter (small nodules), and non-nodules. Small nodules are depicted in Figure 1. Such nodules are difficult to discern.

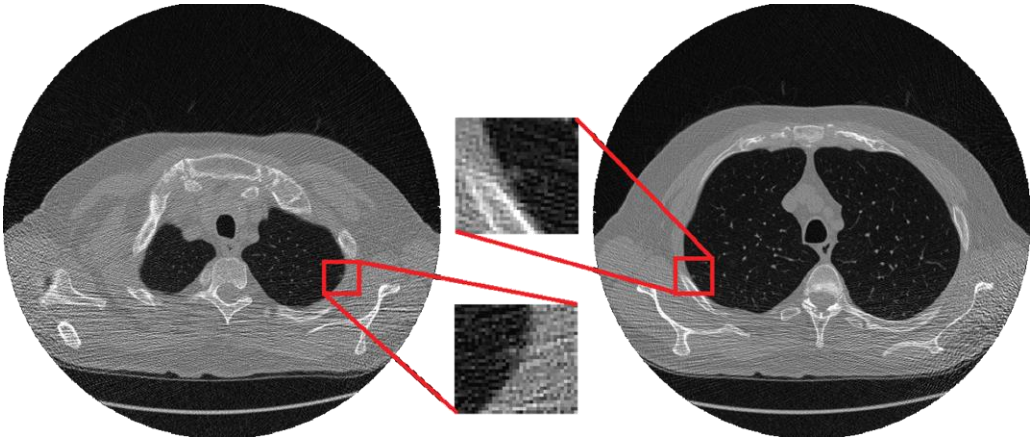

Figure 1. CT scans containing pulmonary nodules

In this study, large and small nodules were included only if they were identified unanimously by each of the four reading radiologists. The resulting dataset was comprised of 564 large nodules and 368 small nodules. Nodules not identified unanimously by each of the four readers were excluded for this study, however, some of these excluded cases will be included in future models.

For each large nodule, expert radiologists who were annotating the data also assigned a malignancy value. Malignancy values were not assigned for small nodules, as these nodules are typically not reliable for diagnosis of lung cancer. Malignancy values and other lesions identified by the radiologists but were excluded here will be used for future work. Beyond the malignancy values, eight more metrics describing each large nodule was reported by each reading radiologist. For example, one of the metrics included is 'sphericity,' which is independently related to the probability of a nodule's malignancy. Some of these further metrics can also be explored when developing the malignancy classification model for the proposed hierarchical CADe and CADx

system. Figure 2 shows large nodules that were exported as 216x216 pixel images but treated as 36 channel 36x36 images for training purposes.

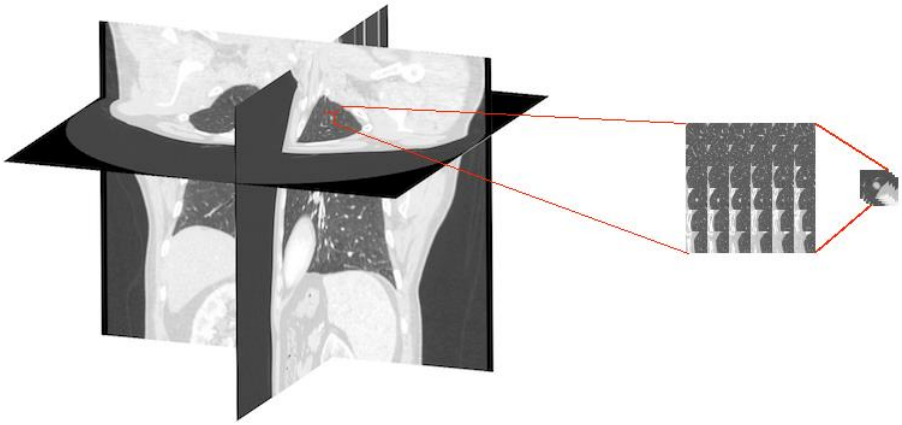

Figure 2. A localized volume around each nodule, comprised of 36 36x36 voxel slices, was extracted.

Methods

Initially, the 932 nodules were split into train and test datasets (80% and 20%, respectively). Typically, DNNs perform better with larger datasets. Due to the limited number of confirmed nodules for training, two methods were used to enlarge the dataset. Nodules for training were localized by readers, and localization of candidate nodules was assumed for the purposes of this preliminary study. Using the averaged centroid value from the readers' notes, a 36x36x36 voxel cube was extracted from the corresponding CT scan for each nodule. Assuming nodule morphometry independent of gravity, 48 unique perspectives of each nodule were extracted from each voxel cube. Each of these perspectives was exported as a 6x6 sheet of 36x36 images. This technique increased the size of the training dataset from 746 to 35,808. Random cropping of 36x36 image slices, uniformly for each slice in a sheet, into 34x34 image slices further increased the training dataset to 465,504.

Caffe (Jia, et al. 2014), an open source deep neural network solver from Berkley Vision and Learning Center, was used in this study to model DNNs. Caffe is typically considered one of the fastest options for modeling DNNs, heavily utilizing GPUs. A high performance Nvidia GPU was used for all cases in this study, however, entry level gaming GPUs are capable of running cases for all of the network architectures reported.

The neural network architectures examined were inspired by elements of two well known and well performing network architectures (Krizhevsky, et al. 2012; LeCun, et al. 1998). Components of each of these architectures were incorporated due to their success in two significantly different visual classification tasks (*i.e.* handwritten character recognition and general image classification). Each of the four network architectures differ only in the neurons per layer and the number of convolution layers, thus, a detailed architecture description for each of the neural networks will not be given. All of the networks included three max pooling layers, two fully connected layers, and a softmax output. Two max pooling layers were always included following the first and second convolution layers while the third max pooling layer was always introduced following the last convolution layer. Rectified linear units were used as the activation function, a common practice in deep learning

applications. Weights were initialized using Xavier initialization (Glorot et al. 2010), more appropriate for the grayscale images used in training and testing the models. For all networks, the first fully connected layer contained 36 neurons while the second fully connected layer contained only six neurons. Dropout layers were included with the fully connected layers to reduce over fitting.

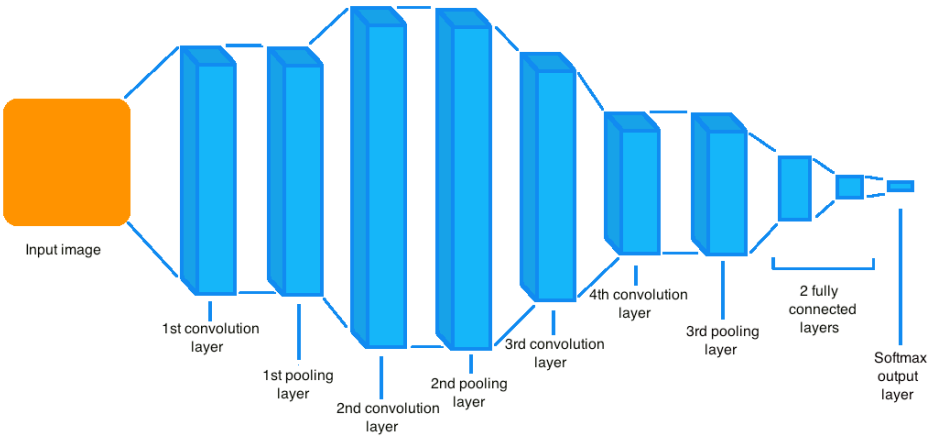

Figure 3. The simplest deep neural network architecture examined, containing four convolution layers, three max pooling layers, and two fully connected layers.

Both a two-dimensional method and a limited three-dimensional approximation were used during training. The scope of the method and results reported and described in this study is limited strictly to the better performing three-dimensional approximation technique. This technique was inspired by 2.5D (Roth, et al. 2014) and 3D (Turaga, et al. 2010) methods used to train DNNs with other volumetric medical data. In application data was loaded for each slice of a voxel cube as a separate channel on a single image. These channels were treated in the same fashion as color encoding channels (*e.g.* RGBA) when training with color images (Krizhevsky, et al. 2010). Only the closest five channels were included with kernels. Kernel size for the first two convolution layers, and for the final convolution layer, were constant for each of the four network architectures; a 5x5 kernel was used by the first convolution layer, a 4x4 kernel was used by the second convolution layer, and a 3x3 kernel was used by the final convolution layer. Kernels for the max pooling layers in each of the network architectures were also constant; a 3x2 kernel for the first max pooling layer and 2x2 kernels for the following max pooling layers.

Proposed Hierarchical Models

The method described and examined in this work is one of three models that makeup a proposed diagnostic tool for CADe and CADx of lung cancer. The proposed hierarchical model is set forth as an alternative to successive segmentation based methods that can be used for false positive reduction. Figure 4 depicts the detection and diagnosis process of the proposed tool with a hierarchy of models. The first layer of this hierarchy would function to identify anomalous objects in 2D, transverse slices of each CT scan. The model localizes these objects, and the locations with the highest probabilities are used to generate volumes of interest (VOIs) for processing by the model in the second layer. This model (an extended version of the model demonstrated in this study) will determine if the object within the volume of interest is a large nodule or any of a variety of benign objects. After detection of the large nodules in a scan, the final layer will use a model trained to diagnose the likelihood of malignancy.

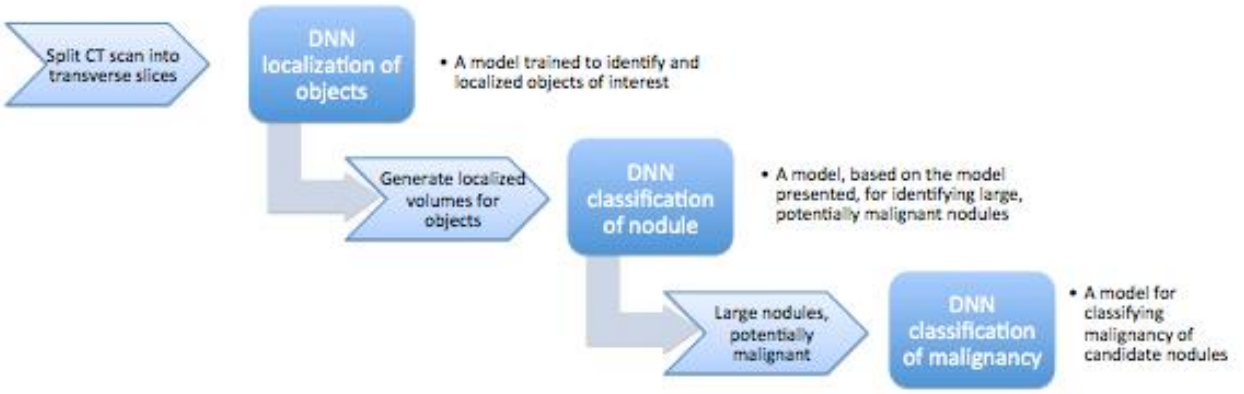

Figure 4. The proposed image processing and deep neural network classifiers to be used for detection of pulmonary nodules and diagnosis of lung cancer.

Results

DNN architectures for each of the four convolution layer variations examined all demonstrated accuracy of greater than 81% for the binary classification task. The results for the best performing models identified for each of the convolution layer variations are reported in Table 1. The peak accuracy observed among the four variants ranged only 1.02%. The ranges of sensitivity and specificity among the variants were slightly more significant at 3.18% and 3.72%, respectively.

The best performing architecture was comprised of five convolution layers (ten layers total). This model achieved an accuracy of 82.1% with a sensitivity of 78.2% and a specificity of 86.1%. Both the three and four layer networks performed very closely, as did the two five and six layer networks. Similarly, sensitivity and specificity for the same network pairs were observed to be much closer than with networks in the alternate pair.

Iterations	Convolution layers	Classification accuracy	Sensitivity	Specificity
8700	3	81.08%	75.77%	87.58%
6600	4	81.02%	74.61%	89.36%
7300	5	82.10%	78.19%	86.13%
9800	6	81.50%	78.11%	85.64%

Table 1. A Comparison of Observed Maximum Accuracy Over a Range in the Number of Convolutional Layers

Discussion

The 1.02% range of peak accuracy among the four reported neural network architecture variants, although small, is an important performance improvement from practical perspective. On the other hand, given the low magnitude of this range and the challenges posed by tuning optimization parameters for the complex network architectures, it is relatively early to confirm any benefit from

further increasing the convolution layers. However, it does appear that the gap between sensitivity and specificity is smaller for networks with five and six convolution layers.

The performance of deep neural networks is highly dependent upon the choices for various optimization parameters. Typically, many combinations of parameters and network topologies must be tested to achieve maximal accuracy. Due to this, and the computational costs associated with running each model, optimization is a difficult and uncertain process. The results depicted here do, however, represent relatively extensive experimentation with parameter and network optimization. Furthermore, they represent the most extensive convolutional neural networks used for pulmonary nodule classification.

The study has some limitations. These results cannot be compared directly with previous studies that have used traditional segmentation methods because they do not represent a complete CADe system. These studies are able to report the number of false positives per scan, a metric of particular significance in this study that is unquantifiable without a complete CADe system. The DNNs presented here are currently being tested on sets of candidate nodules generated using existing CADe systems. These forthcoming results will potentially provide further validation for using DNNs for false positive reduction in pulmonary nodule detection and diagnosis.

The results of this study can, however, be compared to results reported in a similar pulmonary nodule binary classification task (Shen, et al. 2015). In this study, the malignancy values reported by radiologists for all identified nodules were split into benign (1-2) and malignant (4-5) classes. This is important to note because all nodules with an average malignancy value of three were discarded. The accuracy reported for this classification task was 84%. After consideration of these curation choices for both the test and training datasets, the 82% accuracy in this study was determined to be satisfactory.

Conclusion

Results from this preliminary study were consistent with expectations from the review of literature. This study provides initial validation and implementation of a novel analytics technique, *Deep Learning*, for application in the medical image recognition domain. This application has wide applicability in other areas of vision recognition, neuroscience, etc.

The classification accuracy is low relative to existing studies employing segmentation-based methods. The accuracy of deep neural network classifiers typically increases with the size of the training dataset, however, and the dataset used in this study is not sufficiently large to achieve maximal accuracy. Demonstration of classification accuracy at the levels reported here can be largely regarded as a successful initial attempt based on the size of the training dataset. This study provides initial validation for our effort to expand upon the existing dataset and fine tune the algorithm for improving the classification accuracy to levels that are more commonly associated with deep learning and image classification.

REFERENCES

Society, A.C. 2016 February 8, 2016 [cited 2016 February 28, 2016]; Available from: http://www.cancer.org/cancer/lungcancer-non-smallcell/detailedguide/non-small-cell-lung-cancer-key-statistics.

Gurcan, M.N., et al., Lung nodule detection on thoracic computed tomography images: preliminary evaluation of a computer-aided diagnosis system. *Medical Physics*, 2002. 29(11): p. 2552-2558.

Murphy, K., et al., A large-scale evaluation of automatic pulmonary nodule detection in chest CT using local image features and k-nearest-neighbour classification. *Medical Image Analysis*, 2009. 13(5): p. 757-770.

Messay, T., R.C. Hardie, and S.K. Rogers, A new computationally efficient CAD system for pulmonary nodule detection in CT imagery. *Medical Image Analysis*, 2010. 14(3): p. 390-406.

Firmino, M., et al., Computer-aided detection (CADe) and diagnosis (CADx) system for lung cancer with likelihood of malignancy. *Biomedical engineering online*, 2016. 15(1): p. 2.

Tajbakhsh, N., M.B. Gotway, and J. Liang, Computer-aided pulmonary embolism detection using a novel vessel-aligned multi-planar image representation and convolutional neural networks, in *Medical Image Computing and Computer-Assisted Intervention*--MICCAI 2015, Springer. p. 62-69.

Van Ginneken, B., et al. Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. in *12th International Symposium on Biomedical Imaging* (ISBI), 2015 IEEE.

Kumar, D., et al., Discovery Radiomics for Computed Tomography Cancer Detection. arXiv preprint arXiv:1509.00117, 2015.

Shen, W., et al. Multi-scale convolutional neural networks for lung nodule classification. in Information *Processing in Medical Imaging*. 2015. Springer.

Krizhevsky, A., I. Sutskever, and G.E. Hinton. Imagenet classification with deep convolutional neural networks. in *Advances in neural information processing systems*. 2012.

LeCun, Y., et al., Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998. 86(11): p. 2278-2324.

Sermanet, P., et al., Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229, 2013.

Szegedy, C., et al. Going deeper with convolutions. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

LeCun, Y., Y. Bengio, and G. Hinton, Deep learning. *Nature*, 2015. 521(7553): p. 436-444.

Ciresan, D., et al. Deep neural networks segment neuronal membranes in electron microscopy images. in *Advances in neural information processing systems*. 2012.

Prasoon, A., et al., Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network, in *Medical Image Computing and Computer-Assisted Intervention*–MICCAI 2013. 2013, Springer. p. 246-253.

Clark, K., et al., The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *Journal of digital imaging*, 2013. 26(6): p. 1045-1057.

Armato Iii, S.G., et al., The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Medical Physics*, 2011. 38(2): p. 915-931.

Jia, Y., et al. Caffe: Convolutional architecture for fast feature embedding. in *Proceedings of the ACM International Conference on Multimedia*. 2014. ACM.

Glorot, X. and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. in *International conference on artificial intelligence and statistics*. 2010.

Roth, H.R., et al., A new 2.5 D representation for lymph node detection using random sets of deep convolutional neural network observations, in *Medical Image Computing and Computer-Assisted Intervention*–MICCAI 2014. 2014, Springer. p. 520-527.

Turaga, S.C., et al., Convolutional networks can learn to generate affinity graphs for image segmentation. *Neural computation*, 2010. 22(2): p. 511-538.