Deep Learning for abnormality detection in Chest X-Ray images

Christine Tataru (CS231n student)
Stanford University: Computer Science
ctataru5@stanford.edu

Darvin Yi
Stanford University: Computer Science
darvinyi@stanford.edu

Archana Shenoyas
Stanford University: Computer Science
shenoyas@stanford.edu

Anthony Ma
Stanford University: Computer Science
akma327@stanford.edu

June 13, 2017

Abstract

Heart and lung failure account for more than 500,000 deaths annually in the United States and are most commonly screened for using plain film chest x-rays(CXR). The time constraints imposed on radiologists by their massive workload severely impede communication with direct-care physicians and result in unnecessarily long and deleterious time-to-treatment periods for patients. A computer-aided triage system would mitigate these issues by communicating directly to the primary care physician when more tests are necessary, but the development of such a system has been hindered by immense variance of CXR images and lack of labeled data. Deep learning, a technique that has seen considerable growth in recent years, allows for classification of highly heterogeneous images given a sufficiently large dataset. The Stanford Normal Radiology Diagnostic Dataset (NeRDD) provides us with more than 400,000 CXR cases that have been expertly labeled as either normal, non-normal, or extremely high-risk(emergent). To apply deep learning to the novel CXR dataset, we will 1. develop a simple preprocessing pipeline using digital image processing techniques and expert radiologist advice 2. create a pipeline that can apply three neural network architectures that have proven successful in classification tasks, GoogLeNet, InceptionNet, and ResNet, on our cohort of CXR images 3. use neural network visualization techniques to understand what type of features our model weights most heavily. With a computer aided triaging system that is able to visualize its results, we will not only be able to streamline higher risk patients to get the immediate help they need, but also communicate how the network works to clinicians, ultimately improving the national standard of care.

1 Introduction

In radiology, “turn-around time is king” [1] with radiologists being evaluated based on their turn-around time rather than the quality of their reports. Especially in rural areas where direct care providers rely on teleradiology for their CXR interpretation, emphasis on turn-around time can result in sub-standard reports, confusion, misdiagnosis, and gaps in communication with primary care physicians. All of these severely negatively impact patient care, and can have life-changing consequences for patients [2]

A computer-aided triage system would mitigate these issues in several ways. Firstly, it would allow radiologists to focus their attention immediately on higher-risk cases. Secondly, it would allow radiologists more information to help them correct potential misdiagnoses. Lastly, it would provide the primary care physician with immediate information about the patient’s condition and risk level to allow them to order more diagnostic tests without delay and to ask appropriate, informed questions of the interpreting radiologist. The input to our algorithm will be .tiff CXR images along with a label of normal/abnormal. We will then use a CNN to output a classification of normal/abnormal per image. We do not have information of levels of abnormality, so the problem will be strictly binary classification. The development of such a system has so far been hindered by immense variance of CXR images and lack of labeled data. Deep learning, a technique that has seen considerable growth in recent years, allows for classification of highly heterogeneous images given a sufficiently large dataset.

2 Related Work

The concept of computer-aided diagnosis(CAD) for chest x-rays has been around for the past fifty years, and has made dramatic progress from rule-based survival prediction from lung x-rays to machine learning approaches to, now, deep learning [3] [4] [5]. Ginneken et al. make the argument that CAD in radiology is imperative, as the workload of radiologists is quickly becoming unmanageable [6]. .

First, computer-aided diagnosis was sought for other types of diagnostic tasks: breast cancer localization by GoogLeNet and skin cancer classification by a network out of Stanford by Esteva et al. [7] [8]. Both of these well-designed networks, along with others, have proven that convolutional neural networks can be used very successfully in not just natural image classification, but also medical image classification and segmentation [9] [10].

In the world of CXR classification tasks, Rajkomar et al. used GoogLeNet along with image augmentation and pre-training on ImageNet to classify CXR images as either frontal or lateral with 100 percent accuracy [11]. While this is not directly clinically relevant, it is an important proof on concept of the use of deep learning on CXR images. Anavi et al. sought to create a network that could, given a query image, rank the other CXR images in its database by similarity to the query. They found that a 5 layer convolutional neural network was much more effective than similarity based on image descriptors [12]. Such a network could be used to help clinicians search for past cases easily and help inform their current or future diagnosis. In 2016, Shin et al. used a CNN to detect specific diseases in CXR images and assign disease labels. They then used an RNN to describe the context of the annotated disease based on the features of the CNN and patient metadata [13]. They were only able to achieve validation accuracy of .698 on this ambitious task. This performance may be largely due to their relatively small data set size of 7470 images, the challenges of multi-class classification, and incorporating textual data from patient records. Most recently, in 2017, Wang et al. successfully designed a CNN to diagnose specific diseases by detecting and classifying lung nodules in CXR images with high accuracy [14].

While these approaches function as very good proofs of concept, and may in and of themselves be useful for assisted diagnosis, they are not easily generalizable to different diseases as new training data and labels must be obtained to retrain models for each specific sub-question. Our work will be the first study that classifies a chest X-ray as normal versus abnormal to assist primary care physicians and radiologists to move more quickly and efficiently rather than render radiology obsolete. Given that we are not aiming to classify images to specific categories of disease, we will have a much more versatile framework that is applicable to a much larger patient population.

3 Dataset

Data was acquired from Stanford Radiology, which maintains a large database of patient x-rays and includes 400,000 chest x-rays among many other types. For this project, 50,000 images were used for logistical reasons. Some patients may have more than one image associated with their file; in this case each image was treated as a separate patient for purposes of training and prediction. Each image is labeled 0 (normal), 1 (abnormal) or 2 (emergent). The distribution of the 3 image types is 65:35:0.1, and in practice, no emergent cases were included in our subset of data for this project. Original images are 3000 x 3000 pixels.

4 Methods

4.1 Preprocessing

There are many sources of variance in the CXR data which may negatively affect the performance of downstream classification tasks using feature-based methods or neural networks. Major sources of variance include contrast variance, positional variance and view angle variance (eg. Anterior-Posterior vs Medial-Lateral). In this first iteration, we processed all images with histogram equalization to increase contrast within each CXR image. Guided by radiologist advice, we deemed it useful to enhance the difference between the bone and empty space or tissue depicted in x-rays so as to make relevant information more prominent. All image processing steps were carried out using the python scikit-image library [15].

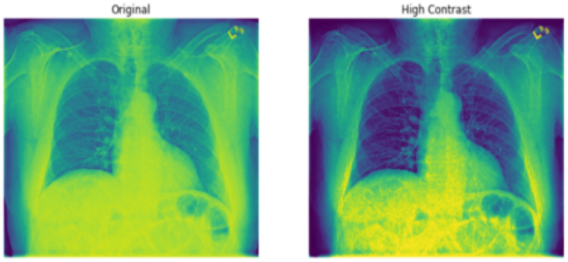

Figure 1: Left: unprocessed image Right: image processed using histogram equilization

4.2 Data Augmentation

To take advantage of our large dataset, we implemented basic data augmentation techniques to prevent overfitting in our model while still making use of all the data available. Each training image, before being input into a neural network, was flipped 0, 90, 180, or 270 degrees. Additionally, each image was either flipped left to right or not. Lastly, each image had some random small amount of gaussian noise added to each pixel value.

4.3 Network Architectures

All CNN models were run on a server with 4 NVIDIA Tesla P40s. The data were split into training and test sets using a 90-10 split. Each CNN model performance was assessed using a held out test set of 10% of our data. Overfitting was assessed by comparing the cross entropy loss and accuracy on training vs test data. Loss was calculated as follows:

$$L = -(1/n) \sum_{i=1}^{n} log(P(CorrectClass)$$

Default hyperparameter settings were as follows: Learning Rate = 0.001, Regularization = L2, Batch Size = 256, and Drop Out = 0.5. These parameters were chosen as those previously optimized on ImageNet.

4.3.1 GoogLeNet

GoogLeNet[14] relies on the concept of the inception module, whose invention aimed to alleviate two main problems: 1. A large network is more accurate and better at classification but also more prone to overfitting and 2. Increasing a network in size dramatically increases the computational power necessary to train that network. The intuitive solution to both these problems is to use sparsely connected architectures, however, our current hardware is much more apt at doing dense matrix operations than it is at doing sparse matrix operations. The inception module is meant to find optimal local sparse structure and represent it by dense components. To this end, the module consists of different sized filters that are each passed over the same input image whose outputs are all concatenated together, along with a max pooling on the original image, to produce the final module output. GoogLeNet implements an additional dimensionality reduction step in each module to limit the number of dimensions being input to each module to save computational power. This implementation of GoogLeNet contains 22 layers.

4.3.2 Inception V3

InceptionV3 [16] uses the same concept of inception modules as does GoogLeNet so as to allow the depth and model complexity to expand without undue computational cost. It does this by using factorized convolutions and aggressive regularization to create a network that is more than double the size of GoogLeNet at 48 layers.

4.3.3 ResNet

Residual Networks [17] reformulate the concept of learning to learn residuals with respect to each layer rather than functions directly. This means each layers try to learn RES where RES is the final loss minus the output of the previous layer. These networks are easier to optimize than traditional ones, and our implementation attains an incredible 152 layers without experiencing classical problems like vanishing gradients.

5 Results

Application of histogram equalization produced CXR images with enhanced contrast that were used as inputs. Each training image also underwent the data augmentation procedure described above.

In this paper, we test the efficacy of 3 qualities: dataset volume, model complexity and depth, and input data resolution. For some experiments, loss curves are shown rather than accuracy for logistical reasons. Accuracy and F1 scores are reported for the final model, GoogLeNet, in section 5.4.

5.1 Dataset Volume

Typically CNN models have a high demand for data in order to perform classifications effectively. For logistical reasons, we were interested in exploring how data volume would affect model performance and predictive capacity. After randomizing our data, we choose a subset of 3000, 10000, and 50000 images respectively and fed that into the GoogLeNet model. The following figure highlights our findings. While training accuracy increased slightly with the larger data sets, from 0.89 in 3000 image case to 0.95 in 50000 image case, the validation accuracy stayed rather consistent at 0.8 maximal prediction accuracy. However, more training images allowed the model to train with fewer iterations, and allowed for large enough batch sizes that the validation accuracy stayed more stable. This slight increase in stability suggests that, for this task, 50,000 images is sufficient to achieve generalizable performance. All further experiments were therefore conducted with 50,000 images.

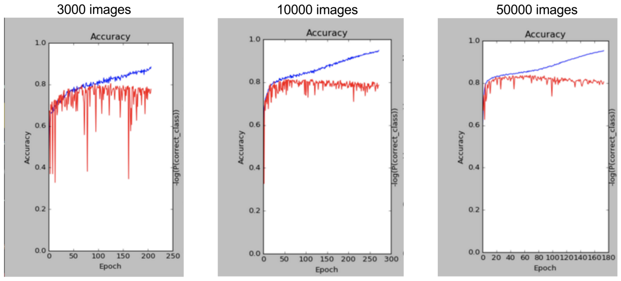

Figure 2: Accuracy curves as GoogLeNet trains on datasets of different sizes

5.2 Network Architecture Comparison

Initially, GoogLeNet was selected for its superior performance on the ImageNet challenge, as well as its success at localizing breast cancer in images[18] [7]. We also examined the more complex yet similar Inception V3 architecture, as well as a manifestation of the Residual network family. As demonstrated in Fig. 3, complexity of the network does not to improve performance of the model significantly. From this, it was determined that further fine-tuning architectures, while potentially effective at increasing performance slightly, would not significantly change the robust results.

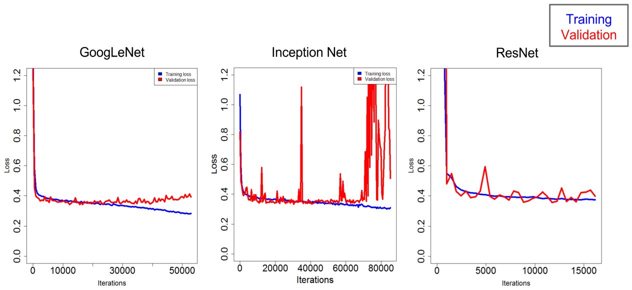

Figure 3: Loss curves as models train over same training/validation data. Models increase in depth from left to right

5.3 Input Data size

Original images were 3000 x 3000 pixels, a prohibitive size for a deep learning network. Images were downsampled by uniformly at random selecting pixels to create 512x512 and 224 x 224 pixel images. Here, we show that image resolution does not seem to play a role in prediction accuracy. This may be due to the relatively small difference in pixel density between the two resolutions selected. Alternatively, it may be that increased image resolution simply does not provide significantly different information.

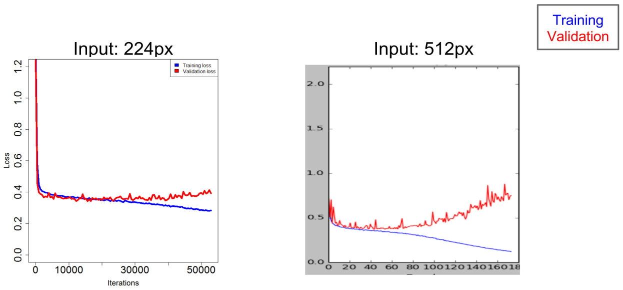

Figure 4: Loss curves for 224x224 and 512x512 pixel images as GoogLeNet trains on same training images

5.4 Performance

Although neither more complex networks nor larger input size increased performance significantly, initial performance by GoogLeNet on 50,000 images at 224 x224 pixels was significantly better than random. Although the breakdown of the training data is around 65:35 normal to abnormal, the validation set used does not reflect this ratio at 27: 73. Nonetheless, the network is able to attain an accuracy of 0.8 and an F1 score of 0.66 with the following confusion matrix.

		Predicted	
		Normal	Predicted
Truth	Normal	138	55
	Abnormal	88	430

The network does not systematically classify either label incorrectly, however, it does have a slight preference for predicting ”normal”, most likely due to the original class imbalance.

Note that the above confusion matrix is reported using validation rather than test images, due to a malfunction of the server making the test data inaccessible.

5.5 Visualizations

In order to gain insight into the relevant features learned by the above-mentioned deep learning algorithms, we sought to create images that were either quintessentially normal or abnormal. By backpropagating the loss of the opposite label back to the image pixels rather than the weights of the network, we were able to modify input images such that the network would classify them very confidently as the opposite label.

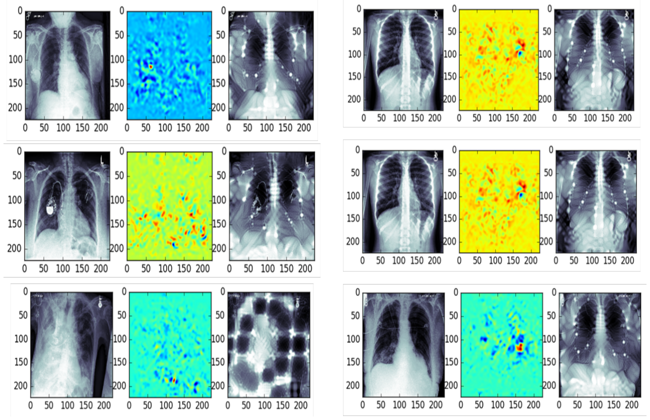

Figure 5: Input abnormal images as they would look for the network to confidently classify them as normal

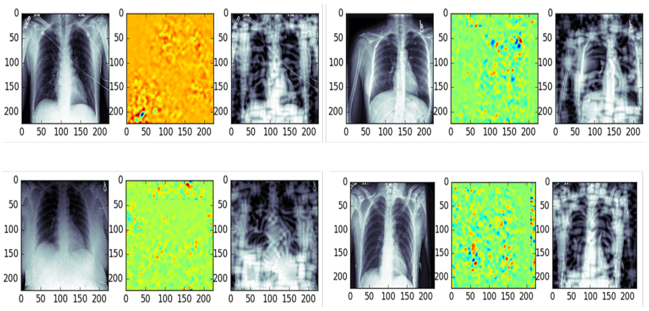

Figure 6: Input normal images as they would look for the network to confidently classify them as abnormal

We consulted with expert radiologists at Stanford Radiology, who agree that the major feature the network seems to learn when creating normal images from abnormal ones is symmetry on a macroscopic level. Though the created ”normal” images do not look exactly like original normal images, they all have this symmetry in common. The images that were made to look very abnormal have no obvious interpretation.

We performed the same process beginning with random noise rather than original images, with similar results.

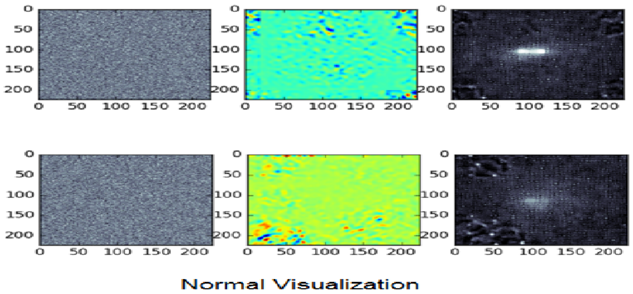

Figure 7: Input random noise images as they would look for the network to confidently classify them as normal

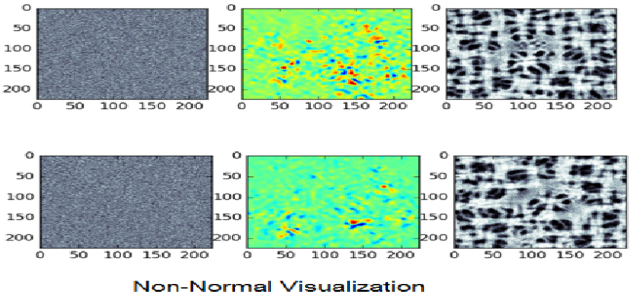

Figure 8: Input random noise images as they would look for the network to confidently classify them as normal

6 Conclusion

We conclude that GoogLeNet, a well-designed architecture of sufficient complexity, achieves significantly above-random classification accuracy when distinguishing between normal and abnormal chest x-ray images. The model is robust to the two image sizes tested(224X224 and 512X512 pixels) and the neural network architecture used. Furthermore, network visualization demonstrates that macroscopic features are learned effectively by the model. In particular, symmetry appears to be a salient feature of normal CXR images detected by the model. As deeper network architectures did not change performance, we expect the same feature to be prominent in InceptionV3 and Residual Network architectures. The above model may be improved in a number of ways: 1. increased preprocessing such as cropping lungs from images or cropping edges of the CXR images to highlight the lung regions 2. weighting of examples such that sensitivity rather than specificity is emphasized, to orient toward clinical usage 3. integration of level of uncertainty in physician ground truth diagnosis 4. integration of a segmentation component such that network can learn small, specific features rather than just macroscopic features and 5. inclusion of more abnormal examples, which are by nature very variant and also less common. Though this model is not ready for clinical adoption, it promises a future functional classification network that can classify normal vs. abnormal chest x-ray images and provide primary care physicians and radiologists with valuable information to significantly decrease time-to-diagnosis and greatly improve the current standard of care.

References

[1] W. L. Jackson. In Radiology, Turnaround Time Is King. *Practice Management*, Nov 2015.

[2] K. Eban. Is a Doctor Reading Your X-rays? Maybe Not." . *NBCNews.com*, Oct 2011.

[3] Bram Ginneken. Fifty Years of Computer Analysis in Chest Imaging: Rule-Based, Machine Learning, Deep Learning. *Radiological Physics and Technology*, 10:23–32, Feb 2017.

[4] GS Lodwick. Computer-aided diagnosis in radiology. A research plan. *Invest Radiology*, 10:115–118, 02 2017.

[5] Kunio Doi. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computational Med Imaging Graph*, 31:198–211, 2007.

[6] Bram van Ginneken, Cornelia M. Schaefer-Prokop, and Mathias Prokop. Computer-aided diagnosis: How to move from the laboratory to the clinic. *Radiology*, 261(3):719–732, 2011. PMID: 22095995.

[7] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck. Deep Learning for Identifying Metastatic Breast Cancer. *ArXiv e-prints*, June 2016.

[8] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 02 2017.

[9] Schmidhuber J. Deep learning in neural networks: an overview. *Neural Netw.*, 61:85–117, 2015.

[10] Hinton G. LeCun Y, Bengio Y. Deep learning . *Nature.*, 521(7553), 2015.

[11] Alvin Rajkomar, Sneha Lingam, Andrew G. Taylor, Michael Blum, and John Mongan. High-Throughput Classification of Radiographs Using Deep Convolutional Neural Networks. *Journal of Digital Imaging*, 30:95–101, Feb 2017.

[12] Y. Anavi, I. Kogan, E. Gelbart, O. Geva, and H. Greenspan. A comparative study for chest radiograph image retrieval using binary texture and deep learning classification. *Conf Proc IEEE Eng Med Biol Soc*, 2015:2940–2943, Aug 2015.

[13] Kirk R. Le L. Dina D.F. Jianhua Y. Shin, H. and Ronald M. S. Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation. *Conference on CVPR.*

[14] Changmiao Wang, Ahmed Elazab, Jianhuang Wu, and Qingmao Hu. Lung Nodule Classification Using Deep Feature Fusion in Chest Radiography. *Computerized Medical Imaging and Graphics: The Official Journal of the Computerized Medical Imaging Society*, 57:10–18, Nov 2017.

[15] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[16] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[18] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.