SIIM2018 ANNUAL MEETING
INSPIRING COLLABORATION
May 31 – June 2 | Gaylord National Resort | National Harbor, MD

Automated Detection of Clinically Significant Pneumothorax on Frontal Chest X-Ray using A Deep Convolutional Neural Network.

Andrew Taylor, MD, PhD, University of California, San Francisco; Clinton Mielke, PhD; John Mongan, MD, PhD

Introduction

Pneumothorax can constitute a medical emergency due to collapse of the lung and subsequent respiratory distress. It is often associated with trauma but may be due to iatrogenic injury within the hospital setting, or can occur spontaneously[1,2]. Pneumothorax of a clinically-significant size is often well seen with standard frontal plain film radiography; however, treatment is reliant on timely review of acquired films, both by the radiologist and the referring physician. Since current practices may result in long worklists of films to be read, particularly films acquired overnight or without accompanying clinical suspicion of a significant problem, an automated method of screening chest x-rays and prioritizing studies with positive findings for rapid review may improve the speed with which pneumothorax is addressed and treated.

Deep convolutional neural networks, a type of machine learning model that has found widespread application in computer vision and image classification tasks[3], are increasingly being utilized in radiology and medical image analysis. While these models can produce highly accurate results, they require large and well-curated datasets in order to achieve acceptable performance on tasks where there is significant visual heterogeneity, as one might expect in a sample of chest x-rays obtained in clinical settings varying from outpatient clinics to inpatient intensive care units, and with patients suffering from myriad illnesses.

Our objective was to create a large, human-annotated dataset of chest X-rays, relatively enriched in the finding of clinically significant pneumothorax, and to use this set to train a deep convolutional neural network to identify these pneumothoraxes. Such a network could be deployed to analyze images at the time of acquisition to prioritize studies that appear to contain a high-acuity finding such as presence of a moderate or large pneumothorax.

Hypothesis

Our hypothesis is that a deep convolutional neural network, trained on a suitably large annotated data set, will be able to accurately detect clinically significant pneumothorax when presented with a set of frontal chest X-rays representative of images obtained in an acute care hospital.

Methods

This retrospective study was approved by our institutional IRB.
A dataset containing chest x-ray images (both with and without pneumothorax present) was selected from the clinical PACS archive based on a search of the imaging report database. Studies were anonymized and converted to 8-bit JPG. They were then re-interpreted and annotated for the presence or absence of pneumothorax by trained radiologists, yielding a dataset comprising 11,032 images (981 containing moderate- and large-size pneumothoraxes, 10,051 with no pneumothorax). Relevant image data and annotations for each study were stored in a SQL database.
JPG images were downsized to 512x512 resolution and separated into an 80/20 training/test split. Training minibatches were generated from the available training dataset in a balanced fashion (50% positive for pneumothorax, 50% negative). On-the-fly augmentation with a number of basic transforms (horizontal flipping, zoom, shear and rotation) were used to expand the original human-annotated dataset.

---

[1] Gupta thorax 2000
[2] Light Pleural diseases 1995
[3] LeCun Nature 2015

Convolutional deep neural network models were implemented using the Keras deep learning library (keras.io) on top of TensorFlow (Google, Inc.). A number of network architectures were tested, including VGG16/19[4], Xception[5], Inception[6] and ResNet[7]. Models were tested using initialization from random weights as well as using transfer learning with models pre-trained on the ImageNet Large Scale Visual Recognition Challenge datasets[8]. A hyperparameter optimization strategy was employed during training of the different models using the Future Gadget Laboratory (https://github.com/Kaixhin/FGLab) framework. Keras training scripts were parameterized with a range of parameters and a random grid search was performed in several rounds of experimentation (see Table 1). Hyperparameter optimization was evaluated using the unbalanced validation set (approximately 10.4% prevalence of pneumothorax) every 10 training epochs. We computed AUC, sensitivity, specificity, and PPV of the entire validation set.
Following hyperparameter optimization, multiple experiments were performed using our top-performing model with repeated random reshuffling of the training/test sets, to ensure stability and generalizability of resulting trained models. During each individual training run, separation was maintained between the training and test data, resulting in multiple parallel developments of the model which showed good performance correlation that was not dependent on the specific images within the training and test sets.

Table 1

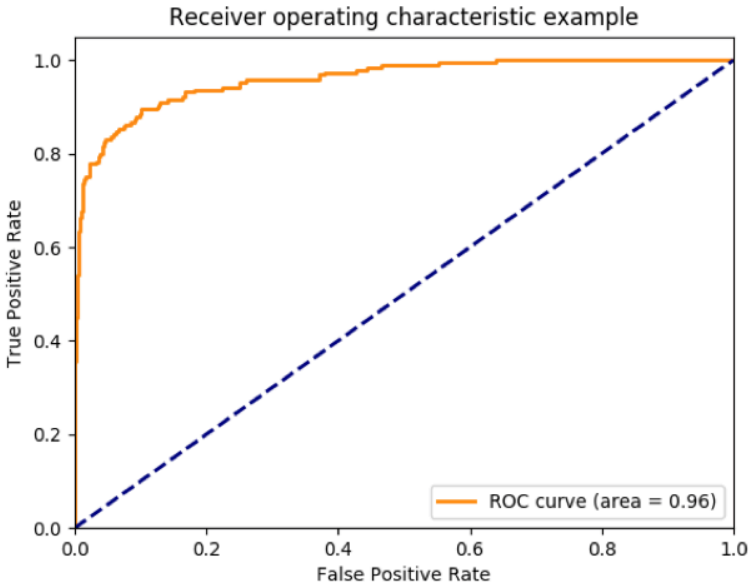
Parameter	Values	Description
Arch	VGG16, VGG19, Resnet50, Xception, inception	Pretrained architecture on ImageNet
pooling	Global Average, Global Max, Flatten	Pooling method after final filter layers
fc1	4, 8, 16, 32, 64, 128	Neuron count for first fully-connected layer after pooling
fc2	0, 4, 8, 16, 32, 64, 128	Neuron count for the second fully connected layer
LR	0.001, 0.005, 0.01, 0.02	Learning rate
LR schedule	constant, cyclic, plateau	Experiments with dynamic learning rates
batch size	4, 8, 16, 32, 64, 128	Batch size for the training
dropout	0, 0.25, 0.5, 0.75	Dropout setting applied to fully connected layers
augmentation zoom	0, 0.25, 0.5, 0.75, 1.0	Maximum fractional zoom range for images. 1.0 = 100% increase in size
augmentation shear	0, 0.1, 0.3, 0.5	Fractional affine shear for image augmentation generator
augmentation rotation	0, 30, 45, 60, 90	Maximum rotational angle in degrees for image augmentation
optimizer	sgd, adam, nadam, adadelta, rmsprop	Optimization algorithm used for training
Batch Normalization	Yes / no	A batch normalization layer was optionally inserted before the pooling layer

[4] Simanyan 2015 ArXiv
[5] Chollet 2016 arXiv
[6] Szegedy 2015 IEEE CVPR
[7] He 2015 arXiv
[8] http://www.image-net.org/challenges/LSVRC/

Results

The best performing model for this task achieved an AUC of 0.96, with a sensitivity of 0.87 and specificity of 0.91 (figure 1). This model was an Inception v3 network using a standard stochastic gradient descent optimizer. We built several custom convolutional architectures with initial random weights but found poor training performance. Transfer learning with a variety of architectures pre-trained on ImageNet performed far better on this task. Among all of the pretrained architectures tried, Inception consistently produced the best validation performance for this particular task.

Figure 1



	Predicted Negative	Predicted Positive
Control	1839	184
Positive (Moderate or Large PTX)	24	165

Although this model was trained on a dataset containing only pneumothoraxes classified as moderate- or large-sized by human radiologists, we characterized its performance using a test set that also contained pneumothoraxes classified as small. In this setting, model AUC dropped from 0.96 to 0.82. However, the sensitivity for moderate and large pneumothorax remained high at 0.86, while the overall specificity of the model was 0.87 (figure 2).

Figure 2

Annotation	Count	Labelled Class	Predicted Neg	Predicted Pos	Sensitivity	Specificity
Control	2023	Neg	1853	170		0.92
Small PTX	418	Pos	250	168	0.40	
Moderate PTX	138	Pos	25	113	0.82	
Large PTX	51	Pos	2	49	0.96	

Conclusion

Using a dataset of over 11,000 frontal chest x-rays visually annotated for the presence of clinically significant pneumothorax, we have successfully trained a convolutional neural network that obtains sensitivity of 0.87 and specificity of 0.91 (AUC 0.96) on a dichotomous task of recognizing images that demonstrate moderate- and large-sized pneumothorax as opposed to control images without significant pneumothorax.

The algorithm is trained using a carefully curated set of images, all of which were primarily annotated based on a de novo visual review by a radiologist specifically for this task. This method of annotation allows greater accuracy and consistency in labeling as compared to relying on the clinical report, since x-ray reports are generally free-text unstructured data and have great variation in both vocabulary and syntax based on the reading radiologist. Although primary re-annotation comes at the cost of significant time and effort, the result is a high level of confidence that the labeling is accurate, and may be less subject to potential sources of bias in interpretation as may occur during the clinical interpretation when the reading radiologist has access to clinical history, prior studies etc.

Further, the best-performing model was produced through performance of thousands of experiments to optimize input image characteristics, overall model architecture, and hyperparameter selection. This degree of flexibility in experimentation was made possible through a purpose-built database used for image and metadata management and dynamic minibatch creation, as well as a modular architecture for model modification and training. Coupled with current GPU compute technology, a very large array of experimental parameters was explored in a reasonable timeframe, producing the results described above. Once the best-performing model was produced, additional experiments demonstrated the stability and reproducibility of the model in both multiple iterations of training with a fixed dataset, as well as training from scratch while shuffling the individual images that comprised the training and test sets to further demonstrate generalizability.

Our result shows sensitivity/specificity and AUC characteristics for detection of pneumothorax that are significantly higher than other recently reported deep learning results based on other datasets[9,10], and is also substantially superior to older reports using traditional feature-detection algorithms[11].

Statement of Impact

We have successfully created a high-throughput, widely deployable deep convolutional neural network for detection of pneumothorax on frontal chest x-rays. The area under the curve for this model was 0.96 on our test set, suggesting that this model would serve as a useful tool for screening chest x-rays for pneumothorax, and could be used to prioritize studies for more rapid review in the hope of improving time to treatment for this potentially life-threatening problem.

Keywords

pneumothorax, chest pathology, convolutional neural network, deep learning, machine learning, computer vision, medical imaging, chest X-ray, plain film, radiography

References

1. Gupta et. al. *Thorax* 2000;**55**:666-71
2. Light et. al. *Pleural Diseases, 3rd edition*: Williams & Wilkins, 1995: 242-77
3. Lecun et al. *Nature* 2015;**521(7553)**: 436-44
4. Simanyan & Zisserman *arXiv* 2015; 1409.1556v6 [cs.CV]
5. Chollet *arXiv* 2016; 1610.02357 [cs.CV]
6. Szegedy et al. *IEEE CVPR* 2015; 1-9
7. He et al. *arXiv* 2015; 1512.03385 [cs.CV]
8. http://www.image-net.org/challenges/LSVRC/
9. Wang et al. *arXiv* 2017; 1705.02313 [cs.CV]
10. Rajpurkar et al. *arXiv* 2017; 1711.05225v2 [cs.CV]
11. Sanada et at. *Medical Physics* 1992; **19(5)**: 1153-60

[9] Wang arXiv 2017

[10] Rajpurkar arXiv 2017

[11] Sanada 1992 Medical Physics