Journal of Machine Learning Research 1 (2019) 1-6 Submitted 05/08; Published 05/08

CheXpertNet

Hui Wei DAVIDWEI@NYU.EDU
Department of Computer Science
New York University
New York, NY 10003, USA

Neil Jethani NEIL.JETHANI@NYULANGONE.ORG
Predictive Analytics Unit
NYU Langone Health
New York, NY 10010, USA

Editor: NA

Abstract

In this paper we expand upon the work conducted by Irvin et al. (2019), and utilize the CheXpert Datasset in order to build a set of models that interrogate the nuances of chest radiographs. Specifically, we illustrate the effect of including lateral view chest X-rays, tailoring the imaging normalization to the statistics of grayscale radiography images, as well as training our model from scratch without the pretrained ImageNet-derived parameters. We show that using out tailored normalization statistics improves performance, while using uninitialized weights showed the potential for increased performance pending additional training time. We also empirically prove the benefit of lateral view imaging, supporting the sourcing of this data. Finally, we employ class activation mapping in order provide visual explainability for our models. We present examples in support of potentially using this method to build trust in our predictions.

1. Introduction

Chest radiography is the most common imaging based examination in the world, often part of standard clinical workflows (Raoof et al., 2012). They are an essential part of screening protocols and inform the diagnosis, management, and treatment of numerous high mortality disease. The large volume of chest radiographs requires manual review by highly trained radiologist, resulting in significant time and expense. Instead, computer-aided radiograph interpretation would allow for more efficient clinical workflows, clinical decision support tools, and high throughput screening. Deep learning practitioners have developed powerful algorithms that allow for the detection and segmentation of objects within images. However, we must make concerted effort to tailor these tools towards radiography images as well as consider the reliability and interpretability of these models for clinical deployment.

Deep neural networks, specifically Convolutional Neural Networks (CNNs) (Krizhevsky et al., 2012), have shown a remarkable ability to detect objects by generating intermediate representations from conserved feature extractors. Most state of the art CNNs are made available upon being trained on the ImageNet dataset (Shin et al., 2016) for object detection. Many researchers utilize the discriminative feature exaction of these pre-trained model for applications involving similar tasks. However, it is unclear how well the imageNet

object detection task can inform object detection from radiologic images. Specifically, chest radiographs have far more conserved features across images, where the differences between pathologies are more subtle. Additionally, whereas, ImageNet images are full color images, radiograph images exist in grayscale with unique pixel intensity ranges. We exploit and quantify the effects of these fundamental differences.

Datasets containing large numbers of silver-standard labeled chest radiographs, Wang et al. (2017) released ChestX-ray14 dataset, which contains frontal-view chest X-ray images. However, it has been shown that 15% of accurate diagnosis requires lateral views (Rajpurkar et al., 2017), so this dataset cannot simulate the real situation of radiography diagnosis. To solve this problem, Irvin et al. (2019) released a larger dataset CheXpert including both frontal and lateral view X-rays. We utilize the larger CheXpert dataset in order to determine how the inclusion of lateral view X-ray effects pathology identification at the patient level.

The task establish by CheXpert is to predict the probability of 14 different pathological observations. The observation labels for training were obtained from radiological reports using natural language processing methods, while the validation set contains 200 radiologist labeled studies from 200 pateints. Irvin et al. (2019) showed for select pathological observations, their model was able to outperform radiologist. Incrementing upon prior performance through experimentation is essential towards equivocating the reliability of these models with that of domain expert radiologist.

Trust in these models is also dependent on the explainability of the models classification. While, deep learning methods are often considered black boxes, class activation mapping (CAM) has been used to roughly characterize the importance of various parts of the image towards a given classification Zhou et al.. We can overlay this information upon the original radiologic image to help explain to patients and physicians why the model made a give prediction. While greater interpretability is still necessary, we can have some confidence in predictions that adhere to our intuition.

In this paper, we propose to build upon existing models trained on the CheXpert dataset by interrogating the effects of using models and processing methods developed for general object detection. Also we determine the potential benefit of introducing multiple chest radiograph viewing angles. We will evaluate the performance and interpretability of these models with the potential for clinical deployment in mind.

2. Data

As mentioned, we will use the CheXpert Irvin et al. (2019) dataset. CheXpert consists of 224,316 chest radiographs of 65,240 patients, including both front and lateral view X-rays. The task of it is to predict 14 diseases as positive, negative and unclear, respectively. The sample is of size 389×320 for frontal views and 320×320 for lateral views. The training set consisted of automatically extracted labels from radiology reports, while the validation set utilizes 200 studies from 200 unique patients with radiologist determined gold-standard labels.

3. Evaluation Metrics

To compare the performances of different models, we propose to use the ROC-AUC score to evaluate the model performance. Since each patient has multiple images, we evaluate our model on the the basis of patients instead of images.

4. Methods

4.1 Model

Following the best Convolutional Neural Networks (CNN) on ImageNet, our backbone model is based on DenseNet (Huang et al., 2016). Compared to other state-of-the-art models on ImageNet, such as ResNet (He et al., 2015), DenseNet adds more shortcut connections between all layers in the block, which makes it easier to train and be able to get the better performance. To be specific, DenseNet 121 is used in all our experiments. Instead of classifying 1000 classes in ILSVRC (Russakovsky et al.), our task is to assign 14 diseases into negative (0), positive (1) and uncertain (u) respectively in the training set. The output of our network is the softmax probability for each label of each disease.

4.2 Training Process

We use CrossEntropy loss for each disease and then the total loss function is the average mean over 14 diseases. To be specific, the loss function is as follows:

$$TrainingLoss = -\frac{1}{N}\sum_{i\in[1,N]}\frac{1}{14}\sum_{j\in 14 classes}(w_0^i y_0^{ij} logp_0^{ij} + w_1^i y_1^{ij} logp_1^{ij} + w_u^i y_u^{ij} logp_u^{ij})$$

Where, $w_\eta = \frac{|N|+|P|+|U|}{\eta}$ and η can be the number of negative $|N|$, positive $|P|$ and uncertain $|U|$ cases for each disease in the training set.

For the training details, the model is initialized with the pre-trained parameters on ImageNet, and all layers are fine-tuned. To augment the dataset, images are first resized to 364×364, and then randomly cropped to 320×320 pixels, finally, randomly horizontally flipped. We use the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and the learning rate 1×10^{-4} and fix it for the whole training procedure. Batch size of 16 are used and the whole model is trained for 3 epochs.

4.3 Validation Process

Different from the training task which has three labels for each disease, the validation set does not have uncertain label, which might be easier for the model. Therefore, for calculating the validation loss and getting the predictive probability, we mask out the uncertain label, and apply the softmax only to the first two labels for each disease. Since one patient can have multiple images, which can be frontal or lateral view, we calculate the AUC score based on each PATIENT, instead of each image.

4.4 Interpretability

In order to provide insight into which regions of the radiographs are driving a given prediction, we implemented class activation mapping (CAM). CAM can be applied to any model architecture that employs a global average pooling layer prior to classification. Via global average pooling the each channel from the final convolution layer is averaged and a weight is assigned to each resultant value to generate the prediction. As the final convolutional layer represents the features extracted from each part of the image, where each coordinate represents a receptive field of the original image. The final convolutional layer can be upsampled to represent the size of the input image and the weights associated with a given class can be used to synthesize the contribution of each receptive field to the final prediction. We used this technique in order to provide interpretability to our predictions as well as validate that the model agrees with human intuition.

5. Results

5.1 Baseline Model

NF	ECM	CM	LO	LL	Ed	Con	PN	AT	PNX	PlE	PlO	F	SD
0.915	0.635	0.811	0.907	0.402	0.904	0.894	0.738	0.812	0.898	0.929	0.980	??	0.883

NF=No Finding, ECM=Enlarged Cardiomediastium, CM=Cardiomegaly, LO=Lung Opacity, LL=Lung Lesion,
Ed=Edema, Con=Consolidation, PN=Pneumonia, AT=Atelectasis, PNX=Pnuemothorax,
PlE=Plueral Effusion, PlO=Plueral Other, F=Fracture, SD=Support Devices
Mean AUC = 0.8236 ?? means all patients have the same ground-true label

Table 1: AUC Score for Baseline Model

From the table and all the ROC curves, we can see that except for Lung Lesion, for other diseases, our model can achieve a good classification performance. Figure 2 (Apendix A.) illustrates the corresponding curves as well as the learning curve.

5.2 Ablation Studies

5.2.1 Frontal Views vs. Both Frontal and Lateral Views

Lateral views X-rays can show some body parts in better resolutions, compared to the frontal views. Thus, for human doctors, bilateral views will greatly improve their diagnosis accuracy. However, deep learning model is kind of different from human doctors, and the lateral views might add more noise to it, resulting in a worse performance. Therefore, whether the lateral view is able to benefit the deep learning model to do the diagnose remains unclear.

To uncover this problem, two experiments are done: (1) train and validate the model using both frontal and lateral views (2) train and validate only using the frontal views. The first experiment is the same as we did in for the baseline model. For the second one, to exclude the influence of reduced total image number, after removing all the lateral views in the dataset, for the patient who has lateral view images, we randomly select the same number of frontal view images of that patient as the substitution. Therefore, we can keep the total number of images the same, without adding extra information for frontal views.

NF	ECM	CM	LO	LL	Ed	Con	PN	AT	PNX	PlE	PlO	F	SD
0.936	0.562	0.856	0.894	0.302	0.920	0.900	0.767	0.783	0.892	0.933	0.879	??	0.874

NF=No Finding, ECM=Enlarged Cardiomediastium, CM=Cardiomegaly, LO=Lung Opacity, LL=Lung Lesion,
Ed=Edema, Con=Consolidation, PN=Pneumonia, AT=Atelectasis, PNX=Pnuemothorax,
PlE=Plueral Effusion, PlO=Plueral Other, F=Fracture, SD=Support Devices
Mean AUC = 0.7867

Table 2: AUC Score using only Frontal Views

Mean AUC of using only frontal views is 0.7867, which is much lower than that of using both views, 0.8236. To our surprise, the AUC score for Lung Lesion in above table is much lower than the baseline model. To see that overall performance decline is not only due to this disease, the mean AUC score of diseases except this one is calculated for both experiments, 0.8588 for bilateral views, and 0.8497 for only frontal view.

The red scores of the above table indicate diseases whose AUC score are higher than the model trained using bilateral views. It is obvious that their improvements are trivial. We speculate that for those 6 diseases, the frontal view already has contain enough information for the model to do the diagnosis. Nonetheless, for most of 14 diseases, the lateral views can benefit the deep learning model to a large extend.

5.2.2 Statistics on ImageNet vs. CheXpert

In CheXNet paper (Rajpurkar et al., 2017), they normalized input X-ray images using the mean and standard deviation of images in the ImageNet training set. However, the intuition tells us that the statistics for those two different datasets are different, since X-ray images are usually gray, while the images in ImageNet are colored.

NF	ECM	CM	LO	LL	Ed	Con	PN	AT	PNX	PlE	PlO	F	SD
0.925	0.560	0.832	0.902	0.688	0.917	0.899	0.733	0.784	0.838	0.932	0.985	??	0.900

NF=No Finding, ECM=Enlarged Cardiomediastium, CM=Cardiomegaly, LO=Lung Opacity, LL=Lung Lesion,
Ed=Edema, Con=Consolidation, PN=Pneumonia, AT=Atelectasis, PNX=Pnuemothorax,
PlE=Plueral Effusion, PlO=Plueral Other, F=Fracture, SD=Support Devices
Mean AUC = 0.8383

Table 3: AUC score for Normalizing Images using Statistics on CheXpert dataset

For 8 out of 13 diseases, the model trained with images normalized on the CheXpert dataset performs better than those normalized on the ImageNet dataset. Combining with the fact that the mean AUC of the first case is higher than the second one (0.8383 vs. 0.8236), it is clear that, at least for CheXpert dataset, normalizing based on the specific dataset the model is working on is better than just using the statistics of ImageNet training set.

5.2.3 Training from Scratch vs. Using Pretrained Weights

Although we can see from the above results that initializing with the pre-trained parameters is able to get the good result, can the model also get the similarly good result without using it? To verify this, as in the training process of the baseline model, a DenseNet model initialized with random parameters, and the same model with pre-trained parameters but fine-tuning all layers, are trained for 3 epochs.

NF	ECM	CM	LO	LL	Ed	Con	PN	AT	PNX	PlE	PlO	F	SD
0.897	0.558	0.796	0.862	0.206	0.851	0.880	0.782	0.796	0.753	0.897	0.965	??	0.737

NF=No Finding, ECM=Enlarged Cardiomediastium, CM=Cardiomegaly, LO=Lung Opacity, LL=Lung Lesion,
Ed=Edema, Con=Consolidation, PN=Pneumonia, AT=Atelectasis, PNX=Pnuemothorax,
PlE=Plueral Effusion, PlO=Plueral Other, F=Fracture, SD=Support Devices
Mean AUC = 0.7675

Table 4: AUC score for Freezing all Conv Layers

Obviously, the mean AUC score of the model trained from scratch is much lower than the model initialized using the pre-trained parameters (0.7675 vs. 0.8236). However, there are already 4 diseases, shown in blue and red color, whose AUC score approximates to or higher than those from the model using the pre-trained parameters. Another interesting observation is, most (10 out of 13) diseases gain the best AUC score at epoch 3. We can be certain that if it is trained for a longer time, the model trained from scratch will have the similar final performance with the model using the pre-trained weights.

5.3 Class Activation Maps

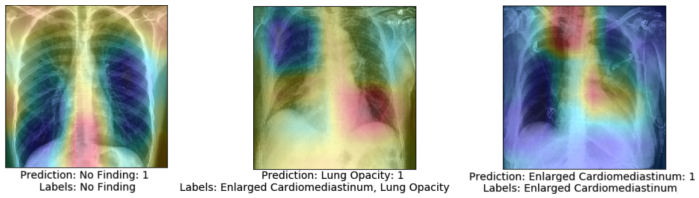

Figure 1: Class Activation Maps of Sample Correctly Identified Chest X-Rays.

Figure 1 presents three randomly selected correctly classified images overlaid with a heatmap representing the class activation mappings. While these interpretations have not be validated by a radiologist we can make a few apparent observations without validation. The first radiograph has evidence of lung opacity and the CAM appears to attend to the left lower lobe of the lung. The second radiograph was deemed to have no findings and the CAM seems not to attend to any part of the lung. Thirdly, the final radiograph with an enlarged cardiomediastinal appears to attend to area surround the heart as well as another area technically within the mediastinum.

6. Conclusion

In this work, we investigate the possibility to utilize deep learning model for solving the Chest X-ray classification problem. Our model can achieve a competitively good performance. More importantly, we successfully show the heatmap for different diseases, which greatly enhances the interpretability of the model, and we hope it will provide a significant reference for the radiologists to diagnose. In addition, through different ablation studies, we show whether different strategies can contribute to improving our performance or not, which is the basis for the future works.

Appendix A.

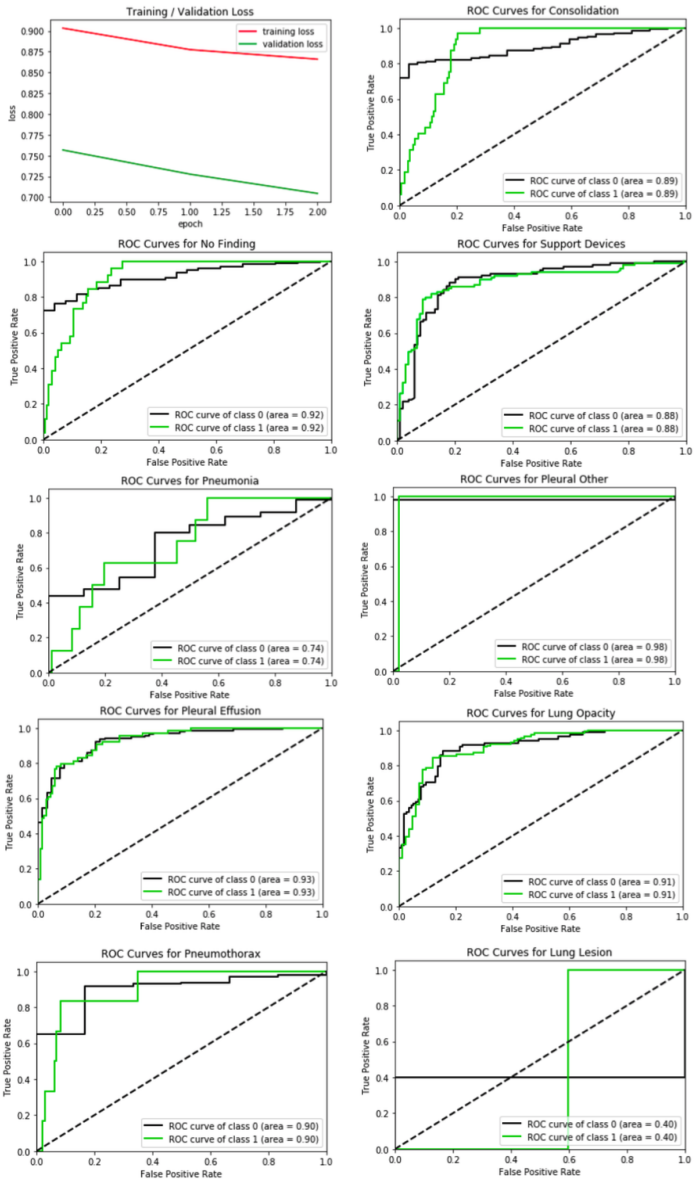

Figure 2: Learning Curve and Reciever Operactor Curves (ROC) for the Baseline Model.

References

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. 12 2015. URL `http://arxiv.org/abs/1512.03385`.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. 8 2016. URL `http://arxiv.org/abs/1608.06993`.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. 1 2019. URL http://arxiv.org/abs/1901.07031.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks, 2012. URL http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networ.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. 11 2017. URL http://arxiv.org/abs/1711.05225.

Suhail Raoof, David Feigin, Arthur Sung, Sabiha Raoof, Lavanya Irugulpati, and Edward C. Rosenow. Interpretation of Plain Chest Roentgenogram. *Chest*, 141(2):545–558, 2 2012. ISSN 00123692. doi: 10.1378/chest.10-1302. URL http://www.ncbi.nlm.nih.gov/pubmed/22315122https://linkinghub.elsevier.com/retrieve/pii/S0012369212600968.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, Li Fei-Fei, O Russakovsky, J Deng, H Su, J Krause, S Satheesh, S Ma, Z Huang, A Karpathy, A Khosla, M Bernstein, A C Berg, and L Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. Technical report. URL http://image-net.org/challenges/LSVRC/.

Hoo-Chang Shin, Holger R. Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M. Summers. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298, 5 2016. ISSN 0278-0062. doi: 10.1109/TMI.2016.2528162. URL http://ieeexplore.ieee.org/document/7404017/.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. 5 2017. doi: 10.1109/CVPR.2017.369. URL http://arxiv.org/abs/1705.02315http://dx.doi.org/10.1109/CVPR.2017.369.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning Deep Features for Discriminative Localization. Technical report. URL http://cnnlocalization.csail.mit.edu.