CS230

# CheXNet2: End to End Improvements For Chest Disease Classification

**Liam H. Neath**
Department of Electrical Engineering
Stanford University
liamn@stanford.edu

## Abstract

Accurate and timely diagnosis of medical images continues to receive substantial academic and industry attention. More recently, the National Institute of Health (NIH) released a substantial number of chest x-rays, which in turn led to the development of an algorithm named CheXNet[6]. In the spirit of developing a true medical device, we developed and tested a myriad of techniques aimed at improving classification performance. These techniques, which included energy localization, binary relevance and chi-squared informed sample reduction, may be referred to as CheXNet2. Although performance on the original NIH dataset remained unchanged, out-of-sample abnormality classification on a Chinese X-ray dataset improved by 12%.

## 1 Introduction

Medical imaging has undoubtedly improved the diagnostic abilities of Doctors, which in turn has provided timely and proportionate medical action. Unfortunately, inferring afflictions from images (x-rays) can not only be subjective but can also reduce a Doctor's time given the shear number of images, from multiple angles, that he or she may find necessary to review. CheXNet [6], a deep neural network which classified x-ray images for 14 common diseases has showed great performance. In an effort to build on this foundation, we have incorporated normalization techniques and architecture changes to move the existing algorithm closer to a stage where it may be deployed in a medical device.

The core neural network architecture of our system (CheXNet2) takes gray-scale images and a reference batch of x-ray images. Localized energy based normalization [5] is then applied to the images to normalize them in a similar manner to the images that the network was trained on. These normalize images are in-turn passed to a stacked network that incorporates fixed binary relevance pathways and a more general CheXNet pathway. Finally, a multilabel classification is then made among 14 classes. In an effort to achieve increased generality, a separate component of the algorithm is targeted to identifying out-of-sample abnormalities by way of chi-squared significance retraining of the network. This component retains much of the original ChexNet [6] architecture[1] but retrains the model in an informed manner.

[1] Due to financial constraints, the custom ChexNet2 architecture was not applied to this task.

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

## 2 Related work

CheXnet [6], which inspired this project, was certainly not the first published approach to classifying X-rays into classes of diseases. Lakhani and Sundaram [4] released a study that reported astounding performance (AUC of 0.99) on the classification of pneumonia using X-ray images. Interestingly, the authors evaluated ResNet and GoogleNet in their classification task, a nod to the somewhat empirical nature that underlies system development. Unfortunately, their dataset contained only 1007 images, roughly 1% of the size of the NIH dataset. Hence, their training set size severely undermines confidence in the extent to which their trained networks can generalize. This issue has also affected text and categorical based approaches to chest disease classification. Orhan et al [1] used PCA and feed forward networks to classify pneumonia based on medical reports for 150 patients. Once again, although their performance was commendable, the training size was far too small.
The large release of image data by the NIH, in 2017, provided the space for fresh research into this classification problem. Given the nature of the data, convolutional neural networks are a clear choice in system development. ChexNet [6] and ChexNet2 both employ DenseNets for their underlying convolutional neural networks (CNNs). This network architecture, introduced by Huang et al [2], achieves a similar accuracy to ResNet on the ImageNet dataset but with half the number of parameters and roughly half the number of FLOPs. This efficiency not only decreases the training time but also improves architecture search. Undoubtedly, the aforementioned characteristics provided motivation for Rajpukar et al [6] to utilize this model in the original CheXNet algorithm. Although this algorithm had cutting edge performance on the National Institute of Health's (NIH) X-ray dataset, classification on existing, yet smaller, chest x-ray datasets was not considered.
The aforementioned weakness motivated the need for this project as any medical device, targeted to disease classification, should be resilient to variations in the image intensity of the x-rays provided. Interestingly, variations in intensity may originate from the brand of imaging machines as well as the technicians who are supervising the scan. Philipsen et al. (2015) [5] proposed a normalization technique to account for this discrepancy. Their technique, which showed a 15% improvement in lung segmentation performance, has not been applied to neural network classification and thus this project provides a more modern evaluation of their technique.
Lastly, there have been no investigations into the application of multi-label classification techniques on chest disease classification. This is surprising given the strong marginal dependence as common diseases occur together. In considering an out-of-sample classification task, this paper acknowledges the importance of this phenomenon.

## 3 Dataset and Features

The NIH provides one of the largest public chest X-ray datasets in the world[8]. This dataset, which contains 112,120 frontal-view chest X-ray images, mulit-labeled for 14 chest diseases served as the underlying dataset for CheXNet and is used extensively in CheXNet2. The NIH proposed data split of 70%/10%/20% (train, validation and test respectively), was used to train the multi-label classifier components of ChexNet2. All images in the NIH dataset are 1024 x 1024 in gray-scale png images which were scaled to 224 x 224 images prior to normalization and incorporation in the model.
The NIH dataset also provides information on the gender and age of the patient associated with a particular image. Unfortunately, the data is inherently gender and age biased which led to the exclusion of this data in the architecture [2].

The Chinese X-ray dataset [3] originates from a hospital in Shenzhen and includes 662, 3000 x 3000, images that are labeled for the presence of tuberculosis and location of said disease. This dataset also includes the gender and age of the patient.
An energy normalization feature was extracted from the NIH dataset by randomly selecting 50 images and computing the energy localization matrix outlined in Section 4. This derived feature was then applied to all NIH and Chinese images and the resulting images used to train and test a clone of the neural network for pre-processing evaluation purposes.

[2] A breakdown of gender and age in the data can be found in *DataExplore.ipynb*

# 4 Methods

## 4.1 Pre Processing

The pre-processing localized energy normalization stage is one of the major steps that separates CheXNet2 from CheXNet. Localized energy normalization involves the decomposition of an image into B channels. Channel normalization and reconstruction of the image then follows. Philipsen et al., (2015) [5] decompose the image in the following manner:

$$I_i(x) = I_{i-1}(x) - L_{i-1}(x; \sigma_{i-1})$$
$$L_i(x; \sigma) = I_i(x) * G(x; \sigma)$$

Where $I_i(x)$is the image intensity in region *x* for level *i* and $G_i(x; \sigma)$, is a Gaussian filter in region *x*.

This is essentially a modified Laplacian decomposition of the image. Given that numpy contains an optimized implementation of laplacian pyramids, Laplacian decomposition was adopted in ChexNet2's normalization and a vectorized variation of the full algorithm developed. Figures 1 - 3 below demonstrate the effect of normalization.

Figure 1: Normal

Figure 2: $\lambda = 1, \beta = 5$

Figure 3: $\lambda = 2, \beta = 5$

## 4.2 Chi-Squared Sampling

In an effort to improve out-of-sample detection of abnormalities, a re-sampling step, for the NIH data, was also included to develop an abnormal/normal detection network that exists separately from the multi-label architecture. Given knowledge on the similarity of a particular disease with one of the 14 known classified diseases, CheXNet (with proper normalization) was retrained to identify images that were positively labeled for that similar disease and 4 diseases that had the highest Chi-squared values (degree of similarity in labeling). This methodology is inspired by Chekina et al. (2010) [7] and aims to glean more discriminating information by utilizing marginal dependence of labels. Notably, these chi-squared values were computed from the NIH dataset for each pair of labels.[3]

## 4.3 Neural Architecture

The second major architecture change in CheXNet2 involves using a pretrained densenset for *each* individual class. Hence the 14 node output from a trainable DenseNet121 pathway was concatenated with the output from 14 individually trained networks which was then fed forward to 14 nodes corresponding to each class. Lastly, a sigmoid layer was used to clamp values between 0 and 1 and a Binary Cross Entropy loss function used during the optimizing stage for both the individual networks and the final composite network.
The insight behind this architecture originates from the fact that learning to classify each class on its own ignores marginal dependence of these classes [9]. In combining these class specific classifiers, the network can in turn learn a function combining the input $x$ and a universe of small classifiers $H(x)$ to ultimately predict the true class $C$.

$$C = g(x, H(x))$$

[3]Due to the size of the plot and limited space, the resulting pairwise values can be found in *DataExplore.ipynb*.

There have been many proposed ways to generate $H(x)$ including classifier chains and pruned binary relevance models. CheXNet2 employs simple stacking; however, this could be improved with the use of the aforementioned techniques at the cost of training time.

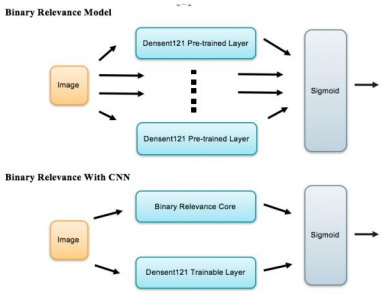


Figure 4: AUC for Transfer Anomaly Detection on Chinese X-ray Dataset

# 5 Results

## 5.1 Hyper Parameters

Training of all model variations involved using batch gradient descent (for tractability) and a variable learning rate. A batch-size of 16 was chosen in order to full utilize the RAM on the GPU while maintaining

The learning rate was set arbitrarily to 0.001 and would be shrunk by a factor of 10 on each epoch if there was no improvement in the cross-validation loss. Interestingly, this approach provides some reassurance against over-fitting the training data as training would end if there was no cross-validation accuracy improvement in 3 epochs.
The normalization step also involved a hyper-parameter choice in that the number of Laplacian channels and normalization steps could be set. Given existing work by Philipsen et al. (2015) [5], the number of channels $B$ and the number of steps $\lambda$ were both varied between 1 and 10.

## 5.2 Transfer Learning for Anomaly Detection

Using a single DenseNet121 network which was trained to classify the presence of *any* chest disease. The model was trained on various normalization parameters and evaluated on its performance to detect abnormalities in the Chinese X-ray dataset by way of an AUC metric (Figure 5). The baseline AUC, without normalization, was 0.64852; however, with normalization, the AUC improved to 0.744. This represent a 12% improvement in accuracy and demonstrates the importance of image normalization despite the depth of the network being used.

## 5.3 Chi-Squared Sampling

Another improvement on this task came in the form of the chi-squared informed re-sampling. Given the fact that the Chinese X-ray set was labeled for tuberculosis, a disease that often occurs with *pneumonia* using the chi-squared table it was found that training on these positively labeled examples should be sufficient due to their high p-values: Pneumonia, Emphysema, Pneumothorax, Pleural Thickening, Fibrosis This approach led to an improvement in the AUC by 4%.Unfortunately, the conclusions of this section are limited by the impact in which small dataset have on an algorithm's ability to generalize. Hence, further steps on evaluating this approach must involve larger test datasets.

## 5.4 Binary Relevance Effect

Surprisingly, the addition of binary relevance models did not provide a noticeable improvement in the AUC with respect to the base model (single DenseNet121 with 14 classes).The table below shows the AUC performance for both models on the various classes:

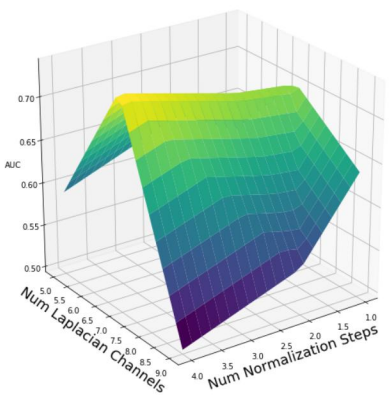


Figure 5: AUC for Transfer Anomaly Detection on Chinese X-ray Dataset

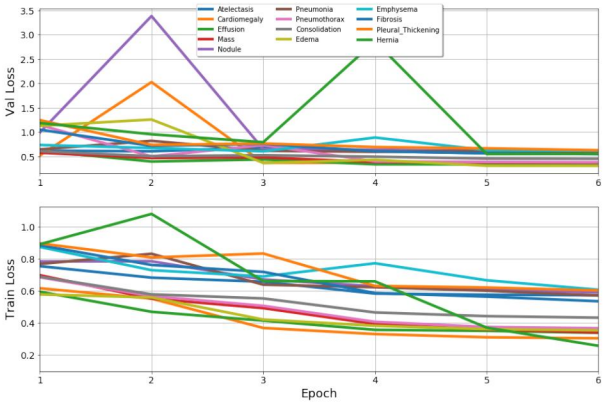


Figure 6: Train and Val Loss for Binary Relevance Models

| disease | Base | Binary Rel + CNN |
|---|---|---|
| Atelectasis | 0.817666 | 0.818333 |
| Cardiomegaly | 0.909232 | 0.899742 |
| Effusion | 0.883482 | 0.885067 |
| Infiltration | 0.714012 | 0.750767 |
| Mass | 0.842527 | 0.837565 |
| Nodule | 0.775551 | 0.775774 |
| Pneumonia | 0.774468 | 0.805006 |
| Pneumothorax | 0.875046 | 0.877391 |
| Consolidation | 0.807895 | 0.824091 |
| Edema | 0.888510 | 0.891687 |
| Emphysema | 0.927198 | 0.914777 |
| Fibrosis | 0.834079 | 0.831855 |
| Pleural_Thickening | 0.783568 | 0.790968 |
| Hernia | 0.924159 | 0.840263 |

Table 1: AUC on Chest Diseases for Base and Custom Architectures

Both models had average AUCs of approximately 0.83. The large reduction in AUC for Hernia may not be a cause of concern as this class is only positively labeled for approximately 200 images. Given this unsatisfactory performance, as the number of calculations was amplified by x14,further work would either need to identify more efficient ways of stacking the relevance models or a finely tuned second-to-last layer that would seek to appropriately weight the combination of binary relevance output and the multi-class CNN.

## 6 Conclusion

Intentioned normalization provided the greatest improvement in chest disease identification. The poor performance noted in the custom binary relevance model may be due to a lack of optimization in the final layer in addition to the approach taken to train the individual relevance models. Further work should not only address this issue but also investigate other approaches to multi-label classifying as diseases tend to have a strong marginal dependence.
Lastly, in order to transition this algorithm from an academic paper to a medical device, there will be many regulatory steps to clear. We envision that improved visualization (GradCam) will aid in the process and should be incorporated in further work.

## 7 Contributions

- Liam Neath
  - Developed all the PyTorch, image normalization and instance orchestration code.
  - Wrote the report

## References


[1] Orhan Er, Nejat Yumusak, and Feyzullah Temurtas. Chest diseases diagnosis using artificial neural networks. *Expert Systems with Applications*, 37(12):7648 – 7655, 2010.

[2] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016.

[3] Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J. Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative Imaging in Medicine and Surgery*, 4(6), 2014.

[4] Paras Lakhani and Baskaran Sundaram. Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*, 284(2):574–582, 2017. PMID: 28436741.

[5] R. H. H. M. Philipsen, P. Maduskar, L. Hogeweg, J. Melendez, C. I. Sánchez, and B. van Ginneken. Localized energy-based normalization of medical images: Application to chest radiography. *IEEE Transactions on Medical Imaging*, 34(9):1965–1975, Sept 2015.

[6] 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. *CoRR*, abs/1711.05225, 2017.

[7] Lena Tenenboim, Lior Rokach, and Bracha Shapira. Identification of label dependencies for multi-label classification. pages 53–60, 01 2010.

[8] 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. *CoRR*, abs/1705.02315, 2017.

[9] Y. Zhang, Y. Li, and Z. Cai. Correlation-based pruning of dependent binary relevance models for multi-label classification. In *2015 IEEE 14th International Conference on Cognitive Informatics Cognitive Computing (ICCI*CC)*, pages 399–404, July 2015.