Predicting Thorax Diseases with NIH Chest X-Rays

Joy Hsu, Peter Lu, Kush Khosla
Department of Computer Science and Department of Mathematics
Stanford University
{joycj, peterlu6, kkhosla}@stanford.edu

Abstract

The use of diagnostic imaging has increased dramatically in recent years. A substantial number are chest x-rays used to diagnose a plethora of conditions. These diagnoses are still primarily done by radiologists manually poring over each scan, with no automated triaging or assistance. We aim to use deep learning to predict thorax disease categories using chest x-rays and their metadata with greater than first-pass specialist accuracy. Our problem can be cast as a multiclass image classification problem with 15 different labels. The paper provides a proof of concept of an automated chest x-ray diagnosis system by utilizing the NIH dataset. Deep learning is used to improve the multiclass classification accuracy of thorax disease classification, measured against a baseline of softmax regression.

1. Introduction and Motivation

The number of X-rays performed in the US each year has been increasing steadily over the past decade. Of these, a substantial number are chest X-rays used to diagnose a plethora of conditions including lung cancer, emphysema, and atelectasis. These diagnoses are still primarily done by radiologists manually poring over each scan, with no automated triaging or assistance. We aim to use deep learning to predict thorax disease categories using chest X-rays and their metadata. A dataset from the NIH was released recently in this year, and we wish to improve the f1 score of thorax disease classification through a image classification problem [1]. This dataset contains over 110,000 gray scale identically-sized images from over 30,000 unique patients corresponding to 14 common thorax disease types.

Although in general doctors are quite good at coming up with a diagnosis, mistakes can happen, and details can indeed be left out. One study found that when seeking a second opinion, 66% of the time a second doctor gave a refined opinion of the original diagnosis, in only 12% of the time was the diagnosis confirmed, and 21% of the time the diagnosis completely changed from the original [2]. A model that can predict diseases based on X-rays would provide a reasonable sanity check in order to help achieve more accurate diagnoses.

Our inputs are 110,000 X-ray images of size 1024×1024, as well as metadata on age, gender, and number of visits to the hospital. We feed these features into a softmax regression as well as a modified residual network to output predicted probability of various thorax diseases, which leads to multi label classification, ranging from a healthy X-ray scan to diagnosis one to many diseases.

2. Related Work

Deep learning has been used extensively in the field of medical research. For example, groups like [3] have been able to use deep generative models to diagnosis patients, while [4] uses deep neural networks to identify when patients have cancer, just by looking at slide images.

In this paper, we cast this problem as a multi-class, multi-label, image classification challenge. Previous work mainly focused on single class diseased X-ray classification [5], as well as specific disease classification [6], whereas we chose to maximize potential of our prediction by utilizing all parts of the NIH dataset.

3. Problem Statement

The three main objectives this paper tackles are: 1) Increase accuracy rate of this 15-class, multi-label prediction to aid doctor's diagnosis; 2) Create new neural net architecture to incorporate image as well as categorical data and compare results with softmax regression / random classifier; 3) Analyze correlation between thorax disease and different patient traits (e.i. age, gender, number of visits to hospital).

4. Data

The released NIH dataset includes 110,000 gray scale X-ray images of 1024×1024 pixels from 30,000 unique patients, each of which originally has corresponding information on patient age, gender, ID, and number of follow up visits to the hospital.

The actual labels can be one or many of 14 common thorax disease types, which include Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion, Emphysema, Fibrosis, Hernia, Infiltration, Mass, Nodule, Pleural Thickening, Pneumonia, and Pneumothorax. No finding is also a category for non-diseased patients.

We chose to utilize the X-ray scan as well as its corresponding patient age, gender, and number of visits to the hospital as our input features. The 1024×1024 image with one gray scale channel is used directly in softmax regression and downsampled in the neural net to predict disease categories. We decided to tackle the complex problem of categorizing a X-ray scan with multi-label thorax diseases of 14 classes to use the dataset to full potential.

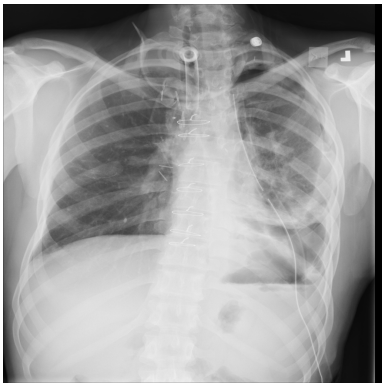

Figure 1: An example x-ray image as input.

5. Methods

We analyze the data using two different models. The first, as a baseline, is a simple softmax regression. This regression assigns probabilities to each of the 15 classes in the dataset for a given image. The second model is a residual neural network that was modified to also take into account metadata from the dataset, such as age, gender and how many times a patient has visited the hospital to get their condition checked.

5.1. Classification from Probabilities and Accuracy Measure

Before talking about the core of the models, it is important to discuss our approach for dealing with a multi-class, multi-label dataset (for any datapoint (x, y), we have $y \in \{0, 1\}^{15}$ and $0 \leq \sum_i y_i \leq 15$, meaning a patient could have any number of diseases). In particular, after obtaining probabilities for each of the 15 classes, both from softmax and our neural network, how we decided which categories would be tagged as positive, and which would be tagged as negative.

Given a datapoint $(x, y) \in \mathbb{R}^{1024^2} \times \{0, 1\}^{15}$, our model predicts $\hat{y}$ in the following way: let $p \in \mathbb{R}^{15}$ be the probabilities assigned to each category for x. $\hat{y}$ was given the value of one at each index i where $p[i]$ was in the top k values of p, where $k = \sum_i y_i$, and zero at the other $15 - k$ indices. Our accuracy for a prediction $\hat{y}$ for a label y was then defined to be $\hat{y} \cdot y/k$ (how many of the positive categories we correctly identified, divided by the total number of positive categories for that example).

It's important to note that in practice it would be hard to use this prediction method, since given a new patient, it is not known *a priori* how many diseases they have. Nonetheless, we believe this to be a rigorous accuracy measure that allows us to train our models well.

In practice, we have tested a thresholding prediction strategy, whereby all classes with a probability larger than some t are marked as 1, and all other classes as 0. This gives the model the opportunity to spread softmax probability over multiple classes equally and have them all be tagged appropriately. In practice, using $t = 0.15$ led to a similar accuracy to *a priori* tagging.

5.2. Softmax Regression

As a baseline, we implemented a simple softmax regression. $p \in \mathbb{R}^{15}$ were obtained through the matrix calculation $p = Wx$ where $W \in \mathbb{R}^{15 \times 1024^2}$. Denoting W as

$$W = \begin{bmatrix} \text{---} & \theta_1^T & \text{---} \\ \text{---} & \theta_2^T & \text{---} \\ & \vdots & \\ \text{---} & \theta_{15}^T & \text{---} \end{bmatrix}$$

W was calculated through optimization of the following cost function [7]:

$$J(W) = \frac{-1}{m} \left[\sum_{i=1}^{m} \sum_{j=1}^{15} 1\{y^{(i)}[j] = 1\} \log \frac{\exp(\theta_j^T x^{(i)})}{\sum_{l=1}^{k} \exp(\theta_l^T x^i)} \right].$$

Which has the following gradient for θ_j: $\nabla_{\theta_j} J(W) =$

$$\frac{-1}{m} \sum_{i=1}^{m} x^{(i)} (1\{y^{(i)}[j] = 1\} - p(y^{(i)} = j | x^{(i)}; \theta_j))$$

5.3. Modified Residual Network

Residual nets were created originally for the ImageNet challenge to allow training of deeper nets while minimizing the difficulty of transforming the data through so many layers. [8] ResNet-50 is a modified version of the first 152 layer net, but shares the same architectural characteristics, with direct transfer of activations to the next layer to ensure good features learned early in the network are not skewed or lost in subsequent layers.

Residual nets are particularly suited for images, which often give rise to many derived and complex features that

are crucial to classification. After deciding on this architecture, we decided to use pretrained weights from the original ImageNet challenge to produce a better initialization. We confirmed this accuracy increase over short runs of 50 epochs over an randomly initialized resnet.

We also transformed the input data with two convolutional layers, with strides 2×2 and filters 32 and 3. This served multiple purposes that boosted final classification accuracy. First, the data was transformed into a 256×256 image with 3 channels, simulating the 224×224 RGB input that ResNet generally receives, rather than feeding in high resolution 1024×1024 greyscale x-rays. This decrease in dimensionality also permitted greater batch sizes for faster training. The additional filters in the first convolutional layer permits more complex features to be captured before the network squeezes the data back down into 3 layers.

Uniquely, this dataset also includes some metadata tags for each image. For example, age, gender, and the hospital visit number are listed for each image. These features were added to the last hidden layer in Resnet-50, prior to the final softmax output.

Our final model architecture includes two initial convolution layers with drop out, concatenated with a 50 layer residual net, with additional layers to add in categorical features. The initial convolution layers with 2×2 strides aim to downsample the 1024×1024 image to 256×256, which is closer to the 224×224 dimensions for the canonical ResNet structure. The residual layers then learn the subtraction of features from input with shortcut connections, as the net directly connects the input of the (n) th layer to that of the (n + x) th layer. The last merged layer consists of the original outputs to the residual net with a concatenated array of other additional data (age, gender, number of visits to hospital). See Figure 2.

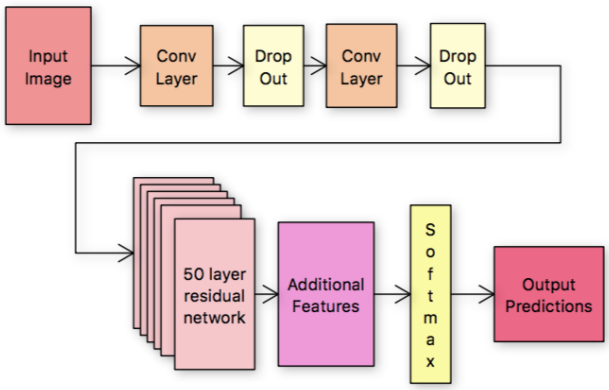

Figure 2: Final Modified Residual Network Architecture

6. Results and Experiments

We analyze the quantitative and qualitative results of the softmax regression and modified residual network model, with the clear conclusion that our neural net performs drastically better than current doctor diagnosis, random weight matrix regression, and softmax regression.

6.1. Quantitative Results

Our final modified residual network architecture achieved an accuracy of 0.58 with our top k classes accuracy metric. Compared to our other strategies- using a random weight matrix with softmax regression, the accuracy measure returned an overall accuracy of .04. The trained softmax model did not perform much better, with an accuracy of 0.08. It's important to note that this is not a binary classification problem, but instead a multi label classification problem with 15 categories, meaning this .58 value is significantly better than an untrained classification.

On average, each image had 1.3 positive labels. We compute our confusion matrix and detailed accuracy:

	Correct	Incorrect
Positive	679	488
Negative	11845	488

Table 1: Distribution of correctly identified labels.

Metric	TP	FN	FP	TN
%	58.2	41.8	4.0	96.0

Table 2: Our confusion matrix.

and F1 score: $\frac{2\cdot TP}{2\cdot TP+FP+FN} = 71.8\%$.

Although this is not a full ROC curve, we can perform preliminary comparisons to state of the art AUC RUC scores computed previously on this dataset. Previous papers focused on binary classification by disease rather than multiclass work, and achieved AUC ROCs ranging from 0.7345 to 0.9248 depending on the disease [6].
Our problem of multiclass multilabel classification achieved a lower overall accuracy score, but is solving a more generalized problem.

6.2. Interpretation of Softmax Regression

Softmax Regression on our data has 0.08 accuracy, and the low accuracy rate can be contributed to the fact that it is a linear model, which cannot model complex correlation.

6.3. Resnet Hyperparameter Tuning

We used two Tesla GPUs on Google Cloud to compute our model, and increased our batch size to the 32, which

is the maximum size we could use with limited space and computing power. We also increased the filter size of our convolutions to stride 32 and 5, and we saw a positive correlation between stride size and accuracy, which unfortunately we could not explore more due to the same reason.

We added 0.5 dropout after each of our initial convolution layers to prevent overfitting, and also used softmax, linear, and relu activation on different layers after experimenting. The model utilizes the Adam optimizer instead of stochastic gradient descent, as it is computationally efficient, has less memory requirements, and is better suited for noisy gradients.

6.4. Interpretation of Resnet

Because we know which features in the second to last layer of our modified ResNet contain metadata tags, we can granularly examine relative magnitudes of weights leading from each of these tags to the corresponding ailment. This allows a basic understanding of which features positively or negatively correlated with the associated condition, from the point of view of the neural network.

In Figure 3, we can see that the strongest signals came from gender for edema and pneumothorax. In particular, a female gender label served as a positive signal for edema, and a male gender label served as a positive signal for pneumothorax. We can sanity check this method of weight examination by observing that the number of hospital visits correlated with nearly all symptoms, which makes logical sense in the context of serious chest conditions.

Atelactasis	Cardiomegaly	Consolidation	
-2.22E-02	-6.12E-01	-4.51E-02	Age
4.98E-01	-4.40E-01	1.48E-01	Gender
2.00E-01	2.36E-01	2.00E-01	# visits
Edema	Effusion	Emphysema	
-4.44E-02	-1.21E-02	-2.83E-02	Age
-1.45E+00	-2.93E-01	-2.46E-01	Gender
2.36E-01	2.16E-01	1.76E-01	# visits
Fibrosis	Hernia	Infiltration	
-1.89E-02	-5.13E-01	-3.16E-02	Age
-3.61E-01	-4.04E-01	-8.78E-03	Gender
1.36E-01	-1.20E-01	1.96E-01	# visits
Mass	No Finding	Nodule	
-7.96E-01	-3.54E-02	-3.36E-02	Age
1.02E-02	-1.55E-01	3.60E-01	Gender
-4.70E-01	1.67E-01	1.60E-01	# visits
Pleural Thickenin	Pneumonia	Pneumothorax	
-5.21E-02	-3.40E-02	-4.86E-02	Age
5.16E-01	-1.10E-01	9.04E-01	Gender
2.13E-01	1.88E-01	2.10E-01	# visits

Figure 3: Weights from Resnet Final Layer

6.5. Model Analysis

Our modified residual network is currently overfitting, as through debugging, we realized that our loss for training is larger than validation, and training for more epochs is reducing training loss but not validation loss. To combat high variance, we added dropout after initial layers of our neural net.

Overall, we believe that with more space and computing power, we can increase the accuracy of our model significantly. Our modified residual network currently trains on 5,000 images, but we have 105,000 more we can utilize so it won't overfit too heavily on our current X-rays.

7. Conclusions

Residual net structures demonstrate considerable potential for modeling X-ray symptom diagnosis. By including basic metadata such as age, gender, and number of hospital visit into a hidden layer of the ResNet50, we were able to successfully predict multiple diagnoses for a given patient roughly 60% of the time, which is significantly more accurate than the 12% confirmation rate currently seen in practice. This prediction accuracy would increase even further with more complex metadata, such as natural language from clinical notes, other lab tests, or other scans.

Additional experimentation is needed with a custom architecture specific to high resolution X-ray images rather than RGB images of a more diverse set of objects. There is also work to be done in testing different ways to integrate the metadata into the network (i.e. introducing it earlier in the network rather than the second to last layer).

With further testing, it is conceivable that an automated multiclass multilabel diagnostic system with accuracy approaching current state of the art binary classification networks for scans could serve as a triaging layer at large hospitals, or a smart assistant to guide busy physicians.

8. Contributions

Joy built the ResNet 50 structure, and the convolutional neural network. Kush extracted the data for reading, defined the accuracy measure and implemented softmax regression. Peter optimized hyperparameters for ResNet, computed statistics and analyzed the results.

References

[1] N. I. of Health, "Nih clinical center provides one of the largest publicly available chest x-ray datasets to scientific community." https://www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-ava 2017.

[2] E. Zimmermann, "Mayo clinic researchers demonstrate value of second opinions." https://

newsnetwork.mayoclinic.org/discussion/
mayo-clinic-researchers-demonstrate-value-of-second-opinions/,
2017.

[3] S. Zhang, P. Xie, D. Wang, and E. P. Xing, "Medical diagnosis from laboratory tests by combining generative and discriminative learning," *CoRR*, vol. abs/1711.04329, 2017.

[4] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, "Deep learning for identifying metastatic breast cancer," June 2016.

[5] M. L.-J. Yann and Y. Tang, "Learning deep convolutional neural networks for x-ray protein crystallization image analysis," *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, vol. AAAI-16, 2017.

[6] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "Chexnet: Radiologist-level pneumonia detection on chest x- rays with deep learning," *CoRR*, vol. abs/1711.05225, 2017.

[7] V. Vryniotis, "Machine learning tutorial: The multinomial logistic regression (softmax regression)." http://blog.datumbox.com/
machine-learning-tutorial-the-multinomial-logistic-regression-softmax-regression/,
2013.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.