Anomaly Detection for Safe Chest X-ray Interpretation with Confidence Estimate and Variational Autoencoder

Nguyet Minh Phu
Stanford University
450 Serra Mall, Stanford, CA
minhphu@stanford.edu

Jun Li
Stanford University
450 Serra Mall, Stanford, CA
junli530@stanford.edu

Abstract

We build a model that can detect fourteen different pathologies from chest radiographs, then teach the model to refrain from making a prediction when given low-quality or bogus inputs.

1. Introduction

Deep learning based chest radiograph interpretation models could provide substantial benefits in many medical settings. However, such models currently have the undesirable property of making predictions of pathology on all given input images, including those that medical doctors (representing the gold standard) are unable to safely make a prediction on, such as (1) chest X-rays that are under/over-penetrated, (2) chest X-rays that are blurry or (3) medical images of non-chest X-rays. A reliable model for clinical use must be able to identify these bad inputs and alert the user to retake the chest X-ray instead of proceeding to make a prediction. The purpose of this project would be to build a model for detecting such images that should not be input into chest radiograph interpretation models.

2. Related Work

Inspired by the work of DeVries and Taylor [3], our first method involves teaching the model to output a confidence estimate in addition to pathology predictions, then using this confidence estimate as a proxy for how likely the model should attempt to make a prediction on the example. Specifically, we use $p' = c \cdot p_i + (1 - c) \cdot y_i$ to obtain our adjusted prediction probability which is an interpolation between the original prediction and the target probability by the confidence c, then we can compute the final loss as a balanced sum of the task loss and the confidence loss, that is $\mathcal{L} = -\sum_{i=1}^{M} log(p'_i)y_i - \lambda log(c)$.

The confidence branch involves combining the output of the last layer of the network to get confidence score $\mathcal{C}$ and can be added onto any existing network architecture as seen in figure 7 in Appendix. This confidence score signifies how confident the network is with regards to its prediction. The more confident the network is, the smaller $\mathcal{C}$ and the less the effect of the correct label y_i is on our adjusted prediction probability. In practice, the network can adjust $\mathcal{C}$ to be 0 to use the correct label y_i for prediction. That is why we add $-\lambda log(c)$ into the loss function to penalized the network for doing so. Thus, the network is prompted to ask for hint through $\mathcal{C}$ only when it really needs.

3. Problem Statement

We develop methods to detect X-rays that are unsuitable to be input into a chest radiograph interpretation model. Since two of our three proposed methods (baseline and confidence estimate) are closely integrated with the base chest radiograph interpretation model, we first developed a model to concurrently detect the presence of 14 different pathologies from frontal-view chest radiographs. The input is a frontal chest X-ray and the output is 14 labels of 0 or 1 corresponding to whether each of the 14 pathologies is present.

Next, we develop methods to detect unsuitable images for the X-ray interpretation model. The input is an image and the output is a score $\mathcal{C}$ ranging from 0 to 1 indicating how suitable the image is to be input into the chest X-ray interpretation model. Low-quality inputs such as under-penetrated or blurry X-rays should have low $\mathcal{C}$ while high-quality frontal chest X-rays should have high $\mathcal{C}$.

4. Methods

An obvious approach is to build a detection model to differentiate good-quality chest x-rays (in-distribution examples) from all bad inputs (out-of-distribution examples). However, this is difficult because there is no comprehensive dataset of "all bad examples that should not be fed into a chest radiograph interpretation model." Thus, we develop and compare a baseline and two methods to detect bad inputs that *do not require bad inputs during training*.

4.1. Baseline: Probability Threshold

Our baseline method is inspired by Hendrycks & Gimpel's idea of using a threshold on the predicted softmax class probability. They observed that correctly classified examples tend to have greater maximum softmax probabilities than erroneously classified and out-of-distribution examples, allowing for their detection [4]. In our case, each of the 14 pathologies is predicted independently, and a sigmoid function is used to convert the score to a probability P ranging from 0 to 1. We note that if the model is more confident about its prediction, then P should be further away from 0.5. A very high P indicates that the model is very confident that the pathology exists and a very low P indicates that the model is very confident that the pathology does not exist. Thus, we develop the confidence score $\mathcal{C}$ as seen in formula (1), where we take the absolute value of the difference of the average of P_i for all 14 pathologies and 0.5 as a metric for measuring how confident the model is/ how suitable the example is to be used for making prediction. The multiplication by 2 is to make the number range from 0 to 1, allowing a probabilistic interpretation.

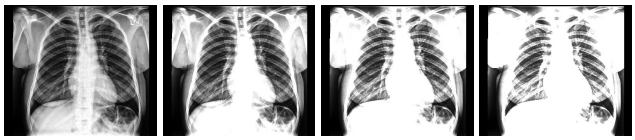
$$\mathcal{C} = 2 \times |\frac{1}{14}(\sum_i P_i) - 0.5|, \text{ where } i \in \text{pathologies} \quad (1)$$

4.2. Variational Autoencoder

Inspired by the anomaly detection method proposed by An and Cho [2], we use the reconstruction probability from the variational autoencoder (VAE) that takes into account the variability of the distribution of variables to build a generative model of chest X-rays of satisfactory quality. The reconstruction probability makes a more principled anomaly score than the reconstruction error, which is used by autoencoder and principal components based anomaly detection methods. Previous experimental results on this proposed method show that it outperforms autoencoder based and principal component based methods. Utilizing the generative characteristic of the variational autoencoder enables deriving the reconstruction probability of the data to analyze the underlying cause of the anomaly.

Specifically, we will use a probabilistic encoder and decoder that paramterize an isotropic normal distribution in the latent variable space and the original input variable space respectively. We will train these using our in-distribution training data.

During testing, for each of the sample we draw from the encoder, the probabilistic decoder outputs the mean and variance parameters, which we use to generate the probability of the image coming from the original distribution. If the probability is below a certain threshold, we classify the example as out-of-distribution.



(a) Original (b) Level 1 (c) Level 2 (d) Level 3

Figure 1: Example on levels of under-penetrated X-rays

4.3. Generative Adversarial Network

5. Dataset

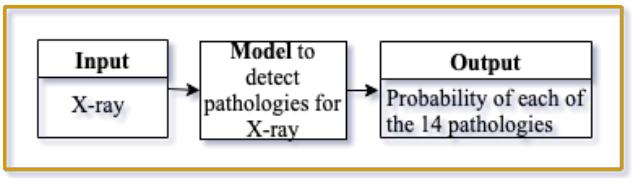
5.1. In-distribution Data

For in-distribution data, we leveraged CheXpert [6], a dataset consisting of 224,316 chest radiographs of 65,240 patients published by the Stanford Machine Learning group. We additionally filtered the dataset for only frontal chest X-rays to train our model. We used the given validation set for validation, and similarly filtered the validation set for only frontal chest X-rays. We randomly sampled 10,000 chest X-rays from different patients for testing, and excluded X-rays from those patients in the training set to prevent the problem of patient overlap.

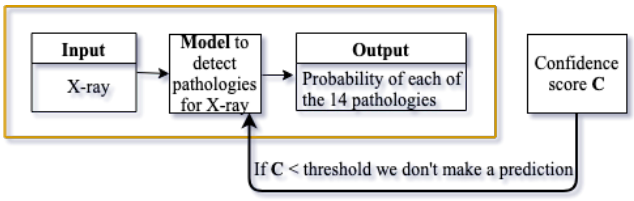
5.2. Out-of-distribution Data

For out-of-distribution test data, we used three datasets corresponding to the three suggested out-of-distribution cases above. Specifically, (1) original CheXpert X-rays test examples with contrast and brightness changed to simulate under/over-penetration[1], (2) original CheXpert X-rays test examples with Gaussian noise filters applied to to simulate blurriness, (3) CheXpert lateral chest X-rays (in contrast to frontal chest X-rays used for training) to simulate X-rays of a different category. For each of the cases except for the last, we used 10,000 test examples and applied the respective transformation to produce the corresponding synthetic out-of-distribution dataset. We used three levels for each transformation. Level 1 corresponds to mild transformation where the X-ray is distorted to a small extent such that doctors should still be able to make a confident prediction. Level 2 corresponds to moderate transformation, where diagnosis confidence should drop. Level 3 corresponds to severe distortion where even experienced doctors should not be able to make any diagnosis. We manually inspected the data and tuned the hyper-parameters of the transformations for each of the three levels. Examples are given in figure 1, 5, 8 and 9. (Figure 8 and 9 are in Appendix.)

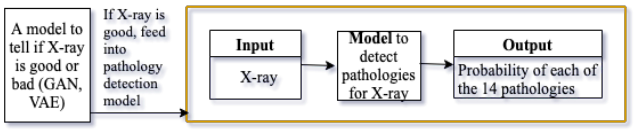
[1]Penetration is the degree to which X-rays have passed through the body. Assessment of penetration is traditionally a standard part of assuring chest X-ray quality.



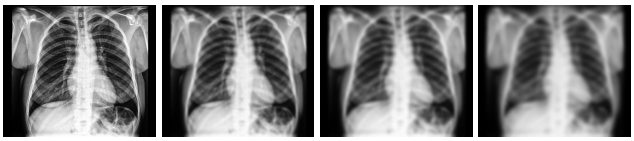
(a) Base Model



(a) Probability Threshold



(a) Cascading Model



(a) Original (b) Level 1 (c) Level 2 (d) Level 3

Figure 5: Example on levels of blurred X-rays

6. Experiments/Results/Discussion

6.1. Train a model to detect 14 pathologies

We first trained a 121-layer Densenet architecture as the baseline model for which we will use the class output probabilities to compute $\mathcal{C}$ for our baseline. We chose Densenet as it is a state-of-the-art model that also requires less computation to achieve high performance [5]. Moreover, DenseNet is also the best performing single model on the leaderboard for the pathology identification competition based on CheXpert, our in-distribution dataset [1].

We used Adam optimizer with a learning rate of 0.0001 and weight decay of $1e^{-5}$ and trained for 3 epochs. We mapped the uncertainty label in the dataset to 1. This corresponds to the U-One model used by Irvin et al [6]. Our model is developed using PyTorch.

	Original	Level 1	Level 2	Level 3
Under-penetrated	0.7334	0.7302	0.7024	0.6664
Over-penetrated	0.7334	0.7119	0.6758	0.5717
Blurred	0.7334	0.7043	0.6935	0.6819
Lateral	0.7334	0.6109	N/A	N/A

Table 1: Mean AUROC for 14 pathologies

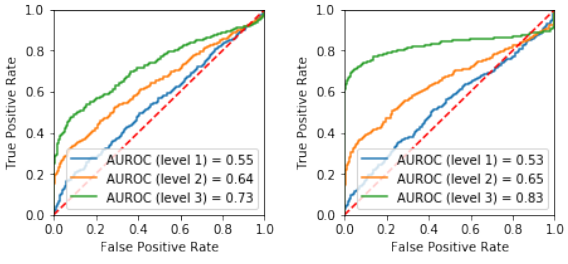
The AUROC scores for detecting 14 different pathologies are reported in table 2 under column titled "Original" in Appendix. The mean AUROC score is 0.7334 on the test set.

Next, we tested the best checkpoint on the different out-of-distribution datasets. The AUROC scores for 14 pathologies for under-penetrated data across 3 levels are reported in table 2 in Appendix. The mean AUROC scores for 14 pathologies for all out-of-distribution datasets across 3 levels are reported in table 4. As we can see, the AUROC scores drop when going from the original to distorted data, and also the higher the level of distortion, the bigger the drop in AUROC is. This shows that our deep learning model may not be able to perform well on poor-quality inputs, and it indeed should refrain from making a prediction. That is, when given a poor-quality image or an image of the wrong type, the model should say "I don't know" instead of proceeding to make a prediction.

6.2. Baseline Result: Differentiating good and bad inputs

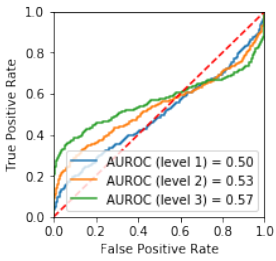
Next, we implemented the baseline using confidence score computed on class probabilities as in formula (1) to differentiate between good-quality frontal chest X-rays (in-distribution data) and poor-quality frontal chest X-rays and lateral chest X-rays (out-of-distribution data) for each of our out-of-distribution datasets. We plotted the ROC curve for this task across different datasets in figure 6. As we can see, the AUROC score for differentiating between good and bad inputs increases as the level of "badness" increases. This is sensible since the worse the out-of-distribution examples are, the easier it is for the model to differentiate between the in- and out- of-distribution examples, and the higher the AUROC score should be for this binary classification task. Thus, our results show that our baseline confidence score $\mathcal{C}$ computed using class probabilities could be effective in capturing how "bad" or unsuitable an input is to be safely used in a chest X-ray interpretation model.

Moreover, we also see that the AUROC scores for differentiating good and bad inputs vary between the differ-

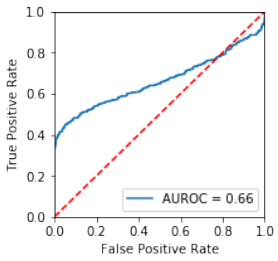


(a) Under-penetration

(b) Over-penetration



(c) Blur



(d) Lateral

Figure 6: AUROC for differentiating good and bad X-rays with **Probability Threshold**

ent types of "badness." For instance, for over-penetrated X-rays, our baseline confidence score $\mathcal{C}$ can be used to effectively distinguish between good and bad inputs at level 3 with a high AUROC score of 0.83, while for blurred X-rays, the corresponding AUROC score is only 0.57. This shows that while our baseline method could be effective for one type of bad inputs, it is not good for all types of bad inputs. A more advanced model will be needed.

7. Conclusion

References

[1] Chexpert: A large chest x-ray dataset and competition. `https://stanfordmlgroup.github.io/competitions/chexpert/`. Accessed: 2019-04-23.

[2] J. An and S. Cho. Variational autoencoder based anomaly detection using reconstruction probability. 2015.

[3] T. DeVries and G. W. Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018.

[4] D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[5] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[6] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *arXiv preprint arXiv:1901.07031*, 2019.

Appendix

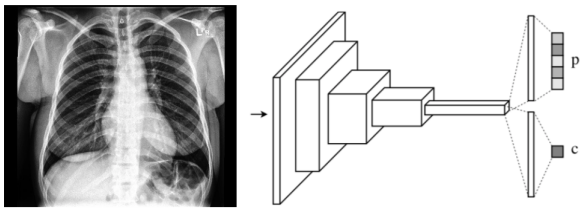
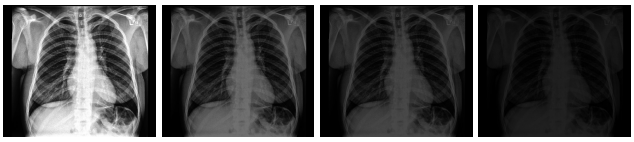
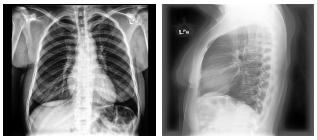


Figure 7: Neural network augmented with a confidence estimate branch



(a) Original (b) Level 1 (c) Level 2 (d) Level 3

Figure 8: Example on levels of over-penetrated X-rays



(a) Original (b) Lateral

Figure 9: Example of frontal-view and lateral-view X-rays

	Original	Level 1	Level 2	Level 3
Mean AUROC	0.7334	0.7302	0.7024	0.6664
No Finding	0.7860	0.7869	0.7737	0.7094
Enl. Cardio.	0.6394	0.6558	0.6176	0.5538
Cardiomegaly	0.8272	0.8020	0.7476	0.6944
Lung Opacity	0.7176	0.7076	0.6834	0.6282
Lung Lesion	0.6647	0.5916	0.6594	0.6582
Edema	0.8593	0.8430	0.8136	0.7571
Consolidation	0.6653	0.7214	0.7353	0.6969
Pneumonia	0.6487	0.6185	0.6302	0.6028
Atelectasis	0.7332	0.7460	0.7156	0.7009
Pneumothorax	0.7960	0.7851	0.7025	0.6378
Pleural Effusion	0.8120	0.8061	0.7745	0.6996
Pleural Other	0.6592	0.7223	0.6306	0.6675
Fracture	0.6000	0.5905	0.5449	0.5760
Support Devices	0.8590	0.8461	0.8051	0.7468

Table 2: AUROC for 14 pathologies - Under-penetration

	Level 1	Level 2	Level 3
Under-penetrated	0.7302	0.7024	0.6664
Over-penetrated	0.7119	0.6758	0.5717
Blurred	0.7043	0.6935	0.6819
Lateral	0.6109	N/A	N/A

Table 3: **GAN** AUROC for differentiating good and bad X-rays

	Level 1	Level 2	Level 3
Under-penetrated	0.8291	0.9454	0.9725
Over-penetrated	0.9395	0.9913	0.9931
Blurred	0.4851	0.4903	0.4415
Lateral	0.7021	N/A	N/A

Table 4: **VAE** AUROC for differentiating good and bad X-rays

The following defining terms are used in AUC and ROC curve:

$$TPR/Recall/Sensitivity = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{TN + FP}$$