Disease Type Prediction

Zhen Yan[1], Yumin Sun[1], Yuchang Zhang[1], and Xiaojing Li[1]

[1]Technical University of Munich

Abstract

Chest X-ray is a widely used medical imaging procedures. A large hospital typically produces over 40,000 chest X-rays per year. However, lacking qualified radiologists to review these X-rays is a major challenge. Reviewing chest X-rays heavily depends on the experience of radiologists and many images are difficult to read when the lesions are in low contrast or overlap with large pulmonary vessels. Since CNN shows an ability to extract hight level features, we explore the possibility of designing computer aided diagnosis for chest X-rays using deep learning approach. We work on 18,000+ images provided from Hackerearth Deep Learning Challenge Cup(HDLCC) to predict single-disease. At last we reach 0.41 weighted f1 score on final off-line test set, with which we locate in the 20th place in the HDLCC (rank 20 / 151 solved / 7079 participants).

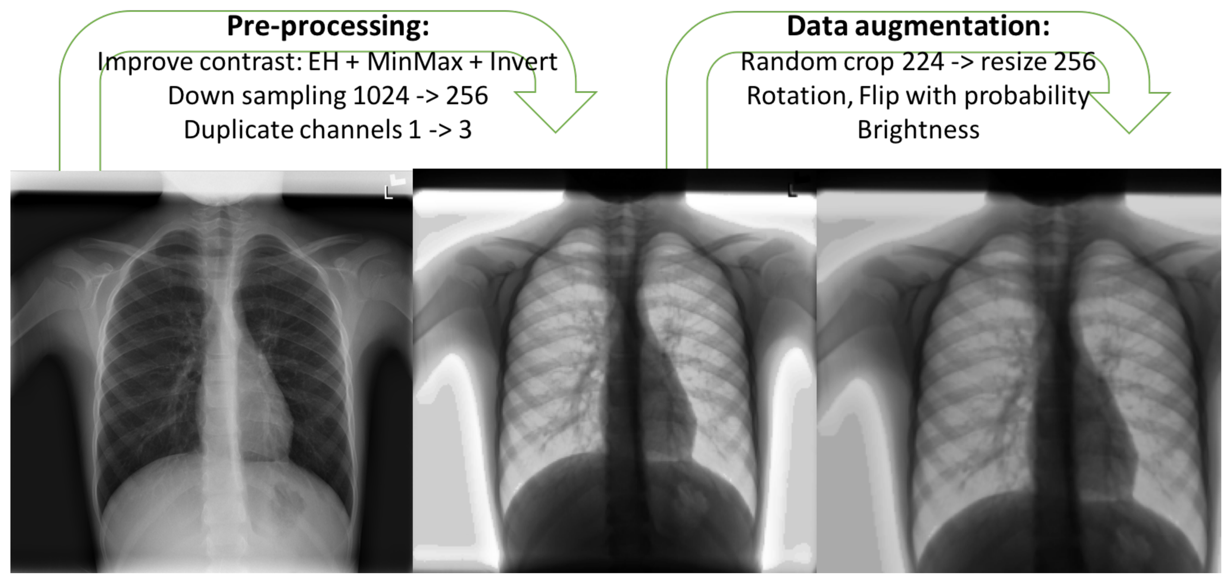
Data pre-processing and augmentation

•Datasets

The dataset includes 18,000+ chest X-ray images of size 1024*1024*1. There are in total 14 diseases and the dataset is strongly unbalanced with 5000+ images from most likely diseases and 50+ images from most unlikely diseases.

•Pre-processing and augmentation

Due to the low contrast in details in X-ray images, image pre-processing is an essential step to gain features. We first improve contrast using equalized histogram and then saturate the darkest and lightest of 0.5% gray values. We also did data augmentation to avoid over-fitting. And we find out even rotation and flips also help to avoid over-fitting, though intuitively those augmented images could never happen in nature.



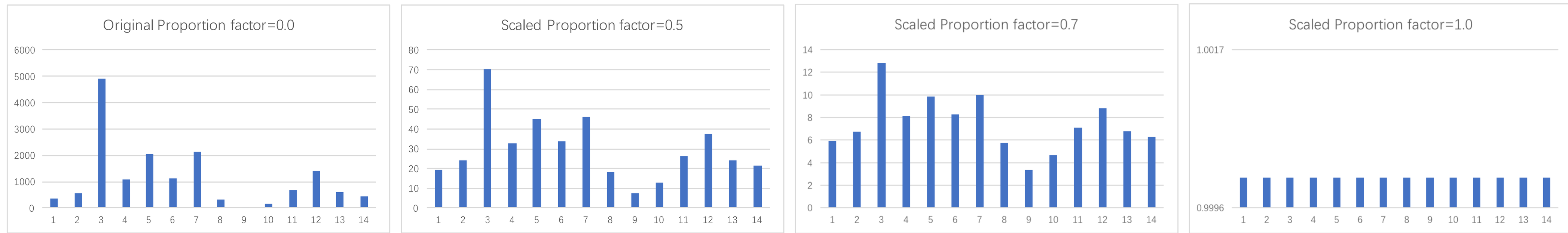
Architecture

•Scalable random weighted sampler

We designed scalable weighted sample to balance the difference of numbers of different types of dieseases.

$$y = \alpha x_i^{\beta}, \ \alpha = \sqrt{\max(x_i)}$$

where x_i is original number of images of i^{th} disearse, $\beta \in [0.5, 1.2]$ is paramter, y_i is number of images of ith disearse after weighted.



•Learning-rate scheduler

Learning rate scheduler is used to accelerate the convergence speed of neural network.

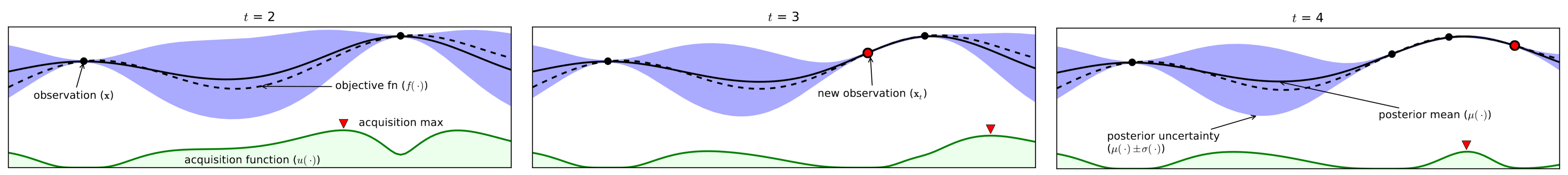
•Bayesian optimization

Optimatize a blackbox, where we input the range of paramters and try to reach better accuacy with the respective paramters.

$$x^* = arg\max a_{LCB}(x; x_n, y_n, \theta)$$

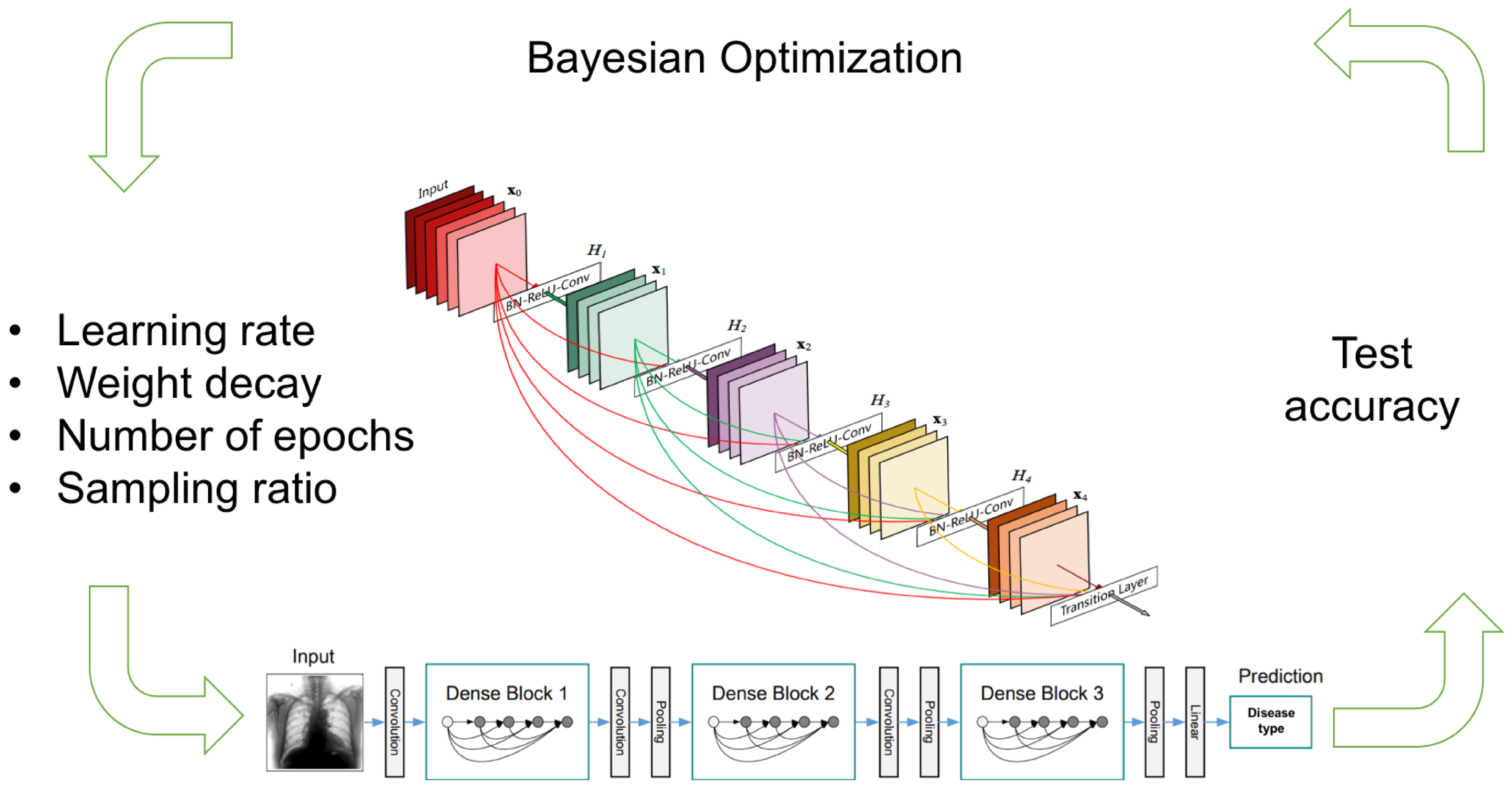
where

$$a_{LCB}(x; x_n, y_n, \theta) = \mu(x; x_n, y_n, \theta) - \kappa\sigma(x; x_n, y_n, \theta)$$



•Network

We treat the denset121 network as the blackbox.



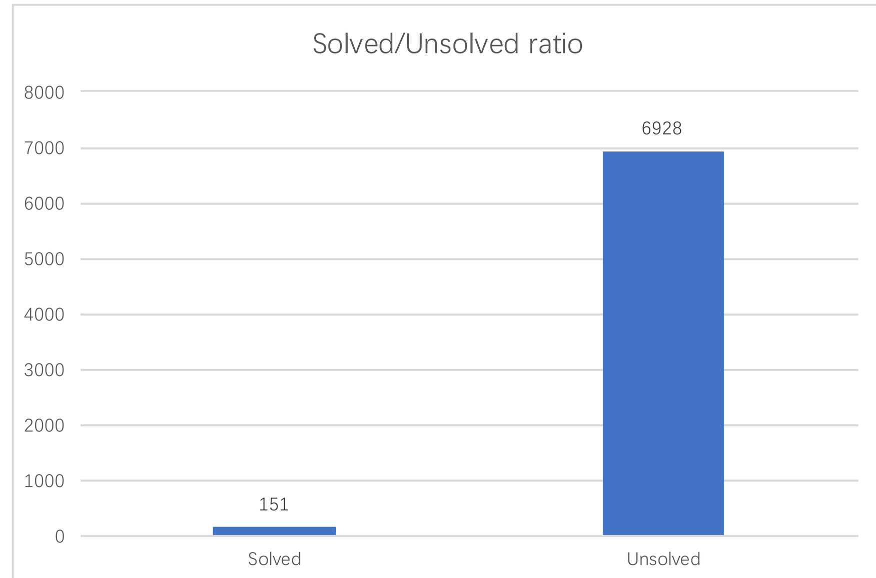
Results

•Our Position

—F1 Score

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

•Solved/Unsolved Ratio



Conclusion

In this project, 18000+ X-ray images of total 14 diseases are trained on alexnet, vgg16, desnet121 and resnet18-101. It shows that they all produced similar results with variance of depth and models. However, training on the pre-trained feature extraction layers usually yields better results than freezing these layers. We therefore conclude that X-ray images indeed have different structures than nature images, so that the general used CNNs are not able to transfered directly to medical images. Considering the highly unbalanced dataset, scalable random weighted sampler and data augmentation are applied. To accelerate the training process, learning rate scheduler and bayesian optimization are introduced. Due to limited image quantity, we could not reach a considerable high accuracy and the final result reaches weighted F1 score of 0.41.