Project Report
CS 330 - Fall 2020
Henry Wang
henryfw@stanford.edu

Multi-Task Training on X-Ray Images

Abstract

This project explores using multi-task learning on x-ray chest images for the detection of pneumonia and pulmonary tuberculosis (TB). The precision and recall statistics for training single-task models, a multi-class model, and a multi-task model are compared. Whether or not the similarities between the two diseases can help improve training accuracies in a multi-task setting is the main topic of the project.

The dataset for pneumonia is about a magnitude bigger than TB. Because of this, the focus is on finding if the results for the smaller dataset can be improved in a mutli-task setting.

The findings show that a multi-task model is able to greatly improve the precision and recall for the smaller TB dataset. The TB single-task precision and recall are 81% and 78% respectively. In the multi-task setting, the precision and recall are 99% and 98%. Even after reducing the degrees of freedom in half in the multi-task setting to address concerns of a larger overall model, the precision and recall are improved to 89% and 81%.

During training all pneumonia images are labeled as false for TB and vice versa. This vastly increased the negative labels for the TB dataset. During prediction, the outputs from the pneumonia images are ignored during calculation. The suspected reason for the improvement to the TB results in the mutli-task model is the increased negative labels as the result of labeling all of the pneumonia dataset as negative during training for the TB branch of the multi-task model.

Code for this project can be downloaded at https://github.com/henryfw/cs-330.

Introduction

The objective of this project is to explore using multi-task learning on x-ray chest images for the detection of pneumonia and pulmonary tuberculosis. Two labeled databases [1, 2] will be used, one for each disease. The precisions and recalls of single-task models and multi-task model are compared. In addition, another model that use a single task but with multiple classes will be compared. Whether or not the similarities [3] between the two diseases can help improve training accuracies in a multi-task setting is the main objective.

Dataset

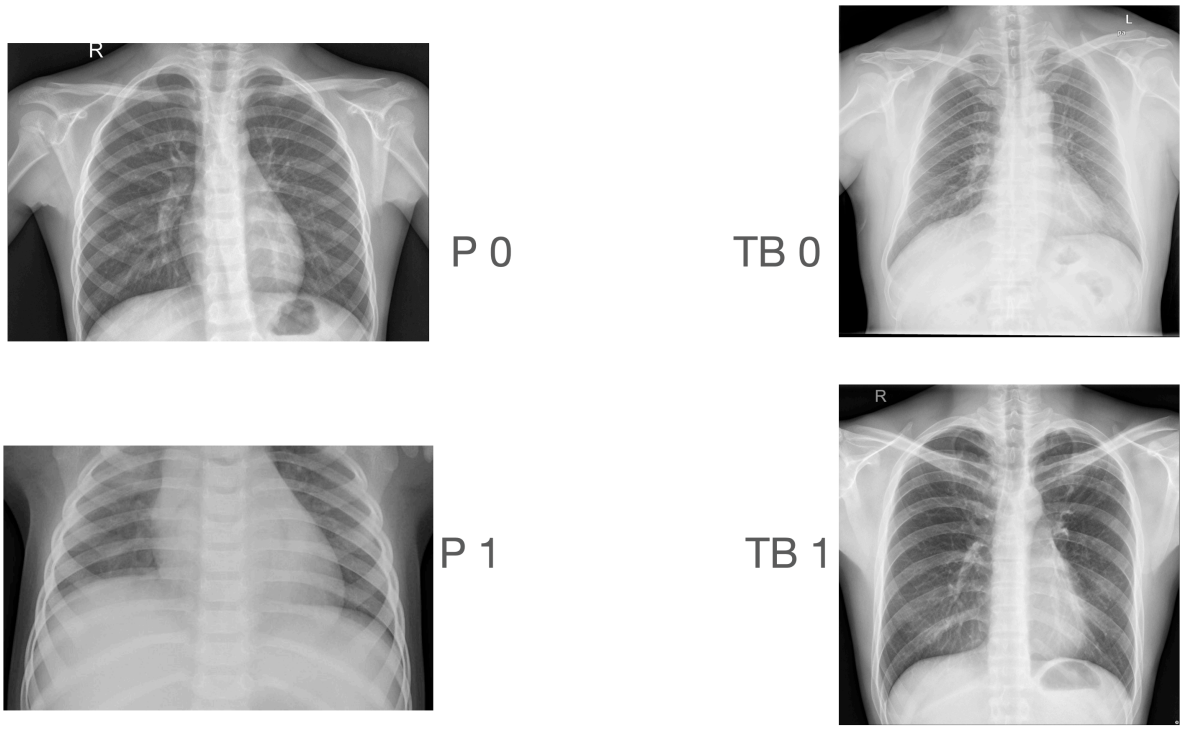

Figure 1. Four samples images from the datasets are shown.

The pneumonia dataset contained about 1600 normal images and 4200 affected images. The TB dataset contained 326 normal images and 336 images of manifestation.

Images are resized to 256x256 grayscale.

Method

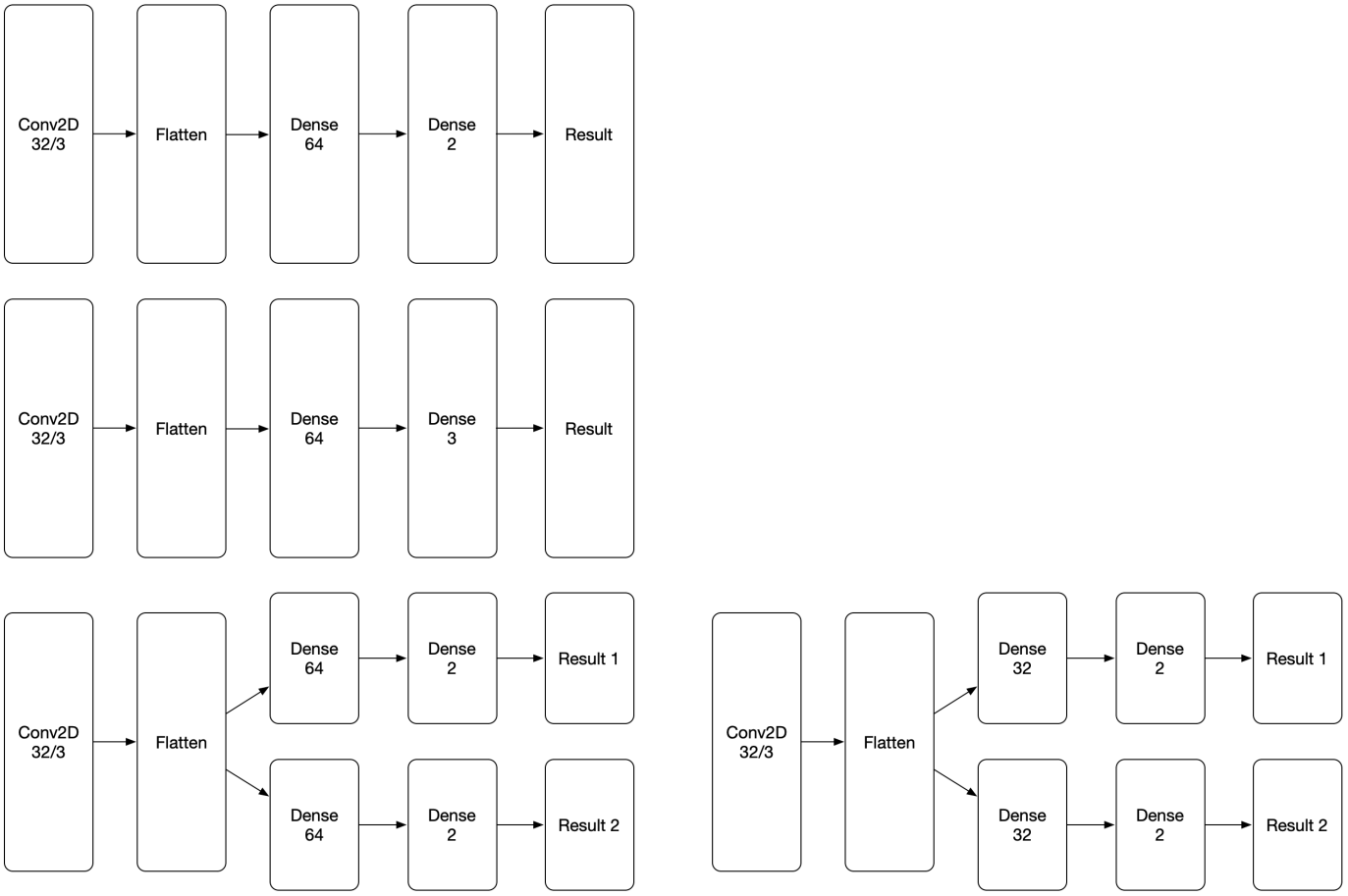

Figure 2. The model architectures are shown. All models use a 2D convolution layer, followed by two fully connected layers. In the The first row is the single-task model. The second row is the multi-class model. The last row is the multi-task models.

The model architectures are shown in Figure 2.

1. Single-task model using a convolutional layer, a sized 64 fully connected layer, and a final output layer for detected and not detected.
2. A multi-class model, similar to the single-task model, with a final output layer of three for pneumonia detected, TB detected, or nothing detected.
3. A mutli-task model with two branches with each branch mirroring the single-task models after branching from the first convolutional layer. Two versions are examined: one with a sized 64 fully connected layer, and one with a sized 32 fully connected layer. The latter is to address possible concerns of having a larger overall degree of freedom.

All training used 80% of the data for training and 20% for testing over 30 epochs. The results are from the 20% of the data used for testing.

All input data are shuffled randomly before each experiment. This may lead to some minor variations in reproduction.

During mutli-task training, all the inputs from the other dataset is labeled as false. This adds additional false labels to each original dataset. This method affects the smaller TB dataset more. When calculating precision and recall for one of the tasks, only the predictions from the associated inputs are looked at. In other words, we ignore the TB prediction in statistics if the input image is from the pneumonia dataset and vice versa.

Result

	P Precision	P Recall	TB Precision	TB Recall
Single	95.7	97.3	**81.0**	**78.3**
Multi-Class	96.1	95.5	**85.7**	**77.1**
Multi-Task 64	99.5	99.3	**98.7**	**97.5**
Multi-Task 32	94.0	96.0	**88.9**	**81.2**

Table 1. The percentages for precision and recall are displayed. The results for the TB experiments are in bold. The multi-class model improved on the single-task model. Both the sized 64 fully connected layer and 32 fully connected layer performed much better than both of the other two for TB.

The results for the larger dataset for pneumonia did not improve in the multi-class setting. However, in the sized 64 fully connected layer multi-task model, the results are much better, increasing from 95.7% to 99.5% for precision and from 97.3% to 99.3% for recall. The handicapped sized 32 fully connect layer performed slightly worse than the single-task model.

For the smaller TB dataset, precision improved from 81% to 85.7% in the multi-class setting. The unhampered mutli-task model improved precision vastly further to 98.7% and recall to 97.5%. The handicapped mutli-task model also outperformed both of the other two architectures.

Discussion & Conclusion

When using the sized 64 fully connected layer version of the multi-task model, the results were much better for both TB and pneumonia. This may be the result of each task in the mutli-task model being able to see more negatively labeled data. In addition, perhaps by sharing the first convolutional layer, the similarities in appearance between the two diseases are captured and utilized.

Pneumonia is less impacted from the mutli-task setting. This is probably because the dataset for pneumonia is a magnitude larger than for TB. Conversely, the enormous improvement for TB probably resulted from the sharing of the much larger pneumonia dataset.

From these experiments, it may be concluded that the increased number of training samples that are marked as false from the pneumonia dataset used during training, but ignored for predictions vastly helped improve the result of the smaller TB dataset in the multi-task setting.

Related Work

A lot of research has been done into using multi-task learning on x-ray images. Farag and et al. [4] explored using multi-tasking learning on x-ray images to accomplish three tasks: diagnosis, segmentation and localization. Alom and et al. [5] use multi-task learning on x-ray and CT scans to detect COVID-19. Eslami and et al. [6] use multi-task learning on x-ray images to perform the tasks of organ segmentation and bone suppression. The work in this project differs from the related works by focusing on using multi-task learning on two different conditions.

References

1. https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia
2. https://www.kaggle.com/kmader/pulmonary-chest-xray-abnormalities
3. https://www.chpso.org/lessons-learned/lessons-learned-tb-misdiagnosed-pneumonia
4. https://arxiv.org/abs/2008.01973
5. https://arxiv.org/abs/2004.03747
6. https://ieeexplore.ieee.org/abstract/document/8999560