Medical Image Diagnostics for Tuberculosis Detection

Balagopal Unnikrishnan, Rajan Dhingra, Savitha Rani Ravichandran and Sravani Satpathy[1]

[1]***Abstract*— This paper explores the usefulness of transfer learning on medical imaging for tuberculosis detection. We show an improved method for transfer learning over the regular method of using ImageNet weights. We also discover that the low-level features from ImageNet weights are not useful for imaging tasks for modalities like X-rays and also propose a new method for obtaining low level features by training the models in a multiclass multilabel scenario. This results in an improved performance in the classification of tuberculosis as opposed to training from a randomly initialized settings - which we propose is a better way for training in a data constrained setting such as the healthcare sector.**

Keywords: Tuberculosis, detection, classification, X-rays, transfer learning, deep learning, medical imaging.

I. Introduction

Tuberculosis is a bacterial disease primarily affecting the lungs. It is a curable and preventable disease affecting developing nations and is one among the top ten causes of death worldwide. The early identification and treatment of tuberculosis is of great value both in terms of reducing cost of treatment and improving health outcomes [1]. The magnitude of the problem can be inferred from the fact that the World Health Organization has a separate EndTB strategy which aims at reducing deaths caused by tuberculosis and the incidence of tuberculosis by 95% and 90% respectively by the year 2035 [2].

In the detection of tuberculosis, there are two pathways that are relevant. One is the patient-initiated pathway where improved awareness of symptoms among people can help early detection and the second pathway being the screening pathway where low cost screenings are required to be systematically done in patient populations of high risk [3]. In the screening process, medical imaging plays a significant role. Chest X-rays are useful in the useful non-invasive diagnosis and screening tool [4].

Tuberculosis is important to be tackled at a global level and is also important for meeting the UN sustainable development goals and is of great importance to developing nations as they affect mainly the working population. The development of medical imaging techniques and algorithms for tuberculosis detection is hence of great importance - as the reduction in incidence of tuberculosis and its elimination in turn reduces global poverty and improves healthcare outcomes for people of developing nations.

For our work, we seek to look at tuberculosis detection for the large-scale screening from chest X-rays. The systems developed would be able to help in multiple ways. Primarily, by detecting the percentage chance of tuberculosis given a chest X-Ray, it would be possible to determine if more in depth tests is required for the confirmation of the disease. The visualisation makes the models more interpretable and reduces the clinician's overall workload and thereby is useful as a good augmentation strategy in developing nations where healthcare professionals are few in number and stretched on resources.

II. Datasets and Related Work

For the purpose of our experiments, we rely mainly on publicly available datasets. The following are the datasets being considered Montgomery and Shenzhen datasets Jaeger et.al [5] and NIH-14 dataset Wang.et.al. [6] hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases.

The Shenzhen dataset [7] consists of chest X-rays taken from Shenzhen No. 3 People's Hospital which are 4020 x 4892 in dimension. The Montgomery dataset consists of frontal chest X-rays taken from Department of Health and Human Services in partnership with Montgomery County in the United States. These were taken for the purpose of screening of tuberculosis and is similar in dimension to the Shenzhen dataset. For the purpose of the transfer learning experiments we have taken the NIH-14 dataset [6] which is a collection of chest X-rays from clinical PACS databases coming under the National Institutes of Health Clinical Centre. It consists of 112,120 chest X-ray images with labels consisting of 14 common chest pathologies. These do not include tuberculosis as a label and is useful for learning the lower level features while transfer learning is done.

Jaeger et. al [5] extracted several features from lung segmented CXRs and employed various classification methods to benchmark the features. Rajpurkar et al. [8] has previously shown that the NIH-14 dataset can be tackled as a classification problem and has provided useful ways of modelling the problem to be solved by deep convolutional networks. Hwang et. al [9] have previously worked on the Shenzhen dataset. The usage of class activation maps (CAMs) are shown to be useful for visualization of convolutional networks by [10]. For the purpose of training, we have augmented the datasets in a manner that is consistent to the ways in which an X-Ray maybe distorted. Rotations, horizontal flipping and perspective transforms are used as augmentation strategies which mirror the real-world scenarios of image flipping, image skew and flipping which occur while scanning of X-rays. A simple rescaling operation is done to bring the images in the numerical range zero to one is done before feeding to the neural networks. Even though contrast enhancement using Contrast Limited Adaptive Histogram Equalization (CLAHE) was experimented with, as it did not provide any improvement in obtained results.

[1] This research was conducted by the Institute of System Science, National University of Singapore.

Balagopal Unnikrishnan, Rajan Dhingra, Savitha Rani Ravichandran and Sravani Satpathy are the masters students at ISS NUS (balagopal.u@u.nus.edu, rajan@u.nus.edu, savitha@u.nus.edu, sravanisatpathy@u.nus.edu)

III. Experiments

For the detection of tuberculosis, the task is considered to be a binary classification task with the final layer giving the probability that the given X-rays image has tuberculosis. The given set of experiments must deal with answering the following - the effectiveness of using ImageNet weights for transfer learning and methods to improve that performance using pre-trained models. Since the final aim is to develop a system that can be deployable and can be generalized in a much broader sense, in addition to the above, we need to also look at the ideal architectures that can be used, the elements within architectures that are detrimental to the performance for the given task, the selection of loss function such that it can be used for a variety of diseases and finally make the models interpretable for clinical use.

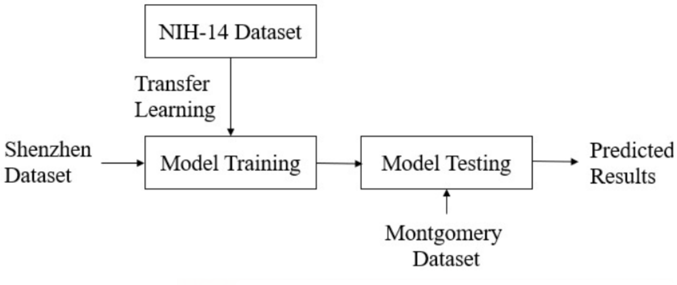

Figure 1: Block Diagram

For comparison of the backbones to be used in the models, we first train models in the following manner - keeping the set of images in Shenzhen dataset for training and validation purposes and testing the models on the Montgomery dataset. In these experiments, keeping the all other parameters such as learning rate, early stopping and the train-validation splits to be constant, we vary the backbone of the architecture from among a set of proven architectures: ResNet-50, VGG16, VGG19, DenseNet-121 and Inception ResNet

Here the models are initialized with ImageNet weights and the training is done for 100 epochs on each model. From this we select the best architecture to be used in the next step of the process. In the next step as well, we keep the train-validation splits same and keep the Montgomery dataset as the test set. In addition, we also require to understand how the connection from the convolution blocks to the dense layers affect the performance. Hence on the best model chosen, we run a set of experiments by connecting the convolution block which gets the image features to the prediction block which is a fully connected layer with a single node. We connect these, first with a set of two fully connected dense layers and then directly by taking the average of the final convolution layers in the following fashion:

$$\mathrm{Res}_{x,y} = \sum_{k\in 1}^{d} \frac{f_{x,y,z}}{d} \;\forall\, x \in X, y \in Y$$

where,

d = Number of feature Maps

$f_{x,y,z} = Value\ at\ position\ x, y, z\ on\ Feature\ map$

X = [1, height of function]

Y = [1, width of function]

On comparing these, we see which elements are prone to overfitting and which is best for generalization in our scenario.

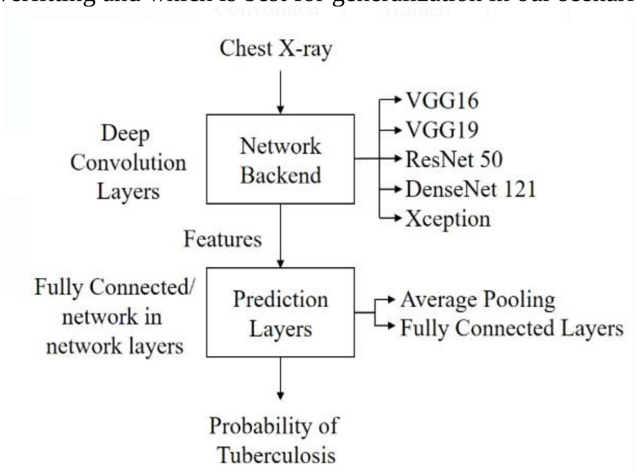

Figure 2: Architecture

Given, the best architecture setting, we see how well the images perform with the ImageNet weights. We also see till whether the features learned on the ImageNet dataset is useful by unfreezing the layers and looking for an increase or decrease in performance. If the model's performance increases as more and more layers are unfrozen, we can conclude that the learned weights on ImageNet are not useful and that the model has to learn a new set of features to better cope with the task at hand. Exploring further to see if another set of pretrained weights can be used, we look at the NIH-14 dataset as an ideal dataset to do the pretraining. The dataset has a variety of diseases (among which tuberculosis is not present) and is from the same modality as that of the data under consideration for tuberculosis. However, the training cannot be done in the same manner as it is no longer a binary classification problem. Hence, we change the last layer to have 14 nodes to indicate each of the diseases and train the images with the same train-test splits as that of [Hwang]. The loss function is now varied to accommodate for multi-class and multi-label classification. This proves to be useful finding the generic loss function that we were looking for - so that the system can generalize to other cases as well. The loss function is now set as:

$$Loss\ Per\ Disease = -y_d \ \log\ P(Presence\ of\ Disease) \\ -(1 - y_d\ \log\ P(Absence\ of\ Disease)$$

$$Total\ Loss = \sum_{i \in presence\ of\ Disease} LPD_i$$

With the metrics remaining the same as before for the tuberculosis classification task. Area under Receiver Operator Curve (AUROC) is used as the evaluation metrics. Here the only difference is that, in a binary setting this is a single value, in a multi-class setting, we take per-class AUCs.

Once the training is done on the NIH-14 dataset, we take the convolution layer weights and revert to our old setting of binary classification setting, now using these weights with the architecture obtained from the initial experiments. Now we compare the effectiveness of pre-training on ImageNet vs. the NIH datasets. Under identical training settings, we look at the AUC scores while we unfreeze more layers for training. If our hypothesis, that NIH weights are better and they help learn the lower layer features better, we should see two trends - the first being that the model with NIH weights should perform better than ImageNet weights and the second being that if the NIH weights are useful, we should see a dip in performance when all layers are unfrozen.

The results are presented in the next section. For interpretability, the team has also presented the class activation masks which have been found to be useful to visualize which portion of the image has contributed to the classification result [10] [11]. This gives the clinicians an added functionality that he can now look at which areas of the X-rays to concentrate more on and also gives an intuitive idea of how the model has interpreted the X-rays.

IV. RESULTS AND DISCUSSIONS

We conducted various test involving the Shenzhen & Montgomery dataset and finalized the structure which is resulting into highest Test Set Accuracy.

A. Backbone Finalization

Firstly, all the backbones were tried on the Model training. Inception ResNetV2 and Dense Net performed equally well on the Test Set (Montgomery).

Model	Initialized weights	Shenzhen AUC	Montgomery AUC
ResNet-50	ImageNet	0.99	0.7
VGG-16	ImageNet	0.5	0.5
VGG-19	ImageNet	0.5	0.5
Inception ResNet V2	ImageNet	0.99	0.8
Dense Net	ImageNet	0.99	0.8

Table 1:

We can notice that performance is better by models with skip connections rather than sequential connections. VGG models do not have skip connections and perform only as good as random guessing as indicated by the 0.5 AUC scores. Also, we can see that ResNet-50 and DenseNet-121 give the best performances. DenseNet is choses as the best among these as it achieves similar performance with three times lesser parameters than ResNet [10].

B. Tuning of DenseNet Models

Secondly, the DenseNet Model architecture is finalized, the layers of the models if finalized as per the table

Model	Initialized weights	Shenzhen AUC	Montgomery AUC
Freeze all	ImageNet	0.65	0.54
Unfreeze 5	ImageNet	0.65	0.7
Freeze all + 2 FCN	ImageNet	0.7	0.69
Unfreeze 10 + 2 FCN	ImageNet	0.8	0.72
Unfreeze All + 2 FCN	ImageNet	0.99	0.74
Global Average Pooling	ImageNet	0.99	0.82

Table 2:

We see that the performance is better while performing the averaging as opposed to fully connected layers. This can be explained by [global-avg-pool paper] as the fully connected layers overfitting to the training set and hence giving poor generalization. This is a known issue of fully connected layers [find paper for this]. Hence, we stick to the second method.

C. Inclusion of Transfer Learning Model Weights

Thirdly, the Transfer Learning Model weights are used, these weights are then tested with freezing different set of layers and results is as per the table.

Model	Initialized weights	Shenzhen AUC	Montgomery AUC
Freeze all	nihc-14	0.85	0.82
Unfreeze 5	nihc-14	0.85	0.81
Unfreeze 10	nihc-14	0.89	0.84
Unfreeze 20	nihc-14	0.89	0.83
Unfreeze all	nihc-14	0.99	0.79

Table 3:

On unfreezing the layers, we see that though the performance initially increases, it then plateaus and then decreases when all weights are unfrozen. This proves our initial hypothesis that these weights are useful and pretraining on the NIH dataset has helped learn useful lower level features. Unfreeze layer up to 10 resulted into best results and it was finalized for the model.

V. MODEL INTERPRETATION

To interpret the network predictions, we also produced heatmaps to visualize the areas of the image most indicative of the disease using class activation mappings (CAMs) [15]. To generate the CAMs, we feed an image into the fully trained network and extract the feature maps that are output by the final convolutional layer. Global Average Pooling [16] in the final feature maps helps us to focus on localization of relevant features and is expressed as class activation map.

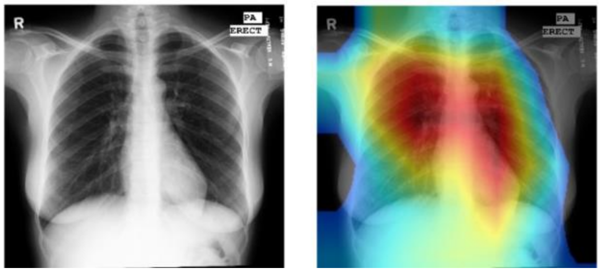

Figure 3: shows example of CAMs on the detection of normal X-rays; it focusses on the clear area in the Lungs.

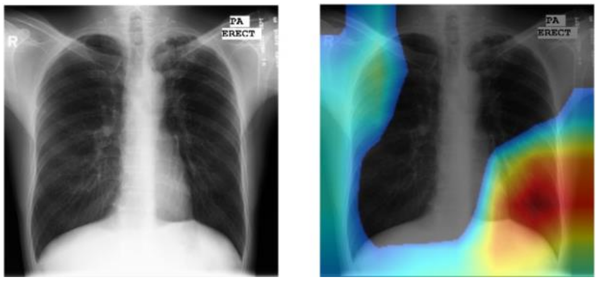

Figure 4: shows example of CAMs on the detection of abnormal X-rays; it focuses on the infected area of the Lungs.

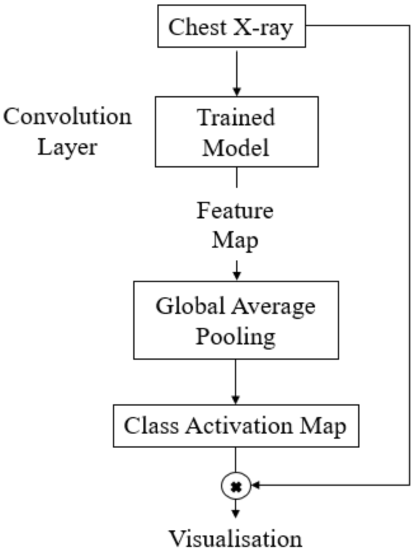

Figure 5: Visualisation Schema

VI. Deployment and Extensibility

The deep learning model was deployed as a python web-app. The pipeline consisted of a python web server. The Bottle framework was chosen for this purpose as it has not external dependencies and has a one file deployability which makes it ideal to package. All code has been written as a combination of TensorFlow and Keras libraries. OpenCV has been used for visualizations and IO tasks while numerical operations and optimizations have been done using numpy. In the interest of teams who might be taking the work forwards, we find the following things to be useful while the model is deployed. I single instance of a model serving results in batches is most efficient as it can use the GPU more efficiently. In the case of one at a time prediction, take care not to reload the models as it just hogs up the memory. The loading of large models to system memory is what takes the most time in these pipelines. Hence as far as possible, avoid duplicate loads or switching between models.

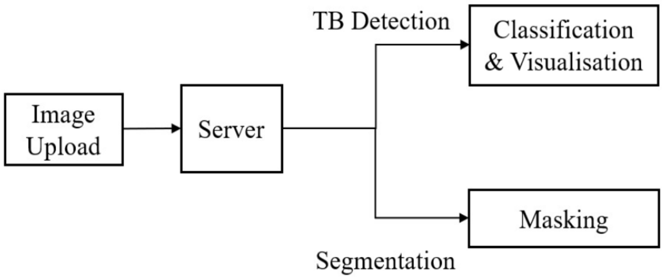

Figure 6: Implementation of Visualization

In certain datasets, we see that the X-Rays contain labels on its side which is indicative of the disease of the patient. This is seen when pre-diagnosed X-Rays are used for data. The model may sometimes consider them as features for prediction and such models will not generalize to the real-world cases. Hence, in this scenario a useful preprocessing technique is to use a trained U-Net to semantically segment out just the area of focus, in this case the lungs and use that for training purpose. This is useful to removing markers like labels and position of certain elements which might cause the model to learn wrong features for prediction. In such cases, it would be best to visualize the model using the modules provided so see where the model focuses on most to get the predictions.

VII. Conclusion and Future Work

The experiments help understand the nature of transfer learning strategies to be used for medical images. In the process, we also develop and application for screening of patients for tuberculosis and have presented a demo consistent with problem statement presented and with appropriate visualizations. We have also suggested the ideal architectures that can be used and the elements that should be avoided to get better generalization. We show that the ImageNet weights are insufficient and the usage of appropriate data for pretraining is important and makes the entire process more efficient.

There can be multiple areas in which this work may be taken forward. One obvious question to be answered is how well the system would perform against human counterparts. A human benchmarking against several of these datasets is required to see both the individual performance and the agreement between various clinicians. From inputs of healthcare professionals, we understand that tuberculosis detection is not just limited to examination of X-rays, but can also take as input various things like patient history, lab reports and tests etc. which help improve the final prediction. An interesting area of research would be to combine these sources together and come up with an interpretable model which would make use of clinical inputs as well as the image data.

Solving the spread and incidence of tuberculosis is one which can give great rewards to developing nations. Though the problem has not been that well studied due to lack of financial motivations for private players, the release of new public datasets has helped in recent years to put this as a prominent research problem.

Acknowledgment

This work is supported by the Institute of Systems Science, National University of Singapore (NUS).

References

[1] Global Tuberculosis Report 2018 by World Health Organisation[http://apps.who.int/iris/bitstream/handle/10665/274453/9789241565646-eng.pdf?ua=1]

[2] Tuberculosis Detection and Diagonsis by World HealthOrganisation[http://www.who.int/tb/areas-of-work/laboratory/early-detection/en/]

[3] Chest Radiography in Tuberculosis Detection [http://apps.who.int/iris/bitstream/handle/10665/252424/9789241511506-eng.pdf?sequence=1]

[4] Global strategy and targets for tuberculosis prevention, care and control after 2015 by World Health Organisation [https://www.who.int/tb/End_TB _brochure.pdf?ua=1].

[5] Wang, Xiaosong, Peng, Yifan, Lu, Le, Lu, Zhiyong, Bagheri, Mohammadhadi, and Summers, Ronald M. ChestX-rays8: Hospital-scale chest X-rays database and benchmarks on weakly-supervised classification and localization of common thorax diseases. arXiv preprint arXiv:1705.02315, 2017

[6] S. Jaeger, A. Karargyris, S. Candemir, L. Folio, J. Siegelman, F. Callaghan, Z. Xue, K. Palaniappan, R. K. Singh, S. Antani et al., “Automatic tuberculosis screening using chest radiographs,” IEEE transactions on medical imaging, vol. 33, no. 2, pp. 233–245, 2014

[7] S. Jaeger et al., "Two public chest X-rays datasets for computer-aided screening of pulmonary diseases", Quant Imaging Med Surg, vol. 4, no. 6, pp. 475-477, Dec. 2014.

[8] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan , Daisy Ding, Aarti Bagul, Robyn L. Ball, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, Andrew Y. Ng: CheXNet: Radiologist-Level Pneumonia Detection on Chest X-rays with Deep Learning.

[9] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In CVPR, 2016

[10] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. arXiv preprint arXiv:1608.06993, 2016a.

[11] Chang Liu, Yu Cao, Marlon Alcantara, Benyuan Liu, Maria Brunette, Jesus Peinado†, Walter Curioso: TX-CNN: DETECTING TUBERCULOSIS IN CHEST X-rays IMAGES USING CONVOLUTIONAL NEURAL NETWORK

[12] Abnormality Detection and Localization in Chest X-rays using Deep Convolutional Neural Networks

[13] A novel stacked generalization of models for improved TB detection in chest radiographs

[14] S. Hwang, H.-E. Kim, J. Jeong, and H.-J. Kim, “A novel approach for tuberculosis screening based on deep convolutional neural networks,” in SPIE Medical Imaging. International Society for Optics and Photonics, 2016, pp. 97 852W–97 852W

[15] Zhou, Bolei, Khosla, Aditya, Lapedriza, Agata, Oliva, Aude, and Torralba, Antonio. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921{2929, 2016.

[16] M. Lin, Q. Chen, and S. Yan. Network in network. International Conference on Learning Representations, 2014.

[17] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In arXiv:1610.02391v3, 2017. 1, 2