Classification of pulmonary tuberculosis lesion with convolutional neural networks

T Karnkawinpong and Y Limpiyakorn
Department of Computer Engineering, Chulalongkorn University, 254 Phayathai Road, Pathumwan, Bangkok 10330, Thailand
Yachai.L@chula.ac.th

Abstract. The concept of computer-aided diagnosis (CAD) for chest x-rays (CXR) has been around for the past fifty years. CAD can help in early diagnosis and reduce the deaths caused by late diagnosis and lack of treatment. Applying deep learning techniques for classification of medical images has seen considerable growth in recent years. Convolutional Neural Networks (CNNs) are a class of powerful generative models well known for image classification and segmentation. This paper has studied three deep neural networks: AlexNet, VGG-16 and CapsNet, for classifying tuberculosis in CXR images. The customized models are created using the datasets acquired from National Library of Medicine and private Thai datasets. Data augmentation with shuffle sampling is used to prevent overfitting in the constructed models. The performance of classifiers has been evaluated with the measures: accuracy, sensitivity and specificity. All model accuracy increases with the augmented dataset. The method of affine transformation has also applied to investigate the model accuracy when predicting the test set contains variant instances unseen in the training CXR images.

1. Introduction
Tuberculosis remains the infectious disease that causes deaths every year worldwide. TB is mostly found in the lung region that is termed as Pulmonary Tuberculosis (PTB). PTB commonly manifests upper lung and it is hard to confirm its presence as it appears different pathological patterns depending on various factors. PTB is a curable disease. However, due to the massive workload of physicians, it inevitably results in long and deleterious time-to-treatment periods for patients. A computer-aided triage system would mitigate these issues by providing the initial CXR interpretation to output a classification of normal/infected per image.

The concept of computer-aided diagnosis for chest x-rays has been around for the past fifty years. Traditionally, CAD systems use machine learning techniques to perform tasks by analysing relationships of existing data. In this study, we have studied deep neural networks, a type of deep learning that employs multiple hidden layers and has been remarkably successful for image classification. The technique is applied for detecting TB lesions on the CXR images. The proposed approach is promising due to its ability to excel with high-dimensional datasets, such as images. Convolutional Neural Networks are a class of powerful generative models with a variety of architectures. They overcome old method performance of image classification without segmentation and manual feature selection. Features are automatically extracted via their own architectures. In literature, researchers have proposed various CNN classifiers such as AlexNet in 2012 [1], VGG in 2014 [2], Inception V3 in 2015 [3], ResNet in 2015 [4], and Xception in 2016 [5]. To overcome the performance of image classification, the latest CNN architecture was proposed by Sabour et al. in
2. Related work
In the world of CXR classification of TB, the first CNN model for TB detection is AlexNet proposed by Hwang et al. in 2016 [7]. In the model, pre-trained was used as the upper layer except in some lower layers where fine tuning was performed. Montgomery and Shenzhen datasets were used as testing data, whereas the model was trained and validated on customized dataset. The AUC performances on test datasets were 0.93 and 0.88 respectively. Caó et al. [8] applied GoogLeNet [9] model for tuberculosis diagnostics on mobile. They used pre-trained model and achieved the accuracy of 89.6%. Hooda et al. [10] used customized CNNs based on LeNet and AlexNet architectures for detection of TB. They used CXR images from Montgomery and Shenzhen datasets. The customized CNNs had 19 layers and consisted of 7 Conv layers, 7 ReLu-layers and 3 Fully-connected layers. Dropout layers were used for preventing the overfitting layer. They compared performance of three different optimizers and found that Adam optimizer overcome the others achieving the accuracy of 94.73% and validation accuracy of 82.09%. Liu et al. [11] proposed TB detection model based on AlexNet and GoogLeNet architectures with different model parameters. The model was trained with a very unbalanced dataset. The shuffle sampling technique was used to augment data resulting in improving the accuracy of AlexNet from 53.02% to 85.68% and improving that of GoogLeNet from 56.11% to 91.72%. Another study by Rajaraman et al. [12] on comparison of deep learning models with CXRs aimed to compare the performance of their customized CNN with five pre-trained CNNs (AlexNet, VGG-16, VGG-19, Xception and ResNet) using Montgomery, Shenzhen, Kenya and India datasets, maintained by National Library of Medicine (NLM) and National Institutes of Health (NIH). The customized CNN model for the binary classifying as normal or TB, consisted of 3 Conv layers, 3 max pooling layers, 1 dropout layer and 2 Fully-connected layers. They used softmax classifier with SGD momentum and L2-regularization optimizer. Each Conv layer was followed by batch normalization [13] and ReLu layer. The performance of customized model achieved accuracy of 0.824 and AUC of 0.900. The comparison results reported that without optimal features, AlexNet achieved the highest accuracy and AUC on all the four datasets. The datasets were randomly split into 80% for training and 20% for testing. However, with optimal features, VGG-16 outperformed the other models.

3. Dataset
Only chest radiographic images of normal and TB lungs were collected from the datasets of National Library of Medicine (NLM) and the Ministry of Public Health, Thailand. All the CXR images contained in four private Thai datasets are deidentified and each of them is confirmed by a radiologist to ensure all images are labelled accurately. The dataset contains totally 1390 images: normal lung 701 images (example as shown in figure 1) and having TB lesion 689 images (example as shown in figure 2). The class label is assigned to value “1” representing TB positive, or “0” denoting TB negative or normal. All images were down-sampled size to 128x128 pixel resolution to suit the input requirements for the customized model.

The method of shuffle sampling is applied for increasing data to prevent overfitting in the constructed models. The size of dataset was increased from 1390 to 1986 images separated into training data of 1588 images (80%), and test data of 398 images (20%). For further investigation of the model performance, another larger dataset was also created using shuffle sampling. The size of dataset was boosted to 3310 images separated into 80% training data (2648 image), and 20% test data (662 images).
4. Architecture
In this work, we use a Windows system with Intel(R) Core(TM) i7-6700k CPU @ 4.00GHz, 8 GB RAM, a Nvidia GTX 960 4 GB graphical processing unit (GPU), Keras with Tensorflow backend, and CUDA 8.0 for GPU acceleration.

We adopted and customized the AlexNet architecture [12] by reducing the number of filters and neurons in fully connected layers for making the structure simpler, that is, it requires less time and fewer number of parameters computed. The customized architecture contains a sequential, 8-layered CNN for classification as TB or normal. It consists of 5 convolution layers (Conv layer), 3 fully-connected layers (FC layer) and 2 dropout layers. The learning rate is 0.001 and Adam [14] is selected as optimizer algorithm.

The customized model of VGG-16 contains a sequential, 19-layered CNN: 16 convolution layer, 3 fully-connected layers and 2 dropout layers. The learning rate is 0.1, the value of learning rate decay is 0.0001, and the optimizer algorithm is SGD.

The CapsNet architecture presented in [6] is applied in this work. The structure is shallow with 2 convolution layers (traditional convolution layer and primary capsules layer) and 1 fully-connected layer (DigitCaps layer). Conv1 is the layer used for extracting local features that are then used as inputs to the primary capsules (PrimaryCaps). The second layer or PrimaryCaps is a convolution capsule layer with 4 channels. Each primary capsule consists of 8 convolution unit (8D capsules) with a 3x3 kernel and a stride of 2.

5. Evaluation
5.1. Evaluation metrics
The performance of the classifiers is evaluated by three measures: 1) accuracy, 2) sensitivity (also known as recall or the true positives rate) and 3) specificity (also known as false positive rate).

Given that TP denotes the number of true positives; FP denotes the number of false positives; TN denotes the number of true negatives; and FN denotes the number of false negatives. Sensitivity, Specificity, and Accuracy are defined as in equation (1), (2), and (3), respectively.

\[
Sensitivity = \frac{TP}{TP + FN} \tag{1}
\]

\[
Specificity = \frac{TN}{FP + TN} \tag{2}
\]

\[
Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}
\]

5.2. Comparison of accuracy of three models
For AlexNet, the training images are divided into 128 batches and the model is trained for 200 epochs. We randomly initialized weights in each layer. Biases were also random. After each epoch, weights were updated by Adam optimizer [14]. We performed tests using different learning rates and found the
best performance learning rate is 0.001. AlexNet achieved the highest accuracy of 92.90% on shuffle sampling size of 3310.

Regarding VGG-16, the training images were divided into 128 batches and the model was trained for 200 epochs. The weights in each layer were randomly initialized. Biases were also random. After each epoch, weights were updated by SGD optimizer. The learning rate is 0.1, and learning rate decay is 0.0001. VGG-16 defeated AlexNet and achieved the highest accuracy of 94.56% on shuffle sampling size of 3310.

For CapsNet, the training images were divided into 64 batches and the model was trained for 200 epochs. The weights in each layer were randomly initialized. Biases were also random. After each epoch, weights were updated by margin loss function [6] and Adam optimizer [14]. The learning rate is 0.001. The best accuracy of CapsNet yielded 90.33% on shuffle sampling size of 3310. Compared to VGG-16 and AlexNet, CapsNet was defeated. The comparisons of the accuracy among three models using non-shuffle sampling, shuffle sampling size of 1986 and 3310 are visualized in figure 3 (a), (b), (c), respectively.

Summary of all the measures of all classifiers is detailed in table 1, including the values of accuracy, sensitivity and specificity. All classifiers achieve higher accuracy, sensitivity, and specificity on larger datasets. The value of sensitivity is lower than specificity for all models on all datasets. Currently, the highest sensitivity and specificity provided by VGG-16 on shuffle sampling (3310) are 92.83% and 96.06%, respectively.

| Table 1. Summary of performance of the three customized models. |
|---------------------------------------------------------------|
| Performance      | Non-shuffle sampling | Shuffle sampling (1986) | Shuffle sampling (3310) |
|------------------|----------------------|-------------------------|-------------------------|
| Accuracy         | AlexNet 75.18%       | VGG-16 76.26%           | CapsNet 73.74%          |
|                  | AlexNet 83.92%       | VGG-16 81.91%           | CapsNet 81.26%          |
|                  | AlexNet 92.90%       | VGG-16 94.56%           | CapsNet 90.33%          |
| Sensitivity      | AlexNet 73.33%       | VGG-16 74.07%           | CapsNet 72.59%          |
|                  | AlexNet 82.45%       | VGG-16 80.42%           | CapsNet 79.37%          |
|                  | AlexNet 92.51%       | VGG-16 92.83%           | CapsNet 86.32%          |
| Specificity      | AlexNet 76.92%       | VGG-16 78.32%           | CapsNet 74.83%          |
|                  | AlexNet 85.40%       | VGG-16 82.30%           | CapsNet 84.69%          |
|                  | AlexNet 93.24%       | VGG-16 96.06%           | CapsNet 93.80%          |

5.3. Affine transformation
Practically, CXR images are not truly vertical, i.e. there is variance in slightly skew and rotation. To investigate the performance of the three classifiers on the dataset containing affine transformations of CXR, the three models were trained using the shuffle sampling dataset of size 3310 without any affine transformation instances. The constructed three models were then tested with the test set containing 300 normal images and 300 affine transformed images, created with slightly random angle from random data in non-shuffle sampling dataset. The result reported that CapsNet outperformed the others with the accuracy of 68.50%, compared to that of 60.33% of AlexNet and that of 58.50% of VGG-16.

6. Conclusion
We evaluate the efficacy of CapsNet for classifying TB in CXR images, compared to AlexNet and VGG-16. The datasets acquired from NLM and the Ministry of Public Health, Thailand were augmented with the shuffle sampling technique. The results show that the accuracy of all the three models is higher compared to that of non-shuffle sampling. Augmenting the dataset could further increase the model accuracy. With shuffle sampling of size 1986, the best-performing classifier is Alexnet, whereas VGG-16 outperforms the other two classifiers on shuffle sampling of size 3310 and non-shuffle sampling datasets. Compared to Alexnet and VGG-16, CapsNet is insignificantly defeated, its training epochs are much less than, though. Moreover, CapsNet achieves the highest accuracy when testing with the dataset that contains variant instances generated by affine transformation. Currently, with shuffle sampling of size 3310, the VGG-16 classifier achieves the highest sensitivity of 92.83%, and the highest specificity of 96.06%. We believe that deep neural networks could accurately classify TB in CXR images. In order to provide valuable information to significantly decrease time-to-diagnosis, further improvement on achieving extremely high sensitivity (confidently classify as having TB lesion) is desired as well as the improvement of the model so that it could accurately predict the variant and unseen images.

References
[1] Krizhevsky A, Sutskever I and Hinton G E 2012 ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, (Lake Tahoe, Nevada: Curran Associates Inc.) pp 1097-105
[2] Simonyan K and Zisserman A 2014 Very Deep Convolutional Networks for Large-Scale Image Recognition
[3] Szegedy C, Vanhoucke V, Ioffe S, Shlens J and Wojna Z B 2016 Rethinking the Inception Architecture for Computer Vision
[4] He K, Zhang X, Ren S and Sun J 2016 Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 770-8
[5] Chollet F 2017 Xception: Deep Learning with Depthwise Separable Convolutions
[6] Sabour S, Frosst N and E Hinton G 2017 Dynamic Routing Between Capsules
[7] Hwang S, Kim H-E, Jeong J and Kim H-J 2016 A novel approach for tuberculosis screening based on deep convolutional neural networks. In: SPIE Medical Imaging: SPIE) p 8
[8] Cao Y, Liu C, Liu B, Brunette M J, Zhang N, Sun T, Zhang P, Peinado J, Garavito E S, Garcia L L and Curioso W H 2016 Improving Tuberculosis Diagnostics Using Deep Learning and Mobile Health Technologies among Resource-Poor and Marginalized Communities. In: 2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), pp 274-81
[9] Szegedy C, Wei L, Yangqing J, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V and Rabinovich A 2015 Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1-9
[10] Hooda R, Sofat S, Kaur S, Mittal A and Meriaudeau F 2017 Deep-learning: A potential method for tuberculosis detection using chest radiography. In: 2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), pp 497-502
[11] Liu C, Cao Y, Alcantara M, Liu B, Brunette M, Peinado J and Curioso W 2017 TX-CNN: Detecting tuberculosis in chest X-ray images using convolutional neural network. In: 2017 IEEE International Conference on Image Processing (ICIP), pp 2314-8
[12] Rajaraman S, Antani S, Candemir S, Xue Z, Kohli M and Thoma G 2018 Comparing deep learning models for population screening using chest radiography
[13] Ioffe S and Szegedy C 2015 Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift
[14] Kingma D and Ba J 2014 Adam: A Method for Stochastic Optimization