Using Deep Convolutional Neural Networks to Optimize Pulmonary Nodule Classification and Localization in Radiograph Imaging for Early Lung Cancer Detection

James L. Wang

Personal Section

I've always had a fond interest in mathematics since I was young. From my parents quizzing me arithmetic on late evening walks to mastering the abacus to learning number theory, mathematics has constantly perplexed me. It always seems as if the amount of mathematics I've learned is never enough. But so are the applications mathematics has in our daily lives, from splitting the bill to Simpson's Paradox to rocket science, the beauty of numbers integrating seamlessly into our natural world is a spectacle to appreciate on its own.

My interests in artificial intelligence didn't emerge until high school when I became fascinated by speculations of how machine learning could "magically" drive cars and predict geological events. That's when I began exploring this field, using open libraries such as Tensorflow to create simple applications such as image recognition. But I didn't understand how and why these applications worked the way they did, and that's when I began to delve into its mathematical principles.

This coincided with my research opportunity at the University of California, Santa Barbara through the Research Mentorship Program (RMP), where I worked with fellow peers and graduate students on computer vision tasks that required machine learning. There, prior to doing any research on our assigned tasks, we spent two to three weeks just learning the mathematical principles that drive the foundations of artificial intelligence. At the time, the highest math I have learned was AP Calculus, but most of the math I worked with involved multivariable calculus and linear algebra. Both subjects had a mildly steep learning curve given the amount of time I had to master them, but that was just the beginning. Each week we would also have literature review sessions, where we would discuss published work relevant to our field, but the math itself was sometimes overwhelming. One main takeaway I had from this experience

was using math as a formal language to generalize concepts. For example, I would attempt to trace out the mathematical operations involved with simple neural networks to grasp a full understanding, but its complexity increased exponentially as new layers and nodes were being added that there were simply too many to keep track of by hand. Nonetheless, this mathematical background in machine learning gave me a concrete foundation to understanding how I would be modifying neural network training rather than a trial-and-error approach. Throughout this experience, it gave me the confidence and resources I needed to take on a machine learning project on my own from coming up with the project to writing a research paper. While machine learning still remains a black box in many ways, the overarching principles are well within my control.

My research for the Regeneron Science Talent Search (STS) was done immediately after I concluded my research at RMP. Inspired by the implications artificial intelligence has and its potential in the medical field, I decided to integrate these two fields together by looking at lung cancer diagnosis, which is often overlooked to more prominent ones such as breast and skin cancer. Following the same mathematical principles as my research at RMP, I found the research process to be much more streamlined, but there were still obstacles I had to overcome. Different ones this time, ones that required me to think in areas of mathematics I haven't thought of before when it came to optimizing algorithms.

All in all, the integration of mathematics with science bears an immeasurable value in modeling our physical world. My advice for those interested in these two disciplines is to go for it. Don't be afraid when you're challenged with math that may appear difficult at first. There are countless opportunities out there synthesizing the two fields together, meaning you can focus on a specific part of math or science you are interested in, whether that's modeling with differential equations or group theory. Regardless of what it is, you'll find your "mathemagical" niche between numbers and the real world.

Abstract

Preliminary diagnosis of lung cancer has led to countless cases of overtreatment due to false positive classifications made by physicians and radiologists. Most commonly, the misclassification of a benign pulmonary nodule (PN) as malignant from chest X-ray images initiates this process for patients. In the advent of promising machine learning and computer vision models, we investigate the optimization of benign and malignant PN classification using deep convolutional neural networks through transfer learning by fine tuning its convolutional layers. Specifically, we look at how fine-tuning the *VGG19* convolutional neural network model differently affects its classification accuracy. With our optimal model, we test its efficacy in localizing and classifying PNs on chest radiographs using a selection search-based scanning method. We found that fine-tuning the last convolutional block yields the highest predictive performance. Using a reserved image test set, our model is able to yield a classification accuracy of 77% compared to published models yielding 68%. This methodology can be easily generalized and applied to other medical imaging tasks.

Introduction

Lung cancer is the leading cause of death and the second most common type of cancer in both men and women [1]. As people become older, the chances of developing lung cancer increases significantly [2]; however, it can be treated if it is caught early, which is typically through a computerized tomography (CT) scan or preliminary chest X-ray imaging. While X-rays are not as effective as CT scans for early detection, it is the most common technique used for patients, especially those at a low risk for lung cancer. X-rays are preferred by patients for its lower radiation exposure and convenience. Additional imaging and tests are usually conducted if abnormalities are found in the preliminary chest X-ray image, making this the first step in lung cancer diagnosis.

From a chest X-ray image, radiologists can determine the presence of pulmonary nodules (PN), which are small focal radiographic opacities [3]. Radiologists need to classify the PNs as either malignant or benign. Because these distinctions are often very subtle [4], radiologists tend to misclassify benign PNs as malignant. However, even before misclassifying a benign PN, it is likely that a radiologist would misidentify false positives from X-ray noise or other artifacts such as malformations and hemangiomas as malignant PNs. To mitigate inaccurate classifications, computer aided detection (CAD) techniques have been developed using traditional image processing techniques such as considering the contrast of the PN with the ribcage [5]. However, since these models are static, they can only recognize malignant PNs based on algorithms manually coded, resulting in a 68% accuracy [6]. These misclassifications almost always lead to further testing, making lung cancer one of the most overdiagnosed and overtreated diseases [7].

As current methods are ineffective, we examine this problem using a computer vision data- driven approach in the advent of machine learning. Literature review suggests there are features in chest X-rays that allow for benign and malignant PN classification [8]. For example, calcification, which can be detected as a lesion with distinct patterning, is correlated with benign PNs. Image preprocessing is also an important step toward

PN classification, including lung field classification, rib segmentation, and rib suppression [9]. This process is able to reduce the noise caused by the superimposed nature of X-ray images.

A more recent investigation in PN classification uses ResNet to improve the detection of PN [10], but there is no significant classification improvement to differentiate between malignant and benign PNs, yielding only 68% accuracy. This research was limited in that only the final network layer of the convolutional neural network (CNN) was retrained. In our research, we look at improving current CAD systems by fine-tuning existing CNN models at different layers for improving benign and malignant PN classification accuracy. CNNs are fine-tuned by retraining the weights of specific layers and using other layers as feature extractors which is known as transfer learning. Without transfer learning, a very large image dataset would be needed to train an accurate neural network model. In the next section, we address the methodology of our approach in more detail.

Methods

Dataset and Image Preprocessing

We used the chest X-ray dataset provided by the Japanese Society of Radiological Technology (JSRT) association, which contains 93 non-PN images, 54 benign PN images, and 100 malignant PN images [11]. PNs are found in varying locations on the chest, ranging in size from 30 to 170 pixels in diameter. Each PN on a radiograph is labeled as either benign or malignant, as well as its size and location in (x,y) image coordinates. Each PN classification has been validated through CT scans and chemical tests. A sample of benign and malignant PN images is shown in Figure 1. As seen, PNs are often superimposed by the rib cage, making detection and classification difficult.

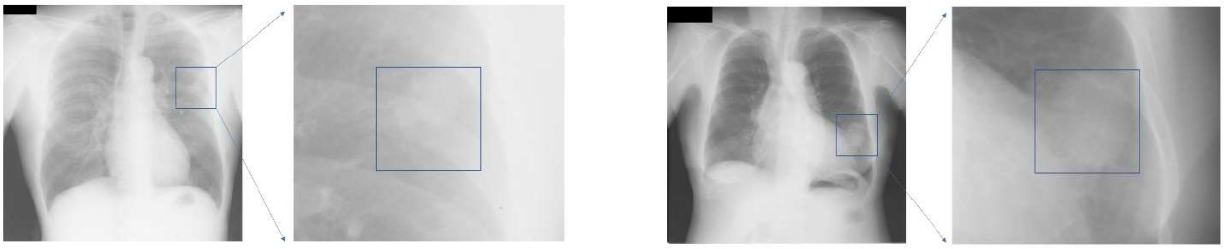

FIGURE 1: Chest radiographs of patients with malignant (left) and benign (right) PNs in different regions of chest. Nodes are boxed on the zoomed-in images.

We used data augmentation techniques to artificially increase its size to 900 total images, with 300 per class. The dataset was then split into respective training, validation, and test sets. The training set is fed into the CNN for transfer learning, the validation set is used to prevent the model from overfitting, and the test set is used to confirm the CNN's predictive power by running it on images that is has not seen before.

Training

Using Tensorflow, we built the *VGG19* CNN. We chose to use the *VGG19* CNN for its deep network of layers, which allows for a high-performing hierarchical representation of visual data [12]. *VGG19* consists of 19 layers: five convolutional blocks are present, with each block consisting of two or four 3 x 3 convolutional layers, followed by a maxpool layer. Maxpooling downsamples an input representation by applying an *argmax* filter on non-overlapping regions of the original representation. By doing this, max-pooling prevents overfitting by providing an abstracted level of representation. The convolutional blocks are connected by 3 fully-connected (FC) layers, followed by a softmax classifier. FC layers connect every node in the CNN and look at the outputs of previous convolutional layers to determine which features correlate most to a class. The final softmax layer is trained by interpreting the outputs of the FC layers as unnormalized log probabilities of the classes and minimizing the cross-entropy loss between them, which is modeled in Equation 1, where for each image *i, fj* is the *j*-th element of the vector of class scores *f* and *yi* is the correct image label [13].

$$\mathcal{L}_i = -\log\left(\frac{e^{f_{yi}}}{\Sigma_j e^{f_j}}\right) \tag{1}$$

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_i + \lambda\sum_{j=1}\left|w_j^2\right| \tag{2}$$

To prevent overfitting, *L2* regularization is used for the softmax classifier. *L2* regularization allows for the cross-entropy loss to favor small weights in the training data. The *L2* regularization parameter is added onto the cross-entropy, as shown in Equation 2 where λ is the regularization strength constant, and *j* represents the *j*-th component of the weights vector *w* [13].

The trained network weights used in this research are from the ImageNet ILSVRC-2014 challenge, which was able to classify a dataset of more than 14 million images into 1000 classes with a 7.0% test error, outperforming other state-of-the-art networks, such as GoogLeNet [12]. In our investigation, we looked at how choosing different layers to fine-tune would affect the model's accuracy. Specifically, we looked at four different cases: removing the last FC layer and retraining the softmax classifier (TF.A); retraining the layers in TF.A and the last convolutional block (TF.B.I); retraining the layers in A and the last two convolutional blocks (TF.B.II); and extracting features from earlier convolutional layers and training a softmax classifier on top of them (TF.C). These four cases are visualized in Table 1.

TABLE 1: *VGG19* transfer learning configurations shown for each implementation [12]. All bolded layers except Softmax layers were fine-tuned using weights from trained models. Other layers were frozen. A rectified linear unit (ReLU) activation layer is placed after each convolutional layer, which is not shown in the table.

CNN Transfer Learning Configuration			
TF. A	**TF.B. I**	**TF.B. II**	**TF.C**
Input (224 x 224 RGB images)			
Conv3-64 Conv3-64	Conv3-64 Conv3-64	Conv3-64 Conv3-64	Conv3-64 Conv3-64
Maxpool			
Conv3-128 Conv3-128	Conv3-128 Conv3-128	Conv3-128 Conv3-128	Conv3-128 Conv3-128
Maxpool			
Conv3-256 Conv3-256 Conv3-256 Conv3-256	Conv3-256 Conv3-256 Conv3-256 Conv3-256	Conv3-256 Conv3-256 Conv3-256 Conv3-256	Conv3-256 Conv3-256 Conv3-256 Conv3-256
Maxpool			
Conv3-512 Conv3-512 Conv3-512 Conv3-512	Conv3-512 Conv3-512 Conv3-512 Conv3-512	**Conv3-512** **Conv3-512** **Conv3-512** **Conv3-512**	**FC-1024** **Softmax**
Maxpool			
Conv3-512 Conv3-512 Conv3-512 Conv3-512	**Conv3-512** **Conv3-512** **Conv3-512** **Conv3-512**	**Conv3-512** **Conv3-512** **Conv3-512** **Conv3-512**	
Maxpool			
FC-1024 **Softmax**	**FC-1024** **Softmax**	**FC-1024** **Softmax**	

Localization Algorithm

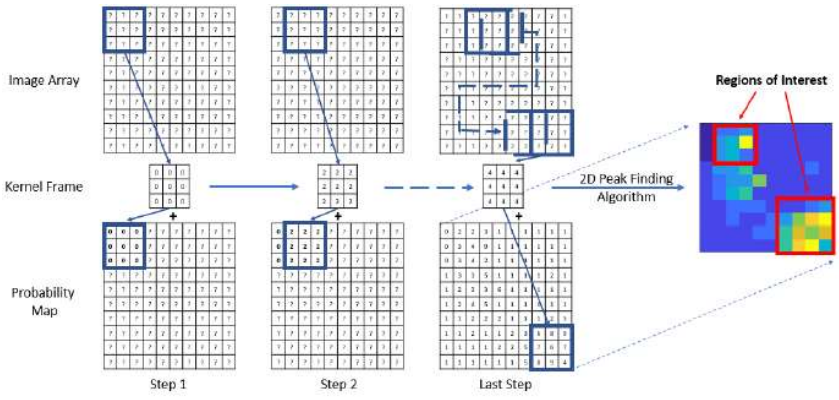

FIGURE 2: Model of our selection search-based scanning method. The image array shows the values that would appear in a one-dimensional grayscale image. Values in the kernel frame show the output value from the trained CNN model, which would be the same in every element of the array, which are added to the current values on their respective probability maps.

After determining which model fine-tuning method yields the highest classification accuracy, we classify and localize PNs on chest X-ray images in the test set. Instead of using the conventional region-based CNN (RCNN) method localization due to its large time and computational expenses for training, we introduce a selection search-based scanning method using a 224 x 224 kernel and convolving it on a chest X-ray image with a padding of 50 pixels, as shown in Figure 2. Since each pixel, except the image borders, overlap each other from the kernel images an equal number of times, we use two probability maps to determine the location of malignant and benign PNs. Two probability maps, one for benign PNs (P_b), and one for malignant PNs (P_m) are first initialized as zero matrices equal to the image size. Then, the kernel frame scans through the entire X-ray image. In each scan, it passes the kernel's values into the trained model, which outputs two values: benign PN and malignant PN probabilities. These probabilities are then added to all elements within the image region the kernel scans on their respective probability maps. After scanning through the entire chest X-ray image, the regions with the highest values in P_m and P_b are identified using a 2D peak-finding algorithm, as higher values increase higher probability of a PN presence. Regions are then boxed for medical diagnosis. As the method above has a complexity of O(n^2), we reduce this exhaustive search through conditional filters so that the kernel does not need to feed every frame through the CNN, such as edge detection.

Results and Discussion

We present the performance of each transfer learning implementation in Table 2, including the model's accuracy on the reserved test set.

TABLE 2: Summary of transfer learning performance for each implementation listed in Methods section.

Transfer Learning Performance Summary

Transfer Learning Implementation	Training Loss	Validation Loss	Training Accuracy	Validation Accuracy	Test Set Accuracy
TF. A	0.007309	0.556430	0.975775	0.678240	0.71
TF. B. I	0.051699	0.507790	0.987297	0.778520	0.77
TF. B. II	0.056787	1.721083	0.982570	0.636682	0.64
TF. C	14.12410	10.87784	0.123711	0.325116	0.33

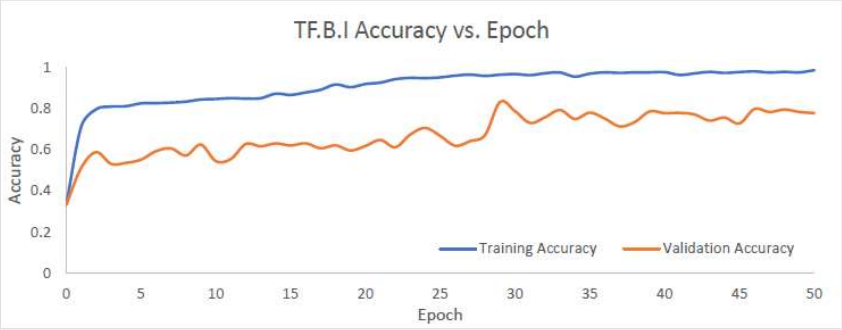

FIGURE 3: The validation accuracy is increasing and is converging with the training accuracy. While a gap is still present between the training and validation accuracy, the rate of convergence suggests that the model is able to generalize the training data.

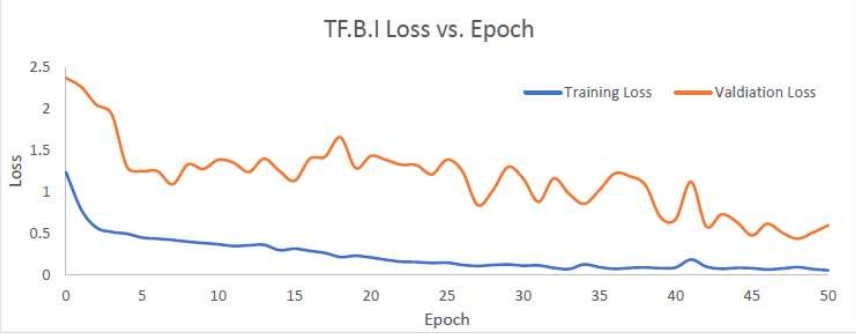

FIGURE 4: Validation loss shows significant downward trend and is converging with training loss. The loss vs. epoch graph provides better insight about the model's ability to generalize data than Figure 3, as the distance between the validation loss and training loss decreases significantly at the end of training.

From Table 2, we find that TF. B. I yields the most optimal solution, with a 77% accuracy for the reserved test image dataset. This is validated by the accuracy and loss plots for TF.B.I (Figures 3, 4), which show

the validation values converging to the training values. This shows that the model is able to generalize the training data

Since TF.B.II yielded the best results, we tested if the model would be able to localize PNs using the selection search-based localization model (Figure 2). Testing this method on all reserved chest X-ray images, we found that the PNs could be localized on a chest X-ray image, but only to its general proximity. In Figure 5, we present the results of localizing PNs on two radiographs, which shows that the model's proposed regions overlaps with the true region, but also includes a large non-PN area. The areas proposed by the model that are outside of the true region originate from the model's classification accuracy limitations and the padding size used. Smaller padding sizes would allow the proposed region to be smaller and more accurate but is computationally more expensive.

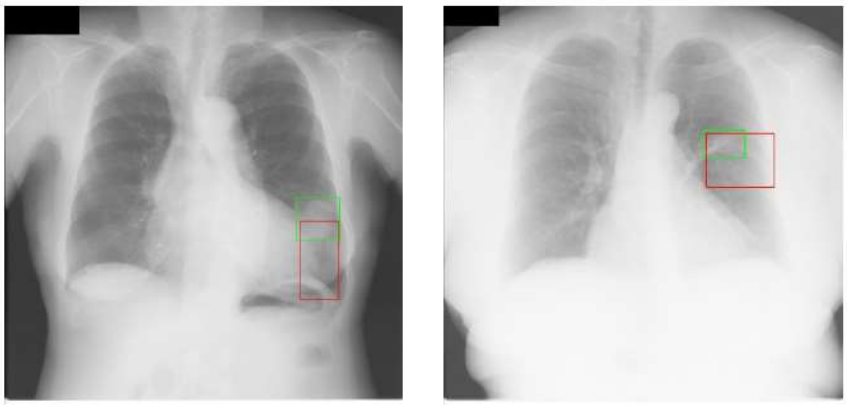

FIGURE 5: Benign (a) and malignant (b) PN radiographs are shown. Green bounding box indicates true PN region and red bounding box indicates region proposed by model. Overlap is present between the true and proposed regions; however, area of proposed region is much larger than true region.

Conclusion

From fine-tuning different transfer learning implementations for optimizing PN classification for chest X-ray images, we conclude that TF.B.I, or fine-tuning the last convolutional block and softmax classifier of the VGG19 model, yields the best performance for this task. From the accuracy and loss values outputted during the training process, the validation loss converges with the training loss, signifying that the trained

model can generalize the training data. Passing the reserved test image set into the trained model yielded a 77% accuracy, which is significant compared to the published models only yielding 68% accuracy [6]. When this model was used to localize PNs using our selection search-based scanning method, PNs were able to be localized to a general proximity of the PN's true region.

Future Work

Future work could involve a larger image dataset which would significantly help reduce overfitting and better generalization with unseen radiographs. With a larger dataset, it is also feasible to fine tune the entire CNN, which would allow for better adaptation to chest radiographs, since many have different high level features from the ImageNet dataset. The training of each CNN model can be improved by using k-fold cross validation technique instead of holdout method. By using the k-fold method, it overcomes the basic drawbacks the holdout method has, such as reserving part of the limited dataset for validation only. For localization, we could decrease the padding of the convolution kernel for a more accurate proposal region. We can also compare the performance of our fine-tuned CNN selection search-based scanning model with a Faster-RCNN.

Bibliography

[1]"Key Statistics for Lung Cancer", Cancer.org, 2017. [Online]. Available: https://www.cancer.org/cancer/non-small-cell-lung-cancer/about/key-statistics.html.
[2]"CDC - Lung Cancer Risk by Age", Cdc.gov, 2017. [Online]. Available: https://www.cdc.gov/cancer/lung/statistics/age.htm.
[3] P. Mazzone, "Pulmonary Nodules", Clevelandclinicmeded.com, 2014. [Online]. Available: http://www.clevelandclinicmeded.com/medicalpubs/diseasemanagement/hematologyoncology/pulmonary-nodules/.
[4] P. Patel, "Solitary Pulmonary Nodule: Overview, Types of Benign Pulmonary Tumors, Etiology", Emedicine.medscape.com, 2017. [Online]. Available: http://emedicine.medscape.com/article/2139920-overview.
[5] S-CB Lo, S-LA Lou, Jyh-Shyan Lin, Matthew T Freedman, Minze V Chien, and Seong Ki Mun, Artificial convolution neural network techniques and applications for lung nodule detection, IEEE Transactions on Medical Imaging 14 (1995), no. 4, 711–718
[6] R T Heelan, B J Flehinger, M R Melamed, M B Zaman, W B Perchick, J F Caravelli, and N Martini, Non-small-cell lung cancer: results of the new york screening program., Radiology 151 (1984), no. 2, 289–293, PMID: 6324279.
[7] Laura J Esserman, Ian M Thompson, and Brian Reid, Overdiagnosis and overtreatment in cancer: an opportunity for improvement, Jama 310 (2013), no. 8, 797–798.
[8] Ali Nawaz Khan, Hamdan H Al-Jahdali, Carolyn M Allen, Klaus L Irion, Sarah Al Ghanem, and Shyam Sunder Koteyar, The calcified lung nodule: What does it mean?, Annals of thoracic medicine 5 (2010), no. 2, 67.
[9] Xuechen Li, Suhuai Luo, Qingmao Hu, Jiaming Li, and Dadong Wang, Rib suppression in chest radiographs for lung nodule enhancement, Information and Automation, 2015 IEEE International Conference on, IEEE, 2015, pp. 50–55.
[10] Isabel Bush, Lung nodule detection and classification, Tech. report.
[11] Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi, Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules, American Journal of Roentgenology 174 (2000), no. 1, 71–74.
[12] Karen Simonyan and Andrew Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).
[13] F. Li, J. Johnson and A. Karpathy, "CS231n Convolutional Neural Networks for Visual Recognition", Cs231n.github.io, 2017. [Online]. Available: http://cs231n.github.io/linear-classify/.
[14] Diederik Kingma and Jimmy Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

All X-ray images used in this report are from the JSRT chest X-ray image dataset.