Lung Nodule Segmentation
Using 3D Convolutional Neural Networks

Research paper Business Analytics

Bernard Bronmans
Master Business Analytics
VU University, Amsterdam

Evert Haasdijk
Supervisor
VU University, Amsterdam

February 2, 2018

Lung Nodule Segmentation Using 3D Convolutional Neural Networks

Research paper Business Analytics

Abstract

Diagnosis of lung cancer caused by malignant nodules in computed tomography (CT) scans is generally performed by pulmonary radiologists in three stages: Nodule localization locates regions of interest within the lung, nodule detection detects if any nodules are present in these regions of interest and nodule segmentation accurately separates any detected nodules from the healthy tissue surrounding them. Recent studies have shown that Convolutional Neural Networks (CNNs) are able to consistently outperform radiologists on both the lung nodule localization and detection tasks. Nodule segmentation has received less attention. In this paper we apply a 3-dimensional CNN model to perform a pixelwise nodule segmentation task. This paper shows that a basic 5-layer 3-dimensional CNN with limited computational capabilities is able to achieve a 87.2% accurate prediction on the pixelwise nodule segmentation task without any feature engineering or extensive preprocessing, suggesting that CNNs might soon also outperform radiologists on the final task.

1 Introduction

Lung cancer is one of the most common types of cancer and the leading cause of cancer related death in developed countries with an approximate 1.6 million deaths reported worldwide in 2012 [1]. High survival rates are strongly correlated with early detection of malignant nodules. Pulmonary nodule detection and localization is done by radiologists spending a significant amount of time analysing X-ray images and 3-dimensional computed tomography (CT) scans of the patient's lungs. Nodules are small pieces of tissue inside the lung which can be benign (noncancerous) of malignant (cancerous). Several charasteristics of the nodule such as size, calcification, lobulation and spiculation of the nodule edges are used as indicators for malignancy. [6]

Originally, computer-aided detection systems (CAD systems) for nodule detection and localization were designed with the intention to reduce workload for radiologists. [11, 13] However, the latest generation of CAD systems has managed to consistently outperform expert radiologists in both the nodule detection and localization tasks [7], suggesting that the CAD systems might completely take over these tasks in the near future.

Although the first Neural Networks designed specifically for lung cancer detection appeared in 1993 [12], the earliest CAD systems using Convolutional Neural Networks appeared in 2005 [15]. Many recent models use a two-step approach in which candidate regions are extracted from the lungs first and subsequently classified on the presence of nodules and expected malignancy [14, 16]. Note that with this approach, the localization and detection tasks are clearly seperated. The extraction of the candidate regions is generally done by a highly sensitive neural network resulting in a high percentage of false positives amonst the candidate regions. The detection task is done by a neural network with high specificity which has the main objective of reducing the number of false positives. The remaining set of candidate regions are manually examined for malignancy by an expert radiologist for a final diagnosis. In the case of a malignant nodule, the radiologist has to manually specify the exact location of the cancerous tissue requiring treatment, which is quite time consuming. Once treatment of a malignant nodule has started, periodic scans are made to monitor progress and the treatment area has to be updated accordingly. This recurring task can be technically labelled as a pixelwise segmentation task, a category in which CNNs have shown significant progress in perfomance lately. A CAD system aiding the radiologist with the segmentation task would decrease the time spent determining the nodule's exact location, resulting in a reduced workload per patient. The segmentation task has so far received relatively little attention by computer scientists compared to the localization and detection tasks.

In this paper we show that with limited computational capabilities and without any feature engineering or preprocessing, a simple CNN is able to achieve a 86.4% accurate prediction on the pixelwise nodule segmentation task.

1.1 Related Work

The latest state-of-the-art models in CAD systems aiding lung radiologists combine the powerful pattern recognition capabilites of Deep Learning with the efficiency of Residual Neural Networks (ResNets) [7] to outperform models with a classic Convolutional Neural Networks setup. Performing the nodule localization and detection tasks on the entire CT-scan at once becomes a feasible option due to the efficiency of ResNets and increased computational capabilities, eliminating the need for extraction of candidate regions.

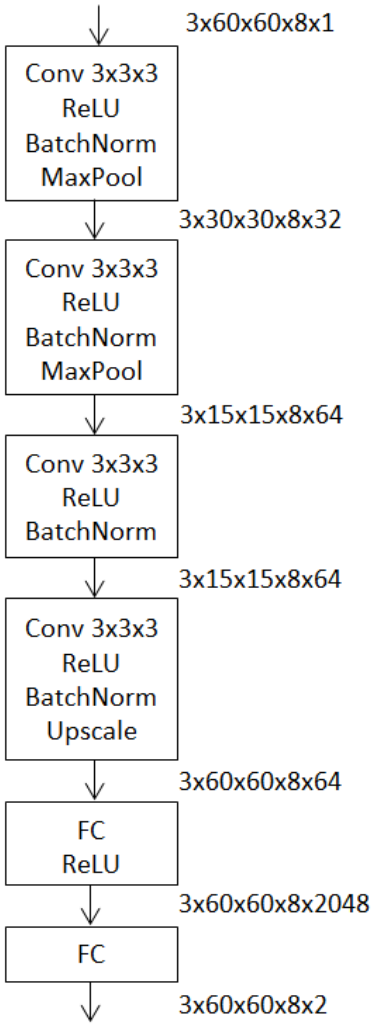


Fig. 1: CNN architecture including data dimension [batch size,x,y,z,features]

2 Methods

Nodule localization is done using a relatively simple fully-connected CNN model implemented in TensorFlow [3], consisting of four convolutional layers followed by a single fully-connected layer, as seen in Figure 1. A single input sample consists of 8 stacked 2D images of 60 by 60 pixels and the batch size of the input is 3. Xavier initialization was used for all weights [5], biases were initalialized at zero. Each convolutional layer applies a 3x3x3 filter on its input with stride of 1 in all three dimensions. Zero padding was applied around the edges of the input to preserve the dimensions of the data. Each convolutional filter is followed up by the ReLu activation function $\max\{val, 0\}$ and subsequently batch normalization. The ReLu allows for more efficient gradient propagation and therefore speeds up convergence of the model. [8]. Batch normalization normalizes the input by the mean and variance of the entire batch to reduce saturation of neurons and improve generalization capabilities [9]. The first two convolutional layers are followed by a 2x2 maxpool

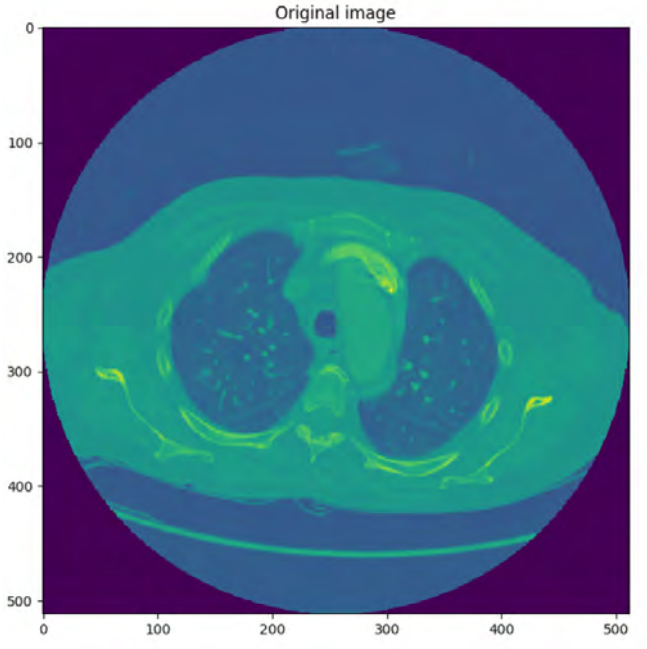


Fig. 2: Slice of CT scan. The colors of the heatmap represent structural density. Dark blue indicates low structural density, e.g. air and fluids; yellow indicates high density such as bone tissue

operation with stride 2 applied to each 2D image in order to reduce computational complexity. The fourth ConvLayer is followed by binary interpolation to upsample the data back to the original input dimensions. The fully-connected layer has 2048 features, which connects to a final fully connected layer with two output channels per pixel which provides the raw logits for pixelwise binary classification.

A softmax activator is applied to the unscaled logits which allows us to interpret the output as class probabilities. Weigthed cross-entropy between this output and the true labels is calculated to measure the model's error. A weighted cross-entropy is used to counteract class imbalance in the data. The average error of all pixels in a sample is used as the loss function to quantify the performance of the model on that sample. Overall performance of the model is quantified by the average loss over all samples. This loss function is minimized by the stochastic gradient descent algorithm Adam [10], using a learning rate of $1e-4$ and an epsilon of $1e-4$. The model was trained on 75% of the available data with 849 original samples artificially augmented to 13,584 variations on the original samples. The model was trained until 100,000 samples had been processed. 135 (11,9%) and 148 (13,1%) original samples were used as test set and validation set respectively.

3 Data

3.1 Image Data

From the Lung Imaging Database Consortium (LIDC) [4], 1018 annotated computed tomography (CT) scans of diagnosed lung cancer patients were used. A single CT scan consists of n images of 512 x 512 pixels, where n is the number of horizontal slices ranging from 80 to 625. Each pixel has one channel of information representing the structural density of the tissue in that pixel. One slice is shown in Figure 2.

3.2 Preprocessing

Each scan in the LIDC database is accompanied by an annotation file containing annotations from up to 4 radiologists. The annotation file header contains several unique identifiers from the corresponding CT scan in order to be able to match them. In the body of the file, each radiologist has differentiated their annotations between three types of nodules: small nodules (smaller than 3 mm in radius), large nodules (larger than 3 mm in radius) and non-nodules (nodules of benign nature). The small nodules were located only by one coordinate representing their centre. Large nodules and non-nodules are located by a set of coordinates representing their edges. Non-nodules are non-malignant and thus treated as regular healthy tissue. The small nodules lack pixelwise labelling by radiologists, making it impossible to extract a valid ground truth on a per-pixel basis. The workaround for this problem is to also label them as healthy tissue. Thus only large nodules are preserved as malignant.

Segmentation of the lungs from the rest of the body is one of the most widely used techniques to remove irrelevant information [2,14]. However, when segmentation fails it also removes the contents of the lungs rendering the sample useless. During manual inspection of the lung segmentation process on 50 scans, an empirical success rate of roughly 80 to 85 percent was observed. Automatic failure detection of the segmentation process proved to be challenging to implement and lung segmentation was therefore removed from the preprocessing pipeline.

Each radiologist has reviewed each nodule on a scale of one to five for the following characteristics: subtlety, internal structure, calcification, sphericity, margin, lobulation, spiculation, texture and level of malignancy. Since these characteristics are not relevant for the aim of this paper, they are ignored.

Inconsistencies between various annotation files, deviating file structures, missing values and erratic naming and populating of variable fields required a robust parser in order to extract as many properly documented nodules as possible. From the 983 scans containing at least one large nodule, 793 were successfully parsed.

3.3 Sampling

Samples of size 8x512x512 were taken from the image data in which at least 4 slices of each sample contain a malignant nodule larger than 3 mm in radius. For each sample, a ground truth for the nodules was created by taking the union of the annotated nodule locations from up to four radiologists. A total of 1132 samples were extracted from the data of which 849 were used for training and 135 and 148 were held out for testing and validation respectively.

3.4 Preprocessing samples

The resolution of the samples were reduced to 8x60x60 using bilinear interpolation to reduce the size of the input to the convolutional neural network (CNN). The pixel values created by various CT-scanners with different configurations were translated to Hounsfield Units to obtain uniform scaling between samples. Sample-wide pixel normalization was applied afterwards to increase performance of the CNN.

3.5 Data Augmentation

Since only 849 samples were available for training the model, various augmentation techniques were applied to artificially increase the number of unique samples used in training by a factor of 16. Every epoch, each of the following operations has a 50% chance of being applied to an original sample:

- being flipped on the horizontal axis
- being flipped on the vertical axis
- rotated 90, 180 or 270 degrees, with 1/3 probability for each option.

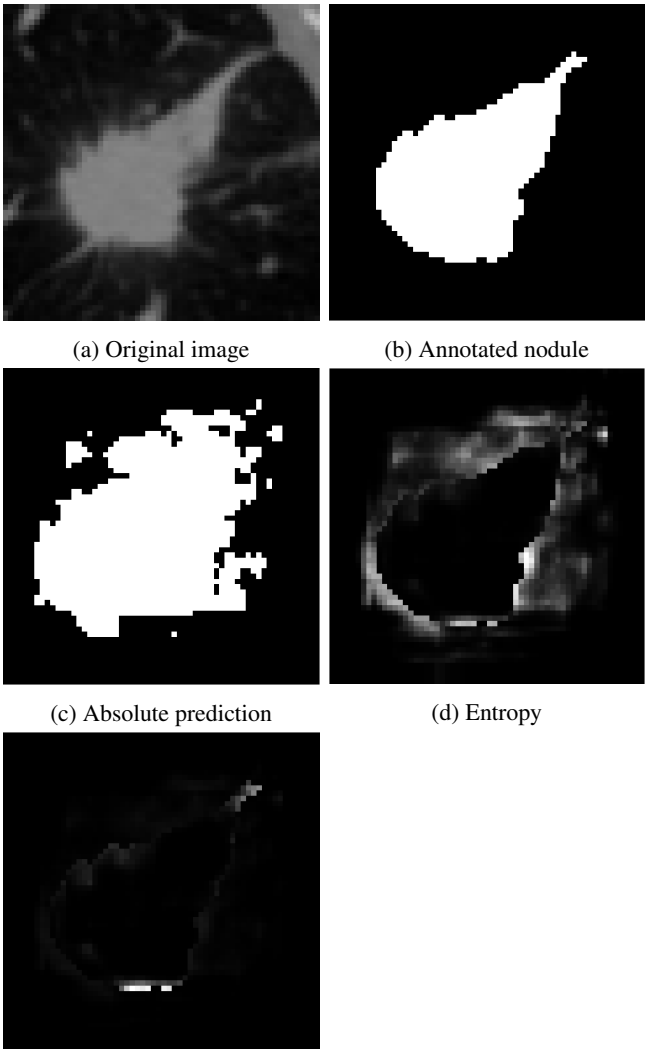
Data augmentation was done on-line instead of off-line because of the small batch size and the fixed order of input of the model. Off-line data augmentation would cause multiple subsequent batches to be (partly) filled with 16 similar samples all originating from one source image. On-line augmentation ensures that each batch is filled with samples originating from different source images.

3.6 Disputed pixels

Each medical scan is annotated by up to four radiologists independently. This means that there are up to four different masks representing the ground truth, causing the problem that pixel-for-pixel consensus is almost never obtained. This was dealt with by defining the union of all positive-labeled pixels as the ground truth for a nodule. Not only is the model trained and evaluated on the resulting mask, but its performance is also be evaluated considering only the pixels that radiologists unanimously agreed on.

Discarding the disputed pixels in the evaluation stage provides a reliable ground truth allowing for a more deterministic analysis of the model's performance but has a drawback: The model has a classification bias towards the positive label as all the disputed pixels were considered to be nodules during training.

The disputed pixels tend to be amongst the most difficult pixels to correctly classify. Besides the strong evidence given by the disagreement between expert radiologists, these pixels are often located around the edge of the nodule. Discarding them decreases the average difficulty of the segmentation task.



(a) Original image (b) Annotated nodule

(c) Absolute prediction (d) Entropy

(e) Loss (weighted entropy)

Fig. 3: Prediction on sample visualized

	% positive pixels	% disputed	% of nodule disputed
Training set	4.39	2.66	60.59
Test set	5.92	3.06	51.69
Validation set	7.06	4.16	58.92

Table 1: Distribution of samples

4 Results

Table 2 shows the performance of the model on the data. Note that the model obtains large differences in loss function for the three data sets but differences in quality of absolute performance are minimal. Figure 4 shows the increasing accuracy as the model was trained. Evaluating the entire training set whilst training significantly slowed down the learning process, therefore the decision was made to evaluate only five of the 849 samples per epoch to gain a rough estimate of the progress. As the high variation of the training set in Figure 4 shows, this was only good enough for an highly uncertain indication. Figure 5 shows that the difference in performance between including or excluding disputed pixels remained stable around 1.16% throughout the training process.

	Loss	Sensitivity	Specificity	Accuracy
Training set				
Including disputed pixels	0.674	0.657	0.881	0.876
Excluding disputed pixels	n/a	0.687	0.881	0.880
Test set				
Including disputed pixels	1.938	0.462	0.904	0.872
Excluding disputed pixels	n/a	0.486	0.904	0.888
Validation set				
Including disputed pixels	1.144	0.680	0.877	0.864
Excluding disputed pixels	n/a	0.721	0.877	0.873

Table 2: Results on train, test and validation set

4.1 Analysis

One critical issue with the dataset is that scans were added in batches by various medical institutions. Data for the training and validation sets was downloaded together at once and samples were assigned to one of the sets at random. The remaining data was downloaded at a later stage to serve as test set. Table 2 shows that the distribution of the data set was most likely not i.i.d. as the test set scores significantly lower in sensitivity and higher in specificity.

The weight ratio of 15:1 in favour of positive pixels in the loss function might have been insufficient in hindsight. Table 2 shows that the model has a relatively low sensitivity and high specificity. The factor 15 was chosen based on manual review of a few samples which resulted in an estimation of 6.67% positive pixels, an overestimation of the number of positive pixels in the entire dataset. The true ratios are shown in Table 1. However, Figure 6 shows after an initial drop in correctly predicted positive pixels, a steadily increasing sensitivitiy. More extensive training might further increase the model's ability to correctly identify nodules.

Due to the extensive duration of the learning proces, only a single set of hyperparameters was applied to the model. Optimizing parameters such as initial learning rate, momentum or number of features used by each layer could further increase the quality of the model.

5 Conclusion

We have seen that a fairly simple model is able to perform decently on the pixelwise nodule segmentation task and is able to reach an accuracy of 86.4%. Although the quality of the segmentation process does not equal that of expert radiologists, it is nonetheless a clear proof of concept and suggests that more complex models with higher computational capacity will be able to outperform human radiologists. It is important to note that this is achieved without the use of feature engineering or extensive preprocessing such as segmenting the lungs from the rest of the body.

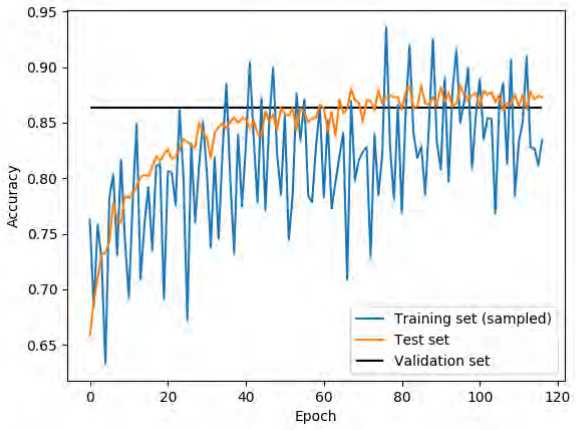


Fig. 4: Accuracy during training.

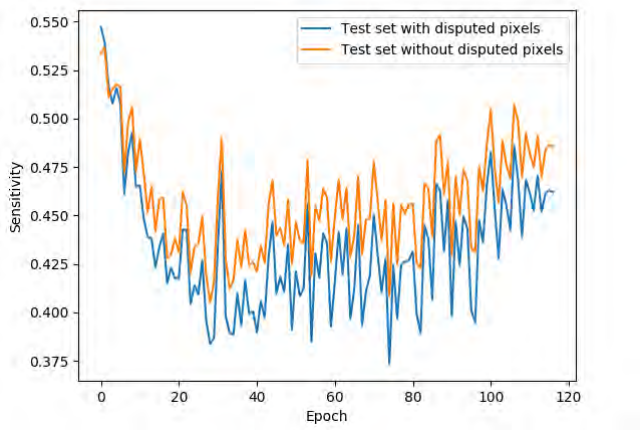


Fig. 6: Sensitivity during training.

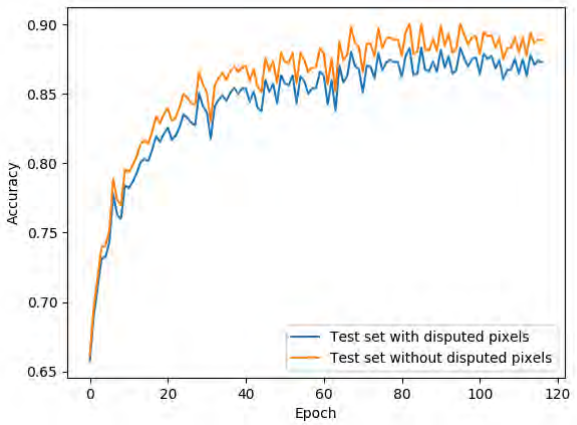


Fig. 5: Stable performance gap due to disputed pixels.

References

[1] Cancer research uk. Retrieved from http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer, accessed June 2017.

[2] Full preprocessing tutorial. Retrieved from https://www.kaggle.com/gzuidhof/full-preprocessing-tutorial, accessed June 2017.

[3] Martín Abadi and co. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[4] S. G. Armato III, G. McLennan, and L. Bidaut. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical Physics*, 38(2):915–931, 2011.

[5] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. *Proceedings of Machine Learning Research*, 9(3):249–256, 2010.

[6] M. K. Gould, J. Donington, and W. R. Lynch. Evaluation of individuals with pulmonary nodules: When is it lung cancer? *Chest*, 143(5 Suppl):e93S–e120S, 2013.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. arXiv:1512.03385, 2015.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on image net classification. arXiv:1512.01852, 2015.

[9] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167, 2015.

[10] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv:1412.6980.

[11] J.-S. Lin, S-.C. Lo, A. Hasegawa, M.T. Freedman, and S.K. Mun. Reduction of false positives in lung nodule detection using a two-level neural classification. *IEEE Transactions on Medical Imaging*, 15(2):206–217, 1996.

[12] S.-C. Lo, M.T. Freedman, J.-S. Lin, and S.K. Mun. Automatic lung nodule detection using profile matching and back-propagation neural network techniques. *Journal of Digital Imaging*, 6(2):48–54, 1993.

[13] S.-C. Lo, S.-L. Lou, J.-S. Lin, M.T. Freedman, M.V. Chien, and S.K. Mun. Artifial convolution neural network techniques and applications for lung nodule detection. *IEEE Transactions on Medical Imaging*, 14(4):711–718, 1995.

[14] B. Sasidhar, D. R. Ramesh Babu, M. Ravi Shankar, and N. Bhaskar Rao. Automated segmentation of lung regions using morphological operators in ct scan. *International Journal of Scientific and Engineering Research*, 4(9):1014–1018, 2013.

[15] K. Suzuki, J. Shiraishi, H. Abe, H. MacMahon, and K. Doi. False-positive reduction in computer-aided diagnostic scheme for detecting nodules in chest radiographs by means of massive training artificial neural network. *Academic Radiology*, 12(2):191–201, 2005.

[16] H. Yang, H. Yu, and G. Wang. Deep learning for the classification of lung nodules. arXiv:1611.06651, 2016.