Lung Cancer Detection with Region Enhanced Multi-Instance Networks

Jay Destories, Jason Fan, Alex Tong

March 2017

1 Introduction and Problem Statement

The problem of image segmentation and structure annotation within the field of biomedical imaging has become a well developed and very active field in the past years. In 2016 and 2015, the LUng Nodule Annotation (LUNA) challenge and SPIE Lungx challenge, asked researchers to develop models to identify pulmonary nodules in lung CT slices. With the 2016 LUNA Challenge, researchers gained access to annotated CT slices that contained segmentation ground truths for abnormal nodules but did not release data about the malignancy of the nodules. In the 2015 SPIE less than 80 CT annotated images of malignant nodules were released to the public.

Finally, with the 2017 Kaggle Data Science Bowl, a larger dataset of 1000+ lung CT images in DICOM format was finally released with cancer/no cancer labels. This has allowed researchers to answer a deeper question about lung CT scans; whether or not there are indicators of malignancy and cancer in a patient's CT scan.

There is, however, one caveat to the Kaggle dataset. Although the presence of malignancy is indicated by global, binary cancer/no-cancer label, the location of malignant nodules and structures are *not* annotated in the training data.

Inspired by recent work in biomedical image segmentation [1], region proposal networks and multiple instance learning for whole mammogram classification [2], We seek to present a novel pipeline for lung cancer detection that enhances multiple instance learning with region proposal.

2 Related Work

Over the past decade there has been a significant amount of work towards computer aided diagnosis of lung cancer [3]. Depending on what kind of data researchers have had access to, previous efforts to identify lung cancer can be categorized by two main approaches.

(1) Pulmonary Nodule Detection methods use image processing techniques to segment and annotate nodules [4, 5, 6]. These methods mimic radiologists by looking for abnormalities in the form of “solitary white nodule-like blob[s]" in a chest x-rays and CT scans. Lung nodules are potential cancer indicators, and as such are an important part in early lung cancer diagnosis. However, not all pulmonary nodules are malignant and many are benign, there will be false positives in diagnosis if we naively associate nodule presence with cancer.

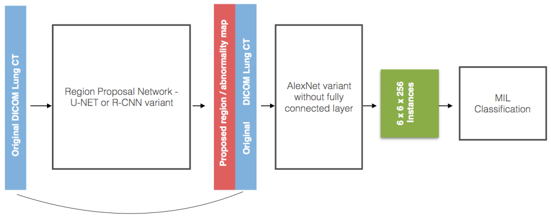
(2) Direct inference and classification methods instead attempts to directly predict the probability of cancer using x-ray and CT images without nodule detection [7, 8].

One challenge that almost all researchers in biomedical image inference face is the problem with datasets being small and weakly labeled. Images labeled with cancer/no-cancer binaries are weakly labeled because imaged tissues only display malignancy locally; not *all* of the tissue an image will have cancer. As Multi-Instance Networks have been used to classify whole mammogram images [2].

3 Method/Algorithm/Pipeline

3.1 Outline of proposed pipeline

Our Lung Cancer detection method uses nodule annotation to enhance direct classification.


Original DICOM Lung CT
Region Proposal Network - U-NET or R-CNN variant
Proposed region / abnormality map
Original DICOM Lung CT
AlexNet variant without fully connected layer
6 x 6 x 256 Instances
MIL Classification

We plan to leverage state of the art region proposal and biomedical segmentation to indicate nodule presence and implicitly weight instances consumed by a downstream MIL classifier.

In our pipeline, a region proposal network trained on the LUNA dataset to propose nodule annotations and abnormal regions in CT slices. This region proposal is then used to annotate nodules in Kaggle dataset. Then a MIL classifier based on the work on mammogram classification by Lou et. al [9], will be trained on the annotated Kaggle dataset.

3.2 Choosing the Region Proposal Network

We have three region proposal/image segmentation techniques we want to investigate. We will tune one segmentation/region proposal network for our final pipeline.

The first segmentation method we will test is U-Net, a CNN for biomedical segmentation developed by Brox et. al [6]. The second segmentation we will test is the fully convolutional neural network developed by Darrell et al add ref (Jay and Jason will be investigating this network as a paper presentation). And the third region proprosal method we will test is Faster R-CNN developed by He et. al [10, 11].

3.3 The Multiple Instance Learner/Classifier

We will adapt MIL techniques developed for mammogram classification developed by Lou et. al add ref. Their implementation feeds the 6 by 6 by 256 output of the last convolutional layer of AlexNet as instances that are then consumed by three different MIL methods/losses which is it? a loss or a method? where is the backprop.

We will use the same architecture and first use 2D instances to look for lung cancer. If time permits, we will adapt AlexNet to instead output 3D, voxel based convolutions as the output instances for MIL methods. We particularly like this AlexNet variant because, amongst the team, we own and have access to two NVidia 970 GPUs.

3.4 Data Augmentation

One major challenge we will face will be the fact that we have a very small collection of data. TODO:...

4 Datasets

The number of lung CT scans available to us is very low. The total number of examples available are many orders of magnitudes smaller than the size of datasets for modern, state of the art classification challenges such as ImageNet or MSCOCO.

Listed below are the datsets we will leverage to train our classification/nodule extraction model.

Dataset	# CT scans	Label Type
Kaggle	1398	Cancer/No-cancer
LIDC-IDRI	1018	Nodule annotation
SPIE	80	Nodule annotation
NLST	?	?

In the dataset above, each CT scan is a stack of approximately 200 2D slices.

The Kaggle dataset will be obtained from the Kaggle Data Science Bowl 2017 competition. The dataset consists of over 1000 CT scans in DICOM format labeled by a cancer/no-cancer binary. We will use this data train our downstream MIL classifier.

The Lung Image Database Consortium (LIDC-IDRI) dataset is the dataset used in the 2016 LUNA challenge.The dataset consists of over 1000 CT scans where each nodule location and radius is annotated for each 2D slice. The dataset was annotated by 4 radiologists and nodules are annotated with varying levels of agreement. We will use this dataset to train our upstream region/nodule proposal network.

The SPIE dataset is a minuscule but have a few important examples where nodules are both annotated and labeled benign or malignant. This dataset may help us validate and visualize the convolutions happening in both the upstream region/nodule proposal and downstream MIL classifier.

We suspect that the National Lung Screening Trial (NLST) dataset will contain about another 1000 CT scans that will either be labeled with the cancer/no-cancer or nodule annotations. We have applied for and are currently waiting for access to the NLST data.

5 Evaluating our results

5.1 Our goals

We have two primary goals in this investigation. First we want to know if region proposal and implicit instance weighting aid multiple instance learning. Second, we want to investigate the how leveraging depth information from our data helps with inference with biomedical imaging.

5.2 How will we know what is going on?

- visualize 2D convolutions in both upstream and downstream nets
- visualize 3D convs if we get to it in the downstream MIL net
- Use SPIE data to see how malignant vs benign tumors are 'seen' by our network

5.3 Quantifying our results

- COMPARE FASTER R-CNN with U-NET and other segmentation methods
- Figure out how the upstream region output effects downstream classifier
- Compare direct MIL classification with region enhanced MIL classification

6 About us

buncha kiddos who have no idea about anything

7 Questions and challenges

1. What will the output of node-extraction be like?

2. What kind of novelty will we bring to the table?

3. How will we deal with the small dataset(s)?

References

[1] Patrick Ferdinand Christ, Mohamed Ezzeldin A. Elshaer, Florian Ettlinger, Sunil Tatavarty, Marc Bickel, Patrick Bilic, Markus Rempfler, Marco Armbruster, Felix Hofmann, Melvin D'Anastasi, Wieland H. Sommer, Seyed-Ahmad Ahmadi, and Bjoern H. Menze. Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3d conditional random fields. *CoRR*, abs/1610.02177, 2016.

[2] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. In *Proceedings of the 1997 Conference on Advances in Neural Information Processing Systems 10*, NIPS '97, pages 570–576, Cambridge, MA, USA, 1998. MIT Press.

[3] M. G. Penedo, M. J. Carreira, A. Mosquera, and D. Cabello. Computer-aided diagnosis: a neural-network-based approach to lung nodule detection. *IEEE Transactions on Medical Imaging*, 17(6):872–880, Dec 1998.

[4] X. Li, L. Shen, and S. Luo. A solitary feature-based lung nodule detection approach for chest x-ray radiographs. *IEEE Journal of Biomedical and Health Informatics*, PP(99):1–1, 2017.

[5] Rushil Anirudh, Jayaraman J. Thiagarajan, Timo Bremer, and Hyojin Kim. Lung nodule detection using 3d convolutional neural networks trained on weakly labeled data, 2016.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[7] Jinsa Kuruvilla and K. Gunavathi. Lung cancer classification using neural networks for {CT} images. *Computer Methods and Programs in Biomedicine*, 113(1):202 – 209, 2014.

[8] Ge Wang He Yang, Hengyong Yu. Deep learning for the classification of lung nodules.

[9] Wentao Zhu, Qi Lou, Yeeleng Scott Vang, and Xiaohui Xie. Deep multi-instance networks with sparse label assignment for whole mammogram classification. *CoRR*, abs/1612.05968, 2016.

[10] Ross B. Girshick. Fast R-CNN. *CoRR*, abs/1504.08083, 2015.

[11] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015.