Paper review
Towards radiologist-level cancer risk assessment in CT lung screening using deep learning. Trajanovski S, Mavroeidis D, Swisher CL, Gebre BG, Veeling BS, Wiemker R et al. Computerized Medical Imaging and Graphics. 2021 Mar; 90: 1010883.

Group 14

October 27, 2021

1 Summary of application domain

With the improvements of healthcare, people don't die from infections or viruses anymore and nowadays cancer is one of the deadliest diseases. Furthermore, in the United States of America lung cancer is the leading cause of cancer with an average five year survival of 18.1 percent mainly because symptoms are only visible at advanced stages of the disease. However, when detected at an early stage, the average five year survival is about 60 to 75 percent. Thus preventive annual check-ups to diagnose the disease early would reduce mortality and prevent more heavy treatments. On the other hand, it will generate much more data to process and analyse for radiologists. Therefore, developing deep learning models capable of assessing lung cancer risk could support radiologists' work and offer better support for the patients.

2 Summary of the used Machine Learning methodology and evaluation metrics

The machine learning framework used in this work consisted of two steps. First two nodule detectors localized all the nodules in the scans. These two were based on hierarchical support vector machines and on semantic segmentation by a deep neural network. Around each nodule a cube was segmented and of each cube three different two-dimensional projections were extracted. This data augmentation was performed to reduce overfitting of the model to the nodule images. Then a neural network based on ResNet estimated the cancer risk based on the ten largest nodules. ResNet (residual neural network) is a neural network that uses skip connections in order to avoid the vanishing gradient problem. This neural network was deep and wide, since it had a high number of both layers and inputs. Next to nodule localization, additional features such as nodule size and confidence score of the nodule detector were added as inputs to the deep neural network. In this research four different datasets were combined which gave a total of 8598 scans. Of these 4807 scans were used as training set, and the remaining 3791 scans were used as test set. Data overlaps between the training and the test sets were removed from the test sets. For all the cancer cases in the datasets the diagnosis was confirmed with a test such as a biopsy, but such a ground truth was not available for the non-cancer cases. The diagnosis of a patient was used to label all the nodule images of this patient, so the data was weakly labelled. The framework was evaluated in three different ways. Its robustness was evaluated by testing the model performance on different datasets. This model performance was reported by the AUC (area under the curve). The performance of the model was tested against the performance of radiologists and against the performance of currently existing models. These results were visualized in a ROC (receiver operating characteristic) curve.

3 Strong and weak points of the methodology and evaluation metrics

3.1 Weak points

- They made certain all patients they labelled as "cancer" actually had cancer, but they say it is possible some patients labelled as "not cancer" may have developed cancer and may therefore be labelled wrongly, inducing a bias in the model leaning towards more false negatives, possibly reducing the sensitivity of the model.

- Only the 10 largest nodules are considered, an 11th cancerous nodule may be overlooked, further reducing sensitivity.

- The model does not tell which nodule was malignant, making intervention based on this model difficult as manual determination of all nodules is still required.

3.2 Strong points

- They made very certain not to have data overlap between training and testing datasets.

3.3 Summary

They made a model that seems to have a very good performance, but the way the model is to be applied is not clear from the paper. The model may have a slight reduced sensitivity bias because of the afore mentioned weak points, so it may not be beneficial to treat only patients labelled as "cancer" as you may miss some. It is neither useful to treat all "cancer" labelled patients and still manually test the rest, because the model does not tell which module was malignant, so that way you still have to manually test all patients.

4 Alternative methodology, evaluation metrics and ideas for improvement

The proposed CAD lung cancer detection system consists of SVM based node detector and ResNet architecture based node classifier. The results of both detection and classification could be further improved by implementing novel object detection algorithms, such as YOLO v3, and new CNN architectures, such as Efficient Net. Some studies have already shown that YOLO outperforms SVM in detection [1], and in general, EfficientNet models achieve higher accuracy and efficiency than existing CNNs [2].

To evaluate the proposed deep learning model, a conventional ROC analysis was performed, which, as mentioned before, ignores information about the location of the nodules. A better alternative is to use the free response ROC (FROC) curves. These curves are similar to the ROC curves except that their horizontal axis indicates the number of false positives per image [3]. The FROC paradigm can account for nodule detection and location on images containing any number of nodules, which increases statistical power [4]. If an indicated location is within a clinically relevant acceptance region of the actual nodule centroid, the event is scored as true-positive. Otherwise, it is scored as false positive [5]. This is more clinically relevant, as it is important not only to detect the presence of nodules, but also to provide further clues to their location.

Finally, there is still room for improvement. For instance, the second stage of the framework uses a 2D convolutional neural network. However, the scans from CT are 3D images, which means that the performance of the model could be surpassed with a 3D approach. Using 3D kernels is more computationally expensive, but would allow the network to learn volumetric features, which in turn could help with nodule detection and classification.

References

[1] Özgür Kaplan and Ediz Saykol. Comparison of support vector machines and deep learning for vehicle detection. In *RTA-CSIT*, 11 2018.

[2] Mingxing Tan. Efficientnet: Improving accuracy and efficiency through automl and model scaling. `https://ai.googleblog.com/2019/05/efficientnet-improving-accuracy-and.html`, May 2019.

[3] Arnau Oliver. Automatic mass segmentation in mammographic images. `http://eia.udg.edu/~aoliver/publications/tesi/node147.html`, 2008.

[4] Xin He and Eric Frey. ROC, LROC, FROC, AFROC: An Alphabet Soup. *Journal of the American College of Radiology*, 6(9):652–655, September 2009.

[5] Dev Chakraborty. The FROC, AFROC and DROC Variants of the ROC Analysis. In *Handbook of Medical Imaging*, volume 1. Physics and Psychophysics, page 774. SPIE Publications, 2009.