Pneumothorax Detection and Localization in Chest Radiographs: A Comparison of Deep Learning Approaches

André Gooßen[1]**, Hrishikesh Deshpande**[1]**, Tim Harder**[1]**, Evan Schwab**[2]**, Ivo Baltruschat**[3]**, Thusitha Mabotuwana**[4]**, Nathan Cross**[5]**, Axel Saalbach**[1]

[1] *Digital Imaging, Philips Research, Hamburg, Germany*

[2] *Clinical Informatics Solutions and Services, Philips Research, Cambridge, USA*

[3] *Institute for Biomedical Imaging, Hamburg University of Technology, Germany*

[4] *Radiology Solutions, Philips Healthcare, Bothell, USA*

[5] *Department of Radiology, University of Washington Medical Center, Seattle, USA*

Abstract

Pneumothorax is a critical condition that requires timely communication and immediate action in order to prevent significant morbidity or patient death. Thus, early detection is crucial. For the task of pneumothorax detection, we measured the performance of three different deep learning techniques: convolutional neural networks, fully convolutional networks, and multiple-instance learning. In a five-fold cross-validation with ROC analysis on a dataset consisting of 1003 chest X-Ray images we measured AUCs of 0.96, 0.92, and 0.93 for the three methods, respectively. We discuss the classification as well as localization performance of these approaches as well as an ensemble of aforementioned techniques.

1. Introduction

In many clinical institutions, the ability to prioritize specific imaging exams is made possible by use of stat or emergent labeling. However, because of overuse and misuse of these labels, a radiologist often has difficulties prioritizing exams with more medically significant findings. As a result, an automated system to triage relevant critical findings should improve the management of patients. In order to prevent significant morbidity or patient death, the timely communication of *Category 1* findings, e.g. pneumothorax, is even endorsed by the American College of Radiology (ACR) (Yarmus and Feller-Kopman, 2012; Larson et al., 2014). In this paper, we investigate and evaluate three deep learning architectures for the detection and localization of pneumothorax in chest X-ray images.

2. Methods

CNN Convolutional Neural Networks (CNNs) are the most commonly employed type of neural network for image classification. They have been successfully used in a broad range of applications from computer vision to medical image processing (He et al., 2016; Wang et al., 2017) and can be optimized in an end-to-end fashion.

In the following, a variant of the ResNet-50 (He et al., 2016) is employed, which was developed explicitly for the analysis of X-ray images (Baltruschat et al., 2018). In order to leverage the higher spatial resolution of X-ray data, e.g. for the detection of small structures, an additional pooling layer was introduced after the first bottleneck block, increasing the input size to 448×448 pixels using one input channel. The network was pre-trained using the NIH ChestX-ray14 dataset (Wang et al., 2017) with 112 120 images from 30 806 patients and annotations for 14 pathologies. For the task-specific training, the existing dense layer for the prediction of the pathologies was replaced by a new dense layer and a softmax layer providing binary pneumothorax classification.

FCN CNNs are considered state of the art for image classification, however, they do not provide any localized information, e.g. position of the pathology in the image. Fully Convolutional Neural Networks (FCNs) allow for a semantic segmentation, i.e. pixel-level classification, and thus, a localization of a pathology. Therefore, as a second approach, a U-Net (Ronneberger et al., 2015) with attention gates (Oktay et al., 2018) and a receptive field of 448×448 pixels was considered. A U-Net consists of a contracting path resembling a CNN, for the integration of context information, and a corresponding expanding path.

In contrast to CNN, the FCN approach requires pixel-level annotations during the training and predicts probability values for each pixel during the application phase. In the scope of this study, we define the area of the detected pneumothorax as a classification measure. Although such measure is biased towards the detection of large pneumothorax, it is conceptually simple and favors the detection of reliable candidates.

MIL Multiple-Instance Learning (MIL) could be seen as an intermediate approach between CNN and FCN, in terms of localization and ground truth training labels. MIL (Dietterich et al., 1997; Maron and Ratan, 1998), provides a joint classification and localization, while only requiring the image-level labels for training. To produce local predictions in the image, the full resolution chest X-ray images are partitioned into overlapping image patches of 448×448 pixels. Using the image-level labels, all of the patches in a non-pneumothorax image will necessarily be negative, whereas at least one of the patches in a pneumothorax image must contain the pathology and therefore be a positive patch. Using only label information at image-level during the training, MIL attempts to learn the fundamental characteristics of the local pathology by automatically differentiating between normal and abnormal parts of the chest X-ray. During inference, the maximum patch score of the patch-level predictions is used to derive the image-level classification.

3. Experiments and Results

The data used in the following experiments consists of DICOM X-ray images, obtained from the University of Washington Medical Center and affiliated institutions. Scanning radiology reports from the last three years, 437 images with pneumothorax and 566 with a different or no abnormality detected were identified. Image-level labels were derived from reports using NLP and pixel-level pneumothorax annotations for 305 of the positive cases were generated manually. For training and evaluation, we divided the dataset into five (patient stratified) cross-validation splits and applied data augmentation to increase the variability of the data.

For CNN and FCN the images were rescaled to 480×480, for MIL we used a resolution of 1120×1120. For all methods we cropped patches of 448×448. In training, we used the Adam optimizer with default parameters, a batch size of 16, and exponentially decreasing learning rate (LR). We performed ROC analysis to assess the model performance (cf. Figure 1).

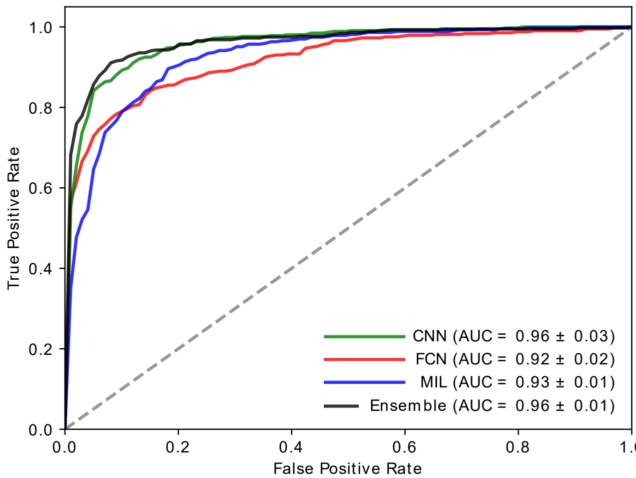

Figure 1: Averaged ROC curves over five splits for all methods as well as an ensemble.

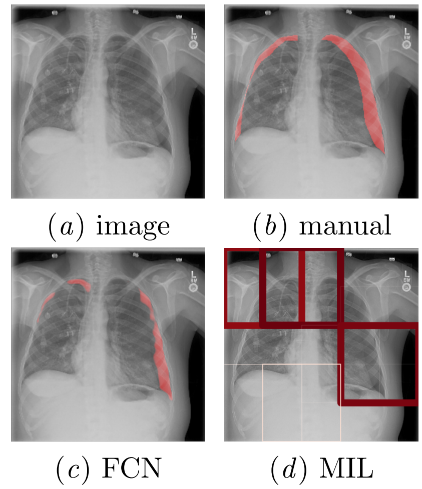

Figure 2: Localization examples for a pneumothorax case.

CNN The pre-trained ResNet-50 was fine-tuned with an initial LR of 10^{-4} for 40 epochs. For testing, a center crop was used for the classification purpose. High and stable AUC values of 0.96±0.03 can be reported.

FCN As pixel-level ground truth annotations were available only for a subset of the images, 871 images in total were used for training the FCN for 400 epochs. As a loss function, a weighted cross-entropy was employed at pixel-level with an initial LR of 10^{-4}. With an average AUC of 0.92±0.02, the overall performance of this method is slightly worse than the CNN. On the other hand, the FCN generates pixel-level probabilities, which indicate the location of the pneumothorax (cf. Figure 2 (c)).

MIL The pre-trained ResNet-50 was also employed as the patch-level classifier within the MIL approach. We chose the binary cross-entropy between the maximum patch score and the image-level label as the loss function. The batch size was selected as the number of 16 patches per image. We trained with an initial LR of 10^{-5} for 30 epochs and achieved an average AUC of 0.93±0.01 using this method. High patch scores (indicated by thicker red frames) give a hint on the location of the pneumothorax (cf. Figure 2 (d)).

Ensemble Learning As the errors made by different architectures do not necessarily coincide, we formed an ensemble of the three constituent methods by linear combination of the individual methods' scores. Here, an AUC of 0.96±0.01 is achieved, where especially working points for very low false positive rates are improved.

4. Conclusion

The three presented techniques provide promising options for the detection and localization of pneumothorax in chest X-ray images. We achieved the best performance in terms of AUC using a CNN, whereas the FCN and MIL provided higher confidence in terms of localization.

This could guide radiologists and simultaneously increase the trust in the proposed deep learning architecture. Future work could elaborate on other techniques to combine the three approaches, e.g. by cascading or merging the architectures into one multi-task network.

Acknowledgments

We would like to thank Christopher Hall for his support and advice. We further thank Tom Brosch and Rafael Wiemker for annotating data and providing valuable input on our network architectures. Finally, we thank Hannes Nickisch for reviewing the manuscript and providing valuable feedback.

References

Ivo M. Baltruschat, Hannes Nickisch, Michael Grass, Tobias Knopp, and Axel Saalbach. Comparison of deep learning approaches for multi-label chest X-ray classification. *arXiv preprint arXiv:1803.02315*, 2018.

Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Paul A. Larson, Lincoln L. Berland, Brent Griffith, Charles E. Kahn Jr., and Lawrence A. Liebscher. Actionable findings and the role of IT support: report of the ACR actionable reporting work group. *Journal of the American College of Radiology*, 11(6):552–558, 2014.

Oded Maron and Aparna Lakshmi Ratan. Multiple-instance learning for natural scene classification. In *ICML*, volume 98, pages 341–349, 1998.

Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y. Hammerla, Bernhard Kainz, et al. Attention U-Net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 3462–3471. IEEE, 2017.

Lonny Yarmus and David Feller-Kopman. Pneumothorax in the critically ill patient. *Chest*, 141(4):1098–1105, 2012.