Interpreting Covid-19 Prediction Models using Information Bottleneck

General Info

Project Title: Interpreting Covid-19 Prediction Models using Information Bottleneck

Contact Person: Ashkan Khakzar, Dr. Seong Tae Kim

Contact Email: ashkan.khakzar@tum.de, seongtae.kim@tum.de

Project Abstract

In order to establish trust in the clinical routine for the use of neural network models for screening Covid-19, it is vital to know how the models reach their predictions. In this project we investigate what parts of the input image are responsible for each network prediction using a state of the art feature-attribution method[1]. Validating that the model is focusing on regions where the pathologies are located reinforces the trust over these models. The feature-attribution method that we use here has the advantage that highlights input regions based on the amount of information contained in that area for the prediction of the model. In this project we use the NIH Chest X-ray[2] dataset for pretraining the network and later use the Covid dataset[3] for finetuning the model.

Background

There are couple of papers which present methods for automatic detection of COVID-19 disease[4][5] or quantifying the COVID-19 severity score[6] from X-ray images.

A huge body of works in neural network interpretability are devoted to understanding what parts of the input are responsible for the predicted output. These methods are known as input feature attribution methods. A well-known method within this category is Class Activation Maps (CAM)[7]. Recently a new feature attribution method[1] grounded on information theory is proposed that can result in more detailed maps, and the highlighted areas are based on the information the model uses for the prediction.

Technical Prerequisites

- Basic understanding of convolutional neural networks
- Very good skills in Python
- Good skills in PyTorch,

---

[1] Schulz, Karl, et al. "Restricting the flow: Information bottlenecks for attribution." arXiv preprint arXiv:2001.00396 (2020).
[2] Wang, Xiaosong, et al. "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
[3] https://brixia.github.io/#get-the-data
[4] Oh, Yujin, Sangjoon Park, and Jong Chul Ye. "Deep learning covid-19 features on cxr using limited training data sets." IEEE Transactions on Medical Imaging (2020).
[5] Ozturk, Tulin, et al. "Automated detection of COVID-19 cases using deep neural networks with X-ray images." Computers in Biology and Medicine (2020): 103792.
[6] Signoroni, Alberto, et al. "End-to-end learning for semiquantitative rating of COVID-19 severity on Chest X-rays." *arXiv preprint arXiv:2006.04603* (2020).
[7] Zhou, Bolei, et al. "Learning deep features for discriminative localization." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

Benefits:

- Working on a Covid19 related project!
- Possible publication (based on the results)

Work-packages and Time-plan:

	Description	#Students	From	To
WP1	Familiarize with the literature. (ChestX ray classification, and feature attribution) and the datasets	all	01.11	
WP2	Familiarize with PyTorch. Come up with a detailed time-plan (Gantt)	all		
WP3	Train a model on NIH ChestX-ray dataset	Group1		
WP4	Fine-tune model on Covid Dataset	Group1		
WP5	Setup Information Bottleneck method	Group2		
M1	Intermediate Presentation II	all	17.12.2020	
WP6	Run Information Bottleneck on NIH ChestXray and Covid datasets	all		
WP7	Share results with clinical partners	all		
WP+	(Optional) Run the framework on private CT dataset	all		
WP8	Documentation	all		
M2	Final Presentation	all	26.02.2021	