Explainability Matters: Backdoor Attacks on Medical Imaging

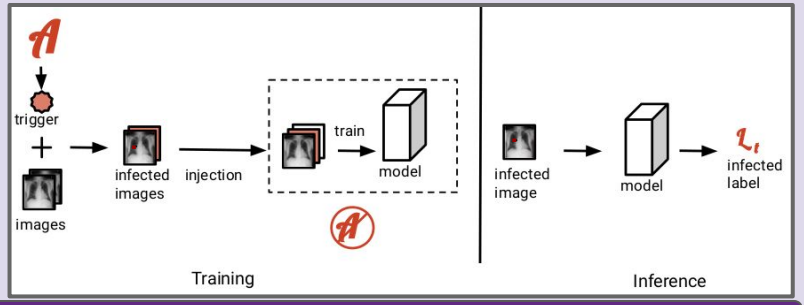
Munachiso Nwadike[1], Takumi Miyawaki[1], Esha Sarkar[2], Michail Maniatakos[1], Farah Shamout[1]
[1]NYU Abu Dhabi, UAE [2]NYU Tandon School of Engineering, USA

Abstract

Deep neural networks have been shown to be vulnerable to backdoor attacks, which could be easily introduced by means of the model training procedure. The exact impact of backdoors is not yet fully understood in complex real-world applications, such as in medical imaging where misdiagnosis can be very costly. Our paper explores the impact of backdoor attacks on a multilabel disease classification task using chest radiography, with the simple assumption that the attacker need manipulate only the training dataset to execute the attack. We show how explainability can be used to identify spatially localized backdoors in inference time.

Methodology

- Dataset: 112,120 Anonymised, HIPAA-compliant chest radiographs from NIH Chestx-ray8 dataset [1]. The true label of each image is a binary vector indicating the presence (or absence) of 14 different diseases.
- Model: DenseNet-121
- Methodology: The attacker has special access to the machine learning dataset. They insert images prior to training, and need not be involved in the training procedure. The user will unknowingly trusts the predictions of the infected model, if it achieves some minimum performance on an independent test set using some metric.



Experiments

A backdoor trigger may be applied to a clean image x by means of function p(x, r, m),

$$x' = p(x, r, m) = x \bullet (1 - m) + r \bullet m$$

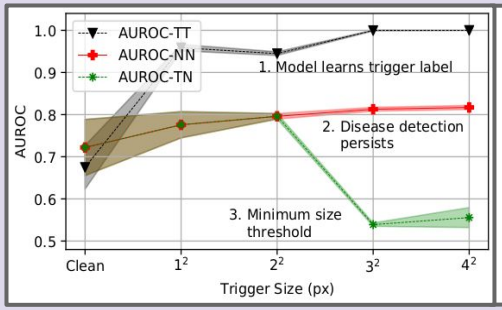
to obtain the infected image x', where r represents the trigger, m denotes the trigger mask that takes a value of 1 at the trigger location and 0 elsewhere, and • is the element-wise product.

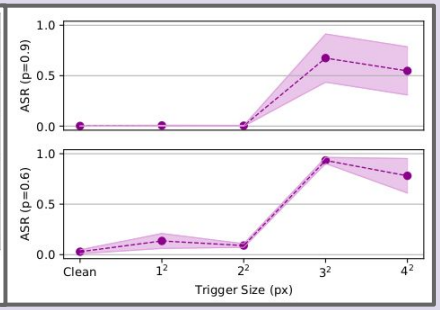
AUROC-NN.'NN' stands for 'Normal image, Normal label'. It measures prediction the true labels of clean images by backdoored model.

$$ASR = \frac{\sum_{x':\mathcal{M}(x')_t \geq p} 1}{\sum_{x':\mathcal{T}(x')_t = 0} 1}.$$

AUROC-TT. 'TT' stands for 'Triggered image, Triggered label'. It measures prediction of the infected labels of infected (triggered) images, but does not provide sufficient information on how well the model misclassifies the infected images, *away from* the true labels.

AUROC-TN. 'TN' stands for 'Triggered image, Normal label'. It measures misclassification, by backdoored network, of infected images *away* from the true (normal) labels. A lower score implies more efficient backdooring.





Explainability

We examined the role of explainability in the context of backdoor attacks using Gradient Class Activation Mappings (Grad-CAM)[2]. Grad-CAM calculates derivative of activations with respect to a convolutional layer of the neural network to compute saliency maps, where more important regions for the classification are indicated by red and less important regions are in blue. We apply Grad-CAM with respect to a middle layer of the DenseNet-121 (layer 207 of 287). We compare with the activations with respect to last conv layer.

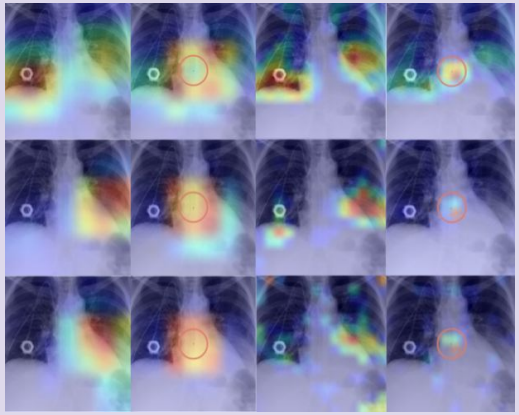


Figure shows the computed saliency maps for the clean and infected versions of an example image, at epochs 1, 4, and 12 (top to bottom). Columns 1 and 2 are taken with respect to the last convolutional layer, while 3 and 4 are taken with respect to the middle convolutional layer. We observe that in clean images of column 1, the heatmap focuses less on the location of the backdoor trigger in column 2. We see in columns 3 and 4 that utilising the middle layer for Grad-CAM gives more fine-grained backdoor trigger localization. The localization heatmap narrows on the center of the image, where the trigger is located in column 4. This is more noticeable at epochs 1 and 4, where there is less overfitting. The increased localization in the middle of the network is understandable since the backdoor trigger pixels can be considered as low level features, and thus may be better detected in the earlier layers of the network. This suggests that explainability can play a complementary role with robustness, since GradCAM shows differences between the saliency maps of clean and infected images, and can help radiologists in questioning model predictions when the saliency maps and predictions seem unreasonable.

Acknowledgements

Many thanks to professor Farah Shamout for her assistance and guidance in this project, as well as to my coauthors. In addition, this work could not have succeeded without the support of thee NYU Abu Dhabi High Performance Computing Team.

References

[1] Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; and Summers, R. M. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2097–2106.
[2] Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision, 618–626.