Comparison of Methods for Interpretability of DNNs in the context of CT images Classification

W. Marchadour[1], J. Maison[1,2], B. Badic[1,3], M. Hatt[1], F. Vermet[4]

1. LaTIM, INSERM, UMR 1101, UBO, Brest, France 2. Aquilab, Lille, France 3.University Hospital of Brest, France. 4. LMBA, CNRS, UMR 6205, UBO, Brest, France

Background

The use of Convolutional Neural Networks (CNN) in medical applications already showed ground-breaking performance, for tasks automation as well as more complex works (e.g. prognosis of patients). While a full integration of such frameworks in clinical routines is already achievable, doctors express a critical concern on their limited transparency (a.k.a. the "black box" effect). In order to improve global understanding and trust, the Interpretability field of work has emerged to identify which input elements are considered important by the networks. However, thorough experiments demonstrated that some of the methods conceived, while giving visually clear results, lost the main purpose of explaining the behavior of networks. The goal of this study is to compare several Interpretability algorithms on two different aspects: the quality of the attribution maps (when compared to ground-truth), and the ability to accurately reflect the functioning of the network.

Material & Methods

Toy Application: **Classification of CT images With or Without Contrast Agent**

1312 patients Chest CT scans (LIDC-IDRI / LNDb databases), application on 290 slices (from as many patients)
ResNet-50 architecture → 99.7% accuracy

5 Gradient-based **Interpretability methods**:

- **Backpropagation**[1] (BP): vanilla gradients w.r.t. the input pixels
- **Integrated Gradients**[2] (IGB): gradients on samples between input image and reference (usually Black)
- **Integrated Gradients Black & White** (IGBW): average of IG with Black, then White reference
- **Expected Gradients**[3] (EG): gradients over modified input image, using random images from train set
- **Deconvolution** (D): keeping activation function during vanilla Backpropagation (ReLU)

+ **SmoothGrad (SG) / SG Squared (SG^2)**[4] on all methods for visual improvement: reducing maps noise by adding noise to the input

Conversion to **Absolute values**, because meaning of sign differs from the methods
Use of **XRAI Segmentation**[5] approach: advanced segmentation of input image
→ **2 types of maps** are evaluated (pixel absolute values / region absolute values)

Manually-crafted **Reference Masks**, reflecting **Human Expert Expectation**

2 Maps Quality Metrics, based on **Comparison between Percent-Occluded Maps and Reference Masks**:

- **Receiver Operating Characteristic** (ROC): Area Under the Curve (AUC)
- **Dice Similarity Coefficient** (DSC): Maximum value (optimal occlusion setting)

1 Maps Representation Metric, based on **Perturbation of the Input of the Network**:

- **Fidelity**[6]: when highlighted input features are replaced, prediction of the network must drop

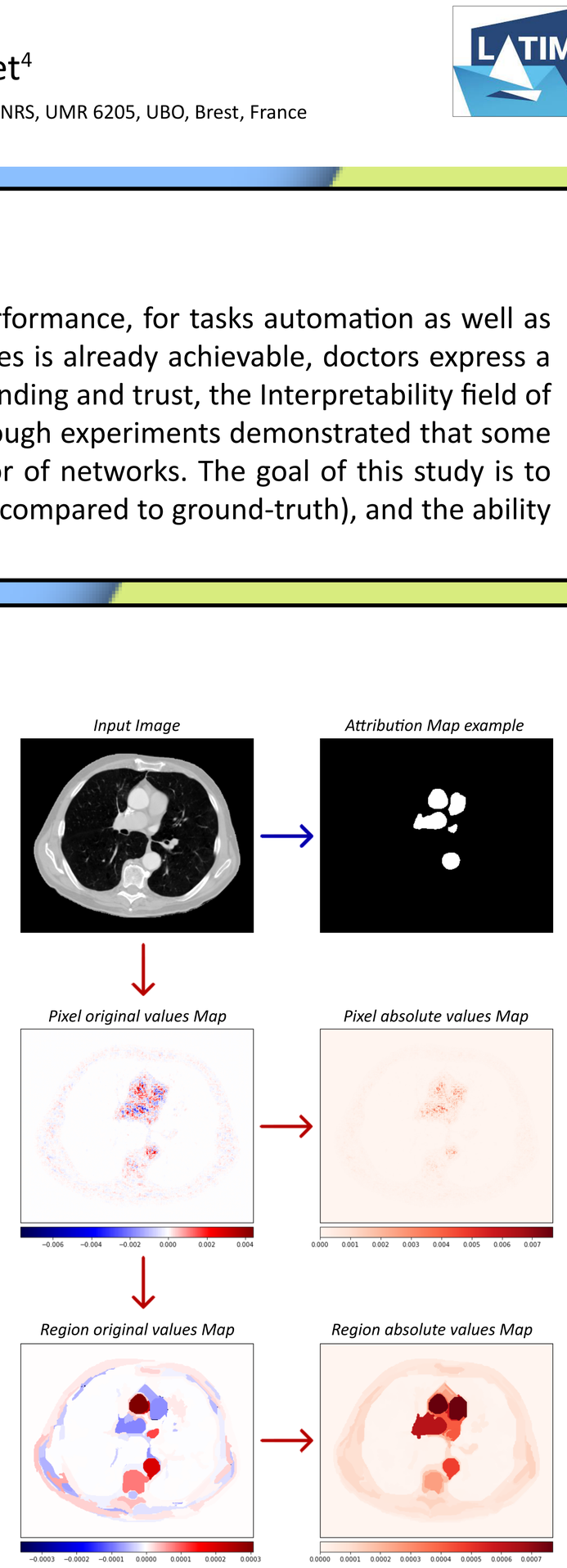

Figure 1: Examples of all types of maps described, along with the Reference mask of important features

Results

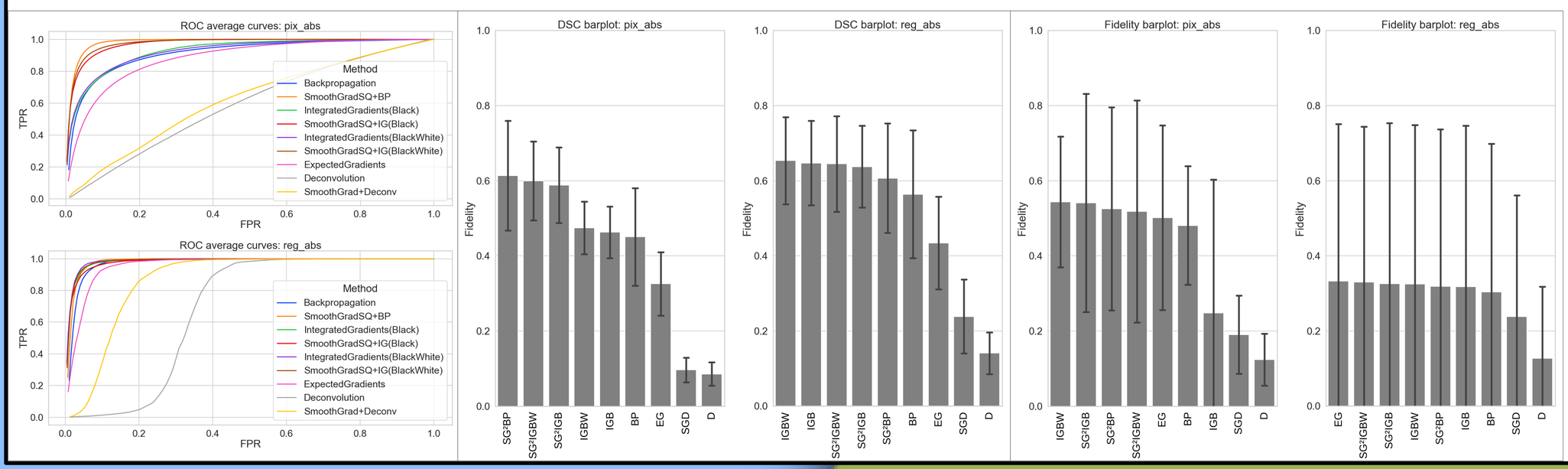

Figure 2:

Left: Area Under the Curve of ROC
Middle: DSC maximum scores
Right: Fidelity scores

In all cases, higher is better

Conclusions

- On the quality of the maps (left and middle results plots), the best scores are obtained when SG^2 is applied, especially on Region Absolute values maps
- On the Fidelity of the algorithms, Pixel Absolute values maps reflect best the network, and many methods seem accurate, except Deconvolution and IGB
- Such results correspond to a specific application, and may differ on other tasks / architectures

[1] Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." arXiv preprint arXiv:1312.6034 (2013).
[2] Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks." International conference on machine learning. PMLR, 2017.
[3] Erion, Gabriel, et al. "Improving performance of deep learning models with axiomatic attribution priors and expected gradients." Nature machine intelligence 3.7 (2021): 620-631.
[4] Smilkov, Daniel, et al. "Smoothgrad: removing noise by adding noise." arXiv preprint arXiv:1706.03825 (2017).
[5] Kapishnikov, Andrei, et al. "Xrai: Better attributions through regions." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.
[6] Brocki, Lennart, et al. "Evaluation of importance estimators in deep learning classifiers for Computed Tomography." International Workshop on ExTrAAMAS. Springer, Cham, 2022.

This work was partly funded by the French Ministry for Research and Higher Education, the French National Research Agency (ANR) through the CIFRE program, the Brittany Region, and University of Western Brittany through the INFORM European Project.