Visualization approach to assess the robustness of neural networks for medical image classification

Elina Thibeau-Sutre, Olivier Colliot, Didier Dormont, Ninon Burgos

Visualization approach to assess the robustness of neural networks for medical image classification

ARAMIS LAB
BRAIN DATA SCIENCE

Elina Thibeau--Sutre[1], Olivier Colliot[1], Didier Dormont[1,2], Ninon Burgos[1]

[1]ARAMIS Lab, ICM, Inserm U1127, CNRS UMR 7225, Sorbonne University, Inria, Paris, France
[2]AP-HP, Department of Neuroradiology, Pitié-Salpêtrière Hospital, Paris, France

elina.thibeausutre@icm-institute.org @AramisLabParis

Deep learning methods have shown a high performance potential for medical image analysis. However, explaining their decisions is not trivial and could be helpful to achieve better results and know how far they can be trusted.

Many methods have been developed in order to explain the decisions of classifiers, but their outputs are not always robust or meaningful [1] and they remain difficult to interpret.

In this study, we adapted to 3D medical images the method of [2] which relies on two visualization methods extensively used: occlusion and saliency maps.

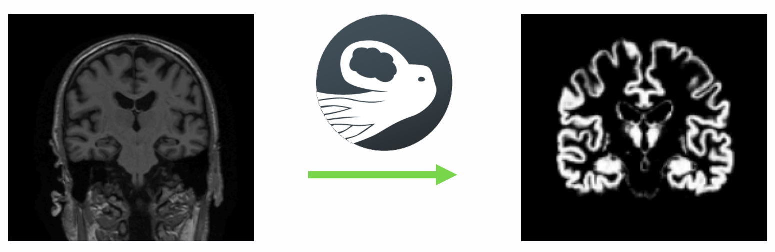
Methods

Materials

Databases:

aibl & ADNI

Modality:

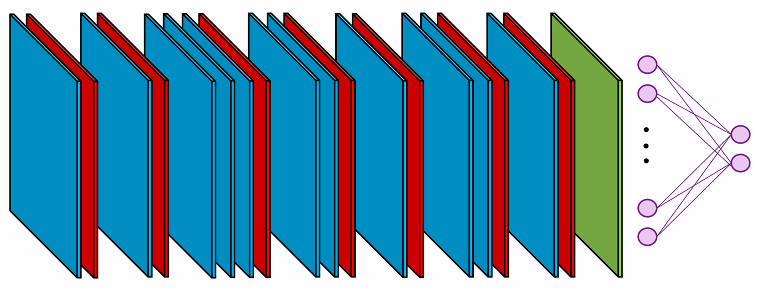


Grey Matter Maps derived from T1-MRI with clinica [3].

- Bias field correction
- Non-linear registration
- Tissue segmentation

CNN architecture and training

Architecture:



Found with Random Search [4] on training + validation.

Performance (balanced accuracy)

- **Training / Validation** (ADNI): 0.89
- **Test** (ADNI): 0.88
- **Test** (AIBL): 0.90

→ No overfitting detected

Mask optimization

Objective:

During training, the weights w were optimized to maximize the score function f_w on a set of images X as follows $w^* = \underset{w}{\operatorname{argmax}} f_w(X)$.

A mask of input size is applied to increase values voxel-wise. The image X'_m masked by m at voxel u is defined as:

$$X'_m(u) = m(u)X(u) + (1 - m(u))$$

The optimal mask covers a minimal amount of voxels in connected parts of the image and transform a set of patients in controls for the CNN.

$$m^* = \underset{m}{\operatorname{argmin}}\, f_{w^*}(X'_m) + \lambda_1 \|1 - m\|_{\beta_1}^{\beta_1} + \lambda_2 \|\nabla m\|_{\beta_2}^{\beta_2}$$

Metrics of evaluation:

- prob_{CNN}: output probability of the CNN for the true class for an input masked by two masks optimized in two different contexts.
  0=similar / 1=dissimilar
- ROI-based: cosine similarity between the vector of the densities of the masks in each ROI of AAL2.
  0=dissimilar / 1=similar

Results

Mask robustness

Figure 1. Grid search on λ_1 and λ_2 hyperparameters

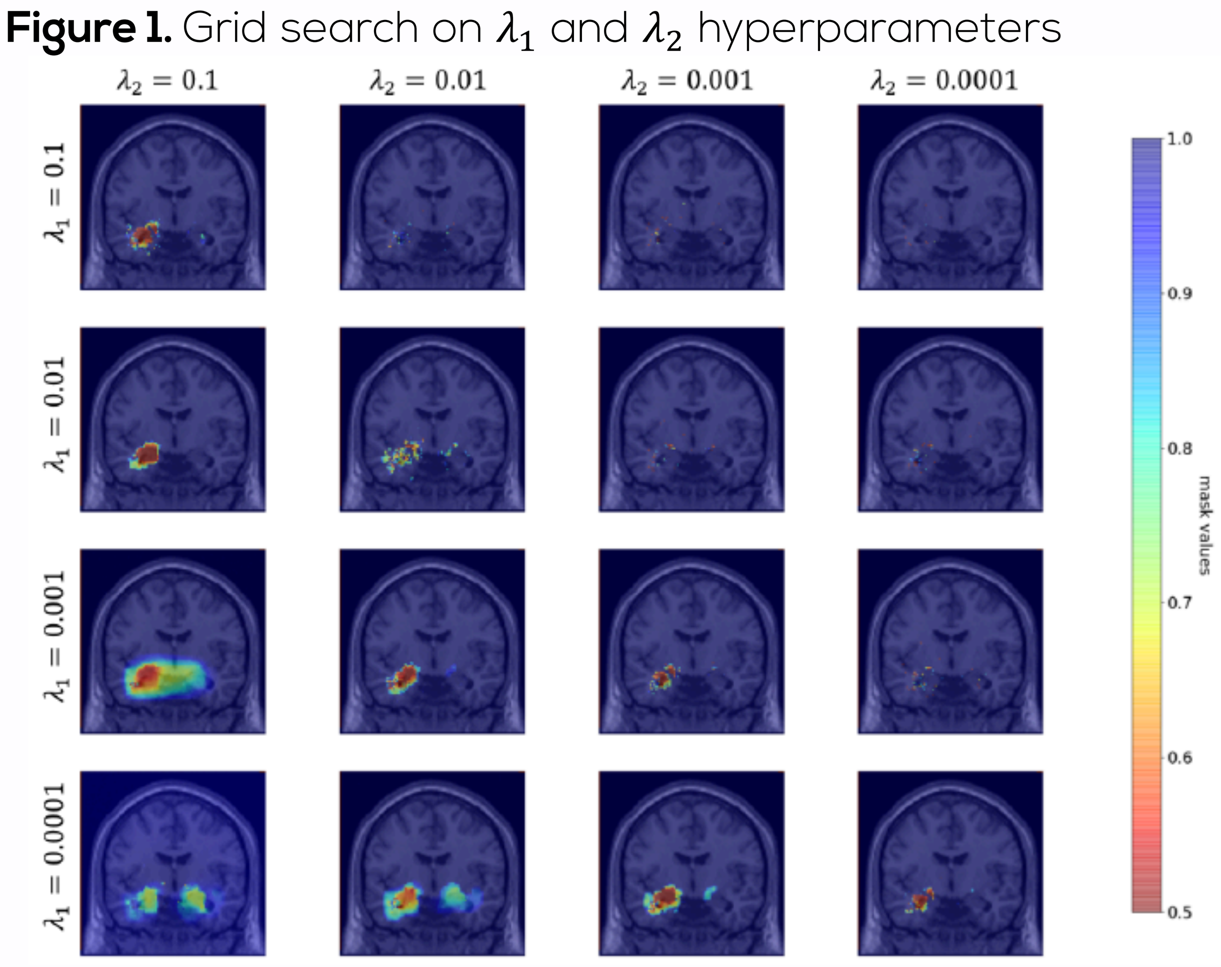


Table 1. ROI-based similarity across different values of λ_1

A	$\lambda_1 = 0.1$	$\lambda_1 = 0.01$	$\lambda_1 = 0.001$	$\lambda_1 = 0.0001$
$\lambda_1 = 0.1$		0.93	0.84	0.83
$\lambda_1 = 0.01$	0.93		0.95	0.91
$\lambda_1 = 0.001$	0.84	0.95		0.91
$\lambda_1 = 0.0001$	0.83	0.91	0.91	

→ Stability across different sets of hyperparameters

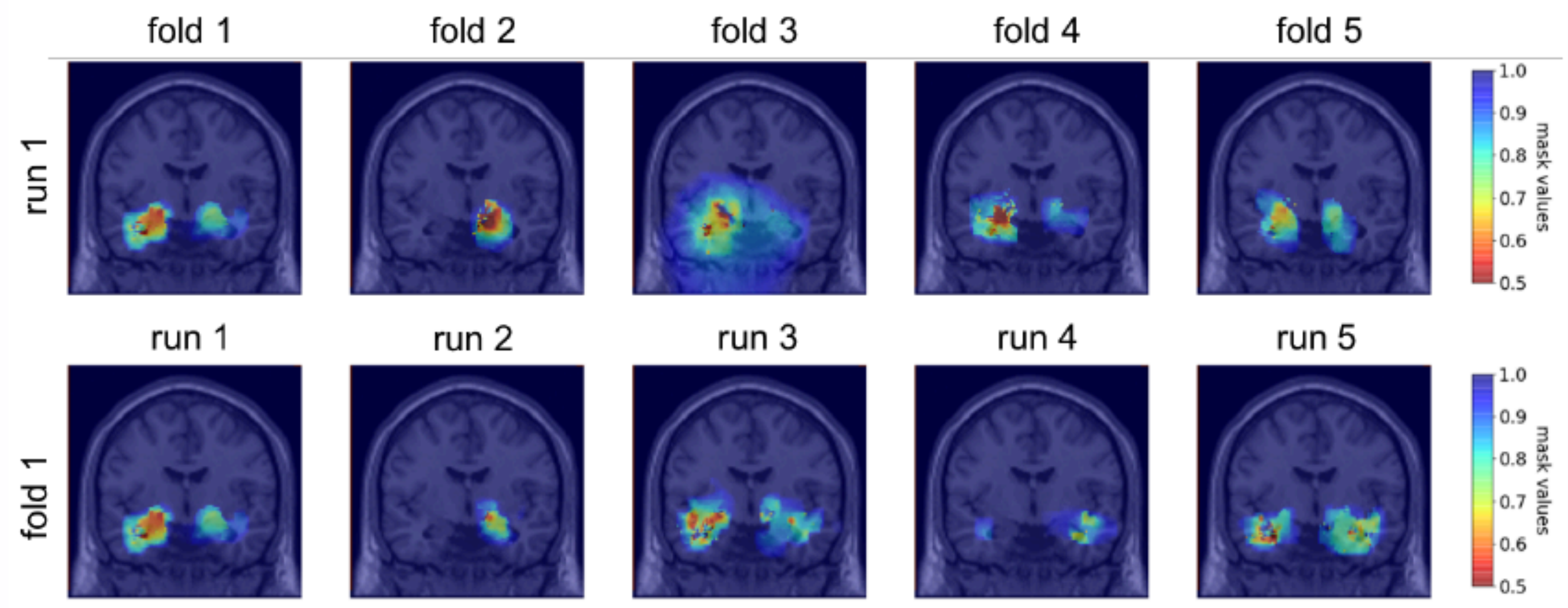
CNN robustness

Evaluation of dissimilarity:

Metric	inter-runs (mean)	inter-folds (mean)
prob_{CNN}	0.82	0.78
ROI-based	0.69	0.65

→ CNN training is not robust towards the regions identified

Figure 2. Masks obtained for the five folds of the CV on the first run (first line) and five runs of the first fold (second line)



Top 5 of more masked ROIs across the 5 folds

hippocampus, parahippocampal gyri, fusiform gyri, amygdalae, putamen, pallidum, temporal gyri, thalamus

→ Coherent with prior knowledge on AD

Conclusion

We demonstrated the robustness of our visualization method by showing the small impact of hyperparameters choice on the resulting mask.

Then we could apply this visualization method to assess the robustness of CNN training and found out that the patterns identified are not robust, though the set of most highlighted ROIs is coherent with previous knowledge on AD.

Bibliography

[1] Adebayo et al, 2018, 'Sanity checks for saliency maps'

[2] Fong and Vedaldi, 2017 'Interpretable Explanations of Black Boxes by Meaningful Perturbation'

[3] Routier et al, 2018 'Clinica: an open source software platform for reproducible clinical neuroscience studies'

[4] Begstra and Bengio, 2012 'Random Search for Hyper-Parameter Optimization'