Visualization approach to assess the robustness of neural networks for medical image classification
Elina Thibeau-Sutre, Olivier Colliot, Didier Dormont, Ninon Burgos

To cite this version:
Elina Thibeau-Sutre, Olivier Colliot, Didier Dormont, Ninon Burgos. Visualization approach to assess the robustness of neural networks for medical image classification. ICM days 2019, Jan 2020, louan, France. hal-03365775

HAL Id: hal-03365775
https://hal.archives-ouvertes.fr/hal-03365775
Submitted on 5 Oct 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Deep learning methods have shown 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.

### Results

#### Mask robustness

**Figure 1.** Grid search on \(\lambda_1\) and \(\lambda_2\) hyperparameters

**Table 1.** ROI-based similarity across different values of \(\lambda_1\)

![Image](https://example.com/image1)

![Image](https://example.com/image2)

> Stability across different sets of hyperparameters

#### CNN robustness

**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)

<table>
<thead>
<tr>
<th>Metric</th>
<th>inter-runs (mean)</th>
<th>inter-folds (mean)</th>
</tr>
</thead>
<tbody>
<tr>
<td>prob(_{\text{CNN}})</td>
<td>0.82</td>
<td>0.78</td>
</tr>
<tr>
<td>ROI-based</td>
<td>0.69</td>
<td>0.65</td>
</tr>
</tbody>
</table>

> CNN training is not robust towards the regions identified

**Top 5 of more masked ROIs across the 5 folds**

*hippocampus, parahippocampal gyrus, fusiform gyrus, amygdalae, putamen, pallidum, temporal gyrus, thalamus*

> Coherent with prior knowledge on AD

#### Bibliography