Identification of unlabeled latent subtypes with saliency maps

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

Identification of unlabeled latent subtypes with saliency maps

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 discover new associations and know how far they can be trusted.

The reliability of interpretability methods is difficult to evaluate, and saliency metrics may not be reliable themselves [1]. Moreover, it is often unclear what the method is proving, and what are its properties or drawbacks [2].

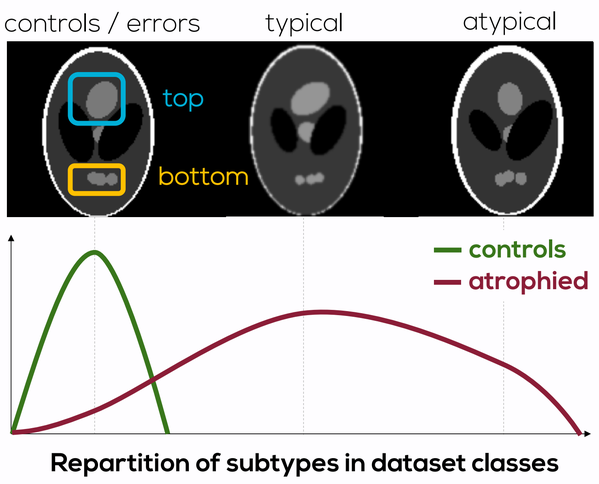
In this study, we used saliency maps on convolutional neural networks, and focused on their ability to accurately highlight patterns of latent subtypes grouped under the same labels in a simulation framework.

Methods

Synthetic data

Clinical cohort modelling

- **Two overlapping classes: controls / atrophied**
- **Composite atrophied label:** typical /atypical
- **Regional atrophy:** top / bottom



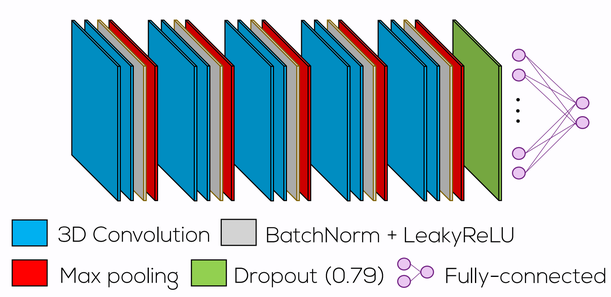
Repartition of subtypes in dataset classes

Datasets

datasets	#samples per class	atrophied composition: typical	atypical	errors
homogeneous	500	85%	10%	5%
heterogeneous	500	65%	25%	10%
large	10,000	85%	10%	5%
test	1,000	50%	50%	-

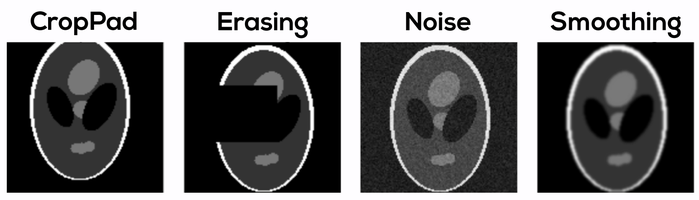
CNN training

Architecture



Task **classification controls vs atrophied**

Data augmentation



Optimization criteria

Comparison to default **cross-entropy (CE)** loss:

- **L1** $|x - y|$
- **L1Norm** $|softmax(x) - y|$
- **SmoothL1** $0.5(x-y)^2$ if $|x-y| < 1$ else $|x-y| - 0.5$
- **MSENorm** $(softmax(x) - y)^2$

Interpretability

Method

A saliency map [3] corresponds to the mean gradients of the **controls** node with respect to the image of an **atrophied** group. Then intensities are related to the changes needed to transform this image into a sample of the **control** group.

Evaluation criteria

3 criteria were designed to evaluate the separability of typical and atypical subtypes based on CNN feature maps or saliency maps of the typical or atypical group:

a. **CNN feature maps subtyping:** mean adjusted rand index between 10 K-means clustering and the true labels.

b. **Saliency maps separability:** (invert for atypical)

$$\frac{intensity_{Bottom}}{intensity_{Top}} * \frac{surface_{Top}}{surface_{Bottom}}$$

c. **Saliency maps specificity:**

$$\frac{intensity_{Top+Bottom}}{intensity_{rest}} * \frac{surface_{rest}}{surface_{Top+Bottom}}$$

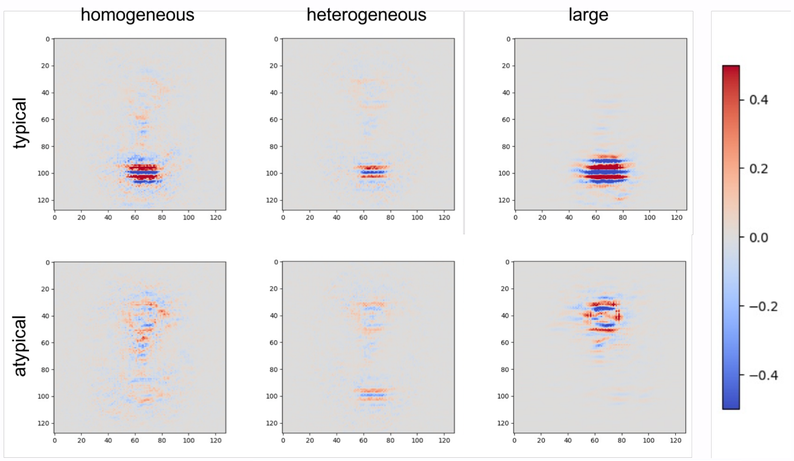
Criteria are evaluated on *test* dataset only.
When a criterion is applied to an atypical saliency map, the symbol † is used.

Values associated to randomness:
a=0.00 ; b=b†=1.00 ; c=c†=1.00

Results

Baseline results

Default training (no data augmentation + CE)



Training data	Evaluation criteria: a	b	c	b†	c†
homogeneous	0.95	12.3	5.3	1.6	4.2
heterogeneous	1.00	8.6	5.1	0.9	4.6
large	0.99	102.5	12.3	12.9	13.7

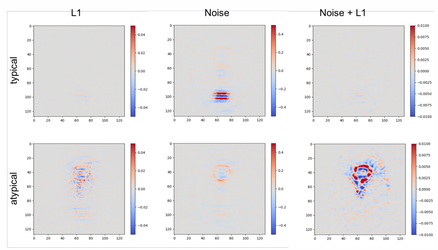
Benchmark

Benchmark on **homogeneous**

Training data	Evaluation criteria: a	b	c	b†	c†
CropPad	1.00	20.60	4.70	2.00	5.09
Erasing	1.00	19.55	5.57	1.79	3.96
Noise	1.00	18.37	9.08	2.32	5.48
Smoothing	1.00	10.36	2.94	1.13	7.88
L1	1.00	11.00	5.81	**3.09**	**5.71**
L1Norm	1.00	16.28	**10.03**	0.02	13.16
SmoothL1	0.80	7.27	3.00	2.01	**6.13**
MSENorm	0.95	12.25	5.32	1.59	**5.62**

Benchmark on **heterogeneous**

Training data	Evaluation criteria: a	b	c	b†	c†
CropPad	1.00	20.60	5.75	**2.56**	**7.29**
Erasing	1.00	13.72	6.28	**1.99**	5.22
Noise	1.00	**17.28**	**8.11**	**2.34**	5.39
Smoothing	1.00	12.11	5.12	1.87	3.93
L1	1.00	11.38	4.44	**2.58**	5.89
L1Norm	1.00	20.02	**9.82**	**2.04**	**11.93**
SmoothL1	1.00	8.23	4.30	**1.51**	5.15
MSENorm	1.00	10.23	6.03	**2.20**	6.25



Coupling best methods: **Noise** + **L1**

Training data	Evaluation criteria: a	b	c	b†	c†
homogeneous	1.00	6.91	2.94	**4.05**	**7.88**
heterogeneous	1.00	6.22	2.95	**1.88**	4.23

Deterioration of results on typical maps

Conclusion

The saliency maps better represent the typical subtype than the atypical one for a small number of samples whereas both subtypes are better represented when the dataset size is large.

Separability and specificity of saliency maps can be improved using data augmentation or by changing the optimization criterion, though these techniques cannot be coupled easily.

Bibliography & code

[1] Tomsett et al., 2020, 'Sanity Checks for Saliency Metrics'

[2] Lipton, 2018, 'The mythos of model interpretability'

[3] Simonyan et al., 2013, 'Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps'

Open source code at:
https://github.com/aramis-lab/AD-DL