Détection des images hors-distribution à partir de l'espace latent des modèles de segmentation d'images médicales

Benjamin Lambert[1,3], Florence Forbes[2], Senan Doyle[1] , Alan Tucholka[1], Michel Dojat[3]

1. *Pixyl, Research and Development Laboratory, Grenoble* ***2.*** *Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK* ***3.*** *Univ. Grenoble Alpes, Inserm, U1216, Grenoble Institut des Neurosciences*

Motivation

Les modèles d'apprentissage profond sont facilement perturbés par des variations dans les images d'entrée qui n'ont pas été observées pendant la phase d'apprentissage, ce qui entraîne des prédictions erronées. Récemment, une nouvelle catégorie de méthodes a émergé pour détecter ces images *hors-distribution*, basée sur l'analyse des activations intermédiaires d'un modèle entraîné. Dans cette étude, nous proposons de comparer 5 de ces méthodes, à partir d'un benchmark comportant 20 types différents d'anomalies, pour un total d'environ 7800 IRMs. Nos résultats montrent que les méthodes qui considèrent les activations de ***toutes les couches intermédiaires*** sont plus performantes que les méthodes se restreignant à une seule couche.

Matériel & Méthodes

- Notre benchmark repose sur la tâche de segmentation de tumeur cérébrale à partir d'IRM T1, utilisant la base de données BraTS 2021 (N=1251). Nous l'avons réparti en 3 groupes : apprentissage (N=651), calibration (N=200), et test (N=400). Pour tester la généralisabilité du modèle de segmentation, nous employons également une base *Control* composée de 74 patients de la base de donnée de glioblastomes LUMIERE. L'ensemble de ces données correspond aux images *en-distribution* (ED).

- Nous proposons ensuite d'utiliser 20 datasets d'images *hors-distribution (HD)*, que l'on peut classifier en 4 catégories : *transformation, diagnostic, modalité,* et *extrême*.

- Nous avons entrainé un modèle de segmentation Attention U-Net 3D [1] à partir des images d'apprentissage. Ensuite, 5 méthodes différentes de détection d'images HD à partir des activations intermédiaires du modèle ont été implémentées : l'analyse spectrale [2], les prototypes de classes [3], la distance de Mahalanobis [4-5], et enfin les Support Vector Machines (OCSVM) [6] à une classe. Ces méthodes sont entrainées à l'aide de la base de calibration.

- **Évaluation** : Chaque détecteur produit un score par image (HD et ED). En les comparant, nous pouvons obtenir des scores de classification (AUPR) pour chaque dataset HD. Nous reportons également la performance de segmentation (Dice) quand la segmentation vérité-terrain des tumeurs est disponible.

Figure 2 : Présentation des 5 détecteurs d'HD testés

Références

1. Oktay et al.: Attention u-net: Learning where to look for the pancreas. Medical Imaging with Deep Learning (2018)
2. Karimi et al: Improving calibration and out-of-distribution detection in deep models for medical image segmentation. IEEE Transactions on AI (2022)
3. Diao et al.: A unified uncertainty network for tumor segmentation using uncertainty cross entropy loss and prototype similarity. Knowledge-Based Systems 246 (2022)
4. González et al.: Distance-based detection of out-of-distribution silent failures for covid-19 lung lesion segmentation. Media 82, 102596 (2022)
5. Çallı, E., et al.: Frodo: An in-depth analysis of a system to reject outlier samples from a trained neural network. IEEE Transactions on Medical Imaging (2022)
6. Wang et al.: Layer adaptive deep neural networks for out-of- distribution detection. Advances in Knowledge Discovery and Data Mining pp. 526–538 (2022)

20 types d'anomalies différentes représentant ≃ *7800 IRMs*

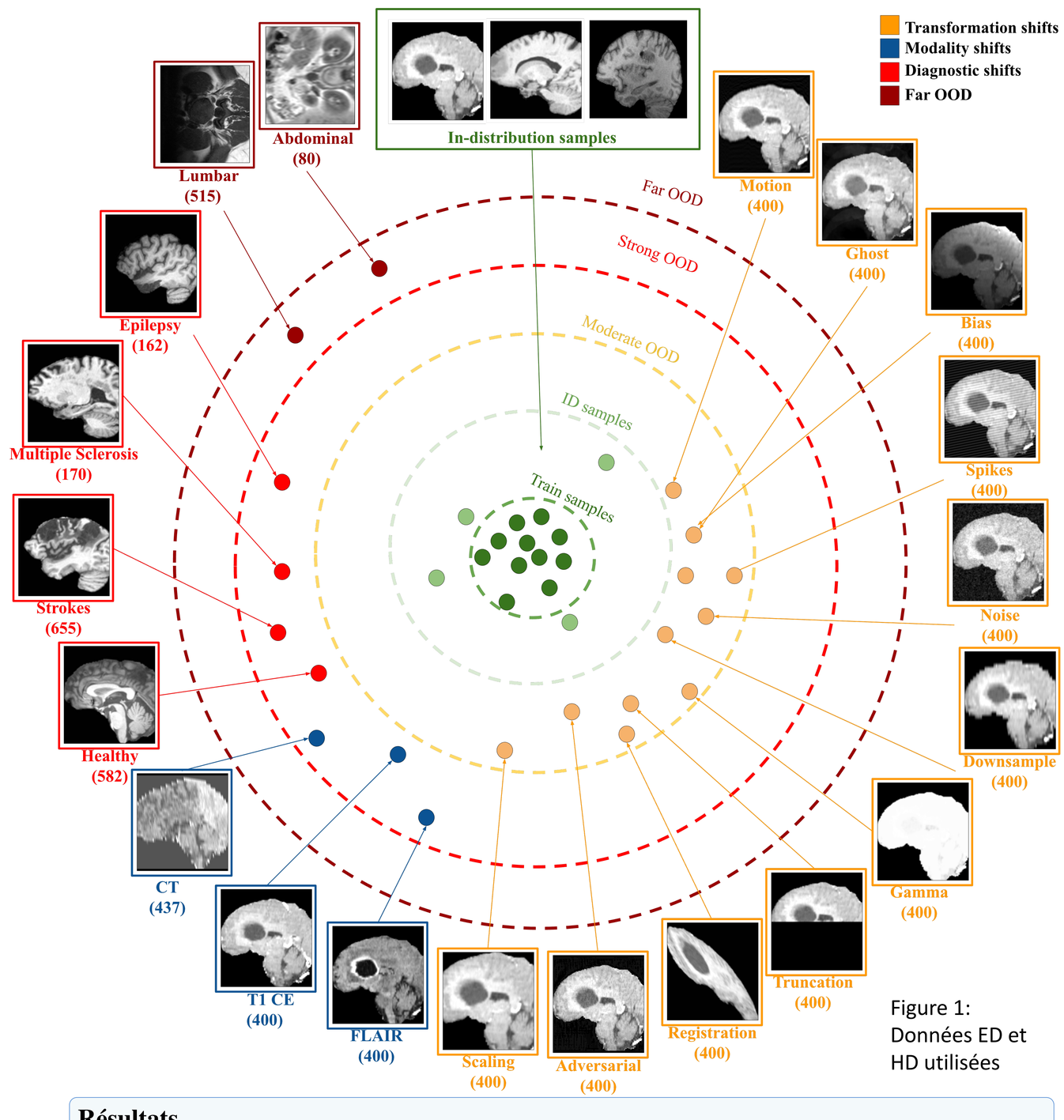

Figure 1: Données ED et HD utilisées

Résultats

- Toutes les méthodes de détection d'HD ont une très bonne performance sur les cas extrêmes (Far OOD), mais leur performance est cependant très variable sur les autres types d'anomalies.

  → Se limiter à des cas extrêmes n'est donc pas suffisant pour bien évaluer un détecteur d'HD.

- FRODO et OCVSM se démarquent des autres méthodes. Ces deux méthodes considèrent les activations de chacune des couches de convolution, contrairement aux autres méthodes (Spectral, Prototype & MD Encoder) qui ne considèrent qu'une seule couche.

  → Les détecteurs qui considèrent toutes les couches intermédiaires sont bien plus performants

	N	Dice	Spectral	Prototype	$MD_{Encoder}$	FRODO	OCSVM	Random
Test ID	400	.83	-	-	-	-	-	-
Control	74	.85	0.18	0.13	0.13	0.14	0.12	**0.16**
Motion	400	.82	0.62	0.49	**0.75**	**0.75**	0.56	0.50
Ghost	400	.80	0.61	0.47	**0.85**	0.80	0.57	0.50
Bias	400	.78	**1.00**	0.86	**1.00**	**1.00**	**1.00**	0.50
Spikes	400	.79	0.89	0.67	0.85	**1.00**	**1.00**	0.50
Noise	400	.81	0.76	0.54	0.73	**1.00**	**1.00**	0.50
Downsample	400	.82	0.65	0.49	0.53	**0.68**	0.55	0.50
Gamma	400	.31	**1.00**	0.95	0.97	**1.00**	**1.00**	0.50
Truncation	400	.61	0.99	0.86	0.99	**1.00**	0.99	0.50
Registration	400	.01	**1.00**	0.93	**1.00**	**1.00**	**1.00**	0.50
Adversarial	400	.58	0.74	0.44	0.84	**1.00**	0.97	0.50
Scaling	400	.77	0.99	0.91	0.99	**1.00**	**1.00**	0.50
Transformation	4400	-	0.38	0.17	0.44	**0.66**	0.39	0.10
FLAIR	400	.10	0.99	0.42	0.99	**1.00**	0.99	0.50
T1Ce	400	.69	0.96	0.79	0.91	0.97	0.90	0.50
CT	437	-	0.98	0.41	0.99	**1.00**	**1.00**	0.52
Modality	1237	-	**0.98**	0.49	0.94	0.97	0.91	0.28
Healthy	577	-	0.57	0.90	0.40	0.97	**0.99**	0.59
Strokes	655	-	0.66	0.88	0.54	**0.88**	**0.88**	0.62
WMH	170	-	0.44	0.77	0.21	**0.87**	0.84	0.30
Epilepsy	162	-	0.40	**0.84**	0.23	0.66	0.70	0.29
Diagnosis	1564	-	0.37	0.71	0.16	**0.83**	0.70	0.23
Lumbar	515	-	**1.00**	0.87	**1.00**	**1.00**	**1.00**	0.56
Abdominal	80	-	**1.00**	0.97	**1.00**	**1.00**	**1.00**	0.17
Far OOD	595	-	**1.00**	0.88	**1.00**	**1.00**	**1.00**	0.44
Overall	7796	-	0.24	0.12	0.12	**0.61**	0.33	0.06

Tableau 1 : Résultat de détection des HDs (AUPR) pour chaque méthode et dataset

Conclusion

Les détecteurs d'images HD se basant sur l'espace latent des modèles de segmentation sont prometteurs, permettant une détection fiable (AUPR=1.00 sur 12 des 20 scénario pour FRODO) et rapide (< 1s). Les méthodes qui se basent sur l'activation de toutes les couches de convolution sont cependant plus robustes que celles qui ne se concentrent que sur une seule couche.