Construction et apprentissage sous contraintes de réseaux monotones pour une classification interprétable et la détection d'anomalies

Valentine Wargnier-Dauchelle, Thomas Grenier, Françoise Durand-Dubief, François Cotton, Michaël Sdika

Building, Training and Constraining Non-Negative Networks to Improve Interpretability

V. WARGNIER-DAUCHELLE[1], T. GRENIER[1], F. DURAND-DUBIEF[1,3], F. COTTON[1,2], M. SDIKA[1]

[1] INSA-Lyon, Université Claude Bernard Lyon 1, CNRS, Inserm, CREATIS UMR 5220, U1294, Lyon, France

[2] Service de Radiologie, Centre Hospitalier Lyon-Sud, Hospices Civils de Lyon, Pierre-Bénite, France

[3] Service de Neurologie A, Hôpital Neurologique, Hospices Civils de Lyon, Bron, France

INTRODUCTION

The lack of interpretability of deep learning reduces understanding of what happens when a network does not work as expected and hinders its use in critical fields like medicine, which require **transparency of decisions**.

AIM

A **healthy vs pathological images classification** model should rely on radiological signs and not on some training dataset biases. We propose to **build, train and constrain a monotonic classifier**, which has some intrinsic explicability properties, such that its decision is based on relevant radiological structures in an unsupervised way. The trained network can be used for **interpretable classification** consistent with high-level clinical knowledge but also as **weakly-supervised pathology segmentation** network.

METHOD

We design the architecture described in Figure 1. The interpretable features space benefits from several intrinsic explicability properties: they are **ordered**, with a **known bound** between healthy/pathological and the **counterfactual examples are more readable** as we can find a **positive** α (see Figure 2).

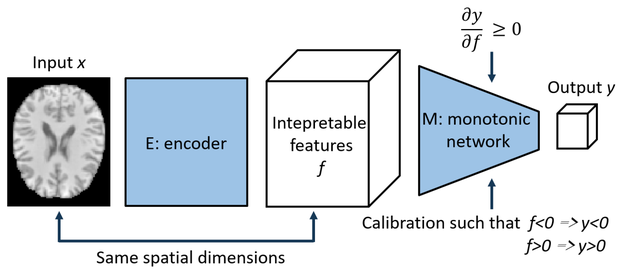


Figure 1: Overall architecture

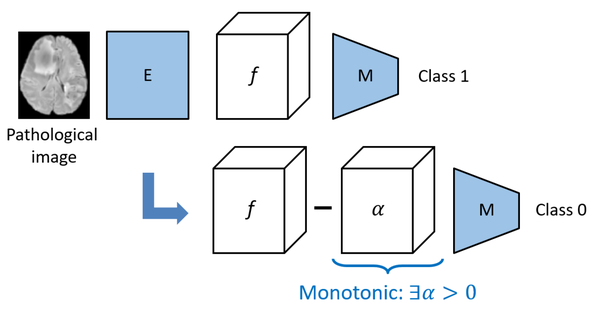


Figure 2: Counterfactual difference generation

• Building

For the M network, **weights are parameterized to be positive**, we **remove biases**, use **convex activations** (r(x)) on half of the features and **concave** (-r(-x)) on the other half and **remove normalization** layers.

• Training

We show that using positive weights strongly **increases the correlation** between random features channels and so state-of-the-art random initializations are not adapted to non-negative networks (Figure 3). We propose a initialization **rescaling each linear layer**, one after another, by its standard deviation.

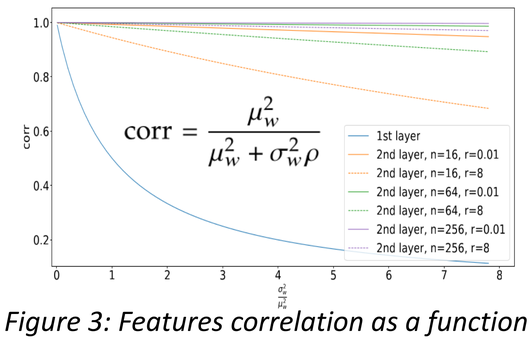


Figure 3: Features correlation as a function of weights standard deviation and mean.

• Constraining

The network is trained with four losses to constrain the interpretable features *f*:

- Classification healthy (0) / pathological (1)
- **Negative** heathy features $f_0 \leq 0$
- **Similar distribution** of the negative characteristics of the two classes $\mathrm{D}(f_{0\_neg}) \approx \mathrm{D}(f_{1\_neg})$
- **Gradient regularization** for healthy class : $\frac{\partial y_0}{\partial f_0} \sim 0$

RESULTS: TRAINING

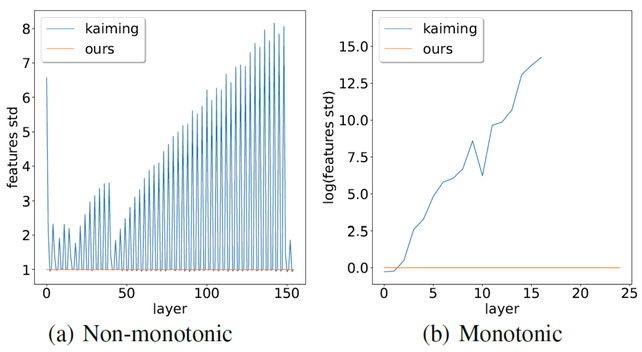


Figure 4: Features standard deviation as a function of the depth in a ResNet152.

• Variance preservation after initialization

The proposed initialization allows a **unit variance** in the whole network whereas, with Kaiming [1] initialization, the variance increases with the layers (Figure 4). For non-negative network, using our initialization is mandatory as the variance tends towards infinity with Kaiming.

Task	Standard		Monotonic	
	Kaiming	Ours	Kaiming	Ours
Healthy vs tumors	**1.00**	**1.00**	0.98	**1.00**
Single Cell	/	/	0.83	**0.92**
MedNIST	**1.00**	**1.00**	NaN	**0.98**
Brain tumors	0.72	**0.73**	NaN	**0.71**

Table 1: Classification accuracy.

• Classification

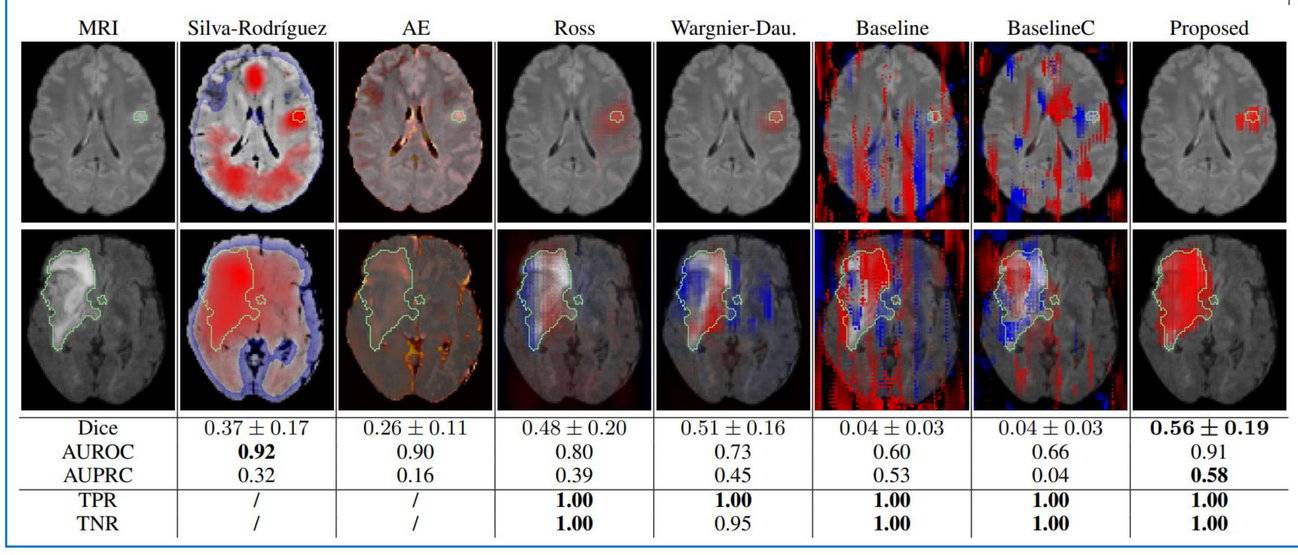
With obtain **similar classification performances** on various tasks with both initialization for standard networks (Table 1). For monotonic networks, only small architectures can be trained with Kaiming initialization due to the variance issue. Replacing a network by its monotonic counterpart does not degrade the classification performances.

RESULTS: SEGMENTATION

• More readable counterfactual examples

We use the counterfactual difference α of Figure 2 to interpret our network (Figure 5). The experiments shows that the difference is **focused on the tumor** even if it is not the case for the features. With the non-negative network, this difference is **more readable as it is positive**, compared to a standard network (Baseline in Figure 6) or standard constrained network (BaselineC).

Figure 5: Interpretable features (top) and corresponding counterfactual difference α (bottom) for the proposed method. Tumor is in green.



	Silva-Rodríguez	AE	Ross	Wargnier-Dau.	Baseline	BaselineC	Proposed
Dice	0.37 ± 0.17	0.26 ± 0.11	0.48 ± 0.20	0.51 ± 0.16	0.04 ± 0.03	0.04 ± 0.03	$\mathbf{0.56 \pm 0.19}$
AUROC	**0.92**	0.90	0.80	0.73	0.60	0.66	0.91
AUPRC	0.32	0.16	0.39	0.45	0.53	0.04	**0.58**
TPR	/	/	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
TNR	/	/	**1.00**	0.95	**1.00**	**1.00**	**1.00**

Figure 6: State-of-the-art comparison: segmentation maps and metrics. Tumor is in green. Blue represents negative attributions/counterfactual difference and red positive ones. For reconstruction methods, reconstruction error scale is from black to yellow.

• Interpretable classification and anomaly detection

Our proposition outperforms both interpretable classification (Ross [2], Wargnier-Dauchelle [3]) **and anomaly detection** (Silva-Rodríguez [4], AE [5]) state-of-the-art methods in terms of Dice and AUPRC for tumors segmentation. It reaches perfect classification performances (Figure 6) .

CONCLUSION

We propose to use a **constrained non-negative network** and the generation of **counterfactual examples** for:

1) A more **pathology-driven classification**
2) A **weakly-supervised segmentation** method outperforming state-of-the-art.

To do so, we propose :

1) A methodology **for transforming any network into a non-negative network**
2) A theoretical analysis to identify the **failure causes of the state-of-the-art random weights initializations** for certain layers
3) An **efficient weights initialization** which maintains **a unit variance** in all the layers of the network
4) **Well chosen constraints** for interpretable features

REFERENCES

[1] HE. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. IEEE ICCV. 2015.

[2] ROSS et al. Right for the right reasons: training differentiable models by constraining their explanations. IJCAI. 2017.

[2] WARGNIER-DAUCHELLE et al. A weakly-supervised gradient attribution constraint for interpretable classification and anomaly detection. IEEE TMI. 2023.

[3] SILVA-RODRIGUEZ et al. Looking at the whole picture: constrained unsupervised anomaly segmentation. BMCV. 2021.

[4] BAUR et al. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. MICCAI BrainLes WS. 2018.

ACKNOWLEDGEMENTS

Labex PRIMES, FLI, : ANR-11-LABX-0063, ANR-11-IDEX-0007, ANR-11-INBS-0006, ANR-10-COHO-002

CONTACT INFORMATION

For more information:
valentine.wargnier@creatis.insa-lyon.fr