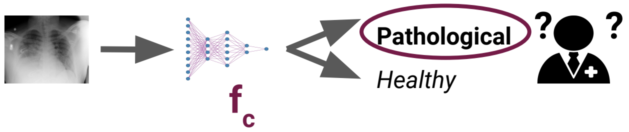
Combining Similarity and Adversarial Learning to Generate Visual Explanation: Application to Medical Image Classification

Martin Charachon[12], Céline Hudelot[2], Paul-Henry Cournède[2], Camille Ruppli[1], Roberto Ardon[1]
[1]Incepto Medical
[2]Université Paris-Saclay, CentraleSupélec, MICS

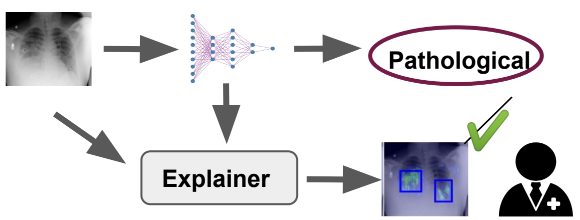
Introduction

Problem: **Explain** decision of black-box classifiers



Objectives:

What are the **discriminative regions** in the image for the classifier? Are they **interpretable** for humans (clinicians) ?



Prior Work

→ Saliency maps [1], Activation maps based [2]
→ **Perturbation-based** [3, 4, 5]

[1] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," in ICLR, 2014
[2] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in ICCV, 2017
[3] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in ICCV, 2017
[4] P. Dabkowski and Y. Gal, "Real time image saliency for black box classifiers," in NIPS, 2017
[5] . Elliott, S. Law, and C. Russell, "Adversarial perturbations on the perceptual ball," ArXiv, 2019

Contributions

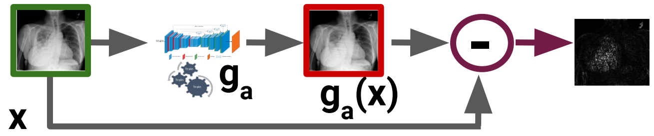
New definition of **visual explanation** through **adversarial** example generation

New Optimization workflow combining the **training** of an **adversarial generator** and a **similar generator**

New **regularization** methods **improving** several explanation methods in terms of **weak localization**

"Naive" Method

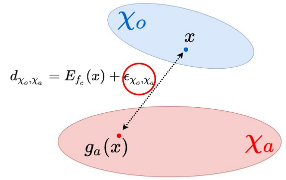
Visual explanation: $E_{f_c}(x) = |x - g_a(x)|$



$f_c(g_a(x)) \neq f_c(x)$

Issues:

→ Different space χ_o and χ_a
→ Reconstruction errors

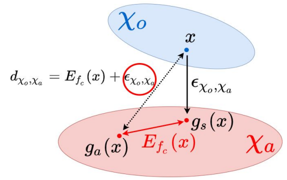


Proposed Method

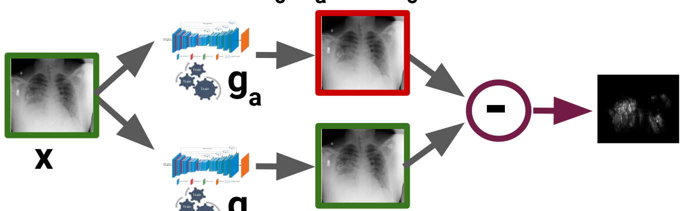
Approach: Generate an adversarial example $g_a(x) \in \chi_a$
Project x in space $\chi_a \rightarrow g_s(x)$

Visual explanation:

$$\boxed{E_{f_c}(x) = |g_s(x) - g_a(x)|}$$



$f_c(g_a(x)) \neq f_c(x)$



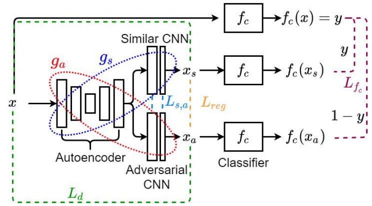
$f_c(g_s(x)) = f_c(x)$

Joint Optimization Problem

Optimization function

$$(\bar{g}_s, \bar{g}_a) = \underset{g_s, g_a}{\operatorname{argmin}} \left\{ \begin{array}{l} \mathbb{E}_x \left(\begin{array}{ll} L_d(x, g_s(x), g_a(x)) & + \\ L_{f_c}(x, g_s(x), g_a(x)) & + \\ L_{reg}(x, g_s(x), g_a(x)) & \end{array} \right) \\ + \quad L_{s,a}(g_s, g_a) \end{array} \right\}$$

Optimization framework



Regularization:
N random geometric transformations ψ_i at test time

$$\overline{E}_{f_c}(x) = \frac{1}{N+1}\left[E_{f_c}(x) + \sum_{i=1}^{N} \psi_i^{-1}\left(E_{f_c}(\psi_i(x))\right)\right]$$

Experimental Results

Adversarial and Similar generation

Great similarity: x, $g_s(x)$ and $g_a(x)$
↘↘ **reconstruction errors** between $g_s(x)$ and $g_a(x)$

Weak Localization

$$IoU_i = \frac{M_{GT} \cap M_{Ei}}{M_{GT} \cup M_{Ei}} \qquad AUC_{Loc} = \sum_i P_i(R_i - R_{i-1})$$

IOU scores at different thresholds of binarization - Comparison to State of the Art Methods without (Top) and with (Bottom) augmentations

Explanation method *Percentile*	IOU 80	85	90	95	98
Gradient [1]	0.203	0.199	0.187	0.152	0.097
	0.256	***0.252***	***0.236***	***0.190***	***0.117***
GradCAM [2]	0.237	0.225	0.195	0.138	0.070
	0.271	***0.263***	***0.244***	***0.190***	***0.105***
BBMP [3]	0.233	0.226	0.204	0.154	0.087
Mask Generator [4]	0.222	0.219	0.208	0.169	0.103
	0.259	***0.264***	***0.259***	**0.221**	**0.137**
"Naive"	0.177	0.173	0.158	0.118	0.064
	0.239	***0.230***	***0.208***	***0.156***	***0.087***
Ours	0.248	0.250	0.232	0.173	0.097
	0.292	**0.292**	**0.272**	***0.206***	***0.115***

Estimated AUC scores for Precision-Recall and Computation time - Comparison to State of the Art Methods without (Top) and with (Bottom) augmentations

Explanation method	Total AUC	Partial AUC	Time (s)
Gradient [1]	0.287	0.189	2.04
	0.374	***0.274***	2.83
GradCAM [2]	0.326	0.235	0.78
	0.397	***0.302***	5.09
BBMP [3]	0.326	0.229	17.14
Mask Generator [4]	0.327	0.226	0.09
	0.404	***0.308***	0.68
"Naive"	0.238	0.145	0.10
	0.325	***0.232***	0.75
Ours	0.339	0.256	0.05
	0.412	**0.328**	0.63