DLinCV Lecture 1.4

GRADIENTS: Saliency maps, adversarial attacks

6.6.2023

Aasa Feragen
afhar@dtu.dk

Today's learning goals

After today's lecture you should be familiar with how gradients are used for:

- producing saliency maps for explainable AI
- performing adversarial attacks
- and you should be aware of some of their un-intuitive properties and how to make proper use of them

NB! See also lecture notes on Learn.

Gradients

- For a function $f : \mathbb{R}^n \to \mathbb{R}$, the gradient

$$\nabla_x f = (\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n})$$

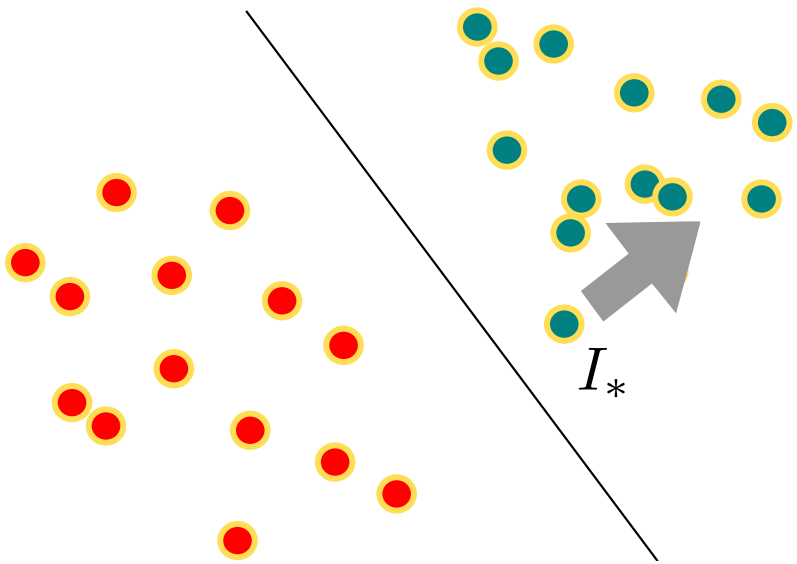
  at $x \in \mathbb{R}^n$ is a vector in $\mathbb{R}^n$ pointing in the direction of maximal slope.
- For an image classifier $f : \mathbb{R}^{H \times W} \to \mathbb{R}$ predicting probability of positive class, and an image $I \in \mathbb{R}^{H \times W}$, what is ∇f?

Gradients

- For a function $f : \mathbb{R}^n \to \mathbb{R}$, the gradient

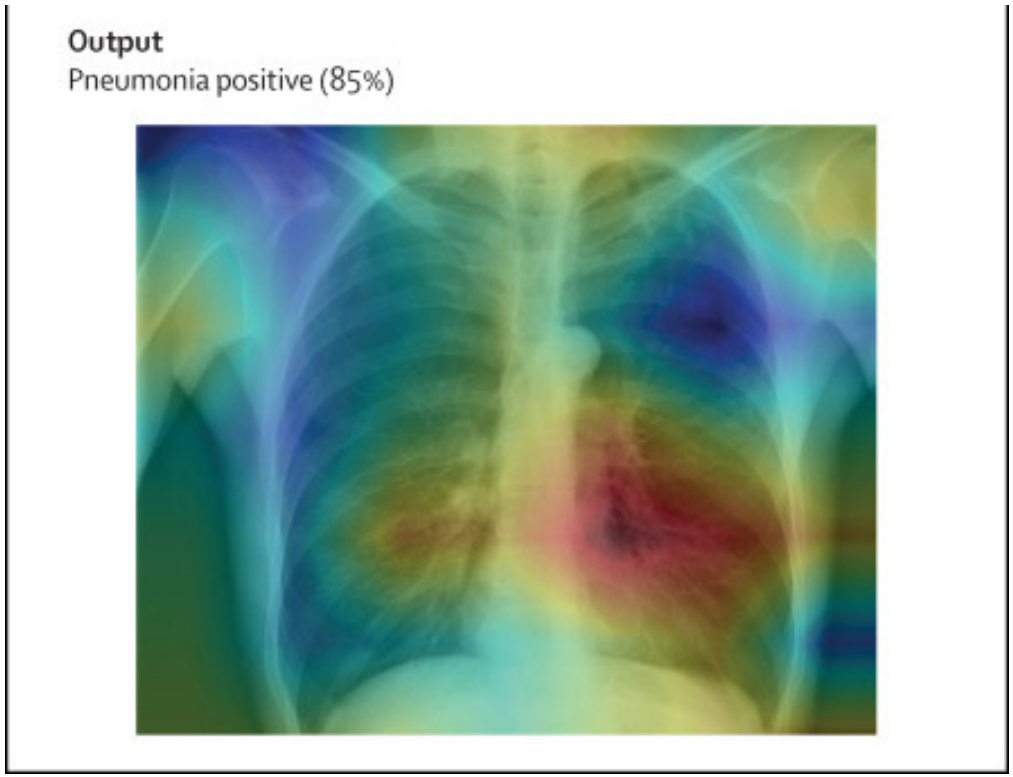
$$\nabla_x f = (\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n})$$

  at $x \in \mathbb{R}^n$ is a vector in $\mathbb{R}^n$ pointing in the direction of maximal slope.
- For an image classifier $f : \mathbb{R}^{H \times W} \to \mathbb{R}$ predicting probability of positive class, and an image $I \in \mathbb{R}^{H \times W}$, what is ∇f?
- A matrix (image) in $\mathbb{R}^{H \times W}$, pointing in the direction of the positive class.



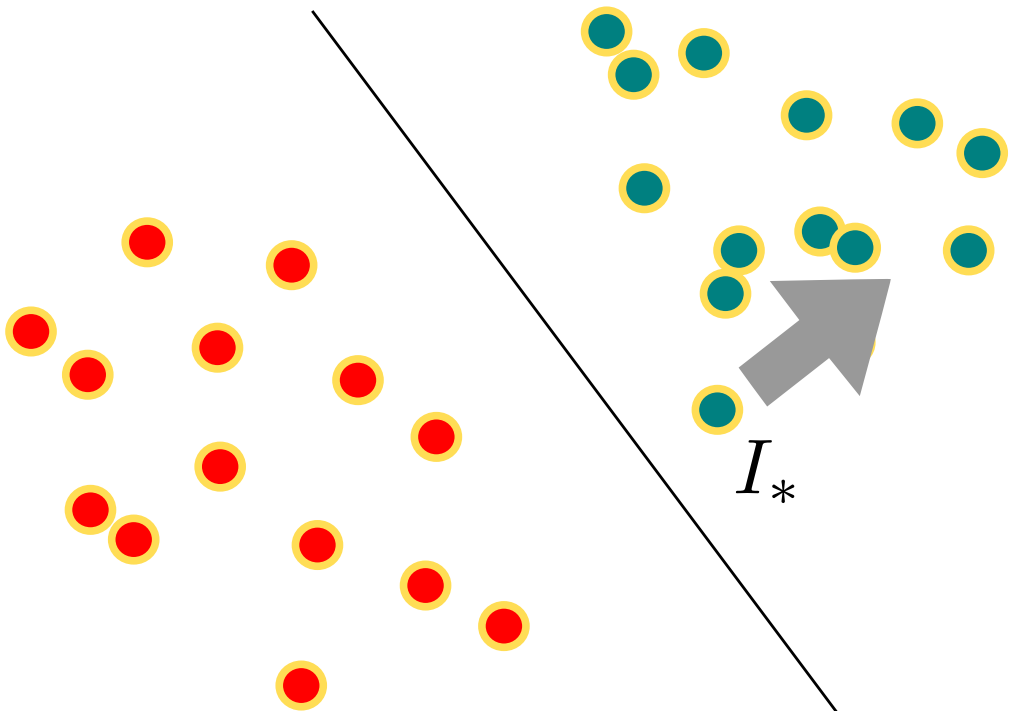
Saliency maps

A *saliency map* is a heatmap in the image plane that aims to emphasize those regions of the image that were the most important – or *salient* – for the made prediction



Question: How would you design such a tool?

Saliency maps – how might you define them?



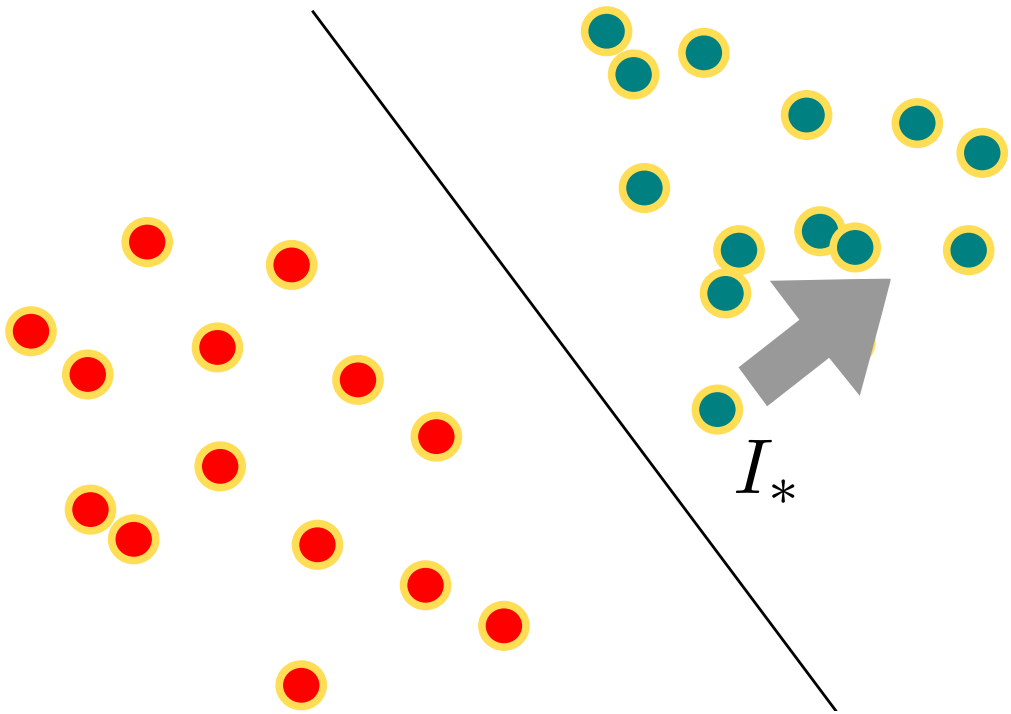
Common solution: Use gradients

Vanilla saliency maps: Gradients

The vanilla gradient saliency map for the image is simply given by the gradient

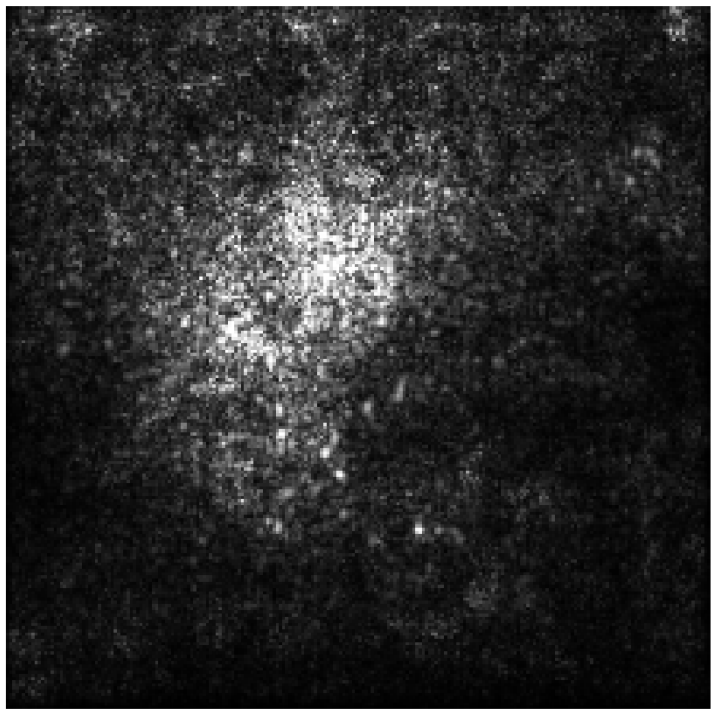
$$\nabla_x f_l(\theta, I_*)$$

of the (usually the correct) l^{th} class prediction of the network f at I_*.



SmoothGrad

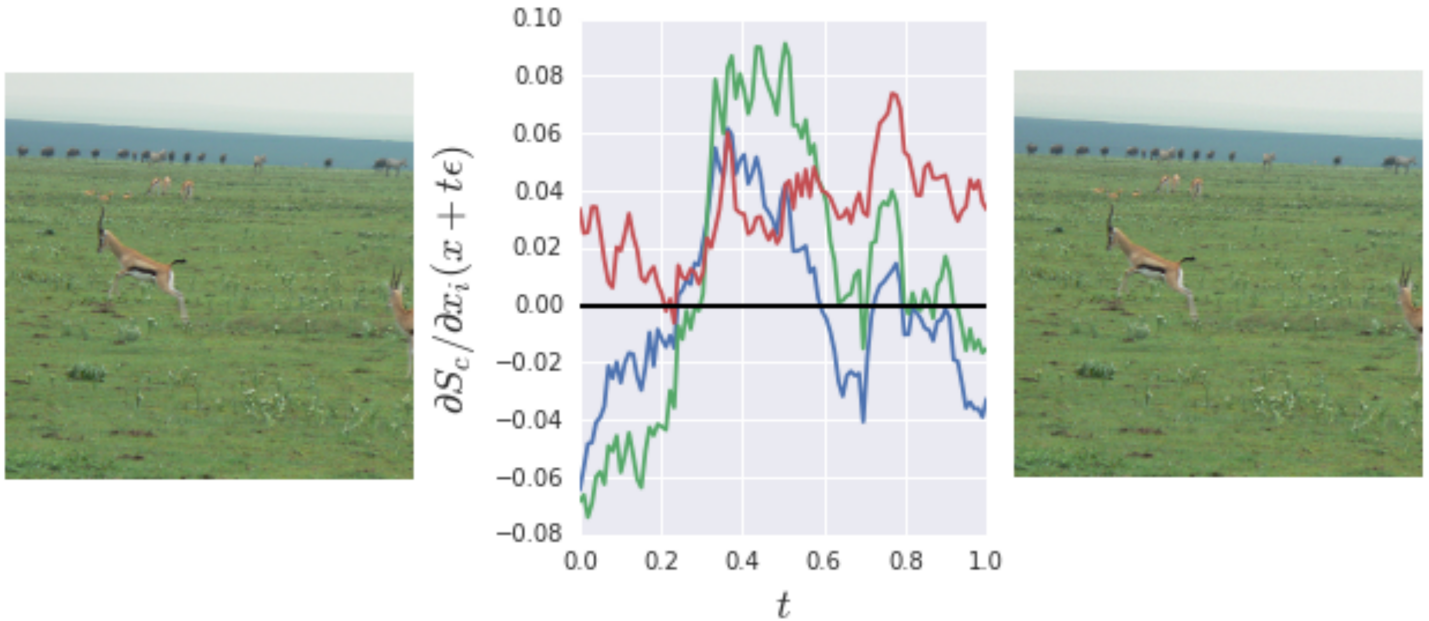
Vanilla gradients tend to look "noisy":



Smilkov et al, SmoothGrad: removing noise by adding noise, `https://arxiv.org/pdf/1706.03825.pdf`, 2017

SmoothGrad

Reason: gradients fluctuating heavily as functions on image space



Smilkov et al, SmoothGrad: removing noise by adding noise, `https://arxiv.org/pdf/1706.03825.pdf`, 2017

SmoothGrad

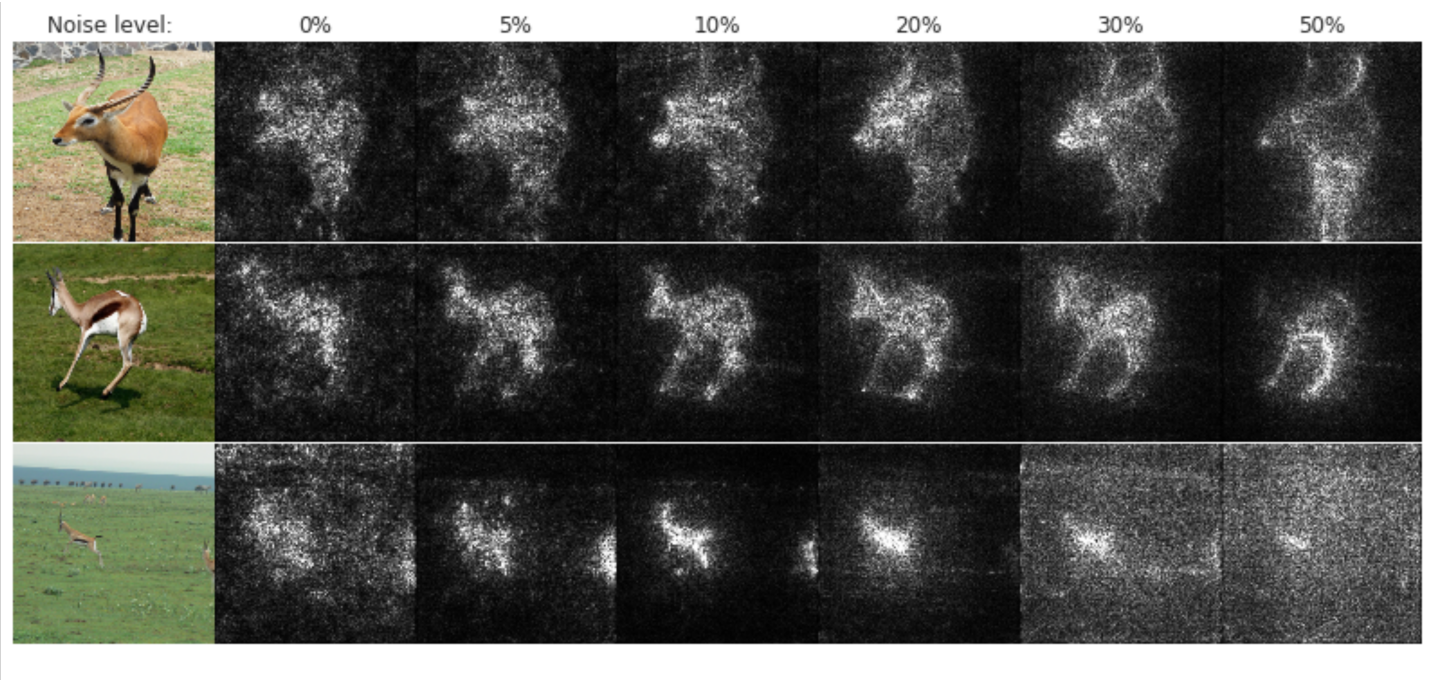
Smoothgrad: Apply Gaussian smoothing in image space – via stochastic approximation:

$$\hat{S}(I_*) = \frac{1}{n} \sum_n \nabla_I f(I_n), \tag{1}$$

where σ is the standard deviation of the Gaussian noise

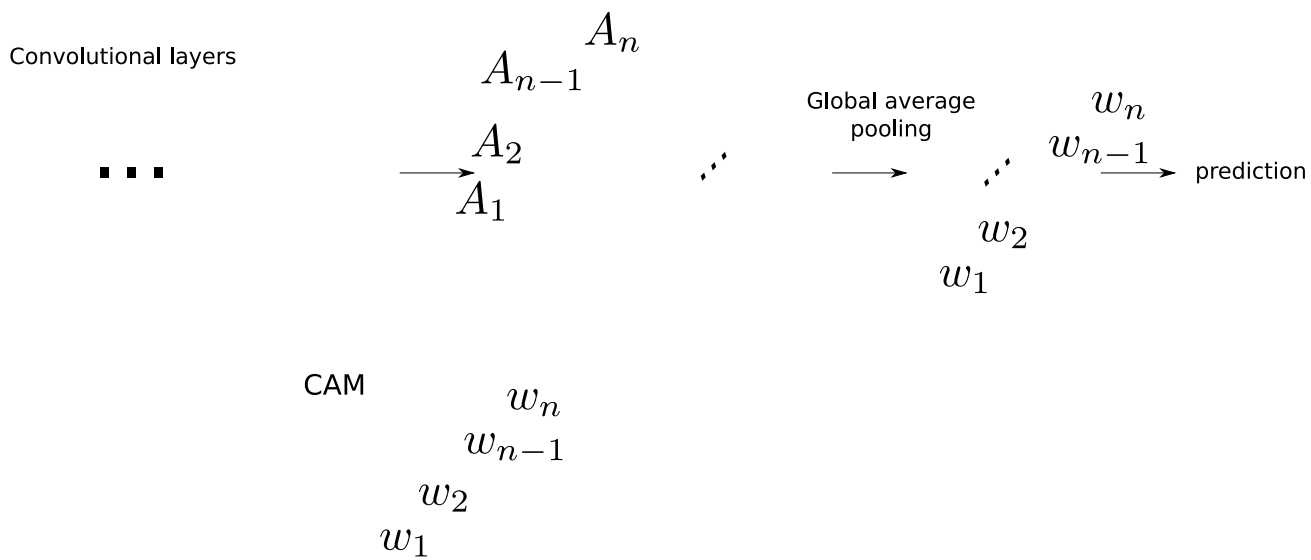
Smilkov et al, SmoothGrad: removing noise by adding noise, `https://arxiv.org/pdf/1706.03825.pdf`, 2017

SmoothGrad



Smilkov et al, SmoothGrad: removing noise by adding noise, `https://arxiv.org/pdf/1706.03825.pdf`, 2017

CAM: Class Activation Mapping



GradCAM

- Post-hoc method
- Activation mapping weights computed via gradients:

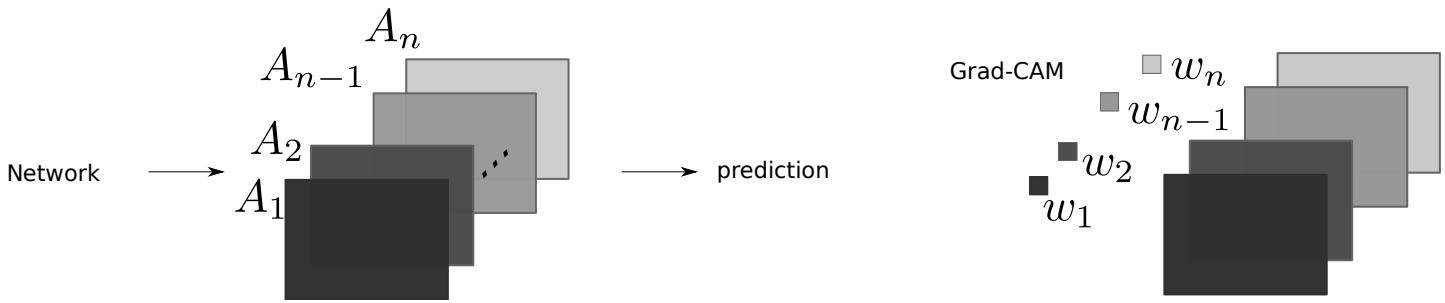
$$w_k = \text{mean}(\nabla_{A^k} f_l(I_*)$$

where $f_l(x)$ is the (before softmax) prediction of the label l for input image I_*, A^k is the features of the k^{th} penultimate channel.

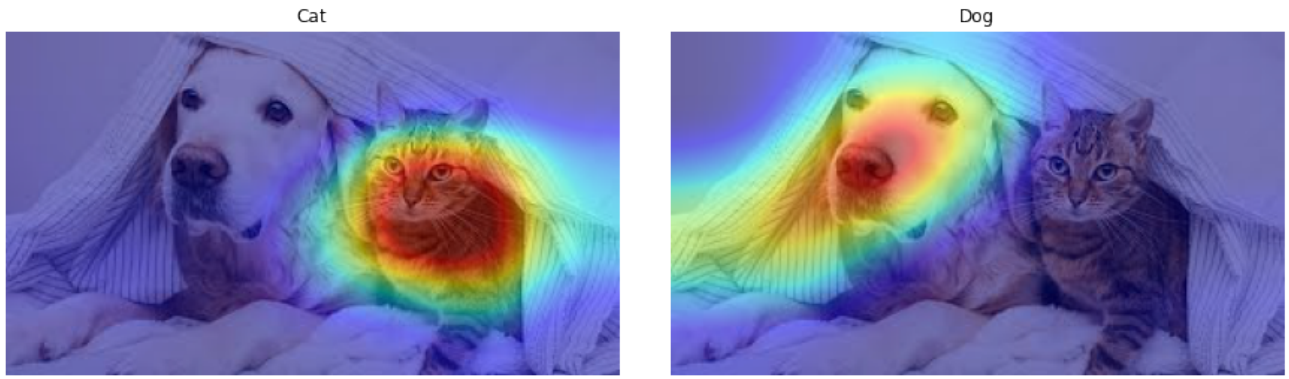
- Saliency map:

$$\hat{S}(I_*) = ReLU(\sum_k w_k A^k),$$

- ReLU turns off negative contributions – more intuitive output



GradCAM



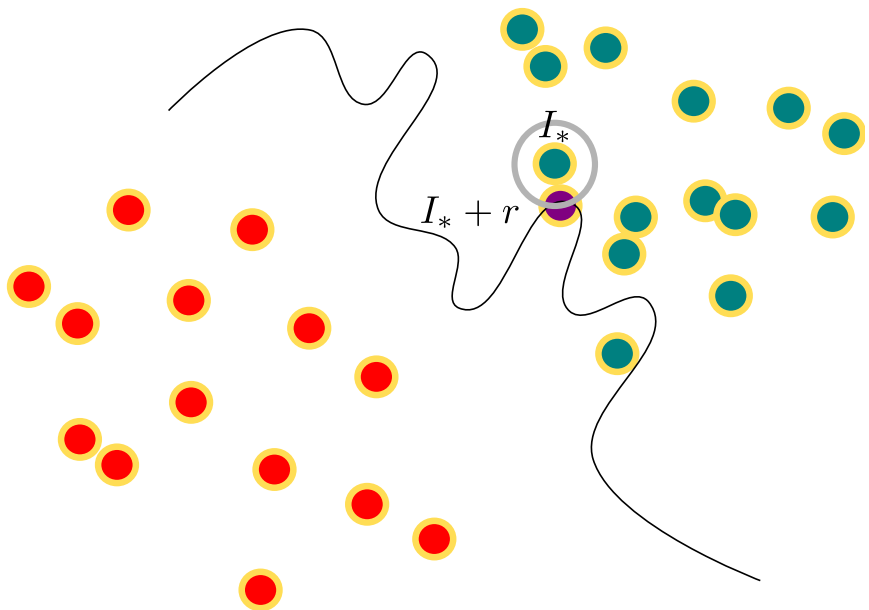
Examples, using cat labels vs dog labels

Adversarial attacks

- Given: Image $x \in \mathbb{R}^{H \times W}$, classifier $f : \mathbb{R}^{H \times W} \to \{0, \ldots, L\}$, and a target class label l
- An adversarial attack on the image x towards the class l: Find the image noise $r \in \mathbb{R}^{H \times W}$:

$$\text{minimize} \|r\|^2 \text{ while } \begin{array}{l} f(I_* + r) = l, \\ I_* + r \in [0, 1]^{H \times W}. \end{array}$$

- **Question:** How would you find such an example?



Szegedy et al, Intriguing properties of neural networks , https://arxiv.org/abs/1312.6199, 2014

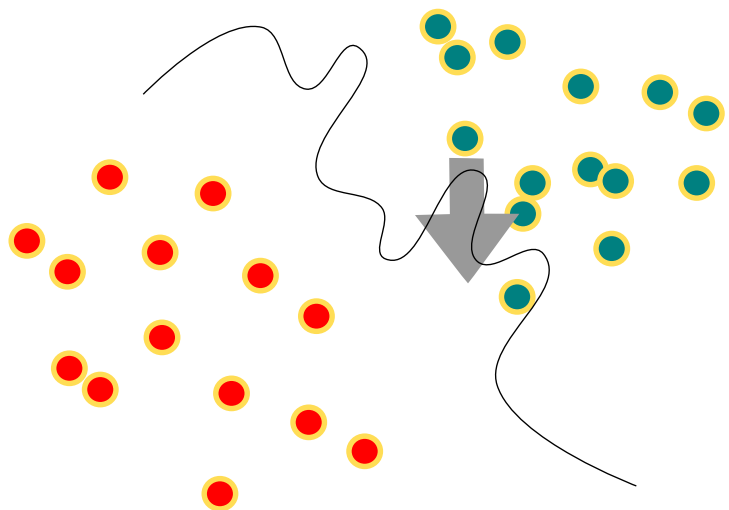
Adversarial attacks: Fast gradient sign method (FSGM), 2015

- The fast gradient sign method obtains adversarial examples via "noise"

$$r = \varepsilon \text{sign}(\nabla_I \mathcal{L}(\theta, I_*, y))$$

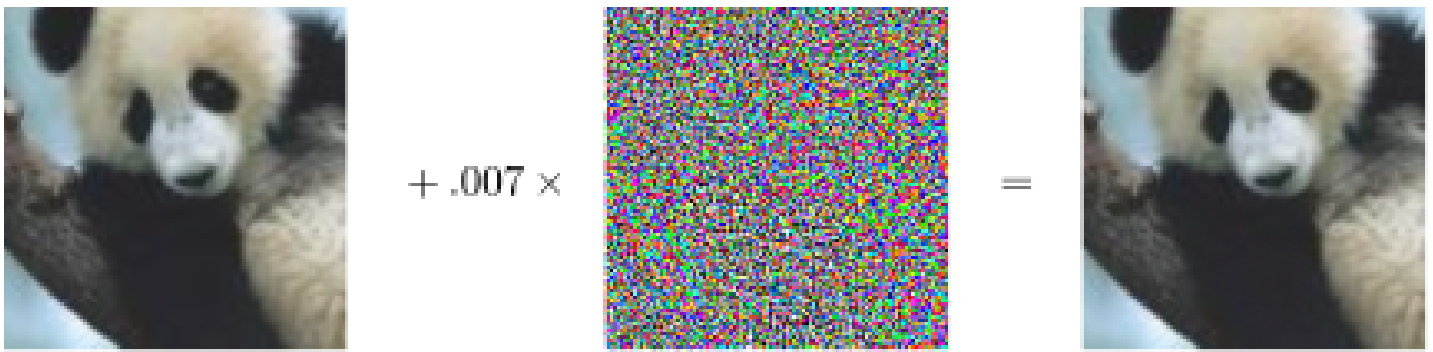
for loss $\mathcal{L}$, image I_*, correct label y, model parameters θ.

- In other words – move away from the correct class by maximizing its loss (linear approximation)



Goodfellow et al, Explaining and Harnessing Adversarial Examples, `https://arxiv.org/abs/1412.6572`

Adversarial attacks: Fast gradient sign method (FSGM), 2015



Goodfellow et al, Explaining and Harnessing Adversarial Examples, `https://arxiv.org/abs/1412.6572`

Adversarial attacks: Basic Iterative Method (BIM)

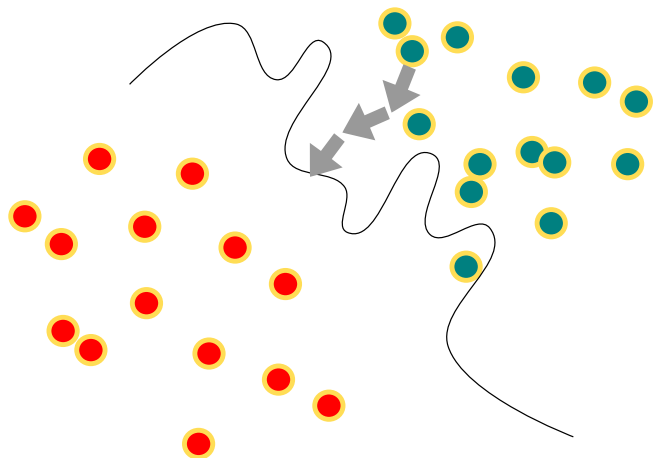
- Let go of the linear approximation by iterative attacking

$$I_0^{adv} = I_*,$$

and

$$I_{n+1}^{adv} = \mathrm{Clip}(I_n^{adv} + \alpha \mathrm{sign}(\nabla_I \mathcal{L}(\theta, I, y))),$$

where Clip cuts off image values that go outside $[0, 255]$, and α is a step (e.g. 1).



Kurakin et al, Adversarial Examples in the Phyiscal World, `https://arxiv.org/pdf/1607.02533.pdf`, 2017

Adversarial attacks: Basic Iterative Method (BIM)

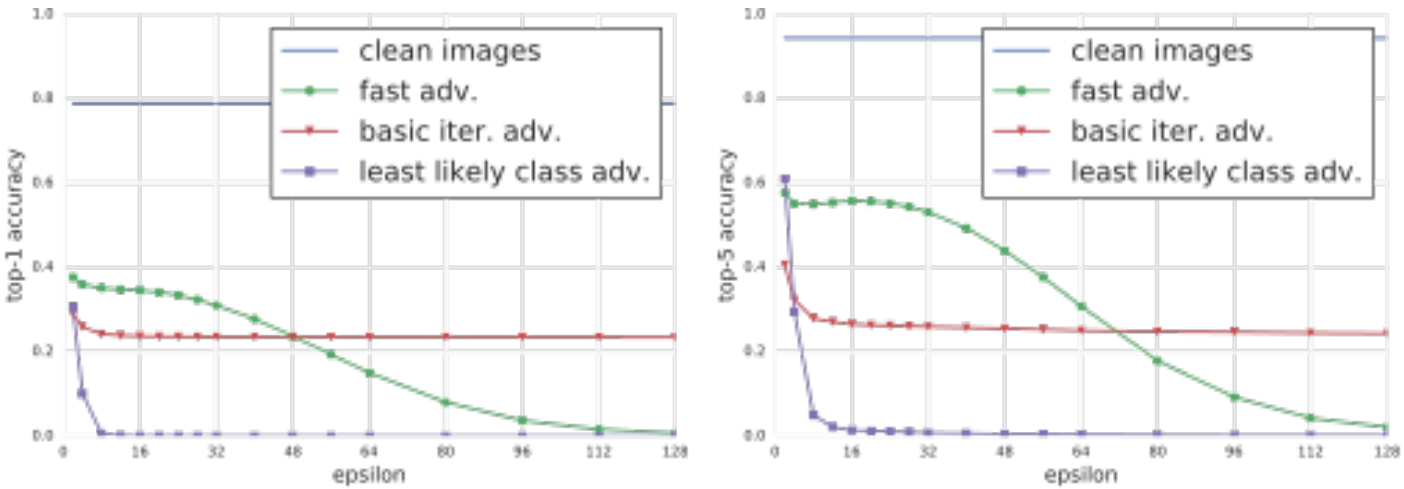


Figure 2: Top-1 and top-5 accuracy of Inception v3 under attack by different adversarial methods and different compared to "clean images" — unmodif ed images from the dataset. The accuracy was computed on all 50‹ 000 validation images from the ImageNet dataset. In these experiments varies from 2 to 128

Kurakin et al, Adversarial Examples in the Phyiscal World, `https://arxiv.org/pdf/1607.02533.pdf`, 2017

Saliency maps:
The # 1 hate object of explainable AI

https://www.thelancet.com/journals/landig/article/PIIS2589-7500(21)00208-9/fulltext#bib16

THE LANCET
Digital Health

VIEWPOINT | VOLUME 3, ISSUE 11, E745-E750, NOVEMBER 01, 2021

The false hope of current approaches to explainable artificial intelligence in health care

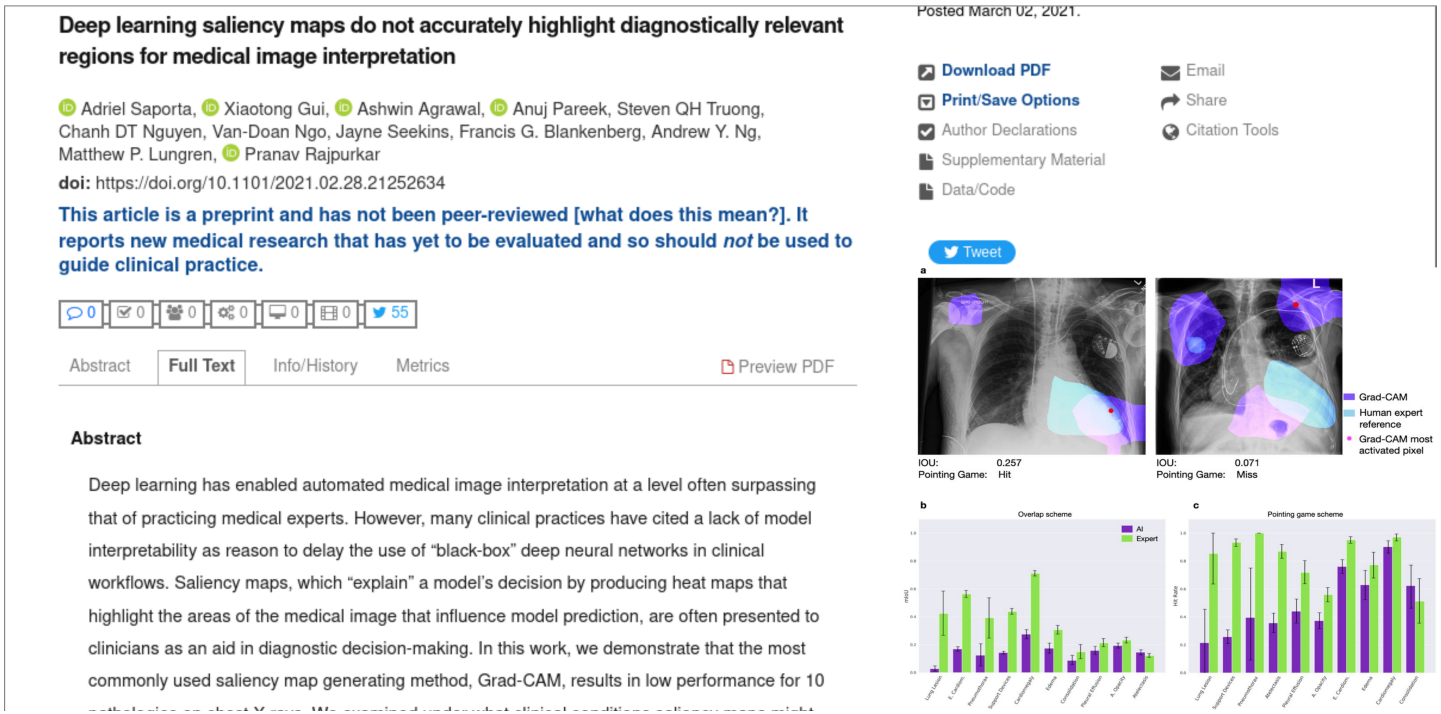
Marzyeh Ghassemi, PhD • Luke Oakden-Rayner • Andrew L Beam, PhD

Open Access • Published: November, 2021 • DOI: https://doi.org/10.1016/S2589-7500(21)00208-9 •
Check for updates

Summary

The black-box nature of current artificial intelligence (AI) has caused some to question whether AI must be explainable to be used in high-stakes scenarios such as medicine. It has been argued that explainable AI will engender trust with the health-care workforce, provide transparency into the AI decision making process, and potentially mitigate various kinds of bias. In this Viewpoint, we argue that this argument represents a false hope for explainable AI and that current explainability methods are unlikely to achieve these goals for patient-level decision support. We provide an overview of current explainability techniques and highlight how various failure cases can cause problems for decision making for individual patients. In the absence of suitable explainability methods, we advocate for rigorous internal and external validation of AI models as a more direct means of achieving the goals often associated

Saliency maps:
The # 1 hate object of explainable AI



Deep learning saliency maps do not accurately highlight diagnostically relevant regions for medical image interpretation

Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven QH Truong, Chanh DT Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G. Blankenberg, Andrew Y. Ng, Matthew P. Lungren, Pranav Rajpurkar

doi: https://doi.org/10.1101/2021.02.28.21252634

This article is a preprint and has not been peer-reviewed [what does this mean?]. It reports new medical research that has yet to be evaluated and so should *not* be used to guide clinical practice.

0 0 0 0 0 0 55

Abstract | Full Text | Info/History | Metrics | Preview PDF

Abstract

Deep learning has enabled automated medical image interpretation at a level often surpassing that of practicing medical experts. However, many clinical practices have cited a lack of model interpretability as reason to delay the use of "black-box" deep neural networks in clinical workflows. Saliency maps, which "explain" a model's decision by producing heat maps that highlight the areas of the medical image that influence model prediction, are often presented to clinicians as an aid in diagnostic decision-making. In this work, we demonstrate that the most commonly used saliency map generating method, Grad-CAM, results in low performance for 10

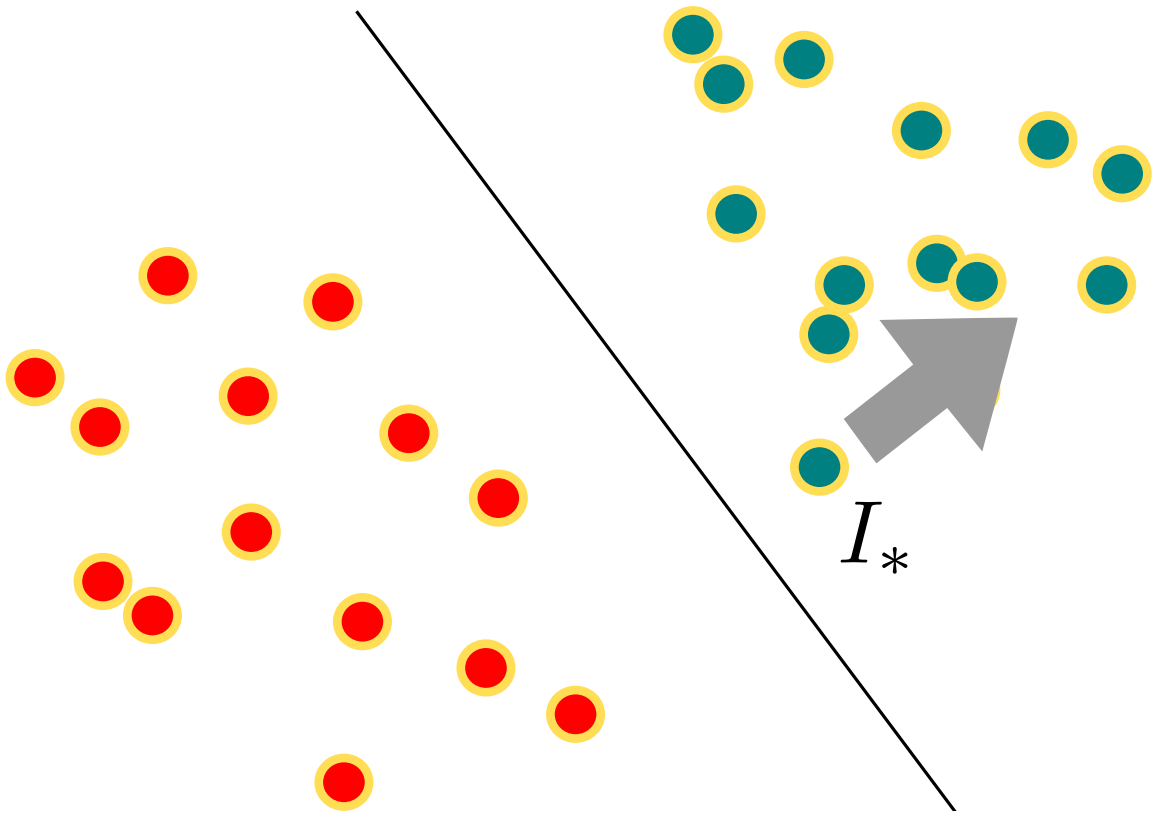
Posted March 02, 2021.

Download PDF | Email
Print/Save Options | Share
Author Declarations | Citation Tools
Supplementary Material
Data/Code

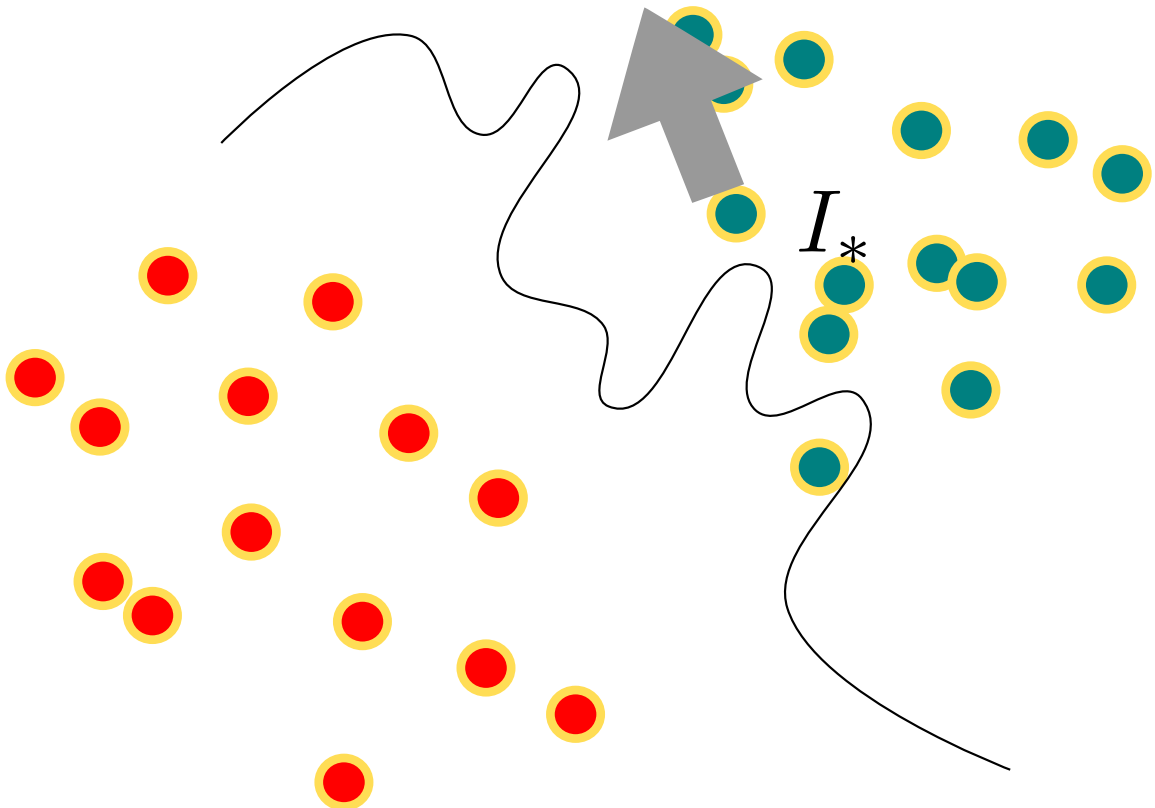
Tweet

Question: What is happening?

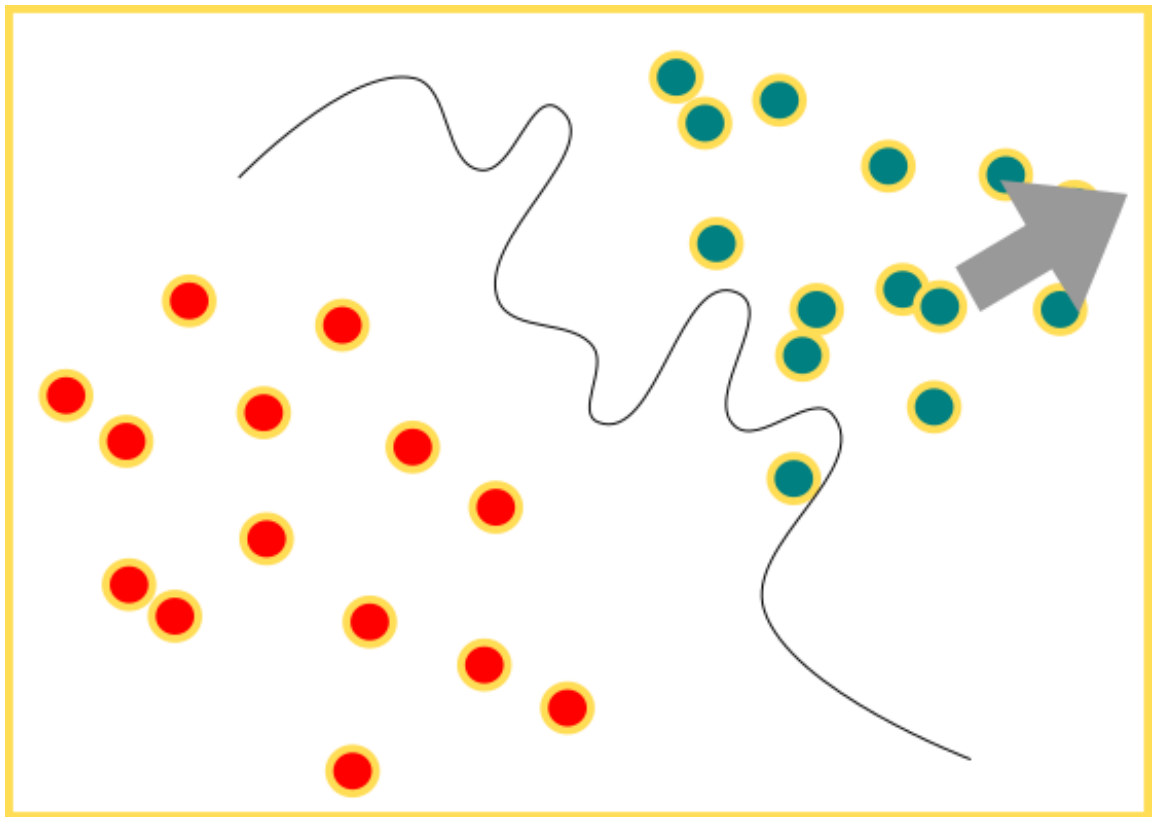
Saliency maps:
The # 1 hate object of explainable AI



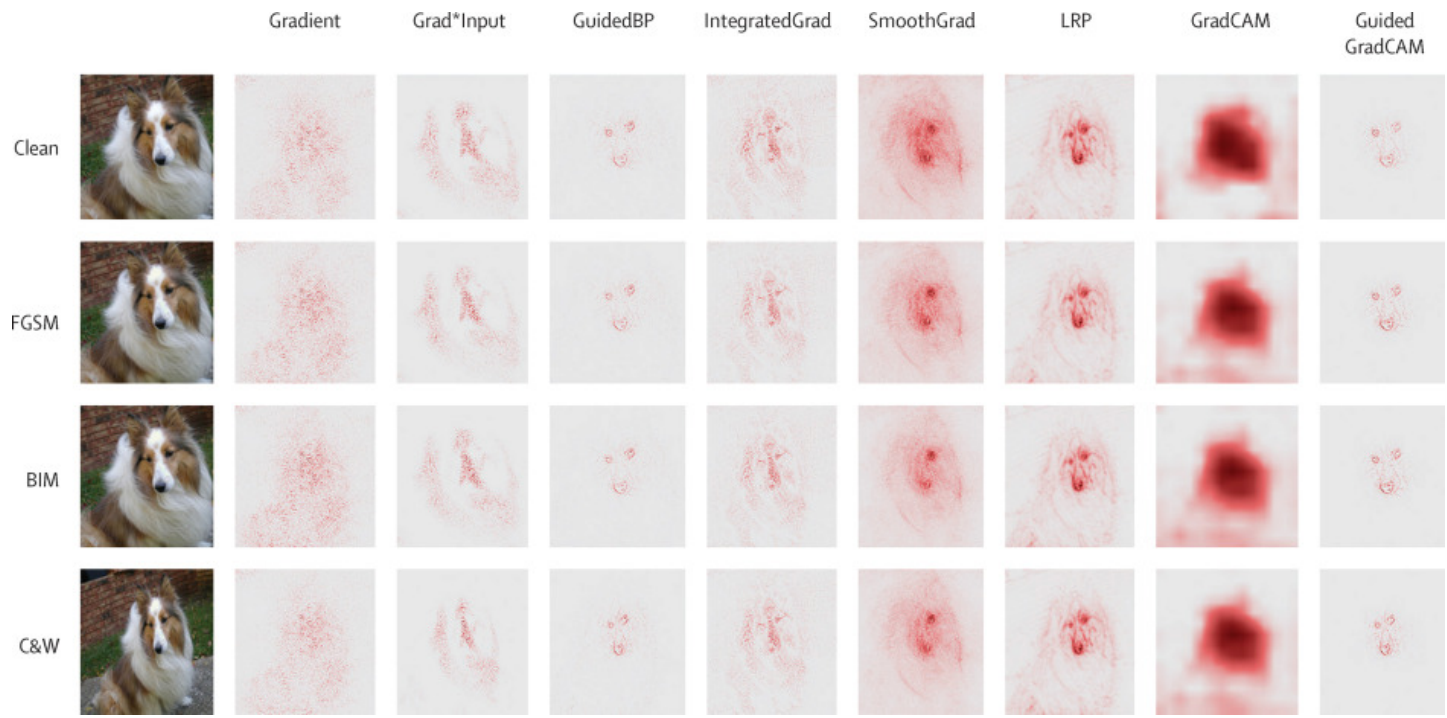
Saliency maps:
The # 1 hate object of explainable AI



Saliency maps:
The # 1 hate object of explainable AI

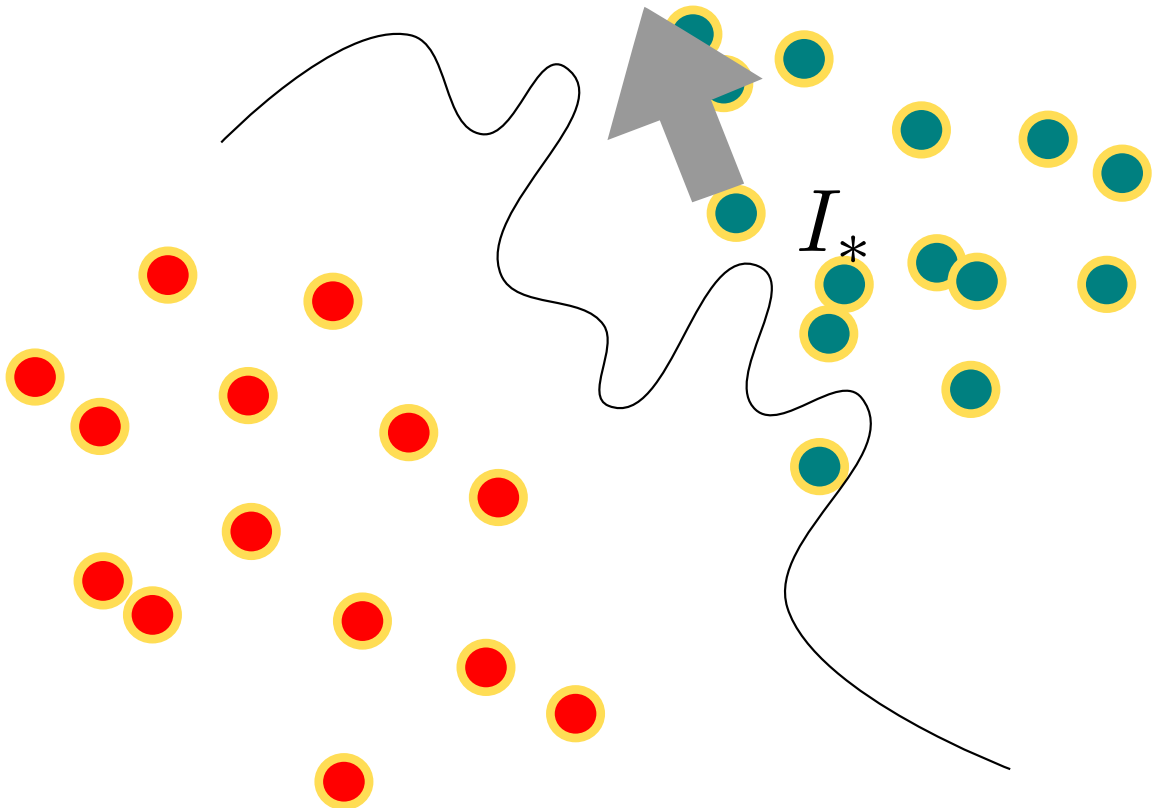


Saliency maps:
The # 1 hate object of explainable AI



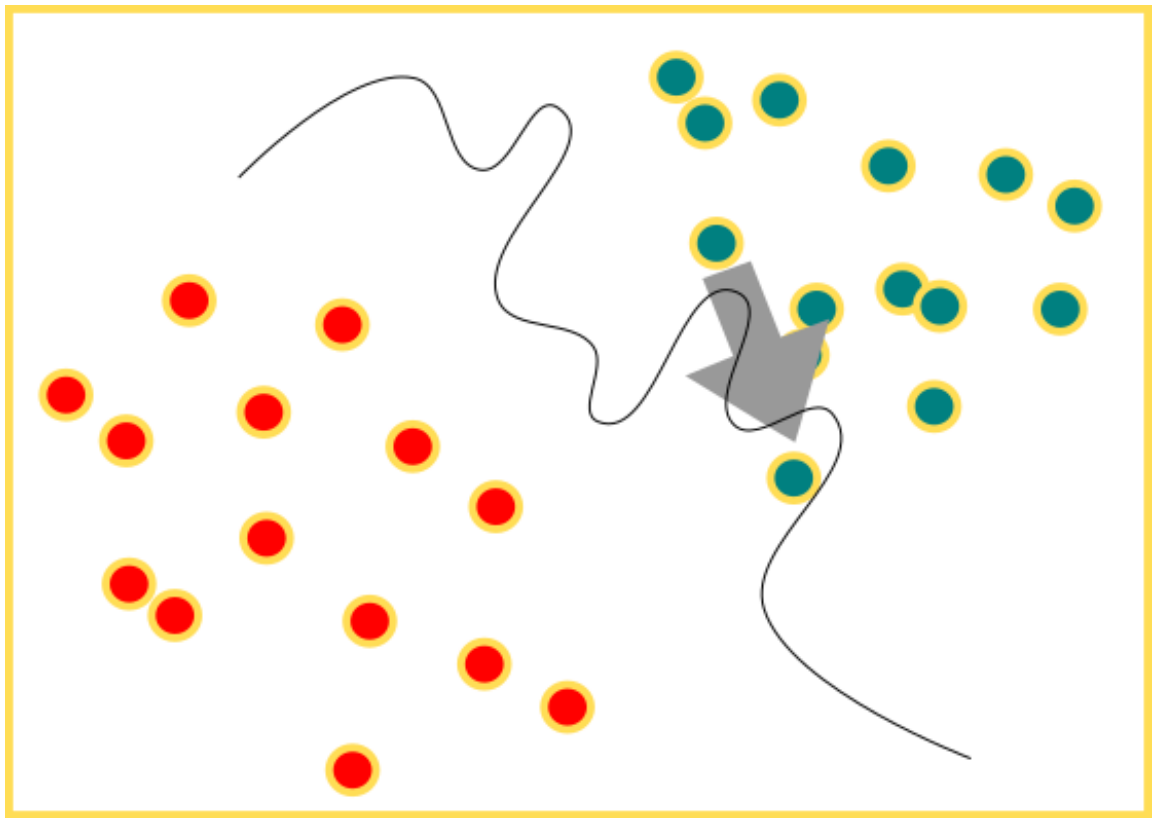
Although the prediction is changed using adversarial attacks, the saliency map looks unchanged!
Question: What is happening?

Saliency maps:
The # 1 hate object of explainable AI



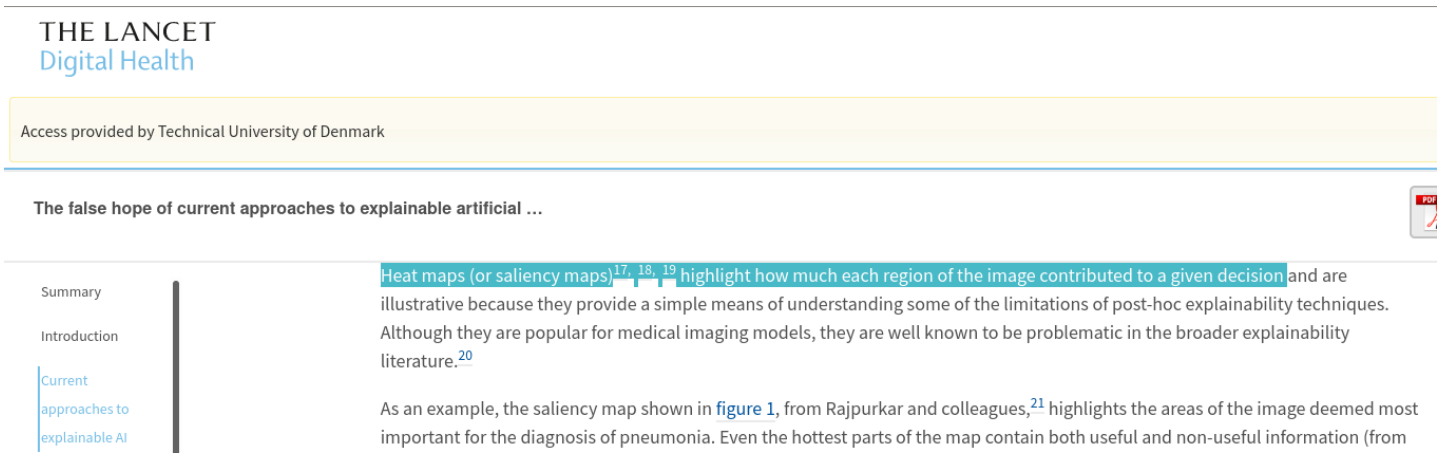
Saliency map...

Saliency maps:
The # 1 hate object of explainable AI



Adversarial attack... we are just seeing the effect of a smooth gradient.

Saliency maps:
The # 1 hate object of explainable AI



THE LANCET
Digital Health

Access provided by Technical University of Denmark

The false hope of current approaches to explainable artificial ...

Summary

Introduction

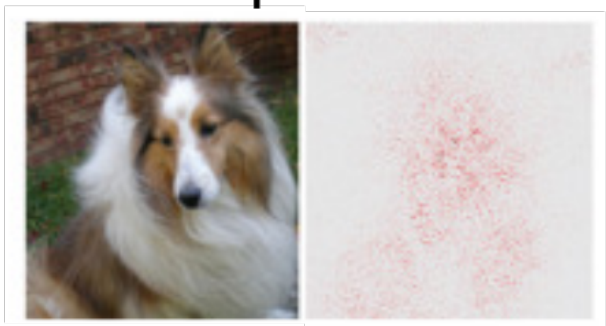
Current approaches to explainable AI

Heat maps (or saliency maps)[17, 18, 19] highlight how much each region of the image contributed to a given decision and are illustrative because they provide a simple means of understanding some of the limitations of post-hoc explainability techniques. Although they are popular for medical imaging models, they are well known to be problematic in the broader explainability literature.[20]

As an example, the saliency map shown in figure 1, from Rajpurkar and colleagues,[21] highlights the areas of the image deemed most important for the diagnosis of pneumonia. Even the hottest parts of the map contain both useful and non-useful information (from

Saliency maps:
The # 1 hate object of explainable AI

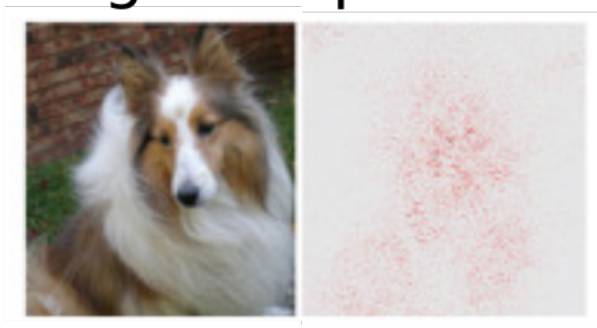
What contributed
to the prediction?



Saliency maps:
The # 1 hate object of explainable AI

~~What contributed to the prediction?~~

How do I efficiently change the prediction?



Summary

You should now be familiar with how gradients are used for:

- producing saliency maps for explainable AI, and
- performing adversarial attacks,
- and you should be aware of some of their un-intuitive properties and how to make proper use of them