FACULDADE ENGENHARIA
Departamento de Informática

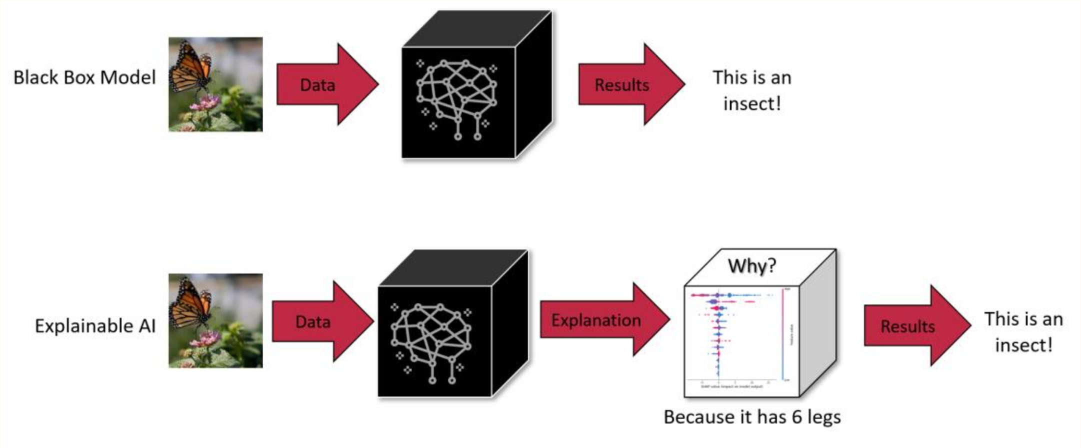
Integrated Gradients

Henrique Jesus M12359
Prof. Dr. Hugo Proença

Visão Computatcional

Introduction

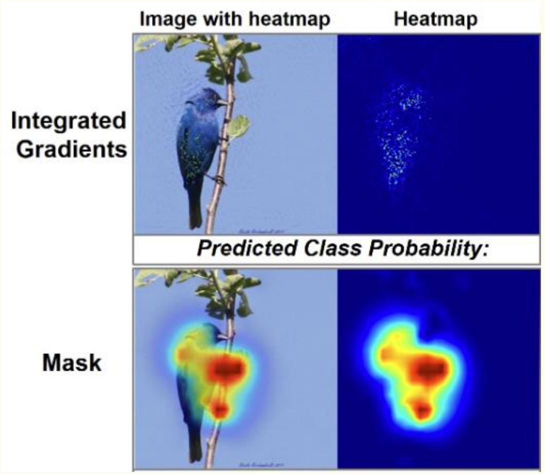
- Explainable AI (XAI)
- Black Box



Source: https://images.squarespace-cdn.com/content/v1/5bce4071ab1a620db382773e/1626294924656-L8ZFLDNLYK7C33DHX4I9/Picture1.png

How to make a model explainable?

- Paper "Axiomatic Attribution for Deep Networks"
- Integrated Gradients



Source: https://media.arxiv-vanity.com/render-output/7235504/IntroductionFinal2.jpg

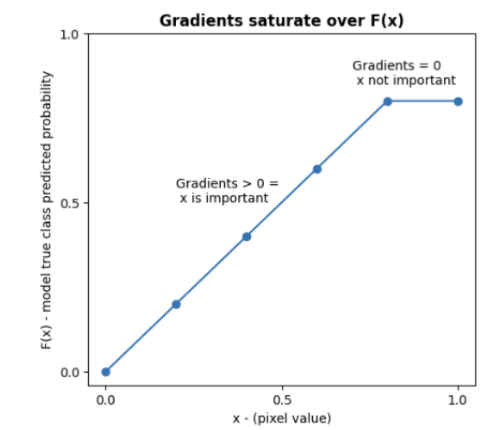
Gradients

- The direction of steepest increase of the function at that point.
- Determine the rate of change of a function in relation to its variables.

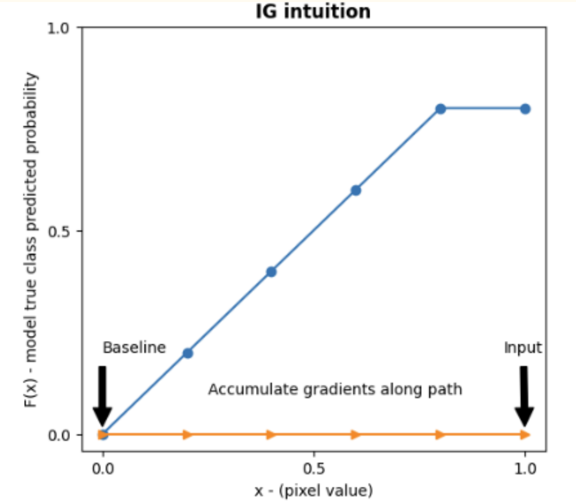
$$f(x, y) = x^2 + y^2$$

$$\frac{\partial f}{\partial x} = 2x \quad \frac{\partial f}{\partial y} = 2y$$

$$\mathrm{grad}(f(1,2)) = (2(1), 2(2)) = (2, 4)$$



Integrated Gradients



$$IntegratedGradients_i(x) ::= (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha$$

where:

$_i$ = feature (individual pixel)
x = input (image tensor)
x' = baseline (image tensor)
α = interpolation constant

$$IntegratedGrads_i^{approx}(x) ::= (x_i - x'_i) \times \sum_{k=1}^{m} \frac{\partial F(x' + \frac{k}{m} \times (x - x'))}{\partial x_i} \times \frac{1}{m}$$

where:

k = interpolation constant step
m = number of steps in the Riemann sum approximation of the integral

$$IntegratedGrads_i^{approx}(x) ::= \overbrace{(x_i - x'_i)}^{5.} \times \overbrace{\sum_{k=1}^{m}}^{4.} \overbrace{\frac{\overbrace{\partial F(\overbrace{x' + \tfrac{k}{m}}^{1.} \times (x - x'))}^{2.}}{\partial x_i}}^{3.} \times \overbrace{\frac{1}{m}}^{4.}$$

1. Generate alphas α
2. Generate interpolated images = $(x' + \frac{k}{m} \times (x - x'))$
3. Compute gradients between model F output predictions with respect to input features = $\frac{\partial F(\text{interpolated path inputs})}{\partial x_i}$
4. Integral approximation through averaging gradients = $\sum_{k=1}^{m} \text{gradients} \times \frac{1}{m}$
5. Scale integrated gradients with respect to original image = $(x_i - x'_i) \times \text{integrated gradients}.$