University of Glasgow | School of Computing Science

THE AWARDS 2020 | UNIVERSITY OF THE YEAR

Integrated Gradients

Dr. Fani Deligianni,
fani.deligianni@glasgow.ac.uk
Lecturer (Assistant Professor)
Lead of the Computing Technologies for Healthcare Theme
https://www.gla.ac.uk/schools/computing/staff/fanideligianni

WORLD CHANGING GLASGOW

Attributions’ Desirable Properties

- Sensitivity
- Implementation invariance
- Completeness
- Linearity
- Symmetry preserving

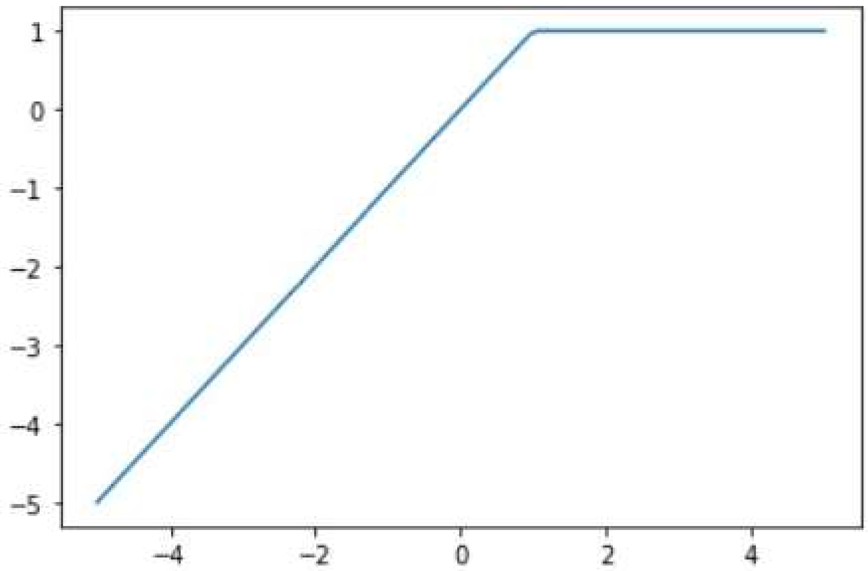
Sensitivity

- If the prediction does not depend on an input feature, then the attribution to that feature should be always zero
- If the prediction of two inputs that differ in one feature is different then a non-zero attribution should be assigned to the feature

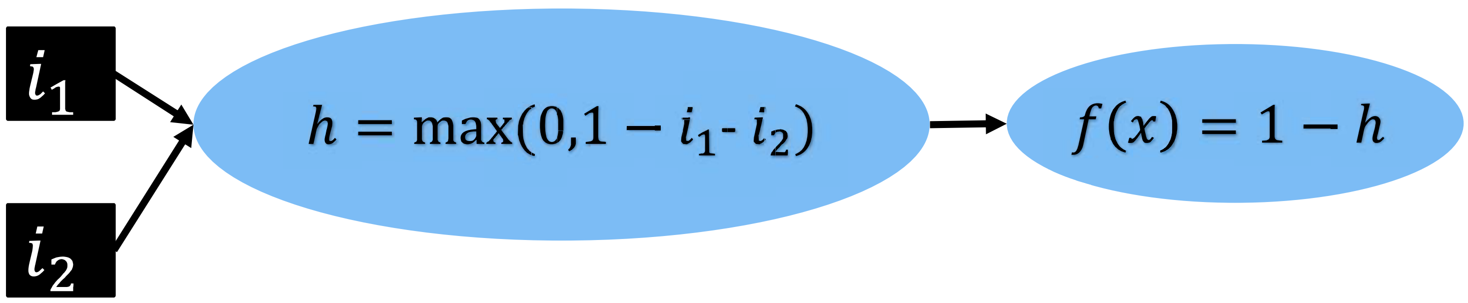
Gradient Methods - Sensitivity

- Gradients violate the sensitivity axiom
- Predictive functions may result to zero despite input values are far from baseline

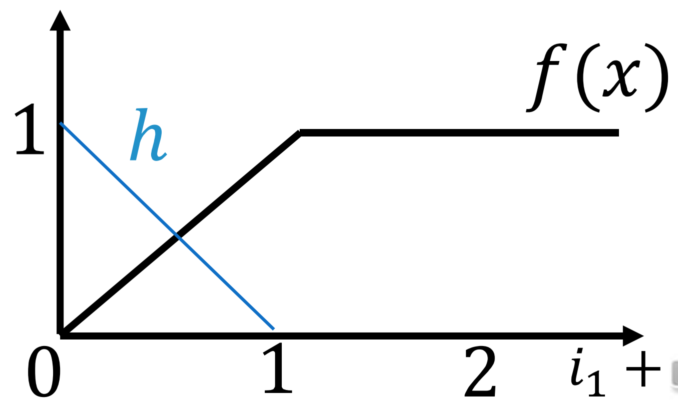
$$f(x) = 1 - ReLU(1 - x)$$



Model Saturation



- Perturbation methods
- Gradient methods



Implementation Invariance

- If the output of two models is always identical, regardless their implementation, their attributions should be always identical
- Gradient methods satisfy this property

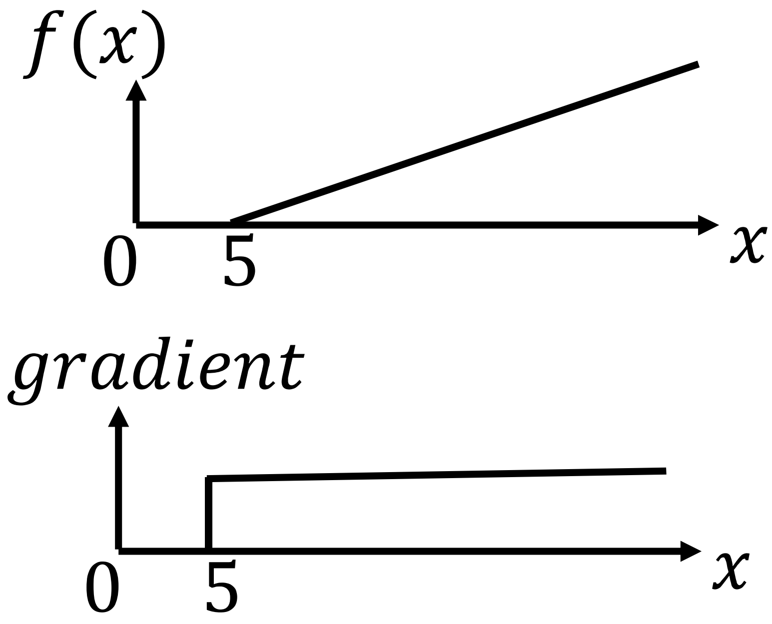
Completeness

- The feature attributions sum to the output for a given sample

Gradient Methods - Completeness

- Gradients violate the completeness axiom
- Importance score may vary significantly over small changes in input

$$f(x) = \max(0, x - 5)$$



Linearity

- For a model that is a linear combination of two submodels:

$$f(x) = af_1(x) + bf_2(x)$$

  the attributions are a linear combination of the submodels' attributions:

$$\varphi(x) = a\varphi_1(x) + b\varphi_2(x)$$

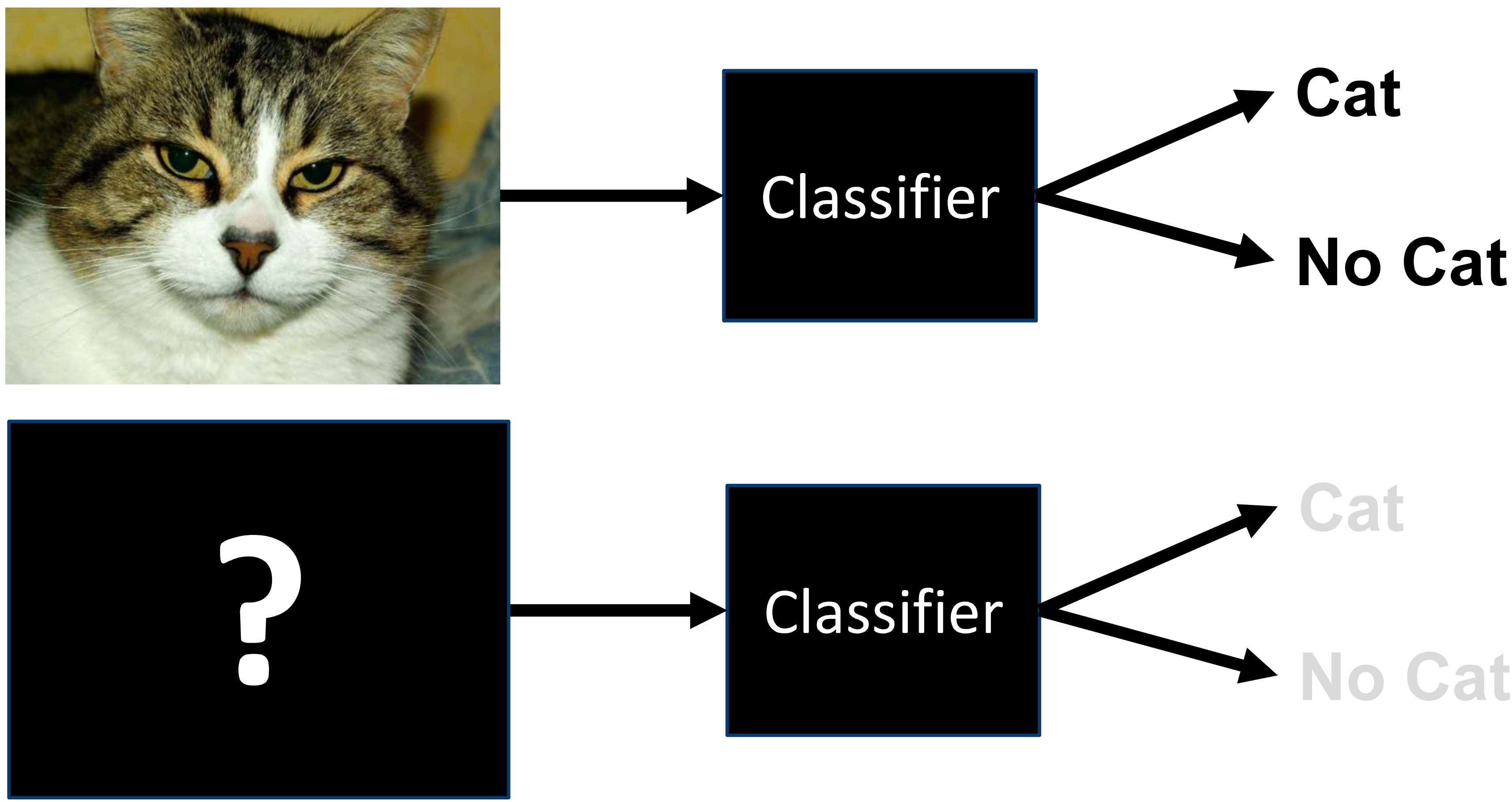
Symmetry Preserving

- Symmetric variables with identical values should achieve identical attributions

Gradient Methods

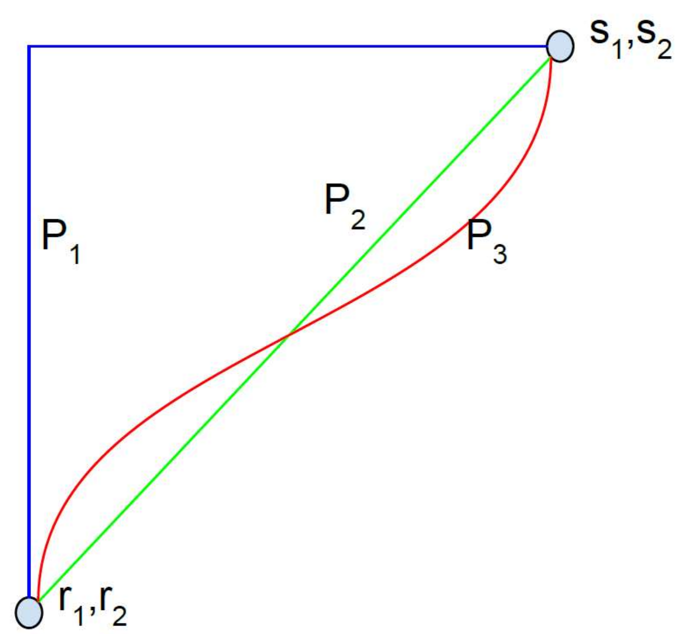
- Gradient methods are implementation invariant ✓
- Gradient methods satisfy linearity condition ✓
- Gradient methods could violate sensitivity ✗
- Gradient methods could violate completeness ✗

Baseline Explanations



Integrated Gradients

- Consider the straight-line path between baseline and input
- Integrate the gradients along this path



Input: (s_1, s_2)
Baseline: (r_1, r_2)

Sundararajan et al. ‘Axiomatic Attribution for Deep Networks’, ICML, 2017

Integrated Gradients (IG)

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

- where:
  - ***i*** = feature (individual pixel)
  - ***x*** = input (image tensor)
  - ***x'*** = baseline (image tensor)
  - ***k*** = scaled feature perturbation constant
  - ***m*** = number of steps in the Riemann sum approximation of the integral
  - $(x_i - x'_i)$ = a term for the difference from the baseline.

Integrated Gradients - Theoretical Properties

- Integrated Gradient satisfy implementation invariant ✓
- Integrated Gradient methods satisfy linearity ✓
- Integrated Gradient methods satisfy sensitivity ✓
- Integrated Gradient methods satisfy completeness ✓
- Integrated Gradient methods satisfy symmetry ✓

Summary

- Conditions/Axioms are needed to define what is a ‘good’ explanation
- Axioms such as sensitivity and implementation invariance have been proposed as some of the properties that need to be fulfilled
- Integrated Gradients (IG) satisfy both the axioms of sensitivity and implementation invariance
- IG is considered a path methods that exploit gradients between the baseline and an input value to provide local explanations of a model’s decision
- IG has become a popular interpretability technique due to its broad applicability to any differentiable model, ease of implementation, theoretical justifications, and computational efficiency

References

- Sundararajan et al. ‘Axiomatic Attribution for Deep Networks’, ICML, 2017.
- Erion et al. ‘Improving performance of deep learning models with axiomatic attribution priors and expected gradients’, Nature Machine Intelligence, 2021.