

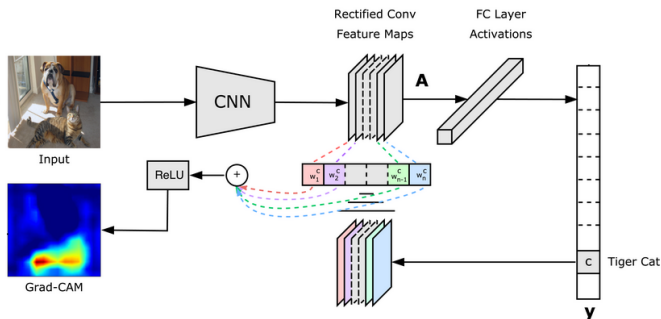
# On Evaluating Explanation Methods for Deep Networks

Miguel Lerma and Mirtha Lucas

October 29, 2021

# Problem Statement - Example

Our network predicts that the image contains a Tiger Cat.  
Our explanation method highlights the location of the Tiger Cat in the image.



Problem: How do we evaluate the performance of our explanation method?

# Problem Statement I

Quoting Samek et al (2015) [2]:

- ▶ Deep Neural Networks (DNNs) have demonstrated impressive performance in complex machine learning tasks such as image classification or speech recognition.
- ▶ However, due to their multi-layer nonlinear structure, they are not transparent, i.e., it is hard to grasp what makes them arrive at a particular classification or recognition decision given a new unseen data sample.
- ▶ Several approaches have been proposed enabling one to understand and interpret the reasoning embodied in a DNN for a single test image.
- ▶ While the usefulness of heatmaps can be judged subjectively by a human, **an objective quality measure is missing.**

## Problem Statement II

- ▶ How to provide objective quality measures for explanation methods of deep networks.
- ▶ In particular, how to evaluate the performance of explanation methods in deep networks.

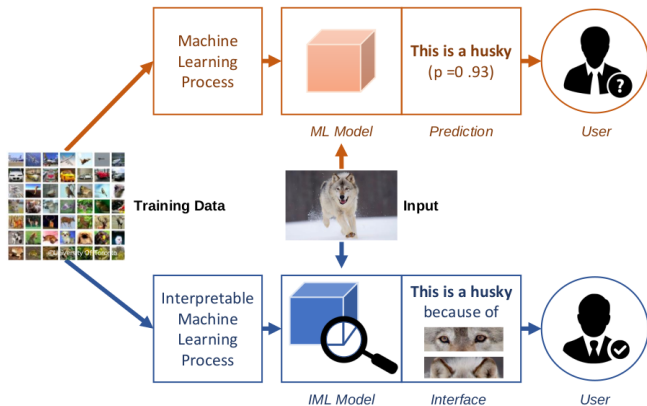
Warning!

We are not going to provide a final solution to the problem!

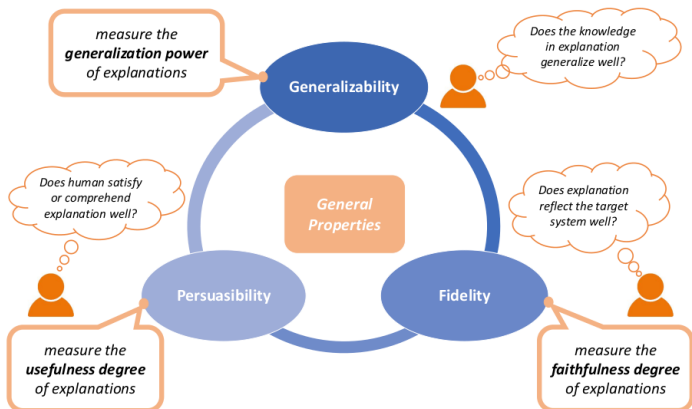
Our goal is only to discuss its difficulty and possible approaches.

# A General Framework for Evaluating Explanations

From Yang et al (2019) [3]

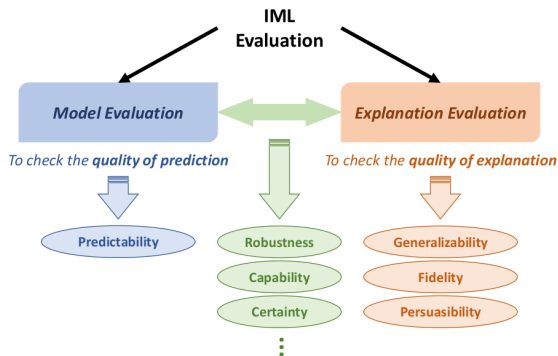


# Three general properties for explanations in IML



# IML Evaluation

IML evaluation can be divided into model evaluation and explanation evaluation.



How good an explanation is if the model predictions are wrong? - Well, it may help to determine why they are wrong and how to improve the model (maybe?).



# What is Evaluation?

- ▶ What are we evaluating?
- ▶ How do we evaluate it?
- ▶ How good our evaluation method is?

Examples of well established evaluation metrics and practices:

- ▶ For a new classifier: accuracy, sensitivity, specificity, etc.
- ▶ For a new scientific hypothesis: empirically tested with experiments.
- ▶ For a new algorithm: theoretical approach (correctness, complexity), empirical approach (test it in practice and apply appropriate evaluation metrics).
- ▶ For a new mathematical theorem: formal proof... Wait! Is Mathematics a science? For that matter, is Computer Science a science?

## Some Context

- ▶ Technology/Development: Apply present knowledge.
- ▶ Science/Research: Add new knowledge.
- ▶ Empirical/Natural Sciences: Based on observation, hypotheses, experiments... E.g.: Physics, Chemistry, Biology..., and Computer Science.
- ▶ Formal Sciences: Based on logic and formal systems. E.g.: Mathematics, Statistics..., and Computer Science!

Note: Computer Science is both empirical and formal science. Computers are physical objects, computer systems can be benchmarked and tested empirically, but “theoretical” Computer Science was born as a branch of Logic - Turing (Computability Theory), Church (Lambda Calculus), Gödel (Recursive Functions)...

# Evaluating Explanation Methods in Machine Learning

Back to our main subject.

- ▶ What are we evaluating?: explanation methods intended to explain predictions generated by machine learning models.
- ▶ How do we evaluate it?: Two approaches.
  - ▶ Formal (axiomatic methods): Using logic and mathematical reasoning to prove that the model satisfies a number of desirable properties.
  - ▶ Empirical: Testing and using appropriate metrics.
- ▶ How good the evaluation method is?:
  - ▶ Formal approach: The mathematical proofs must be correct. This can be checked by direct inspection.
  - ▶ Empirical approach: TBD (no single widely used way of determining the quality of an empirical evaluation of a explanation method).

# Formal Approach I

Paradigmatic example : Integrated Gradients (IG). Quoting Sundararajan et al 2017 [5]:

*[...] we found that every empirical evaluation technique we could think of could not differentiate between artifacts that stem from perturbing the data, a misbehaving model, and a misbehaving attribution method. This was why we turned to an axiomatic approach in designing a good attribution method.*

The authors study the problem of attributing the prediction of a deep network to its input features. They identify two fundamental axioms, *Sensitivity* and *Implementation Invariance*, that attribution methods ought to satisfy.

## Formal Approach II

In IG, a deep network is represented as a function  $F : \mathbb{R}^n \rightarrow [0, 1]$  of its inputs  $x_1, \dots, x_n$ . An attribution of the prediction at input  $x = (x_1, \dots, x_n)$  relative to a baseline input  $x' = (x'_1, \dots, x'_n)$  is a vector  $A_F(x, x') = (a_1, \dots, a_n) \in \mathbb{R}^n$ , where  $a_i$  is the contribution of  $x_i$  to the prediction  $F(x)$ .

The properties verified by IG are the following:

- ▶ Sensitivity: An attribution method satisfies Sensitivity if for every input and baseline that differ in one feature but have different predictions then the differing feature should be given a non-zero attribution.
- ▶ Implementation Invariance: Two networks are functionally equivalent if their outputs are equal for all inputs, despite having very different implementations. Attribution methods should satisfy Implementation Invariance, i.e., the attributions are always identical for two functionally equivalent networks.

## Formal Approach III

The authors then proceed to prove that IG in fact verifies the required properties.

They also claim other properties about IG, in particular

*Completeness*:<sup>1</sup>

*Integrated gradients satisfy an axiom called completeness that the attributions add up to the difference between the output of  $F$  at the input  $x$  and the baseline  $x'$ .*

This basically means  $a_1 + \dots + a_n = F(x) - F(x')$ .

The proofs are mathematical in nature; there are examples at the end of the paper, but no actual metrics are provided.

---

<sup>1</sup>In fact their statement that IG verifies completeness is technically incorrect for reasons that go beyond the scope of this presentation (see Lerma & Lucas 2021 [8]).

# Empirical vs Formal Approaches

Although objections against empirical methods may make sense, formal approaches have also drawbacks, and empirical methods may be preferred for various reasons:

- ▶ The axiom set used in a formal approach may be incomplete.
- ▶ Extra steps must be taken to ensure that implementations correctly instantiate the theoretical model, and the implementation must be empirically tested anyway.
- ▶ For a data-driven science, a quantitative evaluation is often aligned with the goals.

On the other hand, empirical methods must overcome several difficulties, in particular:

- ▶ How to either obtain a ground truth, or develop an evaluation metrics without using a ground truth.
- ▶ How to determine the quality of the evaluation method. In other words, how is the evaluation method evaluated?

# Overview of Empirical Approaches

Without being exhaustive we can mention the following approaches for empirical evaluation methods:

- ▶ Evaluation Methods with ground truth.
  - ▶ Annotations: Inputs with annotations based on ground truth are provided (e.g. bounding boxes enclosing the location of the regions of interest).
- ▶ Evaluation Methods without ground truth.
  - ▶ Occlusion: Remove the areas either highlighted or not highlighted by the explanation method and determine the impact on the scores.
  - ▶ Perturbation: Blur the areas highlighted by the explanation method and determine the impact on the scores.



## Empirical Methods with Ground Truth - Examples

Here we examine two examples of evaluations with ground truth (given by a bounding box).

- ▶ Pixel Energy, defined as  $\frac{\sum L^c_{(i,j) \in bbox}}{\sum L^c_{(i,j) \in bbox} + \sum L^c_{(i,j) \notin bbox}}$ , i.e., the sum of pixel intensities in the part of the heatmap inside the bounding box divided by the total sum of intensities of the heatmap for the entire image (see energy-based pointing game in sec. 4.3 of [7]).
- ▶ Jaccard-based measures: for a given threshold find the region  $R$  occupied by the pixels of the heatmap whose intensities are above the threshold. Then determine how much the region overlaps with the bounding box  $B$  using Intersection over Union  $IoU = \frac{|R \cap B|}{|R \cup B|}$ , Intersection over Bounding Box  $IoB = \frac{|R \cap B|}{|B|}$ , and Intersection over Region  $IoR = \frac{|R \cap B|}{|R|}$ .

## Empirical Methods with Ground Truth - Drawbacks

- ▶ Obtaining the ground truth may be time/resource consuming and unreliable. A technique is *outsourcing*, e.g. the Amazon Mechanical Turk [4], which requires a large number of people to examine the original images and produce annotations.
- ▶ A bounding box may not fit well the region of interest (ROI) - e.g. using a rectangular box for ROIs that are not rectangular.
- ▶ A bounding box may omit relevant information - e.g. images of a polar bear may contain a white background (icy terrain), which may have an impact in the way the network decides what type of animal is in the image, but is not contained within the bounding box.
- ▶ Low metrics may be an indication of bad prediction power by the network rather than bad quality of the attribution method.

# Empirical Methods without Ground Truth - Explanation Maps

Explanation Map: Given an image  $I$  and a heatmap  $L^c$  generated by the network for this image for a class  $c$ , we find an *explanation map*  $E^c = L^c \odot I$ , where  $\odot$  represents the Hadamard (element-wise) product of  $L^c$  and  $I$ .

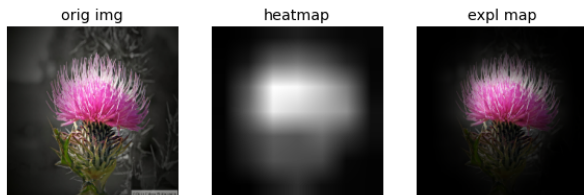


Figure: Original image, heatmap, and explanation map.

This is an example of occlusion method since the explanation map obscures part of the image (the background).

# Empirical Methods without Ground Truth - Explanation Map Metrics

Feeding the network with an image  $I$  we obtain an output  $Y^c =$  predicted probability of class  $c$ . If we feed the network with the explanation map  $E^c$  we will obtain an output  $O^c$ . Then, the following metrics are defined:

▶ Percentage Average Drop =  $\frac{1}{N} \sum_{i=1}^N \frac{\max(0, Y_i^{c_i} - O_i^{c_i})}{Y_i^{c_i}} \times 100$ .

It measures how much the probability predicted by the network decreases when the original image is replaced by the explanation map (lower is better).

▶ Increase in Confidence =  $\sum_{i=1}^N \frac{\mathbb{1}(Y_i^{c_i} < O_i^{c_i})}{N}$ , where  $\mathbb{1}$  is the indicator function with value 1 if the argument is true, and 0 if it is false. It yields the proportion of images for which the explanation map produces a network output larger than the original image (higher is better).

# Empirical Methods without Ground Truth - Drawbacks




- ▶ Without ground truth it is impossible to determine whether the explanation method is faithful, i.e., if the explanation provided really reflects the target system.
- ▶ As a consequence evaluation methods without ground truth do not provide feedback about the quality of the network—e.g. if a network classifying images of dogs and cats misidentifies dogs as cats, in an image containing a dog and a cat, a "good" explanation method will highlight the cat for the output "dog". Without ground truth we cannot tell whether the area highlighted is right or wrong.

# Conclusions




Here are the general conclusions regarding the problem of providing objective evaluation methods for explanations methods of deep networks.

- ▶ Several methods have been proposed, and a few of them have been summarized here.
- ▶ All of them have advantages and drawbacks. There is no single method being consistently used by a majority of researchers (at least judging by the literature we have consulted).

# References I



-  Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. arXiv preprint arXiv:1610.02391 [cs.CV]
-  Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Bach, Klaus-Robert Müller (2015): Evaluating the visualization of what a Deep Neural Network has learned. arXiv preprint arXiv:1509.06321 [cs.CV]
-  Fan Yang, Mengnan Du, Xia Hu (2019). Evaluating Explanation Without Ground Truth in Interpretable Machine Learning. arXiv preprint arXiv:1907.06831 [cs.LG]

## References II

-  A. Sorokin and D. Forsyth (2008). Utility data annotation with Amazon Mechanical Turk. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2008, pp. 1–8, doi: 10.1109/CVPRW.2008.4562953
-  Mukund Sundararajan, Ankur Taly, Qiqi Yan (2017). Axiomatic Attribution for Deep Networks. arXiv preprint arXiv:1703.01365 [cs.LG]
-  Sam Sattarzadeh, Mahesh Sudhakar, Konstantinos N. Plataniotis, Jongseong Jang, Yeonjeong Jeong, Hyunwoo Kim (2021). Integrated Grad-CAM: Sensitivity-Aware Visual Explanation of Deep Convolutional Networks via Integrated Gradient-Based Scoring. arXiv preprint arXiv:2102.07805 [cs.CV]



## References III

-  Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, Xia Hu (2020). Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, p. 24–25, 2020.
-  Miguel Lerma and Mirtha Lucas (2021). Symmetry-Preserving Paths in Integrated Gradients. arXiv preprint arXiv:2103.13533 [cs.LG]