Master Thesis Proposal

Tim Patzelt

October 2021

Motivation

- Increasingly complex models for which XAI can help:
  - building trust in them
  - discovering spurious correlations/bias/Clever Hans Effect
  - debugging
- Potential paradigm shift from:
  - "Method A is better than Method B" to "Method A is better for that Architecture/Problem Domain than Method B"
  - "New Attribution Methods add value to XAI" to "only new kind of explanation methods add value to XAI"
- Provide examples

Survey of Background Literature

Overview

- What is a good Explanation?
  - Qualitative evaluations of explanations are expensive to obtain and not necessarily lead to a general understanding of explanations
  - Quantitative evaluations exist. But since there is no general agreement what a "good" explanation is, the choice of an appropriate metric can be hard, even for professionals.
- There exists a conceptual framework for quantitative evaluation of attribution methods [1]:
  - It is based on the logical concepts of necessity, sufficiency and proportionality
  - Explanation goes beyond "importance" of features
  - It introduces the metrics *Total Proportionality for Necessity (TPN)* and *Total Proportionality for Sufficiency (TPS)*
- The solution space of attribution methods is one real-valued number for each input feature. Although it is easy to describe, it has an uncountably infinite number of elements.
- Assuming that the notion of proportionality for necessity/sufficiency is the desired property of attribution values, each attribution method can be seen as a heuristic to find a good value in the solution space.
- TPN and TPS are functions formulating a mathematical optimization problem. It can be solved by applying well-established optimization techniques to find values maximizing TPN or TPS. Since both TPN an TPS are mutually exclusive, either one or the other can be maximized by a solution.

- The optimization method could replace dedicated attribution methods leading to two possible implications:
  1. All present attribution methods are rendered obsolate since they are only heuristics and we can calculate the exact solution
  2. The benefit of different attribution methods does not lie in the values it produces, but in the method itself (elaborate more)

I think we need to phrase this a bit differently. The attribution methods are not obsolete, they just become one possible heuristic that we expect to find worse solutions for a given metric, such as sufficiency, than explicitly optimizing the metric as an objective.

So the issue is that if the literature is right with the assumption that explanations can be evaluated by using sufficiency (etc.), then specific attribution heuristics do not provide any benefit over plain optimization methods.

From this, the next argument follows (2.) that there might be more to an explanation method than just it's optimal value, i.e., the method that creates the attribution value itself contributes to the understanding.

In this thesis, we want to find out if this is indeed the case, If our empirical results confirm our hypothesis, i.e., we find better solutions to the sufficiency problem than existing attribution methods, then either we can stop the search for new attribution methods or we specify the intrinsic value of attribution methods to the understanding of a machine learning model. -- And a discussion on these two options may be part of the thesis.

Make sure to frame these hypotheses and the problem you investigate well. Make an example from the text domain and illustrate the implication if the

---

Theoretical Background

- Explain Deep Neural Nets, their Complexity, what task they handle and in which domain
- Outline Purpose and Applications of Explainable Artificial Intelligence (XAI)
  - Use-Cases
  - People using it (researcher vs management/decision maker)
- Briefly sketch evolution of XAI methods (perturbation-based vs gradient-based, ad hoc vs post hoc)
- Solution space is well defined for attribution methods:
  - An attribution method A computes a set of attribution scores $s_1, s_2, \ldots, s_n$ for each $x_1, x_2, \ldots, x_n$
  - Constraints can apply, e.g. Proportionality [1] enforcing attribution scores being proportional to the output change.
- Several papers define quantitive evaluation metrics related to sufficiency and necessary as defined in [1]:
  - Proportionality-k for Necessity [1]
  - sensitivity-n [2]
  - Area Under The Curve (Most Relevant Features first or Least Relevant features first) [3]
  - sensitivity (a)(b) [4]
  - Area Over Perturbation Curve [3]
- There are different optimization techniques for maximizing or minimizing a real-valued function like TPN and TPS:
  - iterative optimization methods like Gradient Descent
  - Evolutionary/Genetic Algorithms or Particle Swarm Optimization
- Outline characterics why the metrics are suitable for optimization
- Cover different initialization options for input values
- Explain Performance/Memory Profiling (Landau Notation, Space Usage)

Proposed Methodology

Outline the experimental and theoretical methods specific to the proposed research. Justify the choice of methods as opposed to alternatives. Avoid giving too much detailed information; just outline the general approach and why you chose to use them

- TPN and TPS as evaluation metrics for attributions values
- formulate optimization objective based on them
- apply different optimization methods to objective
- compare results of optimization methods with results of "conventional" attribution methods in terms of computational theory
- ...

Research Plan

State the long - and short -term objectives of your research program. Outline specific projects planned to meet these goals, including the timelines for completion of each stage, methods to be used, and dissemination of results.

Resources

Provide details on the instrumentation and materials needed, along with the estimates of human resources required for each project/activity throughout the lifetime of the project

References

[1] Wang, Zifan, Piotr Mardziel, Anupam Datta, and Matt Fredrikson. “Interpreting Interpretations: Organizing Attribution Methods by Criteria.” *ArXiv:2002.07985 [Cs]*, April 4, 2020. http://arxiv.org/abs/2002.07985.

[2] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks, 2017.

[3] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. Digital Signal Processing, 73:1–15, Feb 2018).

[4] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 3319–3328. JMLR. org, 2017).