Master Thesis Proposal

Tim Patzelt
October 2021

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.
- [1] provide a well-defined framework for quantitavley evaluating attribution methods:
  - based on the logical concepts of necessity, suffiency and proportionality
  - allowing interpretation beyond ~~the sole aspect of~~ "importance" of features
  - defining the metrics *Total Proportionality for Necessity (TPN)* and *Total Proportionality for Sufficiency (TPS)*
- Assuming that the notion of proportionality for necessity/suffiency quantifies all desirable properties of attribution values, each ~~instance of an~~ attribution method can be seen as a heuristic ~~technique~~ to find the best value in the solution space.
- Heuristics are practical methods to find solutions to a problem. The outcome is not guaranteed to be optimal, but the heuristics exhibit other favorable properties over methods calculating the optimal solution. Like better runtime, less computational load or, in the case there is no method producing an exact solutions, the fact that they produce an approximation of the optimal outcome.
- TPN and TPS are real functions which ~~can be used to~~ formulate a mathematical optimization problem. By doing so, ~~we~~ can apply well-established optimization techniques finding values maximizing TPN or TPS (but not both since they are mutually exclusive [1]).
- If that could be achieved, the optimization would work without the need for a dedicated attribution method 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)

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 ~~very~~ related quantitive evaluation metrics:
  - 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:
  - 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)

Relevance/Impact

- 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

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).