B1 Progress Report

“Part Florian”

Overview

- 2 papers on ML, relatively directly related to project proposal

- 2 papers on ML, relatively directly related to project proposal
- 1 paper on models / simulation; spin-off

Papers related to the proposal

Clever Hans project

- Krämer / Zeitnitz / Boge

Clever Hans project

- Krämer / Zeitnitz / Boge
- DNNs exploit features (employ strategies) that are misleading (misguided) for the actual task at hand

Clever Hans project

- Krämer / Zeitnitz / Boge
- DNNs exploit features (employ strategies) that are misleading (misguided) for the actual task at hand
- can be very successful on training / testing cases

Clever Hans project

- Krämer / Zeitnitz / Boge
- DNNs exploit features (employ strategies) that are misleading (misguided) for the actual task at hand
- can be very successful on training / testing cases
- “well-generalizing features in the data” (Ilyas et al., 2019)

Clever Hans project

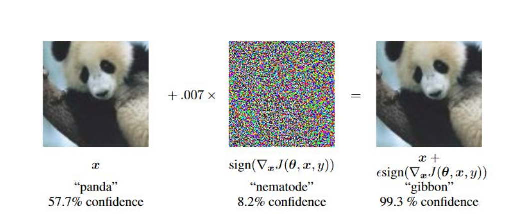
- Krämer / Zeitnitz / Boge
- DNNs exploit features (employ strategies) that are misleading (misguided) for the actual task at hand
- can be very successful on training / testing cases
- "well-generalizing features in the data" (Ilyas et al., 2019)
- non-robust, i.e., "brittle to small adversarial perturbations" (ibid.)

Clever Hans project

natural vs. ‘non-natural’ adversarials

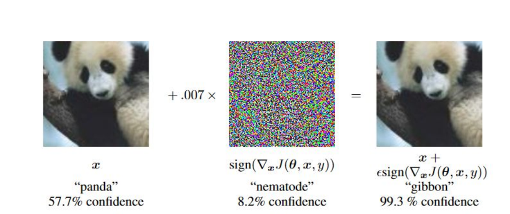
Clever Hans project

natural vs. ‘non-natural’ adversarials



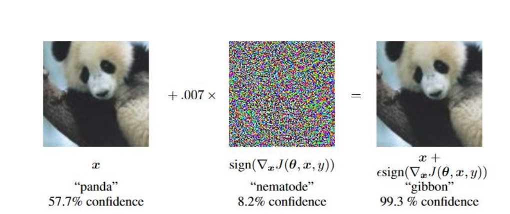
Clever Hans project

natural vs. 'non-natural' adversarials



Clever Hans project

natural vs. 'non-natural' adversarials



anomalies in jet images ~ natural adversarials

Anomalies

Clever Hans project

Anomalies...

... are phenomena (e.g., 1/8000 α-particles scattered back from foil)

Clever Hans project

Anomalies...

... are phenomena (e.g., 1/8000 α-particles scattered back from foil)

... have the power to bring about radical change (Kuhn, Lakatos, Laudan)

Clever Hans project

Anomalies...

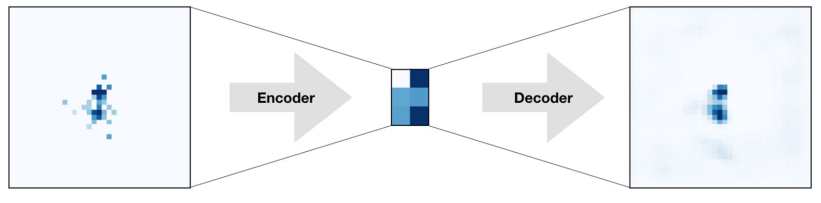
... are phenomena (e.g., 1/8000 α-particles scattered back from foil)

... have the power to bring about radical change (Kuhn, Lakatos, Laudan)

... drive scientific progress (truth / understanding / problem solving)

Model-Independent Searches

Clever Hans project



Clever Hans project

Farina et al., Phys. Rev. D 101, 075021 (2020)

surprisingly, the autoencoder performance is remarkably stable against signal contamination; the performance is barely degraded even if signal is 10% of the training sample.

Krämer et al.

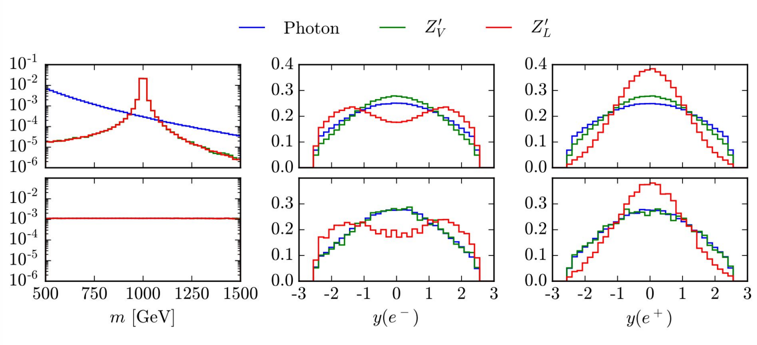
we train the AE on a pure sample of top jets and call it an inverse tagger. While the former setup is designed to perform the well-known task of tagging top jets as anomalies, the latter setup is designed to perform the inverse task, i.e. tagging QCD jets as anomalies in a background sample of top jets. [...] [T]he inverse tagger performs worse than randomly tagging jets as anomalous. [...] explain the [...] failure of the inverse tagger by the interplay between an insufficient AE performance and the different complexity in the images of the two jet classes.

Actually Smart Hans project

why do DNNs *succeed* very often?

Actually Smart Hans project

why do DNNs *succeed* very often?



Actually Smart Hans project

Minds and Machines
https://doi.org/10.1007/s11023-021-09569-4

SI: MACHINE LEARNING: PREDICTION WITHOUT EXPLANATION

Check for updates

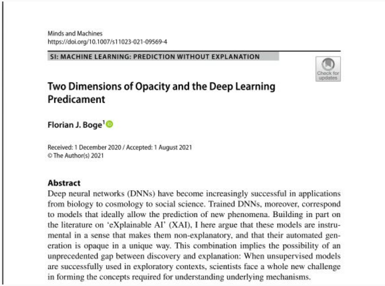
Two Dimensions of Opacity and the Deep Learning Predicament

Florian J. Boge[1]

Received: 1 December 2020 / Accepted: 1 August 2021

Abstract
Deep neural networks (DNNs) have become increasingly successful in applications from biology to cosmology to social science. Trained DNNs, moreover, correspond to models that ideally allow the prediction of new phenomena. Building in part on the literature on 'eXplainable AI' (XAI), I here argue that these models are instrumental in a sense that makes them non-explanatory, and that their automated generation is opaque in a unique way. This combination implies the possibility of an unprecedented gap between discovery and explanation: When unsupervised models are successfully used in exploratory contexts, scientists face a whole new challenge in forming the concepts required for understanding underlying mechanisms.

Actually Smart Hans project



Minds and Machines
https://doi.org/10.1007/s11023-021-09569-4

SI: MACHINE LEARNING: PREDICTION WITHOUT EXPLANATION

Check for updates

Two Dimensions of Opacity and the Deep Learning Predicament

Florian J. Boge[1]

Received: 1 December 2020 / Accepted: 1 August 2021

Abstract
Deep neural networks (DNNs) have become increasingly successful in applications from biology to cosmology to social science. Trained DNNs, moreover, correspond to models that ideally allow the prediction of new phenomena. Building in part on the literature on 'eXplainable AI' (XAI), I here argue that these models are instrumental in a sense that makes them non-explanatory, and that their automated generation is opaque in a unique way. This combination implies the possibility of an unprecedented gap between discovery and explanation: When unsupervised models are successfully used in exploratory contexts, scientists face a whole new challenge in forming the concepts required for understanding underlying mechanisms.

concepts?

Actually Smart Hans project

Manuscript submitted to the *Synthese* TC
"Philosophy of Science in Light of Artificial Intelligence"

Functional concept-proxies and the Actually Smart Hans Problem: What's special about deep neural networks in science

anonymised

Abstract
From a certain vantage point, a deep neural networks (DNNs) are nothing but parametrized functions $\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})$ of some data vector $\boldsymbol{x}$, and their 'learning' is nothing but an iterative, algorithmic fitting of the parameters to data, with a precaution against over-fitting for the sake of generality. Hence, what could be special about DNNs as a scientific tool or model? Following a number of recent approaches, I here argue that DNNs are capable of developing what I call *functional concept-proxies* (FCPs), and that this makes them interestingly different from traditional multivariate methods in statistics. I will illustrate the salient difference by considering the possibility of what I call 'Actually Smart Hans predictors', i.e., DNNs that robustly succeed because they learn to trigger on features connected to the data that are not transparent to human researchers.

Keywords Deep Neural Networks · Concepts · Reasoning · Clever Hans Problem · Automated

Actually Smart Hans project

Manuscript submitted to the *Synthese* TC
"Philosophy of Science in Light of Artificial Intelligence"

Functional concept-proxies and the Actually Smart Hans Problem: What's special about deep neural networks in science

anonymised

Abstract
From a certain vantage point, a deep neural networks (DNNs) are nothing but parametrized functions $\boldsymbol{f_\theta}(\boldsymbol{x})$ of some data vector $\boldsymbol{x}$, and their 'learning' is nothing but an iterative, algorithmic fitting of the parameters to data, with a precaution against over-fitting for the sake of generality. Hence, what could be special about DNNs as a scientific tool or model? Following a number of recent approaches, I here argue that DNNs are capable of developing what I call *functional concept-proxies* (FCPs), and that this makes them interestingly different from traditional multivariate methods in statistics. I will illustrate the salient difference by considering the possibility of what I call 'Actually Smart Hans predictors', i.e., DNNs that robustly succeed because they learn to trigger on features connected to the data that are not transparent to human researchers.

Keywords Deep Neural Networks · Concepts · Reasoning · Clever Hans Problem · Automated

functional concept proxies?

Actually Smart Hans project

functional proxies

x is a *functional proxy* for y *iff* x fulfils all the same causal roles as y, but is otherwise distinguished from y in further defining properties.

Actually Smart Hans project

functional proxies

x is a *functional proxy* for y *iff* x fulfils all the same causal roles as y, but is otherwise distinguished from y in further defining properties.

relative functional proxies

Given a set of contexts, C. Then x is a *functional proxy* for y, *relative* to C, *iff* x fulfils all the same causal roles as y in any $c \in C$, but is otherwise distinguished from y in further defining properties.

Actually Smart Hans project

(relative) functional concept proxies

Given a set of tasks, T. Then x is a *functional concept proxy (FCP)*, *relative* to T, *iff* x fulfils all the same causal roles as does any intrasubjectively stable contentful state, y, that is the basis of a higher congitive process of human reasoning tackling the $t \in T$, but is otherwise distinguished from y in further defining properties, including that x is not connected to conscious mental representations whereas y is.

Actually Smart Hans project

Actually Smart Hans Problem

DNNs may develop FCPs based on features that are (a) non-obvious or even “humanly inscrutable”, (b) well-generalising across data sets, and (c) highly fruitful for scientific prediction and discovery. Human researchers may thus fall behind qua being left without the right concepts to (i) comprehend the reasons for the given DNNs success and to (ii) develop theoretical models of their own to advance science in the ways we’re used to.

Models: Measuring or Cognitive Instruments?

Why Trust a Simulation?

Models, Parameters, and Robustness

in Simulation-Infected Experiments

Florian J. Boge

Abstract

Computer simulations are nowadays often directly involved in the generation of experimental results. Given this dependency of experiments on computer simulations, that of simulations on models, and that of the models on free parameters, how do researchers establish trust in their experimental results? Using high-energy physics (HEP) as a case study, I will identify three different types of robustness that I call conceptual, methodological, and parametric robustness, and show how they can sanction this trust. However, as I will also show, simulation models in HEP themselves fail to exhibit a type of robustness I call inverse parametric robustness. This combination of robustness and failures thereof is best understood by distinguishing different epistemic capacities of simulations and different senses of trust: Trusting simulations in their capacity to facilitate credible experimental results can mean accepting them as means for generating belief in these results, while this need not imply believing the models themselves in their capacity to represent an underlying reality.

Models: Measuring or Cognitive Instruments?

Why Trust a Simulation?

Models, Parameters, and Robustness

in Simulation-Infected Experiments

Florian J. Boge

Abstract

Computer simulations are nowadays often directly involved in the generation of experimental results. Given this dependency of experiments on computer simulations, that of simulations on models, and that of the models on free parameters, how do researchers establish trust in their experimental results? Using high-energy physics (HEP) as a case study, I will identify three different types of robustness that I call conceptual, methodological, and parametric robustness, and show how they can sanction this trust. However, as I will also show, simulation models in HEP themselves fail to exhibit a type of robustness I call inverse parametric robustness. This combination of robustness and failures thereof is best understood by distinguishing different epistemic capacities of simulations and different senses of trust: Trusting simulations in their capacity to facilitate credible experimental results can mean accepting them as means for generating belief in these results, while this need not imply believing the models themselves in their capacity to represent an underlying reality.

"cognitive instruments"

Models: Measuring or Cognitive Instruments?

Morrison, *Phil. Studies* (2009)

[Certain models] [n]ot only [...] allow us to interpret so-called measurement outputs, but [...] the models themselves can function as measuring instruments [...].

Models: Measuring or Cognitive Instruments?

Morrison, *Phil. Studies* (2009)

[Certain models] [n]ot only [...] allow us to interpret so-called measurement outputs, but [...] the models themselves can function as measuring instruments [...].

Parker, *BJPS* (2017)

[simulations] can be embedded in measurement practices in such a way that simulation results constitute measurement outcomes

Models: Measuring or Cognitive Instruments?

- three different arguments (one strawman)

Models: Measuring or Cognitive Instruments?

- three different arguments (one strawman)
- critique of the premises

Models: Measuring or Cognitive Instruments?

- three different arguments (one strawman)
- critique of the premises
- what does it take to be a ‘cognitive’ instrument?

Models: Measuring or Cognitive Instruments?

- three different arguments (one strawman)
- critique of the premises
- what does it take to be a ‘cognitive’ instrument?
- causal contact (literal instrument) vs. inferential connection

Models: Measuring or Cognitive Instruments?

me (following Rowbottom, *The Instrument of Science* (2019))

thus calling models 'cognitive' also means that they are capable of promoting understanding—understanding of a variety that, though not objective in the sense of involving the truth of the relevant model, does imply advanced control over the phenomena. This control can manifest in various ways, including and especially in the ability to use these models as templates for further, even more sophisticated ones that accommodate more empirical data, as evidenced by the converged hadronization models in HEP.