XAI: Stop using Black Boxes

Blog Post by Howard Heaton

July 19, 2021

TL;DR – Unexplainable machine learning can exhibit "bad" behavior/biases and can be replaced with robust guarantees/explainable data-driven optimization.[1]

The Idea. Black boxes[2] are common in machine learning (ML). The notion of simply "throwing AI" at a problem[3] *might* provide decent results on a particular problem, but this is not insightful and may generalize poorly. Consider Amazon's efforts to hire people in a data-driven way based on existing hires, which was found to discriminate against women.[4] That's a problem. What about if a similar AI service was used for auto-loan decisions? Could applicants be told *why* their application was accepted/rejected? Did the decision improperly discriminate based on demographics? If so, is there an appropriate way to correct such biases? This issue becomes even more life-threatening in critical situations like detecting and classifying cancerous tumors in medical scans. Thus, although leveraging big data is tremendously successful, it is imperative to inquire how we can ensure data-driven algorithms (aka ML) execute in a just and explainable fashion.

Explainable AI (XAI) is the notion that solutions generated by a data-driven algorithm can be understood by humans. The aim of XAI is to explain each computed solution, what will be computed next, and uncover information the actions are based on (see this short article). The guiding principles of XAI are transparency, interpretability and explainability. This is all fine and well, but the abstract idea of "understoood" is difficult to quantify and implementation can take many forms. For example, one could train a model-free[5] network and perform analysis on the network *after* training by using statistical analyses with input augmentations. This is a reasonable approach, but it can be problematic for two reasons. First, after the fact analysis may require varying amounts of work to figure out what's going on.[6] Second, if "bad" behavior is found, there isn't necessarily a way to correct the behavior (without significantly affecting performance). These two reasons lead me to believe this approach has potential, but that it may, in general, be difficult/impractical in real-world settings.

L2O for XAI. An approach that has become the focus of my PhD thesis is to use ML that encodes prior knowledge from models by domain experts. Specifically, white-box[7] XAI schemes can be obtained by ML algorithms that "learn to optimize" (L2O). Depending on setup, this can admit explanations at a low-level to domain experts and a high-level for lay persons.[8] One approach is to do regression[9] with data d by modeling a "true" prediction $x_d^\star$ via minimization, *i.e.*

$$x_d^\star \approx \tilde{x}_d = \operatorname*{arg\,min}_{x\in\mathcal{C}_d} f(x;d), \tag{1}$$

for a set $\mathcal{C}_d \subseteq \mathbb{R}^n$ and a function $f : \mathbb{R}^n \to \mathbb{R}$. Loosely speaking, we can write

$$\min_{x\in\mathcal{C}_d} f(x) = \min_{x\in\mathbb{R}^n} (\text{physics}) + (\text{regularization}) \ \ \text{subject to} \ \ (\text{constraints}). \tag{2}$$

Common scientific applications admit a physical interpretation (*e.g.* through minimization of an energy or measurement discrepancy). Regularization encodes prior knowledge (*e.g.* sparsity). Lastly, constraints can be used to ensure "good" behavior (*e.g.* to not discriminate based on certain demographics). Okay, this sounds great. But, even if I can formulate these terms, how does big data fit in?

[1] The explainability ideas in this blog are derived from several "learn to optimize" (L2O) papers.

[2] A black box is a system for which only inputs and outputs can be accessed, and its internal workings are not necessarily understood by humans.

[3] Here I mean the practice of guessing a neural network architecture and doing hyper parameter tuning until the training loss looks "good," without any a priori meaningful insight into *why*.

[4] Check out this 2018 Reuters article.

[5] Model-free approaches rely heavily on sample data to determine how to perform updates (*e.g.* feed forward steps), which do not necessarily resemble any interpretable model of the task at hand.

[6] For example, slight changes to the network architecture or training process can result in a trained configuration that has different sensitivities to inputs and requires more tuning/work to conduct analyses.

[7] White-box models are those for which behaviors can be explained along with how they produce predictions and what the influencing variables.

[8] For example, a credit applicant could find out which pieces of information contributed most significantly toward their credit decision by knowing which contributed the largest value to some credit risk function (without them explicitly knowing the risk function).

[9] In the context of ML, regression problem require generating inferences that are continuous variables (as opposed to, say, classification problems).

The model described in (2) can be thought of as a first-order approximation of a "true" model for the task at hand. Building on this analogy, additional parameters/terms can be included can be included to augment the model (*e.g.* weight the relative importance of each term or identify additional corrections). Thus, we may augment (2) by writing[10]

$$f_\Theta(x;d) = (\text{physics}) + (\text{regularization}) + \underbrace{(\text{data-driven correction term})}_{\text{depends on weights } \Theta}. \tag{3}$$

As a simple example, suppose we want to recover a sparse signal $x_d^\star$ from noisy measurements $d = Ax_d^\star + \varepsilon$ (*n.b.* here ε is noise). We can use the model function[11]

$$f_\Theta(x;d) = \underbrace{\|Ax-d\|_2^2}_{\text{physics}} + \underbrace{\lambda\|x\|_1}_{\text{regularization}} + \underbrace{\|W_1x\|_2^2 + \|W_2d\|_2^2}_{\text{correction}}, \tag{4}$$

where $\Theta = (\lambda, W_1, W_2)$. Given data d, we estimate $x_d^\star$ by the minimizer $\tilde{x}_d$ of f_Θ in (4). A variation[12] of this idea was used for sparse coding problems, which resulted in a roughly 20-fold speedup over classic optimization for sparse coding.[13]

In the "old days," if the number of scalars defining Θ was small, then a domain expert could choose them by intuition or repeatedly trying choices by hand. However, we can encode the weights Θ into a neural network $\mathcal{N}_\Theta$ by defining

$$\mathcal{N}_\Theta(d) = \tilde{x}_d, \quad \text{where} \quad \tilde{x}_d = \underset{x\in\mathcal{C}_d}{\arg\min}\, f_\Theta(x;d). \tag{5}$$

This definition enables one to use data d and weights Θ to define a model problem (via f_Θ and $\mathcal{C}_d$) with solution $\tilde{x}_d$ that is precisely the network inference, *i.e.* $\tilde{x}_d = \mathcal{N}_\Theta(d)$. In terms of explainability, we can determine the relative contribution of the terms to obtain $f_\Theta(\tilde{x}_d, d)$. In the toy problem above, this could reveal measurement noise significantly affected the solution $\tilde{x}_d$ or that it was difficult to obtain a sparse output. In a more elaborate setting like medical CT image processing, f_Θ can instill CT measurement data compliance, total variation regularization and/or an energy used in classic image segmentation (*e.g.* the Chan-Vese model) in addition to a correction term leveraging a convolutional network structure.[14]

The weights Θ of the network $\mathcal{N}_\Theta$ are tuned using a distribution $\mathcal{D}$ of data and a loss function ℓ by solving the training problem

$$\min_\Theta \mathbb{E}_{d\sim\mathcal{D}}\left[\ell(\mathcal{N}_\Theta(d), x_d^\star)\right], \tag{6}$$

where $x_d^\star$ is the true signal corresponding to data d. We won't go into details here about how to evaluate $\mathcal{N}_\Theta(d)$ or update weights Θ, but this can be accomplished using fixed point networks (FPNs), a framework discussed in my prior blogs. We emphasize the network $\mathcal{N}_\Theta$ takes in data d and outputs an inference $\tilde{x}_d$, following the same flow as commonplace neural networks (just now with added structure). Extensions of this L2O-XAI approach can also be obtained with operator theory.[15]

CONCLUSION. The era of big data has offered an explosion of techniques and tools for using data to improving and/or automate tasks. For domains like medicine, defense, finance, and law, XAI is crucial for data-driven algorithms to be considered trust-worthy. To obtain desired transparency and behavior guarantees, XAI is essential to move forward. I do not believe black box approaches offer adequate explanation. However, a promising framework that automatically incorporates XAI into its structure without notably hindering performance is optimization-based ML that "learns to optimize" (aka L2O networks).

[10] The physics and regularization terms can also include dependence on Θ.

[11] Note ℓ_1 regularization is commonplace for encouraging sparsity.

[12] There the authors instead unrolled an optimization algorithm for K steps and used a *different* Θ_k for each layer k.

[13] Check out Gregor and LeCun's seminal, highly cited ICML paper.

[14] This is similar to what we did in the paper *Feasibility-based Fixed Point Networks*.

[15] We omit these due to space limit, but may expound further in a later blog.

Last Updated: July 19, 2021