A NetAI Manifesto (Part I): Less Explorimentation, More Science

Walter Willinger
NIKSUN, Inc.

Arpit Gupta
UCSB

Arthur S. Jacobs
UFRGS

Roman Beltiukov
UCSB

Ronaldo A. Ferreira
UFMS

Lisandro Granville
UFRGS

ABSTRACT

The application of the latest techniques from artificial intelligence (AI) and machine learning (ML) to improve and automate the decision-making required for solving real-world network security and performance problems (NetAI, for short) has generated great excitement among networking researchers. However, network operators have remained very reluctant when it comes to deploying NetAI-based solutions in their production networks, mainly because the black-box nature of the underlying learning models forces operators to blindly trust these models without having any understanding of how they work, why they work, or when they don't work (and why not). Paraphrasing [1], we argue that to overcome this roadblock and ensure its future success in practice, NetAI *"has to get past its current stage of explorimentation, or the practice of poking around to see what happens, and has to start employing tools of the scientific method."*

1. INTRODUCTION

Most deployed networking solutions, be they ubiquitous protocols such as TCP or special-purpose systems such as load balancers, make decisions based on domain-specific heuristics that rely on partial network state information extracted from active or passive network measurements. For more than two decades, networking researchers have been exploring how to improve and automate these heuristics-based decision-making processes with the help of NetAI. In the process, they have enthusiastically embraced the development of new learning models by applying a workflow paradigm commonly referred to as the "standard ML pipeline." Comprised of *(i)* a learning task that is characterized by a model specification, *(ii)* a training dataset, and *(iii)* an independent and identically distributed (iid) evaluation procedure, this paradigm provides a blueprint for producing trained models that "work." Here, the statement "the model works!" is short for "according to the evaluation procedure used, the model has excellent expected predictive performance (*e.g.,* F1-score close to 1) when used for the originally posed learning task."

Like in many other application domains of ML (*e.g.,* computer vision, self-driving vehicles), leveraging this workflow paradigm in the networking domain has also enabled transformational progress, with ML-based solutions frequently and easily outperforming domain-specific state-of-the-art heuristics. However, despite this progress and ensuing promises, NetAI in its current form has largely failed to gain traction among network operators.

In this paper, we criticize the use of the standard ML pipeline that is popular with NetAI researchers. In particular, we show that relying on this widely-adopted ML workflow is fraught with problems that question the scientific foundations of the artifacts it produces and argue for abandoning it altogether in favor of a new generation of ML pipelines. In the process, we elaborate on the urgent need to be able to develop ML models that are either inherently explainable or can be explained post-hoc by applying available global explainability tools. We describe an initial attempt at designing and implementing such a new ML pipeline that is capable of accomplishing this feat, comment on its ability to aid the development of a new generation of learning models that focus on the generalizability and safety of ML models, and discuss some exciting new opportunities that arise as a result of this proposed paradigm shift in ML model development and evaluation.

2. THE "DUMBING DOWN" OF NETWORKING RESEARCH

The main reason why ML-based solutions have not been widely adopted in networking is that the models that the standard ML pipeline outputs are in general black boxes. In effect, such an output forces network operators to blindly trust the resulting learning model, providing them with little to no understanding of how the model works, why it works, or when it doesn't work (and why not).

This dissatisfaction has been compounded by an increasing awareness among researchers that the standard ML pipeline defines indeed a low bar for claiming that the trained models it produces as output "work." In particular, by relying on an evaluation procedure that assesses an output model's expected predictive performance simply on data drawn from the same distribution as the training dataset (*e.g.,* a randomly held-out subset of the training dataset), the standard ML pipeline lacks any means to quantify the effectiveness of the trained models beyond what is captured by commonly-used metrics such as the F1-score. In effect, we argue that NetAI in its current form has contributed to a "dumbing down" of networking research as it has promoted a blind belief in the high-performant black-box models it considers.

We are not alone in criticizing the standard ML pipeline, its widespread and largely uncontested use across differ-

ent application domains of ML, and its overly pragmatic approach to evaluating the resulting trained models solely based on their effect (*i.e.,* "they work!"). For example, there is a growing body of work in the ML literature that is concerned with the surprisingly poor reported performance of many of the trained models that result from an application of the standard ML pipeline as soon as they get deployed in real-world environments [3, 4]. In fact, many of these works identify the fact that modern ML workflows such as the standard ML pipeline tend to be *underspecified* (*i.e.,* return many distinct models with equivalently strong test performance) as a key reason for why the resulting trained models do not generalize (*i.e.,* fail to perform as expected in deployment). Because of an evaluation procedure that relies on held-out data that have the same distribution as the training data, the standard ML pipeline has been shown to be especially prone to this underspecification problem, resulting in the observed poor model behavior in practice.

3. UNDERSTANDING CAUSE VS. EFFECT

As more of the failures and limitations of modern ML workflows such as the standard ML pipeline come to light, it is arguably justified to describe adhering to these workflows as being akin to "explorimentation" [1]. In fact, in the context of NetAI, we agree with the basic sentiment expressed in [1] that *"while appropriate for the early stages of research to inform and guide the formulation of a plausible hypothesis, [the standard ML pipeline] does not constitute sufficient progress to term the effort scientific."* Moreover, as NetAI is trying to overcome the general reluctance of network operators to deploy its ML-based solutions in their production networks, the standard ML pipeline's pragmatic "it works!" approach, typically quantified in terms of high F1-scores, to assessing its output by means of an iid evaluation procedure is no longer sufficient. In fact, to paraphrase [1], to ensure that network operators can begin to understand cause and not just effect of proposed ML-based solutions, *"it will be necessary to get past the stage of explorimentation and start employing tools of the scientific method."*

At the same time, we are not arguing for universally abondoning the use of ML workflows such as the standard ML pipeline and replacing them with "tools of the scientific method." For example, when employing ML-based solutions for low-stakes decision-making (*e.g.,* generic image classification, commercial recommendation systems, spam filtering), understanding how and why the underlying learning models make their decisions or knowing when they work or when they don't work is generally unnecessary or overkill — in such cases, for a black-box model to make a few wrong decisions is fully expected and tolerated, has little to no repercussions (*e.g.,* financial or reputation-wise), and can be fixed in future versions of the trained models. However, the situation is drastically different in cases where ML models are used for high-stakes decision-making (*e.g.,* predicting criminal recidivism risk, child welfare screening, medical treatment recommendation, self-driving vehicles) and where making a wrong decision or using an underspecified model can negatively impact the lives of people or the financial health or public reputation of companies. In such cases, innovative approaches that emphasize understanding "cause" over assuring "effect" which, after all, is the *raison d'être* of science, should be at the forefront of researchers' minds so they can successfully explain a trained model's decision-making process to end users who look for assurances that they can trust a proposed trained model.

In the NetAI domain, we say network operators "trust" a given ML-based solution if they are comfortable with relinquishing control to the model (see [4] and references therein). Given that network operators have remained reluctant to deploy trained models produced by the standard ML pipeline is evidence that they generally don't trust NetAI-based solutions. This applies in particular to trained models proposed for solving network security- or network performance-related problems where the consequences of a wrong decision can range from lost revenues, service contract terminations, customer dissatisfaction, and shutdown of business-critical services. As such, these models are clearly non-starters when it comes to engender trust in ML-based solutions among network operators. In contrast, "white-box" models such as decision trees promise to be ideal vehicles for convincing network operators that they can trust the models. These models not only describe in detail how and why every single decision is made, but domain experts can also examine them to find out when they work or don't work (and why not) and provide a means for scrutinizing the obtained model for indications of potential underspecification issues.

4. THE "OPENING UP" OF NETWORKING RESEARCH

To engender more trust in ML models, recent studies have argued for developing ML workflows that, instead of first creating black-box models and then trying to "explain" them, should generate white-box models such as decision trees that are inherently interpretable in the first place [2]. An attractive property of such workflows would be that they eliminate the need for any post-hoc explainability efforts because the models they output already reveal the underlying process by which they make their decisions and can therefore be directly checked and assessed by human domain experts, at least in theory. They also invite a direct comparison with the decisions domain experts would make when faced with the same data. However, as commented in [2], *"the belief that there is always a trade-off between accuracy and interpretability has led many researchers to forgo the attempt to produce an interpretable model. This problem is compounded by the fact that researchers are now trained in deep learning, but not in interpretable machine learning. Worse, toolkits of machine learning algorithms offer little in the way of useful interfaces for interpretable machine learning methods."*

Networking researchers contemplating developing ML-based solutions for their problems are therefore confronted with a serious dilemma. On the one hand, most modern ML pipelines, including the standard ML pipeline, focus almost exclusively on producing black-box models, but the use of such models in ML-based solutions that involve making high-stakes decisions is being increasingly criticized for the potentially tremendous harm they can inflict. The black-box models' inability to provide understanding in how and why they make their decisions also has the largely unintended consequence of contributing to a continued "dumbing down" of networking research. On the other hand, few, if any, ML pipelines exist today that have been explicitly designed to produce white-box models such as decision trees, even though their use in ML-based solutions that involve making high-stakes decisions is not

only preferred but recommended for the full transparency they provide for potentially life-altering decision-making. Moreover, the white-box models' ability to provide understanding in how and why they make their decisions makes them ideally suited for "opening up" networking research; that is, transforming networking research into a science by means of both a renewed focus on understanding "cause" and an intentional effort towards de-emphasizing "effect".

In an effort to resolve this dilemma, we recently developed and implemented TRUSTEE [4]. TRUSTEE defines a novel ML workflow that takes the trained black-box model (*i.e.,* model specification, training dataset) that results from an application of the standard ML pipeline as input and generates a white-box model in the form of a decision tree and an associated trust report as output. In synthesizing this decision tree, TRUSTEE strikes a balance between model fidelity (*i.e.,* accuracy of the decision tree with respect to the black-box model), model complexity (*i.e.,* the size of the decision tree and its explicitness and intelligibility), and model stability (*i.e.,* correctness, coverage, and robustness of the decision rules or branches of the decision tree). Using the decision tree that TRUSTEE extracts from the given black-box model, networking researchers can examine how or why the trained black-box model makes its decisions for a majority of data samples and can scrutinize it for indications of potential underspecification issues. Moreover, domain experts can use it to compare whether or not the black-box model makes the same decisions they would make when faced with the same data samples, and network operators can inspect TRUSTEE's output to gauge their trust in the given black-box model.

5. 2 STEPS FORWARD, 1 STEP BACK?

Early indications are that TRUSTEE has been a welcome and much-needed addition to the toolkits that researchers in the area of NetAI have relied on. As the use of ML in the networking domain continues to attract large numbers of researchers, TRUSTEE provides a concrete means for questioning some of the exhibited hubris and overconfidence by NetAI researchers, scrutinizing the soundness of NetAI-based solutions that have been reported in the existing literature, and performing some much-needed sanity checks on the myriad of proposed black-box models that have been trained for solving networking-specific problems. For example, by examining more than half a dozen of frequently cited and fully reproducible ML models from the existing networking literature, all of which are the results of using the standard ML pipeline, we found TRUSTEE to be especially good at refuting reported claims of "the model works!" and providing supporting evidence. In refuting these claims, TRUSTEE identified concrete instances of model underspecification issues, including trained models that leveraged shortcut learning strategies (akin to "cheating"), showed vulnerabilities to out-of-distribution samples (akin to "rote learning"), or exploited spurious correlations in the training data (akin to "lucky guesses"). The problematic nature of these findings argues for more caution with respect to the use of black-box models in the field of networking, suggests looking at developments in this area with a highly critical eye, and identifies common pitfalls or "blind spots" of proposed ML-based solutions that prevent operators from trusting them and deploying them in their production networks.

At the same time, while we agree with much of the reasoning in [2] where the author argues why interpretable black boxes should be avoided altogether in high-stakes decisions, our work with TRUSTEE caused us to take a more nuanced view with respect to explainable black-box models. For one, using the decision tree that TRUSTEE extracts from a given black-box model demystifies much of the decision-making process or "inner workings" of black-box models. In fact, this extracted decision tree becomes at once the main vehicle for domain experts to check if the given black-box model makes decisions in accordance with existing domain knowledge. An even more tantalizing application of a TRUSTEE-extracted decision tree is examining it with respect to the given black-box model's ability of teach the domain experts new decision-making strategies. Here, the term "teach" is meant in the sense of carefully inspecting the decision tree to see if it reveals legitimate strategies that the domain experts have been unaware of but upon painstaking examination recognize as meaningful and relevant decisions that deserve to be added to their existing domain knowledge.

6. CONCLUSION

We are presently not aware of any such examples in the NetAI domain where a given black-box model, via its TRUSTEE-extracted decision tree, teaches domain experts novel decision-making strategies. However, the likely existence of such examples suggests a natural "division of labor" in the NetAI domain between machines and humans that achieves the best of both worlds; i.e., leverages the raw computational power and algorithmic capabilities of ML to let machines do the grunt or "dirty" work (*i.e.,* sifting through training data, finding potentially useful patterns, and distilling them in a trained black-box model) and rely on the intelligence and inherently limited computing capabilities of humans to apply reasoning and logical thinking (*i.e.,* determining whether or not the detected patterns are meaningful and relevant for the problem at hand or point to possible underspecification issues with the trained black-box model). As this perspective explicitly argues for the need to keep human domain experts in the loop, it is counter to widely-held beliefs or common myths about the impact of increasingly autonomous technologies in general and NetAI-driven network automation in particular, namely that their wide-spread adoption will ultimately eliminate humans from the loop.In Part II of this NetAI Manifesto [5], we will revisit this perspective and argue why and how developing NetAI-based automation capabilities that work in practice will require keeping humans in the loop.

7. REFERENCES

[1] J. Z. Forde and M. Paganini. The Scientific Method in the Science of Machine Learning. *ICLR Debugging Machine Learning Models Workshop, 2019.*

[2] C. Rudin. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nat Mach Intell 1, 206–215, 2019.*

[3] A. D'Amor et. al. Underspecification Presents Challenges for Credibility in Modern Machine Learning. *Journal of Machine Learning Research 23, 1–61, 2022.*

[4] A. S. Jacobs et al.AI/ML for network security: The emperor has no clothes. *Proc. ACM CCS'22, 2022.*

[5] W. Willinger et al. A NetAI Manifesto (Part II): Less Hubris, More Humility. *Performance Evaluation Review, this issue, 2023.*