Proceedings of Machine Learning Research 176:302–307, 2022 NeurIPS 2021 Competition and Demonstration Track

Exploring Conceptual Soundness with TruLens

Anupam Datta DANUPAM@CMU.EDU
Matt Fredrikson MFREDRIK@CMU.EDU
Klas Leino KLEINO@CS.CMU.EDU
Kaiji Lu KAIJIL@ANDREW.CMU.EDU
Zifan Wang ZIFAN@CMU.EDU
Carnegie Mellon University
Ricardo Shih RICK@TRUERA.COM
Shayak Sen SHAYAK@TRUERA.COM
Truera

Editors: Douwe Kiela, Marco Ciccone, Barbara Caputo

1. Introduction

As machine learning has become increasingly ubiquitous, there has been a growing need to assess the trustworthiness of learned models. One important aspect to model trust is *conceptual soundness*, i.e., the extent to which a model uses features that are appropriate for its intended task. Deep networks have quickly become the face of modern machine learning, with unparalleled success at complex human tasks such as vision and natural language processing. However, deep networks are notoriously opaque, further emphasizing the need for transparency into their internal logic. A large body of work has arisen to address this problem, by providing *explanations* that distill aspects of a model's behavior to be better understood by human practitioners.

In this demonstration,[1] we present TruLens, a new cross-platform library for explaining deep network behavior that implements a general class of gradient-based explanations captured by the "influence-directed" explanation framework of Leino et al. (2018). Throughout our presentation, we will take the unique perspective that to accurately assess the conceptual soundness of a model, an explanation must be *faithful*—i.e., the explanation must be causally related to the model's behavior. By contrast, the literature has often attempted to justify explanations based on their appeal to human intuition. However, this begs the question, as it assumes the model captured human intuition in the first place. Instead, we argue that the utility of an explanation framework comes from its flexibility to faithfully answer a wide range of queries, but *not* from its tendency to produce reasonable, visually-appealing, or intuitive explanations.

Our demonstration will show that faithful explanations can surface erroneous model behavior that may not be manifested in the validation set, and would therefore otherwise go unnoticed prior to model deployment. Thus, conversely, faithful explanations that align with our expectations of conceptually sound predictions, serve as evidence that the model is trustworthy. Finally, we observe that erroneous behavior caused by *adversarial examples* (Szegedy et al., 2014) is indicative of a lack of conceptual soundness. As adversarial examples are ubiquitous in standard deep networks, this observation suggests that robustness to adversarial examples is necessary for establishing conceptual soundness.

1. Full materials for this demonstration are available at https://truera.github.io/neurips-demo-2021.

2. Influence-Directed Explanations via TruLens

When a deep neural network makes a decision, we would like to know whether we can trust it. Did our model make a connection between spuriously-correlated events? Did our model learn a pattern that we overlooked but might find useful? If our model made a mistake, why? Answering these high-level questions requires the ability to make a rich set of queries that help us learn about a model's predictive behavior and assess its conceptual soundness. Primarily, an explanation framework is meant to give us the means to make such queries—its utility comes from its ability to express and accurately answer a wide range queries.

TruLens implements the explanation framework of Leino et al. (2018), which formalizes a general family of axiomatically justified gradient-based explanations through a saliency measure, *Internal Influence*, that assigns an importance score to input features. A key property of gradient-based explanation methods is their causal relationship to the model's behavior, which ensures faithfulness. Many popular gradient-based explanation methods can be viewed as an instantiation of Internal Influence, e.g., Saliency Maps (Simonyan et al., 2014), Integrated Gradients (Sundararajan et al., 2017), and SmoothGrad (Smilkov et al., 2017). In addition, Internal Influence provides flexibility along three key axes not captured by other approaches: (1) the *quantity of interest*, which allows us to specify the aspect of the model's behavior we are interested in understanding—e.g., *why did the model predict class A*? Or, *which features specifically distinguish between class A to class B*? (2) the *distribution of interest*, which allows us to specify how local we want our explanation to be—e.g., *do we want to explain a behavior on a single point*? Or, *do we want to understand a model's behavior more globally in the neighborhood of that point*? And (3) the *network slice* allows us to peer into the internal mechanisms at a specific layer in the network—this lets us understand the model's behavior in terms of the high-level features it encodes.

TruLens allows the user to specify the quantity of interest, distribution of interest, and network slice, offering the unique capability to make highly flexible queries to thoroughly probe a model's behavior. Finally, TruLens is unique in its cross-platform support for both TensorFlow and Pytorch, making it accessible to a wider audience of developers.

3. Assessing Conceptual Soundness

Explanation frameworks are valuable tools for assessing the conceptual soundness of learned models. In particular, explanations can help us detect unsound feature usage that might lead to erroneous behavior when the model is deployed.

As a demonstration[2] of this, consider the following example. Suppose that we train a model to recognize faces from the Labeled Faces in the Wild (LFW) dataset (Huang et al., 2007), which contains faces of several public figures from the early 2000s. Figure 1a displays a sample of training points taken from a subset of LFW containing the five most common public figures. We observe that the picture in the top right corner has a distinctive pink background—in fact, this is one of only two pictures in the dataset with this background; both are of Tony Blair. We hypothesize that a model may overfit by learning to use the pink background as a feature for Tony Blair, as the feature is indeed predictive of Tony Blair on the training set. Of course, despite its coincidental usefulness on the training set,

2. https://colab.research.google.com/drive/1Iswyxd4rorKqqQWkC4kieAwFpfqBS25v?usp=sharing

(a) (b) (c) (d)

Figure 1: **(a)** training instances from the LFW dataset. **(b)** explanation generated with TruLens highlighting the most important features for labeling the shown instance as "Tony Blair." **(c)** test image correctly labeled as "Gerhard Schröder." **(d)** edited image erroneously labeled as "Tony Blair."

the background is clearly not conceptually sound, and is unlikely to be useful on new data. If the model overfits in this way, it will be evident from an inspection of the features that are encoded and used by the model on instances with pink backgrounds.

To test our hypothesis, we train a simple convolutional neural network and use TruLens to probe its behavior. Figure 1b shows the results of explaining the model's behavior on the training instance with the pink background using Internal Influence—the details of the parameters used to generate this explanation with TruLens are not important for our discussion, but they are included in the full demonstration materials. We see that indeed, the pink background is used by the model to identify Tony Blair in this instance, warning us that our model is not conceptually sound.

The fact that the model uses the background as a feature may lead to errors on new points as the model is deployed. This may not be readily apparent without explanations, as our validation set contains no points with similar pink backgrounds. However, the possibility of this scenario is corroborated by the model's behavior on the instances from Figures 1c and 1d, where we see that the model correctly classifies the former as Gerhard Schröder, but when the image is edited to have a pink background, as in the latter, the model incorrectly predicts "Tony Blair."

Quality Explanations Require Quality Models. Prior work has often judged the quality of explanation frameworks based on the degree to which their generated explanations appeal to human intuition. When we admit the possibility of conceptually unsound models, we see that this is a flawed approach. For example, the highlighting of the pink background in Figure 1b can hardly be considered an "intuitive" explanation. However, we saw evidence that the model did use a pink background as a feature for Tony Blair, making it a more faithful explanation. Measures of explanation quality that reward explanations that highlight pixels within human-designated bounding boxes (Zhang et al., 2016), or that are rated as "trustworthy" by human subjects (Chattopadhay et al., 2018) end up penalizing faithful explanations in favor of wishful but fallacious methods that assume conceptual soundness *a priori*. Indeed, many explanation methods that produce visually appealing explanations have been found to fail basic sanity checks for faithfulness (Adebayo et al., 2018). As we will further explain in Section 4, conceptually unsound models are the *norm* in modern machine learning, not the exception. Thus, future work will have to focus on training conceptually sound models before we can expect intuitive explanations.

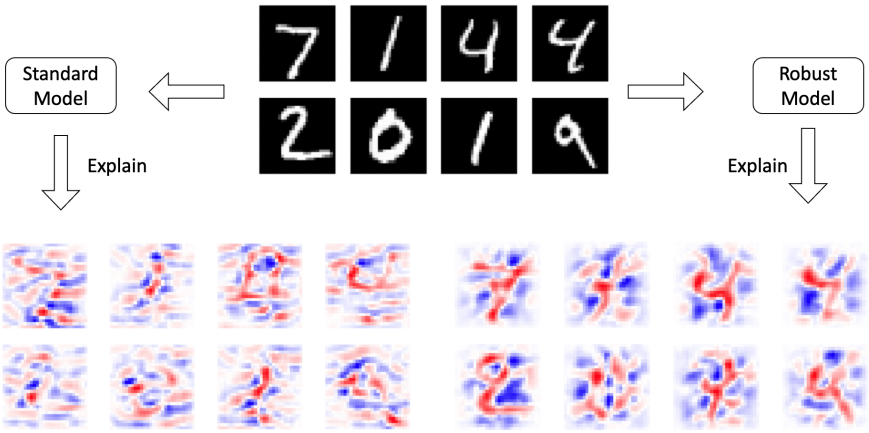

Figure 2: Visualizations of Saliency Map explanations generated with TruLens on MNIST digits for a standard (left) and a robust model (right).

4. Robustness as a Prerequisite for Conceptual Soundness

Deep networks are easily fooled by malicious perturbations to their inputs, termed *adversarial examples* (Szegedy et al., 2014). The ubiquity of adversarial examples in typical networks clearly constitutes a security concern—manifested as unexpected erroneous behavior on seemingly benign inputs. Moreover, adversarial examples establish a network as conceptually unsound. Specifically, the perturbations that produce adversarial examples are semantically meaningless by construction, but are nonetheless causally relevant to changing the model's prediction—meaning that they will be identified by faithful explanations despite their conceptual irrelevance. Indeed, adversarial examples can be understood to arise as a result of models relying on "non-robust" features (Ilyas et al., 2019).

A rich body of literature provides various methods for training "robust" neural networks that are resistant to adversarial examples, ranging from heuristic defenses (Madry et al., 2018; Zhang et al., 2019) to those that facilitate provable robustness guarantees (Croce et al., 2019; Leino et al., 2021; Leino and Fredrikson, 2021; Trockman and Kolter, 2021; Wong et al., 2018). As adversarial examples are a fundamental roadblock for conceptual soundness, robustness is essential for achieving networks that can faithfully exhibit intuitive explanations. This observation is borne out in the literature as the gradients on models trained to be robust have been shown to be more interpretable than those on their non-robust counterparts (Etmann et al., 2019; Tsipras et al., 2019; Wang et al., 2021).

As a demonstration[3] of these findings, we compare explanations generated by TruLens on robust models to those on their non-robust "standard" counterparts (Figure 2). For the robust models in our demonstration we use GloRo training (Leino et al., 2021),[4] which represents the state-of-the-art for provably-robust training methods. It is apparent from Figure 2 that explanations from the robust model are more interpretable than their non-robust counterparts, as the former delineate crisper outlines of corresponding digits.

3. https://colab.research.google.com/drive/196PjI40gjIUtV4hqBMCymgY2B3Urj5zC?usp=sharing
4. Code available at https://github.com/klasleino/gloro.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Neural Information Processing Systems (NIPS)*, 2018.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.

Francesco Croce, Maksym Andriushchenko, and Matthias Hein. Provable robustness of ReLU networks via maximization of linear regions. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Christian Etmann, Sebastian Lunz, Peter Maass, and Carola Schoenlieb. On the connection between adversarial robustness and saliency map interpretability. In *Internaltional Conference on Machine Learning (ICML)*, 2019.

Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts, Amherst, 2007.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Neural Information Processing Systems (NIPS)*, 2019.

Klas Leino and Matt Fredrikson. Relaxing local robustness. In *Neural Information Processing Systems (NIPS)*, 2021.

Klas Leino, Shayak Sen, Anupam Datta, Matt Fredrikson, and Linyi Li. Influence-directed explanations for deep convolutional networks. In *2018 IEEE International Test Conference (ITC)*, 2018.

Klas Leino, Zifan Wang, and Matt Fredrikson. Globally-robust neural networks. In *International Conference on Machine Learning (ICML)*, 2021.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Asher Trockman and J Zico Kolter. Orthogonalizing convolutional layers with the cayley transform. In *International Conference on Learning Representations (ICLR)*, 2021.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations (ICLR)*, 2019.

Zifan Wang, Matt Fredrikson, and Anupam Datta. Boundary attributions provide normal (vector) explanations. *CoRR*, abs/2103.11257, 2021.

Eric Wong, Frank Schmidt, Jan Hendrik Metzen, and J. Zico Kolter. Scaling provable adversarial defenses. In *NIPS*, 2018.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *ICML*, 2019.

Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. In *European Conference on Computer Vision (ECCV)*, 2016.