Explanatory Dialogs: Towards Actionable, Interactive Explanations

Gagan Bansal
Paul G. Allen School of Computer Science and Engineering
University of Washington
Seattle, WA 98195
{*bansalg*}*@cs.washington.edu*

Abstract

Adoption of AI systems in high-stakes domains (e.g., transportation, law, and health care) demands that human users trust these systems. A desiderata for establishing trust is that the users understand the system's decision process. However, a high-performing system may use a complex decision process, which may not be *interpretable* by itself. We argue that existing solutions for generating interpretable explanations have limitations and as a solution, propose developing new explanation systems that enable interactive and actionable dialogs between the user and the system.

Introduction

AI systems continue to improve in accuracy on a variety of perception tasks primarily due to the use of highly expressive machine learning (ML) models trained on parallel computing hardware (e.g., GPUs). Yet these systems struggle to earn their users trust: they fail to provide a human-interpretable rationale for their decisions, making it impossible for users to understand failures or to double-check the systems basis for a decision before adopting it. Human experts on the other hand, earn trust much more readily by engaging in open-ended discussions to explain their decision process and answering questions to reveal their decision-making rationale and limitations. Adopting AI systems for high-stakes applications, where their decisions can put human lives at risk, demands that these systems be accountable by explaining their reasoning. In recent work (Bansal and Weld 2018), we showed that the insight gained from explanations can be useful for discovering novel and a diverse set of high confidence mistakes in ML classifiers that may exists due to a difference in the training and test distribution.

Unfortunately, existing solutions suffer from important weaknesses. First, fully faithful explanations remain too complex for users to comprehend (Lipton 2016). Second, systems such as LIME (Ribeiro, Singh, and Guestrin 2016) and gradient-based explainers (Ross, Hughes, and Doshi-Velez 2017), which approximate local behavior are widely inaccurate when extended to other areas of test distribution, further decreasing user-system trust. Third, none of the systems let users put their newly gained (albeit limited) understanding to use. Overcoming these limitations lies in enabling a unique and previously untried method of explanatory dialogs to occur between users and AI systems.

Explanatory Dialogs

Explanatory dialogs helps humans understand AI system decisions. This new approach provides actions that help users gradually learn about a system's complex decision process and operate on it, for example, to improve its efficacy.

Dialogs and Actions

Explanations by themselves are of little use unless they empower a user to act, like ask a follow-up question, add new training data, correct an erroneous label in existing data, specify new constraints, or augment the explanation vocabulary etc. What follows are two of many possible actions that explanatory dialogs can support.

The ability to ask follow up questions is useful when an AI system uses a complicated reasoning process (e.g., non-linear decision boundary learned by a neural network) because a fully faithful explanation would be too complex for ready comprehension. It is practical for such dialogs to start simply by presenting a qualified explanation and to progressively let users drill down for more detail. For example, Figure 1a shows an image classification task, for which the system first explains its decisions in terms of an easily understood representation, such as super-pixels in the image. Since this explanation is an approximation, the user can then ask the system for more details about alternate representations e.g., smaller super-pixels or about the examples used to train the system (Koh and Liang 2017).

The effectiveness of a drill down procedure would, of course, depend on these representations. For example, instead of super-pixels computed using an automated algorithm (Ribeiro, Singh, and Guestrin 2016), a user may prefer a representation grounded in an intermediate representation used by the system, or an existing external knowledge base (Chen et al. 2015), or multiple modalities (Hendricks et al. 2016). A dialog-based system should also permit human-supplied representations by supporting the interactive definition of new terms and then presentation of new explanations in terms of an augmented vocabulary.

Another useful action *is ability to perturb the system or its inputs and analyze the resultant behavior*. To illustrate

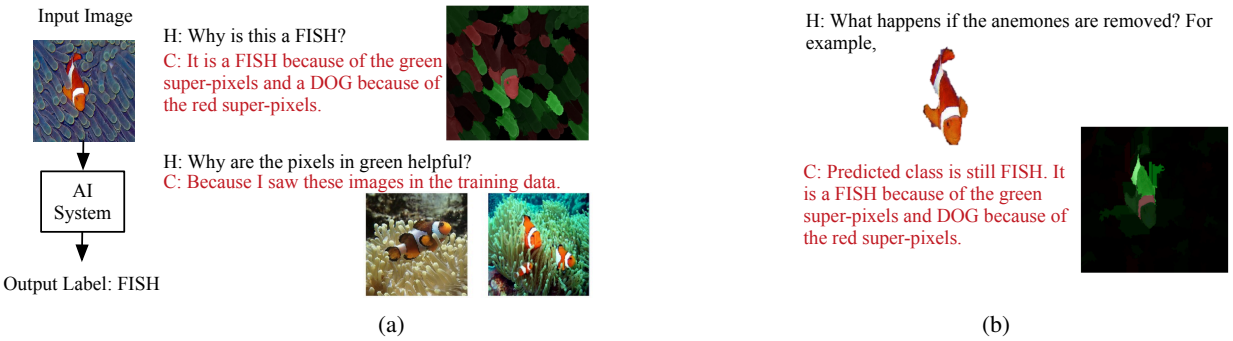

Figure 1: An example of an explanatory dialog for gaining insight into a dog/fish image classifier.

the importance of this action consider the following example. One possible conclusion from the drill down interaction shown in Figure 1a is that the anemones in the background heavily influence the systems decision to predict a fish, whereas the pixels that constitute the actual fish are not useful and, in fact, indicate the other class (dog). In contrast, Figure 1b shows that the explanation of a perturbed version where the anemones have been removed revealing that the system consequently relies on the pixels representing the fish. Without the second explanation, the user may gain an incomplete understanding of the system.

Contributions

- We have identified *actionability* as another desiderata for explainable AI. Much like faithfulness, supporting actionability would provide a critical missing element in making AI systems more explainable and therefore more trustworthy.
- We have defined a basic framework for supporting explanatory dialogs; specifically, the actions that it would support.
- We have implemented a preliminary code base for explanatory dialogs that allows the user to drill down by asking for explanations in terms of alternate representations or by perturbing an example using an image editor.

Finally, to effect a change, in addition to refining and implementing the approach of explanatory dialog and actions considerable work is required on a number of research questions closely related to this approach such as the meta-problem of evaluating trust in an explainer, learning with constraints discovered using explanatory dialog, developing new representations, evaluating effect of dialog on performance.

Previous work: Identifying Unknown Unknowns

A classifier's low confidence in prediction is often indicative of whether its prediction will be wrong; in this case, inputs are called *known unknowns*. In contrast, *unknown unknowns* (UUs) are inputs on which a classifier makes a high confidence mistake. Identifying UUs is especially important in safety-critical domains like medicine (diagnosis) and law (recidivism prediction). Previous work by Lakkaraju et al. (2017) on identifying unknown unknowns assumes that the utility of each revealed UU is independent of the others, rather than considering the set holistically. While this assumption yields an efficient discovery algorithm, we (Bansal and Weld 2018) argue in that it produces an incomplete understanding of the classifier's limitations. Experimental results on four datasets show that our method outperforms bandit-based approaches and achieves within 60.9% utility of an omniscient, tractable upper bound. While finding a UU is important, we currently focus on explaining *why* the classifier erred.

References

Bansal, G., and Weld, D. S. 2018. A coverage-based utility model for identifying unknown unknowns. In *Proc. of AAAI*.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S. K.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft COCO captions: Data collection and evaluation server. arxiv:1504.00325.

Hendricks, L. A.; Akata, Z.; Rohrbach, M.; Donahue, J.; Schiele, B.; and Darrell, T. 2016. Generating visual explanations. In *Proc. of ECCV*.

Koh, P. W., and Liang, P. 2017. Understanding black-box predictions via influence functions. In *Proc. of ICML*.

Lakkaraju, H.; Kamar, E.; Caruana, R.; and Horvitz, E. 2017. Identifying unknown unknowns in the open world: Representations and policies for guided exploration. In *Proc. of AAAI*.

Lipton, Z. C. 2016. The mythos of model interpretability. In *ICML Workshop on Human Interpretability in Machine Learning*.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. Why should I trust you?: Explaining the predictions of any classifier. In *Proc. of KDD*.

Ross, A. S.; Hughes, M. C.; and Doshi-Velez, F. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proc. of IJCAI*.