Abstraction, Validation, and Generalization for Explainable Artificial Intelligence

Scott Cheng-Hsin Yang[1], Tomas Folke[1], and Patrick Shafto[1]

[1]Rutgers University Newark

June 8, 2021

Abstract

Neural network architectures are achieving superhuman performance on an expanding range of tasks. To effectively and safely deploy these systems, their decision-making must to be understandable to a wide range of stakeholders. Methods to explain AI have been proposed to answer this challenge, but a lack of theory impedes the development of systematic abstractions which are necessary for cumulative knowledge gains. We propose Bayesian Teaching as a framework for unifying explainable AI (XAI) by integrating machine learning and human learning. Bayesian Teaching formalizes explanation as a communication act of an explainer to shift the beliefs of an explainee. This formalization decomposes any XAI method into four components: (1) the inference to be explained, (2) the explanatory medium, (3) the explainee model, and (4) the explainer model. The abstraction afforded by Bayesian Teaching to decompose any XAI method elucidates the invariances among them. The decomposition of XAI systems enables modular validation, as each of the first three components listed can be tested semi-independently. This decomposition also promotes generalization through recombination of components from different XAI systems, which facilitates the generation of novel variants. These new variants need not be evaluated one by one provided that each component has been validated, leading to an exponential decrease in development time. Finally, by making the goal of explanation explicit, Bayesian Teaching helps developers to assess how suitable an XAI system is for its intended real-world use case. Thus, Bayesian Teaching provides a theoretical framework that encourages systematic, scientific investigation of XAI.

Hosted file

`applied-ai-letter-submitted-version.pdf` available at https://authorea.com/users/418754/articles/525405-abstraction-validation-and-generalization-for-explainable-artificial-intelligence

Hosted file

`main.tex` available at https://authorea.com/users/418754/articles/525405-abstraction-validation-and-generalization-for-explainable-artificial-intelligence