Neural Response Time Analysis: XAI Using Only a Stopwatch

Eric Taylor[1], Shashank Shekhar[2], and Graham Taylor[2]

[1]Vector Institute
[2]University of Guelph

June 8, 2021

Abstract

How would you describe the features that a deep learning model composes if you were restricted to measuring observable behaviours? Explainable artificial intelligence (XAI) methods rely on privileged access to model architecture and parameters that is not always feasible for most users, practitioners, and regulators. Inspired by cognitive psychology research on humans, we present a case for measuring response times (RTs) of a forward pass using only the system clock as a technique for XAI. Our method applies to the growing class of models that use input-adaptive dynamic inference and we also extend our approach to standard models that are converted to dynamic inference post hoc. The experimental logic is simple: If the researcher can contrive a stimulus set where variability among input features is tightly controlled, differences in response time for those inputs can be attributed to the way the model composes those features. First, we show that RT is sensitive to difficult, complex features by comparing RTs from ObjectNet and ImageNet. Next, we make specific a priori predictions about RT for abstract features present in the SCEGRAM dataset, where object recognition in humans depends on complex intra-scene object-object relationships. Finally, we show that RT profiles bear specificity for class identity, and therefore the features that define classes. These results cast light on the model's feature space without opening the black box.

Hosted file