Title: Deep Neural Network Feature Attribution via Extended Influence Functions

With the advent of deep learning and widely available data, classification models have been able to achieve superhuman levels of accuracy. In areas of high-risk decisions, however, these models have yet to be convincingly implemented in practice. This lack of implementation is partially due to the black-box nature of these algorithms; it is difficult to place trust in a model when one does not know how it works. Several proposals have been made for interpreting the decisions of black-box classifiers; one such proposal is that of influence function approximation. From robust statistics, influence functions can be used to attribute changes in the loss function due to small perturbations in the input features. In this paper, we extend the influence function approximation to deep networks by computing gradients in an end-to-end manner. Under this approach, we pinpoint the changes in loss with respect to individual inputs. We demonstrate the effectiveness of the algorithm on the eICU dataset by comparing the features chosen by the algorithm to those chosen by traditional methods, state-of-the-art deep learning methods, and human experts. The features chosen by influence functions were more like those chosen by human experts than those chosen by all other methods and were computed in less time than the current state-of-the-art.

Author List:

Jacob Epifano (Primary)

Ravi Ramachandran

Sharad Patel

Ghulam Rasool (Senior)

Domain Conflicts:

???

Subject Areas (1 Primary, 3 Secondary)

Primary: (Robust Statistics and Machine Learning)

Secondary: (Accountability, Transparency and Interpretability), (Healthcare), (Supervised Learning)