Detecting and Debugging of Discrimination in Deep Neural Networks

Verya Monjezi
vmonjezi@miners.utep.edu
Department of Computer Science
University of Texas at El Paso

Saeid Tizpaz-Niari
saeid@utep.edu
Department of Computer Science
University of Texas at El Paso

Ashutosh Trivedi
ashutosh.trivedi@colorado.edu
Department of Computer Science
University of Colorado Boulder

Gang Tan
gtan@psu.edu
Department of Computer Science & Engineering
Pennsylvania State University

Abstract —The deep neural networks (DNNs) are increasingly deployed in socioeconomic critical decision support software systems. DNNs are exceptionally good at finding minimal, sufficient statistical patterns within their training data. Consequently, DNNs may learn to encode decisions---amplifying existing biases or introducing new ones---that may disadvantage protected individuals/groups and may stand to violate legal protections. While the existing search based software testing approaches have been effective in discovering fairness defects, they do not supplement these defects with debugging aids---such as severity and causal explanations---crucial to help developers triage and decide on the next course of action.
Can we measure the severity of fairness defects in DNNs? Can such measurements guide us to localize fairness defects in the internal of DNNs? To answer such questions, we present an information-theoretic testing and debugging framework.

The key goal of our approach is to assist DNN software developers in triaging fairness defects by ordering them by their severity. Towards this goal, we propose a quantitative notion of fairness via information-theoretic metrics such as Shannon entropy and min entropy. Hence, we can quantify fairness in terms of protected information (in bits) used in decision making. We present a search algorithm to characterize the amount of discrimination. After finding relevant inputs, we propose a debugging method that localizes layers and neurons with the maximum sensitivity to protected attributes. We showcase on multiple socially critical DNNs that our approach efficiently characterizes the biases, effectively generates discriminatory instances, and compactly localizes layers/neurons with significant biases.