Free, Public Talk

Attribution Based Confidence Metric for Neural Networks

Thursday, October 10, 2024

12:30 pm – 1:45 pm

Hixson-Lied Science Building, Room G-59

Abstract

Dr. Steven Fernandes and his research team proposed a novel confidence metric called the attribution-based confidence (ABC) metric for deep neural networks (DNNs). The ABC metric characterizes whether the output of a DNN on an input can be trusted. DNNs are known to be brittle on inputs outside their training distribution and are hence susceptible to adversarial attacks. This fragility is compounded by a lack of effectively computable measures of model confidence that correlate well with the accuracy of DNNs, impeding their adoption in high-assurance systems. The ABC metric addresses these challenges. It does not require access to the training data, the use of ensembles, or training a calibration model on a held-out validation set, making it usable even when only a trained model is available for inference. Dr. Fernandes and his team provided a mathematical basis for the proposed metric and evaluated its effectiveness. They studied the change in accuracy and the associated confidence over out-of-distribution inputs. The ABC metric appropriately indicated low confidence on out-of-distribution data and adversarial examples, aligning with the low accuracy observed in these scenarios.

Guest Speaker

Dr. Steven Fernandes is an Assistant Professor of Computer Science at Creighton University. He specializes in artificial intelligence, focusing on deep learning, computer vision, and medical image processing. His research primarily revolves around the development and application of novel AI methodologies to enhance the accuracy and reliability of neural network outputs. A key area of his work involves the creation of metrics, such as the attribution-based confidence metric, which assesses the trustworthiness of outputs from deep neural networks.

Hosted by the Creighton Physics Department

Creighton UNIVERSITY