Adoption of deep neural networks (DNNs) into safety-critical and high-assurance systems has been hindered by the inability of DNNs to handle adversarial and out-of-distribution input. State-of-the-art DNNs misclassify adversarial input and give high confidence output for out-of-distribution input. We attempt to solve this problem by employing two approaches, first, by detecting adversarial input and, second, by developing a confidence metric that can indicate when a DNN system has reached its limits and is not performing to the desired specifications. The effectiveness of our method at detecting adversarial input is demonstrated against the popular DeepFool adversarial image generation method. On a benchmark of 50,000 randomly chosen ImageNet adversarial images generated for CaffeNet and GoogLeNet DNNs, our method can recover the correct label with 95.76% and 97.43% accuracy, respectively. The proposed attribution-based confidence (ABC) metric utilizes attributions used to explain DNN output to characterize whether an output corresponding to an input to the DNN can be trusted. The attribution based approach removes the need to store training or test data or to train an ensemble of models to obtain confidence scores. Hence, the ABC metric can be used when only the trained DNN is available during inference. We test the effectiveness of the ABC metric against both adversarial and out-of-distribution input. We experimental demonstrate that the ABC metric is high for ImageNet input and low for adversarial input generated by FGSM, PGD, DeepFool, CW, and adversarial patch methods. For a DNN trained on MNIST images, ABC metric is high for in-distribution MNIST input and low for out-of-distribution Fashion-MNIST and notMNIST input.

Major: Computer Science

Educational Career:
Bachelor’s of Computer Science and Engineering, BS, 2011, Indian Institute of Technology, Dhanbad

Committee in Charge:
Sumit Kumar Jha, Chair, Computer Science
Gary T. Leavens, Computer Science
Sharma Valliyil Thankachan , Computer Science
Mengyu Xu, Statistics and Data Science

Approved for distribution by Sumit Kumar Jha, Committee Chair, on November 3, 2020.

The public is welcome to attend.