Abstract: Using Model Fingerprinting for the Detection of Out-of-Distribution Samples

Machine learning models are used to make predictions and draw conclusions about data from many different real-world sources today; however, it is natural for that data to drift over time. Therefore, it is necessary that models can detect out-of-distribution samples, either introduced accidentally or maliciously. Model fingerprinting determines whether any input to a model is of reasonable input based on the data used to train the model. The activations of the training data for each layer of a layered model are used to train sketches (e.g. autoencoder) for each layer, which are then used to calculate the reconstruction error by running the input data's activations through the autoencoder and determining the root mean square error between the input activations and output values. Unlike previous approaches to outlier detection, model fingerprinting does not assume prior knowledge of out-of-distribution data or require that the model is re-trained every time. By allowing both batchwise and layer-by-layer computations for all steps, this algorithm is memory-efficient and feasible for users to run on any reasonable machine.

To test the effectiveness of the model fingerprint in detecting outlier data, we use both other datasets and generated adversarial data. For the former, we evaluate the model fingerprinting method for a deep convolutional neural network (CNN) trained on the MNIST training dataset using the MNIST test dataset as inlier data and the FASHION-MNIST dataset as outlier data (and vice-versa), and a CNN trained on the CIFAR-10 training dataset using the CIFAR-10 test dataset as inlier data and the SVHN dataset as outlier data (and vice-versa). For adversarial image creation, a deep convolutional generative adversarial network (DCGAN) is trained to generate plausible MNIST and CIFAR-10 images using the entire training sets for the respective target datasets. The generated images are then labeled as outlier data and evaluated using the model fingerprinting algorithm for the corresponding dataset to see if they are correctly identified as out-of-distribution.

We found that the MNIST model fingerprinting method can detect 86% of inliers and 99% of FASHION-MNIST outliers, the FASHION-MNIST model can detect 85% of inliers and 91% of MNIST outliers, the CIFAR-10 model can detect 94% of inliers and 97% of SVHN outliers, and the SVHN model can detect 83% of inliers and 70% of CIFAR-10 outliers. We found that the GAN-generated adversarial images were identified as outliers 68% of the time for MNIST and 88% of the time for CIFAR-10. Because DCGAN is not a malicious method of generating outlier data, these metrics serve as a baseline for future evaluations of adversarial attack detection.

In the future, the model fingerprinting method will be evaluated with more realistic malicious attacks, such as DeepFool and the Fast-Gradient Signed Method (FGSM), which do not make use of all inlier data to generate images.