Transfer learning auto-encoder neural networks for anomaly detection of DDoS generating IoT devices

ABSTRACT

Machine Learning based anomaly detection ap-proaches have long training and validation cycles. With IoT devices rapidly proliferating, training anomaly models on a per device basis is impractical. This work explores the "transfer-ability"of a pre-trained autoencoder model across devices of similar and different nature. We hypothesized that devices of similar nature would have similar high level feature character-istics represented by the initial layers of the autoencoder, while the more distinct features are captured by the innermost layer of the neural network. In our experiments, the centre-most layers of autoencoder models were re-trained with limited new data belonging to a different device. Datasets of seven Mirai infected and nine Bashlite infected IoT devices were used; each dataset also included benign records representing un-infected behaviour. We observed that the model's detection accuracy improved by an average of 9.52% for Mirai and 44.59% for Bashlite. The highest performance improvement of 26.68% and 73.00% was observed when the anomaly model of Ecobee thermostat was tested on other devices before and after transfer learning for Mirai and Bashlite respectively. Additionally, transfer learning took 47.31% and 58.27% less time for Mirai and Bashlite respectively. We further trialed the efficacy of the autoencoder based anomaly model on flow based records of network traffic using the CIC-IDS2017 dataset. It was observed that the model performed best when distinct outliers in the dataset were present, whereas the model failed to perform decently in cases where the malicious activity did not cause significant deviation in network traffic's footprint.