Using Kleisli Morphisms for Malware Detection With Graph Convolution Networks

Colaresi, Landon (School: Pittsburgh Allderdice High School)

Cybersecurity is a major problem in today's world, and malware detection is critical defensive approach. Although numerous antivirus programs exist, the majority of them detect only exact or near-exact copies known malware samples. Machine learning methods for malware detection exist, but most either suffer from the limits of an engineered vector representation or the high computational cost of models that process every single byte. Firstly, I developed a representation for programs which is based around the concept of monadic I/O. Arising from the development of purely functional programming languages, monadic I/O acts as a way to describe a program in terms of both specific actions and arbitrary computations. Under this paradigm, a program can be represented as a graph of both known actions and the Kleisli composition operator (which is a specific formulation of how monadic actions can be composed). Graph Convolution Networks are a powerful type of neural network that accepts graphs as inputs. I used a relatively small GCN to transform these graphs into a prediction of if the sample is malware. I developed a pipeline that disassembles the given executable with radare2, parses this, builds the graph, and trains a machine learning model. I gathered a sample of 387 malicious programs and 387 safe programs, and compared a GCN using call graphs with a GCN using this representation. The new representation reached 83.116% test accuracy after 1,733 epochs. The call graph GCN failed to converge after 24,000 epochs, with a test accuracy of 51.948%.

Awards Won:

Third Award of $1,000

National Security Agency Research Directorate : First Place Award "Cybersecurity"