Network Intelligence and Security – Universal Intelligent Web Crawler

Team Members: Dr. Roberto Perdisci, Spencer King, Irfan Ozen, Karthika Subramani

UNIVERSITY OF GEORGIA

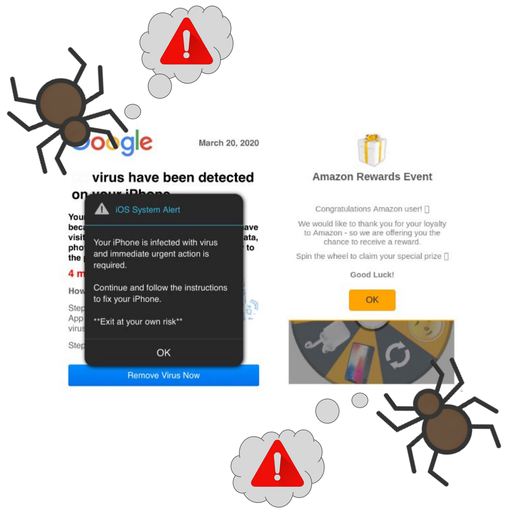
Institute for Cybersecurity and Privacy
UNIVERSITY OF GEORGIA

Award Number 2126641

Can a Computer Surf the Web Like a Human?

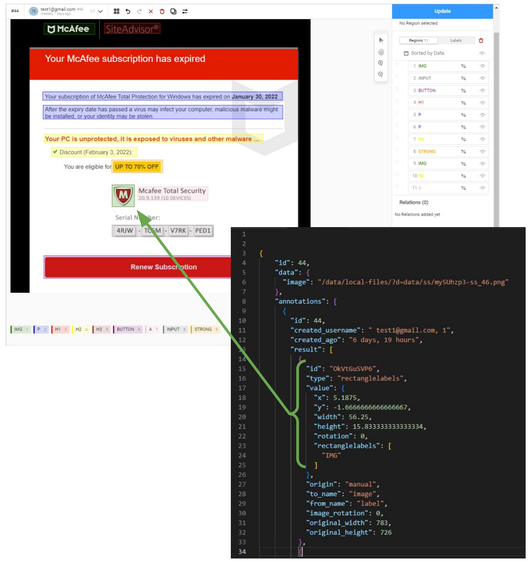
Motivation

- This project is intended to learn how to detect social engineering (SE) attacks in the wild. To do this many attack samples must be collected.
- To crawl through complex web pages (video streaming websites, for example) an intelligent crawler (IC) is needed that understands where to click, mimicking what a human would do.
- This IC can also accept to download advertised software or browser extensions to check if they are malicious or not.



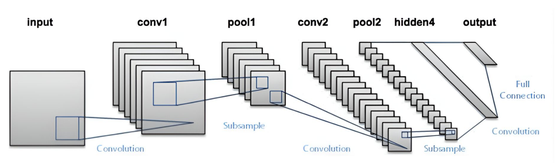
Label Data

- Once the important web page elements are finalized, each web page of interest must be collected as a screenshot. In addition, each element and its position on the page is recorded in the form of a JSON file.
- Each page element of interest must be labeled and recorded into the JSON file.
- Initial labeling is automatically generated, but data pipelining can be employed to allow humans to check and augment data labeling.



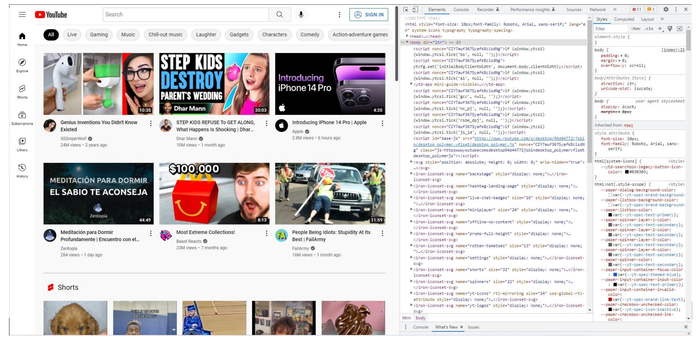
Leverage Deep Learning

- Once a sufficient quantity of data is gathered, deep learning algorithms can be leveraged to perform tasks such as object detection and object localization.
- The idea is to learn the context of web page elements to try to “see” the page as a human user would.



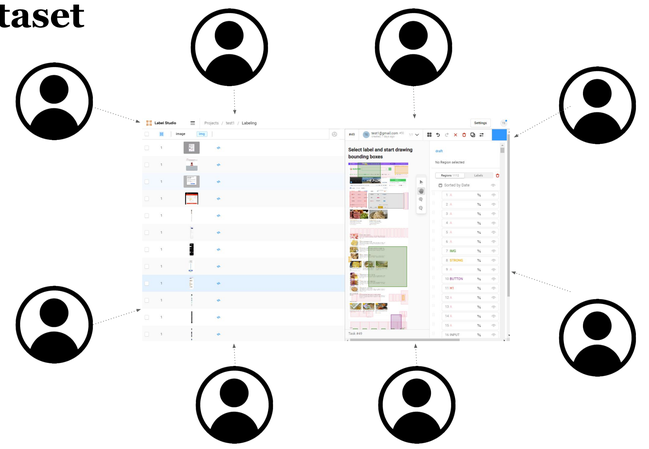
Determine Useful Web Page Elements

- Most modern web pages are made up of hundreds or even thousands of web page elements. Are some elements more important than others? Which ones does the web user interact with?
- These are important questions that must be answered in order to determine how web users behave and how they navigate through web pages.



Build A Large Scale Dataset

- Now that a pipeline to label the data is in place, it must be scalable so that thousands of labeled web pages can be generated from malicious and benign sources.
- To scale, an in-built web application has been developed to quickly allow users to verify and augment webpage labels.



Create an Intelligent Crawler

- Once the data data is analyzed, a full AI system can be put in place to allow a web crawler to interact with a given web page and decide what the next best action is.
- This would allow the crawler to autonomously browse the web with the same logic as a real human user.