Departamento de Automática
Escuela Politécnica Superior
Universidad de Alcalá

ANTEPROYECTO

Presentación de una IA capaz de detectar información oculta en imágenes basada en esteganografía

Marzo - 2022

Autor - ***Sergio Sastre Arrojo***
Director - ***Miguel Ángel Sicilia Urbán***

1. Introducción

En la actualidad existen una gran variedad de ciberataques, y para cada ciberataque hay, a su vez, diversas vías por las que explotarlo. Estas vías se conocen como exploits y, dependiendo del objetivo, se podrán ejecutar o no. Para este trabajo queremos enfocarnos en un exploit en bastante típico que se suele usar para que el atacante pueda espiar a la víctima, recoger su información, infectar su red con un malware... nos referimos a los del tipo "Drive-By Download", más concretamente los "Drive-By Download Browser": este tipo de exploits se centran en ejecutar código desde el propio navegador con el que vulnerar el sistema objetivo.

Existen muchas formas de aplicar este exploit, sin embargo, vamos a basarnos en una de ellas bastante peligrosa (en gran parte por la sutileza que sugiere); nos referimos a la esteganografía, es decir, el arte de la ocultación de información a simple vista. Usando este método, el atacante puede introducir un código malicioso en una misma imagen sin que el usuario se dé cuenta y que éste se ejecute para comprometer el sistema. Para aplicar esta técnica usaremos stegosploits de Javascript. [4]

Una solución que queremos presentar ante este tipo de infortunios es el uso de una IA lo suficientemente entrenada como para detectar estos códigos ocultos en imágenes, de forma que se pueda evitar, en la medida de lo posible, la infección del sistema.

2. Objetivos y campos de aplicación

El objetivo principal de este proyecto es la aplicación de una solución eficaz en la detección de este tipo de ataques. En esencia, se pretende ayudar al usuario para evitar males mayores.

Por otro lado, podemos destacar los siguientes campos de aplicación para nuestro trabajo:

1. Implementación de un sistema de detección de códigos en Javascript en imágenes.

2. Desarrollo de nuevas ayudas en la protección del usuario en páginas web.

En el siguiente apartado, explicaremos de forma breve el proceso que seguiremos para su realización.

3. Descripción del trabajo

A continuación vamos a pasar a explicar las fases sobre las que se va a desarrollar el proyecto. Para ello nos ayudaremos de un diagrama de bloques:

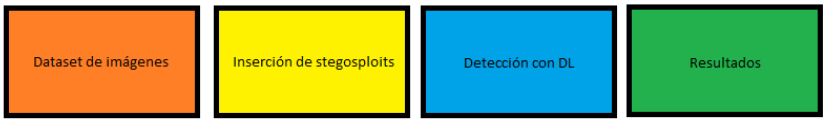

Figura 1: Fases del proyecto

Como se puede ver en la Figura, el trabajo se dividirá en 4 partes:

- **Dataset de imágenes**

  En primer lugar, cogeremos una base de datos con diferentes imágenes para insertar los códigos con los stegosploits. Buscaremos entre las ya conocidas Cityscapes, Mapillary y Coco. [1] [3] [2]

- **Inserción de stegosploits**

  A continuación, introduciremos los stegosploits de Javascript en las imágenes para poder trabajar con ellas en un modelo de Deep Learning.

- **Detección con DL**

En este punto, vamos a trabajar con modelos de Deep Learning para la detección de estos códigos. En primer lugar los entrenaremos debidamente con un set de imágenes para, posteriormente, ejecutar una fase de evaluación y valorar los resultados obtenidos. Previo a todo esto, analizaremos en profundidad los modelos de Deep Learning que definen el estado del arte y, en consecuencia, trabajaremos con los que mejor se adapten a nuestros objetivos.

- **Resultados**

  Por último, concluiremos el trabajo con el análisis de los resultados obtenidos por los modelos y su valoración final para saber con cuanta exactitud se han detectado correctamente los códigos.

4. Fases de desarrollo

Teniendo en cuenta el apartado anterior, vamos a pasar a explicar de forma más detallada las diferentes fases del proyecto, además de su calendarización:

1. Exploración de dataset de imágenes (1 semana)
2. Integración de stegosploits de Javascript en el dataset (1 mes)
3. Exploración de modelos de Deep Learning (1 semana)
4. Entrenamiento de los modelos (1 mes)
5. Realización de experimentos sobre el sistema para comprobar su funcionamiento (1 mes)
6. Realización de memoria en LaTeX (Se realizará a lo largo de todo el cuatrimestre)
7. Conclusiones (2 semanas)

5. Medios disponibles

Dispondremos de las siguientes herramientas:

1. Estación de trabajo con GPU (NVIDIA Tesla) para la realización de experimentos.
2. Acceso a Google Collab y otras nubes para entrenar modelos con GPUs.
3. Acceso a diferentes librerías de Deep Learning:
    - PyTorch
    - Scikit-learn
    - XGBoost
    - LightGBM

Referencias

[1] Cityscapes. The cityscapes dataset. https://www.cityscapes-dataset.com/. 3

[2] Tsung-Yi Lin, Genevieve Patterson, Matteo R.Ronchi, Yin Cui, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, and Larry Zitnik andPiotr Dollár. Coco. common objects in context. https://cocodataset.org/. 3

[3] Mapillary. Mapillary. https://www.mapillary.com/. 3

[4] Saumil Shah. Stegosploit. exploit delivery via steganography and polyglots. https://stegosploit.info/, June 2015. 2