Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Vicepresidencia Primera del Gobierno · Ministerio de Asuntos Económicos y Transformación Digital · Secretaría de Estado de Digitalización e Inteligencia Artificial
Plan de Recuperación, Transformación y Resiliencia
España | digital 2026
incibe_ Instituto Nacional de Ciberseguridad

Gonzalo de Francisco Rodríguez
presenta el resumen del trabajo

Técnicas Básicas Iberseguridad Cuantum

Detección de ciberataques mediante el uso de un modelo de procesamiento de lenguaje natural

L. Gutiérrez-Galeano,
J.J. Domínguez-Jiménez,
I. Medina-Bulo

Actas VIII JNIC, atlanTTic, 77-84, (2023).

28 de marzo de 2024

SERIE DE RESÚMENES EN ESPAÑOL

Palabras clave: Ciberataque, Detección, Fine-Tunning, Procesamiento de Lenguaje Natural, Red Neuronal, T5

Introducción

La ciberseguridad es un área que crece constantemente debido a la continua aparición de nuevos tipos de ataques informáticos, y la mejora de las herramientas y técnicas comunes para combatirlos sigue siendo insuficiente ante la capacidad de los ciberdelincuentes de ir siempre un paso por delante y desarrollar nuevas amenazas. En cuanto a la protección de la red de una organización, la aparición de nuevas ciberamenazas hace necesario el desarrollo de nuevos sistemas que no solo detecten, sino que también puedan inferir nuevos ataques. Una opción para el desarrollo de estos sistemas es el uso de distintas técnicas de inteligencia artificial para la detección de ciberataques, como pueden ser los árboles de decisión o incluso las redes neuronales profundas con algoritmos genéticos. Más concretamente el trabajo original se centraba en un modelo de Depp Learning basado en una red neuronal pre-entrenada para el procesamiento de lenguaje natural (NLP), consiguiendo mediante fine-tunning adaptar su esquema de clasificación para la detección de ciberataques. No se han encontrado referencias previas sobre modelos para la detección de ciberataques basados en el fine-tuning de otros ya existentes que habían sido entrenados para propósitos diferentes al de detección, y los los experimentos realizados en el trabajo original muestran una efectividad superior al 99 % en la detección de ciberataques.

1. Metodología

1.1. Modelo T5

La propuesta para la detección de ciberataques es el empleo de un modelo desarrollado para el PLN, más concretamente el modelo T5, de tipo “encoder-decoder”, es decir, que utiliza redes neuronales recurrentes para la predicción de problemas de secuencia a secuencia relacionados con la clasificación de tipos de palabras, la traducción o reducción de textos. El modelo empleado en este trabajo fue pre-entrenado mediante una mezcla de tareas de aprendizaje supervisado y no supervisado de tipo ”text-to-text”(tanto la entrada como la salida son textos) empleando el conjunto de datos C4, constituido por cientos de gigabytes de texto en inglés obtenido de internet.
El modelo T5 se encuentra disponible en distintos tamaños que varían en función de la cantidad de parámetros de los que conste. Para el desarrollo de los experimentos se seleccionó los tamaños del modelo que podían ser entrenados con la GPU disponible: “t5-small“ y “t5-base”, realizándo dos veces tanto el proceso de fine-tuning como el de experimentación, una por cada tamaño.

1.2. Dataset CIC–IDS2017

El conjunto de datos seleccionado en el trabajo original fue CIC–IDS2017. Consta de 2830743 paquetes de red con 79 características cada uno, y contiene el resultado de la captura del tráfico de cinco días de una red informática que simula el tráfico generado por interacciones humanas realistas, así como tráfico en segundo plano además de una selección de los ataques más típicos y actualizados que se pueden encontrar. Además este dataset contiene tráfico benigno (paquetes que no conforman un ataque) generado mediante el método sistemático llamado BProfile, como son ataques de fuerza bruta a servicio FTP, ataque de denegación de servicio (DoS), ataque Web o infiltración entre otros.

1.3. Proceso de fine-tuning

En este caso, el modelo T5, que había sido preentrenado con el conjunto C4, fue sometido a un proceso de “fine-tuning“ con el conjunto CID–IDS2017. Esta técnica consiste en la utilización de un modelo ya

preentrenado para resolver un problema, y que sirve como punto de partida para resolver otro tipo de problema distinto. Esto se consigue entrenando el modelo pre-entrenado con nuestro conjunto de datos, ajustando los pesos para resolver un nuevo tipo de problema. De esa manera, en lugar de partir con una configuración de pesos aleatoria de la red neuronal, partimos de una configuración de pesos enfocada a la resolución del problema que nos interesa.

1.4. Preprocesado de datos

Los procesos de fine-tuning requieren un preprocesamiento del dataset para que los datos puedan ser usados en dichos procesos y, por tanto, que sirvan como entrada del modelo generado para la detección de ciberataques.

1. Homogeneización de nombres de columnas: se eliminaron todos los espacios en blanco iniciales y finales, se reemplazaron todos los espacios en blanco por un guión bajo y se pasaron todas las letras a mayúsculas.
2. Eliminación de columnas con valores únicos: estas características no son útiles para la construcción de un modelo ya que no aportan información extra. Se eliminaron todas las columnas cuya desviación estándar era igual que 0, lo que quería decir que quería decir que todos los valores eran iguales.
3. Eliminación de columnas con una correlación superior a 0.95.
4. Eliminación de valores infinitos, vacíos y nulos por valores NaN. A continuación, se eliminaron las filas que contenían dicho valor.
5. Eliminación de filas duplicadas, las cuales no aportan mayor información.
6. Adaptación de los datos al modelo T5: hasta este punto se han eliminado un total de 309951 filas, siendo el resultado final un dataset con 48 columnas (características) y 2520792 filas (paquetes). Para conseguir que el modelo, que como entrada solo admite un parámetro correspondiente a una cadena de carácteres, admita los valores correspondientes a 47 características es necesario convertir todas las características a texto y concatenarlas secuencialmente, intercalando como separador la barra vertical —, para que el modelo pueda procesarlas adecuadamente y producir la salida esperada, es decir, el tipo de ataque.

2. Resultados

Ambos modelos fueron entrenados durante 10 épocas. Para el modelo t5-small se seleccionó la época 3 con una tasa de acierto del 97 %, mientras que para t5-base se escogió la época 1 con una tasa de acierto del 99.5 %. Ambos modelos ofrecen resultados similares en la detección de paquetes benignos y en ataques como “DDoS“ y “PortScan“. Sin embargo, en ciertos tipos de ataques, como “DoS Hulk“ y “FTP-Patator“, t5-small muestra predicciones menos precisas que t5-base. Por último, par algunos tipos de ataques como “Web attack - XSS“ o “Heartbleed“ se han reconocido muy pocos paquetes, lo que quiere decir que que el dataset contiene muy poca cantidad de paquetes asociados a estos tipos de ataques, mientras que en otros casos, si el porcentaje de paquetes clasificados es mayor que la cantidad de paquetes de los que disponemos para un tipo de ataque concreto, esto quiere decir que el exceso de paquetes se corresponde con la proporción de falsos positivos obtenidos, como puede ser el caso de “DoS slowloris“ y “DoS Slowhttptest“ para t5-small, o ”DoS GoldenEye”para t5-base.

3. Discusión/Conclusiones

Como hemos visto, el uso del modelo T5 ,diseñado originalmente para el procesamiento del lenguaje natural, ha obtenido resultados muy prometedores en la detección de ciberataques, con tasas de acierto cercanas al 100 % en los dos modelos empleados y con una fase de entrenamiento muy breve en ambos casos. Sin embargo, estos resultados todavía son mejorables con una mayor cantidad de datos existentes para ciertos tipos de ataques, y sería interesante implementar el sistema en una red de datos real, validarlo con otros conjuntos de tráficos de red, además de probarlo con otros tamaños disponibles del modelo T5.

4. Valoración del documento original

El documento presenta un experimento que emplea un modelo de Deep Learning, el T5, inicialmente diseñado para procesar lenguaje natural, para la detección de distintos ciberataques. Entrenado con datos específicos de ataques informáticos, logra una efectividad superior al 99 %. Aunque sugiere mejoras con más datos y pruebas en entornos reales, ofrece un prometedor enfoque en ciberseguridad mediante inteligencia artificial.

Referencias

· L. Gutiérrez Galeano, J.J. Domínguez Jiménez, I. Medina Bulo,*, “Detección de ciberataques mediante el uso de un modelo de procesamiento de lenguaje natural”,* , Actas de las VIII Jornadas Nacionales de Investigación en Ciberseguridad, Vigo (21 a 23 de junio de 2023).