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

Alicia Mayor Bal
presenta el resumen del trabajo

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).

10 de julio de 2024

SERIE DE RESÚMENES EN ESPAÑOL

Palabras clave: Ciberataque, Fine-tuning, PLN, Red neuronal, T5.

Introducción

La ciberseguridad es un área en constante crecimiento, por lo que es necesario investigar el diseño de estrategias que abarquen los **nuevos tipos de ciberataques** que aparecen y, de esta forma, poder proteger el activo más importante: la información.

Las herramientas existentes habitualmente hacen uso de redes neuronales diseñadas desde cero. En este estudio se propone hacer uso de redes neuronales preentrenadas, adaptando su esquema para detecección de ciberataques mediante técnicas de fine-tuning, que permiten modificar los pesos sin alterar la arquitectura inicial de la red.

1. Metodología

El modelo T5 [1] es un modelo que utiliza redes neuronales recurrentes para la predicción de problemas de secuencia a secuencia, preparado para resolver diferentes problemas de PLN.

Se parte de este **modelo T5**, pre-entrenado mezclando aprendizaje supervisado y no supervisado. Está diseñado para resolver diferentes tareas, pero ninguna relacionada directamente con problemas de detección de ataques. Es de tipo "text-to-text", es decir, se alimenta con texto y produce texto.

El **dataset** seleccionado es el CIC-IDS2017, que reúne los ataques más típicos en la actualidad.

El **preprocesado de datos** sigue las siguiente secuencia de tareas:

1. Homogeneizar nombres de columnas.
2. Eliminar columnas con valores únicos.
3. Eliminar columnas con alta correlación.
4. Eliminar valores infinitos, vacíos y nulos.
5. Eliminar filas duplicadas.
6. Adaptar los datos al modelo T5.

Para el **entrenamiento**, se carga el modelo preentrenado, se entrena con el dataset seleccionado y, mediante fine-tuning, se ajustan los pesos para que las salidas se ajusten lo máximo posible a las salidas esperadas.

Se han seleccionado dos tamaños de modelo como punto de partida: el **t5-small (60 millones de parámetros) y el t5-base (220 millones de parámetros)**, y, para cada uno, se llevan a cabo dos etapas de fine-tuning.

2. Resultados

Para evaluar la **eficacia** de los modelos, se comparan los valores de pérdidas del entrenamiento y validación, así como las tasas de acierto. En base a estos datos, representados en gráficas por épocas, se extraen conclusiones para ambos modelos.

2.1. Modelo t-small

Se observa la rapidez con la que aprende el modelo y, se pueden hacer dos claras divisiones por épocas.

De la época 0 a la 5 contienen modelos aceptables, con tasas de pérdida del conjunto de validación menores o iguales a las del conjunto de entrenamiento. Sin embargo, en las épocas de las 6 a la 9 se sobreajustan los datos, dando modelos inaceptables para clasificación de nuevos datos. Se puede destacar la **época 3 como la mejor**, con una tasa de acierto del 97 %.

Para los tipos de ataques DoS slowloris y DoS Slowhttptest se ha producido con este modelo cierta cantidad de falsos positivos.

2.2. Modelo t-base

El modelo aprende con una rapidez análoga al modelo t-small. Sin embargo, en este caso es tal la rapidez que sólo se consideran aceptables las épocas 0 y 1, ya que a partir de estas se sobreajustan los datos. Se destaca la **época 1 como la mejor**, con una tasa de acierto del 99,5 %. A pesar de que hay exceso de falsos positivos para algún tipo de ataque, cabe destacar que para los DoS GoldenEye, los falsos positivos son mínimos, con una cantidad inferior al 10 %.

3. Discusión/Conclusiones

Los resultados obtenidos para ambos modelos son **prometedores**, encontrándose los mejores modelos en etapas tempranas del proceso de fine-tuning. Además, los modelos usados son de tamaño inferior al **T5**, por lo que podrían llegar a mejorar usando tamaños más grandes del modelo.

Se ha conseguido modelos con un **alto porcentaje de acierto**, cercano al 100 %, aunque podría ser mejorable ya que existen algunos tipos de ataques que no han sido reconocidos debido a la escasez de datos para estos ataques.

4. Valoración del documento original

En este paper destaca la organización y buena explicación en la parte de metodología y desarrollo. Se desarrollan claramente los conceptos más importantes para entender el proceso de principio a fin. Además, cabe destacar la subsección del preprocesado de datos, donde, a diferencia de otros estudios similares, se explica paso por paso la secuencia de tareas realizadas sobre los datos para obtener un dataset apropiado.

Referencias

[1] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li and P.J. Liu, .ᴱˣploring the limits of transfer learning with a unified text-to-text transformer", CoRR, vol. 1910.10683, 2020.