CEA Comité Español de Automática

XLV JORNADAS AUTOMÁTICA MÁLAGA 2024

UNIVERSIDADE DA CORUÑA

Jornadas de Automática

Clasificación de capturas de smishing con aprendizaje profundo e IRIS

Blanco-Medina, P.[a,b,*], Carofilis, A.[a,b], Fidalgo, E.[a,b], Alegre, E.[a,b]

[a]*Departamento de Ingeniería Eléctrica y de Sistemas y Automática, Universidad de León, Campus de Vegazana, 24007, Leon, España*
[b]*Investigador Colaborador en INCIBE*

To cite this article: Blanco-Medina, P., Carofilis, A., Fidalgo, E., Alegre, E. 2024. Smishing Screenshots Classification using Deep Learning and IRIS Dataset.
Jornadas de Automática, 45. https://doi.org/10.17979/ja-cea.2024.45.10904

Resumen

El Smishing es una variante del Phishing que utiliza el Servicio de Mensajes Cortos, los smartphones y la confianza de los usuarios en los servicios de mensajería como herramientas de comunicación para poder llevar a cabo actividades maliciosas. Los usuarios suelen informar de estos mensajes a los Equipos de Respuesta ante Emergencias Informáticas a través de capturas de pantalla de sus teléfonos. Estos equipos pueden beneficiarse de una herramienta que clasifique las capturas de pantalla en distintas categorías, antes de extraer su contenido. Comparamos el rendimiento de Redes Neuronales Convolucionales y Vision Transformers, pre-entrenados en conjuntos de datos como ImageNet, para clasificar estas capturas de smishing en dos categorías: texto dividido en múltiples líneas y texto unido. Publicamos un nuevo conjunto de datos, IRIS-244, que contiene 244 capturas de pantalla de mensajes Smishing con URLs de phishing. Combinando estas arquitecturas con técnicas de augmentación, descubrimos que Xception es la arquitectura con el mejor rendimiento, con una precisión media de 78,36.

Palabras clave:
Seguridad, Aprendizaje Profundo, Apoyo a Operadores Humanos, Redes Sociales, Automatización para la Ayuda Internacional

Smishing Screenshots Classification using Deep Learning and IRIS Dataset

Abstract

Smishing, a variant of phishing that uses the Short Message Service, uses smartphones and the trust of people in text messaging as a communication tool to spread more easily. When citizens report these suspicious messages to Computer Emergency Response Teams, they usually do it through a screenshot of their smartphone. Response Teams may find useful an automatic tool that classifies Smishing into different categories before proceeding to further information extraction. We propose to compare the performance of customized Convolutional Neural Networks and Vision Transformers with their pre-trained versions on ImageNet datasets for automatically classifying smishing screenshots into two different categories: joint and separate text. We make publicly available a novel dataset, IRIS-244, containing 244 smishing screenshots with phishing URLs. Combined with data augmentation techniques, we discovered that Xception architecture outperforms the rest of the approaches, with an accuracy score of 78,36.

Keywords: Security, Deep Learning, Human operator support, Social networking, Control and Automation Systems for International Aid

1. Introducción

A pesar del enfoque en seguridad y comunicación de servicios oficiales, los servicios de mensajería cortos (SMS) pueden ser víctimas de ciberataques conocidos como Smishing (SMS + Phishing). A través de este tipo de estafa, los usuarios son incitados a interactuar con un enlace malicioso recibido mediante mensajes instantáneos, a partir del cual el atacante

*Autor para correspondencia: pblanm@unileon.es

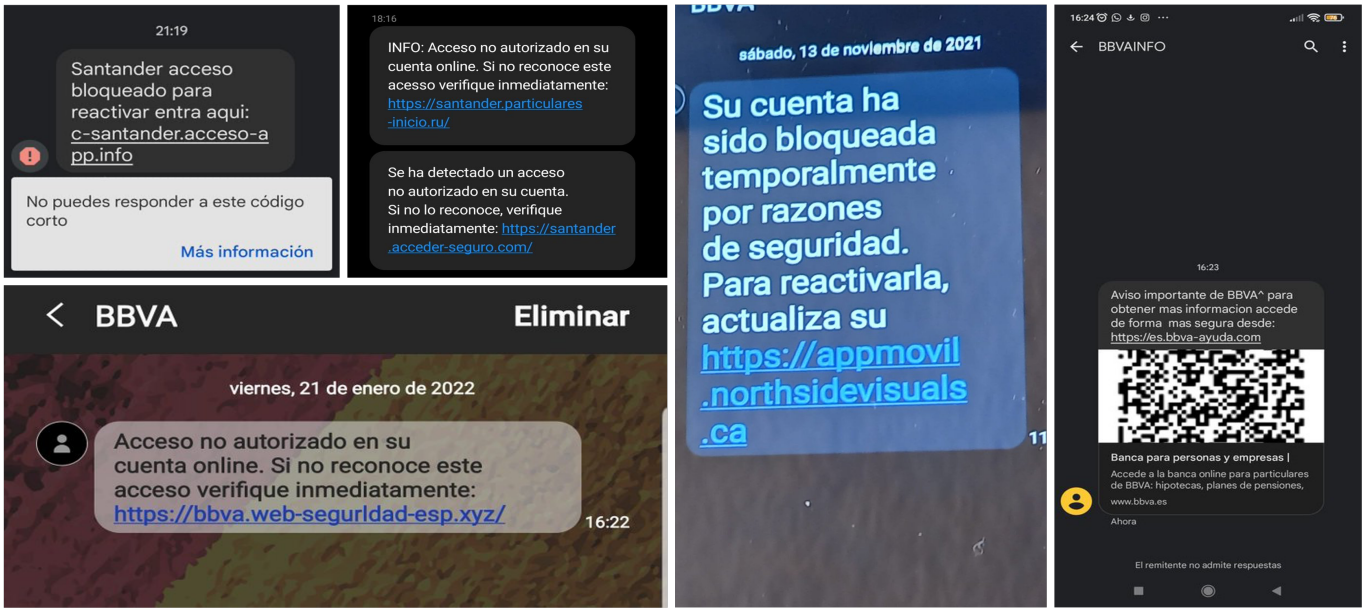

Figura 1: Ejemplos de mensajes Smishing, con múltiples tipos de texto que dificultan su clasificación en categorías y la extracción automatizada de su contenido. Se presentan imágenes con URLs en una única línea y URLs divididas en múltiples líneas, lo cual complica su extracción.

puede robar sus credenciales (Ulfath et al., 2022).

Para evitar este tipo de estafas, los usuarios pueden utilizar aplicaciones de filtrado de Smishing, utilizar listas blancas de contactos, o contratar servicios adicionales de filtrado (Akande et al., 2023). Para difundir este tipo de estafas entre usuarios no expertos, o llevarlos a la atención de organismos oficiales, estos mensajes pueden compartirse mediante la toma de una captura de pantalla en el equipo que los recibe, a través de las cuales no hay riesgo de interacción con el enlace malicioso.

Los Equipos de Respuesta ante Emergencias Informáticas (CERTs en Inglés) buscan alertar con la máxima brevedad posible acerca de campañas activas de Smishing, pero debido al cambio constante en las técnicas de envíos masivos e ingeniería social, los usuarios pueden no ser alertados a tiempo (Rahman et al., 2023).

Recuperar la información de estos mensajes Smishing es sencillo si se dispone del contenido textual ya extraído. Sin embargo, para aquellas campañas compartidas a través de capturas de pantalla, es necesario anotar manualmente el contenido de estos mensajes, lo cual puede resultar muy costoso en cuanto al tiempo que se requiere invertir, debido a la frecuencia de estas campañas, reflejando la falta de conjuntos de datos bien documentados (Mishra and Soni, 2023). La Figura 1 presenta ejemplos de mensajes de Smishing, cuya clasificación puede resultar de interés a los Equipos de Respuesta ante Emergencias Informáticas.

Para automatizar esta tarea pueden utilizarse técnicas de Visión Artificial, de manera similar a la extracción de información en documentos (Coquenet et al., 2023). Sin embargo, en el contexto de mensajes smishing, la tarea de recuperación de texto puede resultar compleja, debido a factores como la disposición de texto según el sistema operativo, letras y tamaños personalizados, u otras modificaciones realizadas a la imagen para proteger la privacidad del usuario.

La rápida clasificación de estas imágenes en categorías bien definidas resulta útil en el ámbito de la ciberseguridad y prevención de Smishing. Una vez categorizadas, puede elegirse el método más apropiado para cada disposición de texto particular, extrayendo la información del mensaje y localizando URLs potencialmente peligrosas.

En este trabajo, proponemos el uso de Visión Artificial, aplicando Aprendizaje Profundo, en la tarea de clasificación de imágenes para dividir capturas de SMS en dos categorías, aquellas con URLs divididas a lo largo del texto y aquellas que contienen URLs en una única línea. Detectar y reconocer una cadena de caracteres dividida en múltiples líneas mediante mecanismos de reconocimiento óptico de caracteres (OCR) es más complejo que una sóla línea, por lo que la clasificación de imágenes en estas categorías nos permite seleccionar los métodos más apropiados para recuperar toda la información localizada en cada imagen.

Estudiamos tanto la aplicación de redes neuronales como de Transformers en un conjunto de datos generado a partir de 244 capturas de pantalla de imágenes Smishing, denominado IRIS-244 (smIshing scReenshots wIth urlS). Analizamos diversas arquitecturas, incluyendo redes neuronales y Transformers, entrenadas desde cero o preentrenadas en grandes conjuntos de datos, y seleccionamos la que mejor rendimiento obtiene en nuestro problema.

El resto del artículo se organiza de la siguiente forma: en la Sección 2 revisamos brevemente trabajos similares aplicados la clasificación de imágenes en Smishing. En la Sección 3 detallamos la metodología de nuestra experimentación, presentando los resultados obtenidos en la Sección 4. Finalmente, la Sección 5 concluye el artículo detallando nuestras conclusiones y futuras líneas de trabajo.

2. Trabajos Relacionados

La tarea de clasificación de imágenes consiste en asociar una o varias etiquetas a una imagen, según los contenidos visuales analizados en ella (Fidalgo et al., 2018). Tradicional-

mente, se han utilizado redes neuronales para extraer características y clasificar imágenes en distintas categorías etiquetadas manualmente, pudiendo aplicarse al campo de la Ciberseguridad (Blanco-Medina et al., 2021). Estas arquitecturas se preentrenan en conjuntos de datos de gran tamaño como, por ejemplo, ImageNet y sus múltiples versiones (Recht et al., 2019), pudiendo ser adaptadas para conjuntos de datos específicos posteriormente (Gangwar et al., 2021).

Desde su introducción en el campo del Procesamiento del Lenguaje Natural (NLP), los Transformers han sido adaptados al contexto de Visión Artifical, tomando el nombre de Vision Transformers (Dosovitskiy et al., 2020), siendo comparables o incluso superiores en diversos campos a las redes neuronales clásicas (Bai et al., 2021). De manera simultánea, su desarrollo ha motivado una retro-alimentación de las redes neuronales, motivando el desarrollo de versiones mejoradas y más eficientes (Liu et al., 2022).

Existen diversos estudios que comparan el rendimiento de los Transformers con redes neuronales como ResNet y VGG (Maurício et al., 2023), desde la precisión en la tarea de clasificación de imágenes hasta el coste computacional de sus implementaciones. Entre los estudios revisados, no se han encontrado trabajos que se centren en la clasificación de imágenes de Smishing. Los trabajos más cercanos estudian la aplicación de Transformers en el contexto del análisis del texto en Smishing (Kumarasiri et al., 2023), no en las dificultades de su extracción ni en las múltiples disposiciones del texto.

3. Metodología

3.1. IRIS-244

Para realizar la experimentación, hemos construido un conjunto de datos de 244 imágenes de Smishing, recopiladas con ayuda del Instituto Nacional de Ciberseguridad (INCIBE), y que contienen un total de 262 URLs. Tras revisar las imágenes, analizando la distribución del texto y tipos de URLs encontradas, dividimos nuestro conjunto de datos en dos categorías finales: URLs divididas y URLs completas, conteniendo un total de 117 y 127 imágenes respectivamente. Estas categorías nos permiten distinguir componentes como la longitud o el tamaño del texto, antes de seleccionar el método de recuperación de texto más adecuado a cada tipo de imagen. Denominamos a este conjunto de datos como IRIS-244 (smIshing scReenshots wIth urlS).

Nuestro conjunto de datos contiene una cantidad inferior de imágenes respecto de otros conjuntos de datos del estado del arte (Recht et al., 2019). Para mitigar esta desventaja, utilizamos técnicas de augmentación de datos (Shorten and Khoshgoftaar, 2019). Tomando las imágenes de nuestro conjunto de datos, aplicamos zoom y rotaciones aleatorias, entre otras técnicas, para aumentar el tamaño del conjunto de datos de entrenamiento. La Figura 2 presenta gráficamente el resultado de incrementar nuestro conjunto de datos utilizando augmentación.

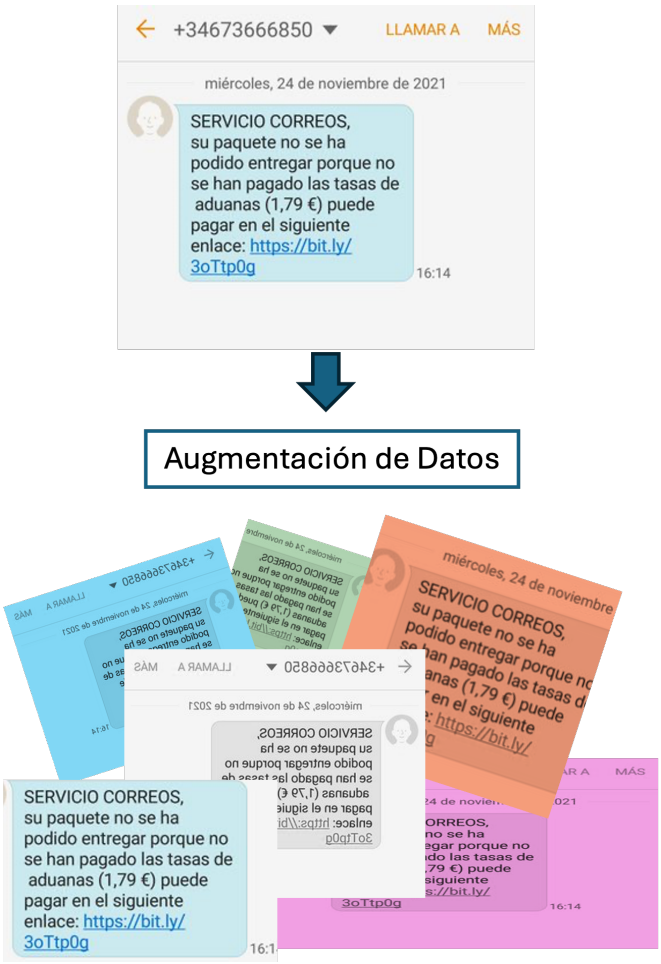

Figura 2: Augmentación de datos aplicada a mensajes de Smishing. Aumentamos el conjunto de datos inicial mediante rotación y zooms en las imágenes de IRIS-244.

3.2. Arquitecturas

Con el objetivo de estudiar la mejor arquitectura para el problema de Smishing, seleccionamos arquitecturas de aprendizaje profundo. Elegimos las redes neuronales Xception (Chollet, 2017) y ConvNext (Liu et al., 2022) basándonos en sus resultados en la tarea de clasificación y novedad de la arquitectura respectivamente. En el caso de esta última, seleccionamos tanto el modelo base como un modelo “large” para comparar su rendimiento en la clasificación de Smishing. Comparamos su precisión frente a los Vision Transformers, implementando todas las arquitecturas mediante la librería Keras (Chollet et al., 2015).

En el caso de los Transformers, implementamos los modelos preentrenados de VIT-Keras [1]. Para el Vision Transformer personalizado, definimos una arquitectura compuesta de capas Multi-Atención aplicadas a la secuencia de parches divididos de la imagen, el cual separa la imagen en secuencias, con un total de 196 parches por imagen. Esta información se procesa en bloques antes de clasificarse en una de las dos categorías definidas. La Figura 3 presenta una visualización del procesado de la imagen mediante Transformers.

[1] https://github.com/faustomorales/vit-keras

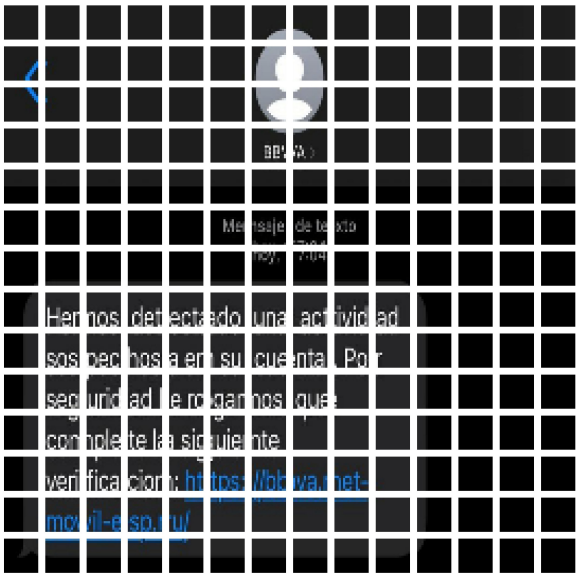

Figura 3: Representación gráfica de una imagen del conjunto de datos IRIS-244 antes de su procesado en el Transformer propuesto.

Para la arquitectura CNN personalizada, tras revisar arquitecturas utilizadas en problemas similares, diseñamos una red que contiene una capa de re-escalado, tres capas convolucionales seguidas cada una de una capa maxpooling, finalizando la arquitectura con dos capas densas. En el caso de las redes preentrenadas, sustituimos las últimas capas con los pesos ya entrenados por una capa de Global Average Pooling seguida de una capa densa, tras la cual se obtiene el resultado final. La Figura 4 presenta un resumen de la arquitectura definida en nuestra experimentación.

4. Experimentación

4.1. Configuración

Toda nuestra experimentación ha sido realizada en un contenedor con 128 GB de RAM, dos procesadores Intel Xeon E5-2630v3 de 2,4GHz, y dos GPUs NVIDIA Titan X.

Los parámetros utilizados para entrenar tanto las redes neuronales como los Transformers fueron 50 epochs, con un batch size de 32, un learning rate de 0,001,y un weight decay de 0,0001. Todas las imágenes se re-escalan a dimensiones de 224×224 antes de iniciar los modelos de clasificación. Entrenamos las arquitecturas elegidas utilizando validación cruzada con K-folds, con un tamaño de K igual a 5, el cual divide nuestro conjunto de datos en cinco grupos para evitar un sobreajuste. Reportamos tanto la precisión individual por fold como la precisión media de todos los métodos entrenados (Pal and Patel, 2020).

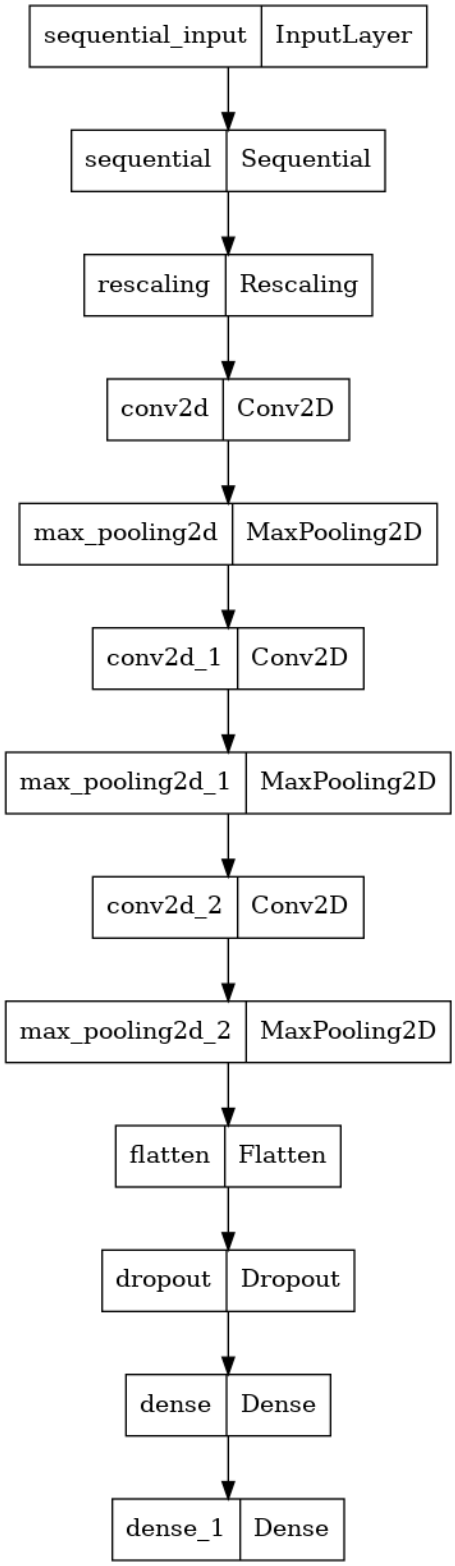

Figura 4: Arquitectura propuesta de red Neuronal Convolucional para clasificar las URLs del conjunto de datos IRIS-244.

4.2. Resultados

En la Tabla 1 presentamos los resultados de la experimentación con CNNs y Transformers personalizados. Observamos que los Transformers obtienen un mejor resultado medio en nuestro conjunto de datos, con una precisión media de 57,79 %, mientras que la red neuronal diseñada obtuvo un 53,32 %.

Destacamos también su menor valor de pérdida con 0,67 frente al 0,77 de la red neuronal, indicando mayor optimización. Aun así, la configuración de la CNN fue el modelo con la mejor precisión utilizando K-Fold, con un 62,50 % en el fold número 5, superando la mejor precisión de 61,22 % obtenida por el Transformer.

Tabla 1: Resultados de Transformers y Redes Neuronales customizadas para la clasificación de Imágenes Smishing. Indicamos tanto el valor de pérdida final (L) como la precisión (P) de cada uno de los 5 Folds en la fase de entrenamiento, presentando además los valores medios de cada arquitectura.

Fold	Transformer		Red Neuronal	
	L	P	L	P
F1	1,11	53,06 %	0,73	44,90 %
F2	0,65	61,22 %	0,87	53,06 %
F3	0,54	61,22 %	0,66	55,10 %
F4	0,55	55,10 %	0,87	51,02 %
F5	0,50	58,33 %	0,75	**62,50 %**
Media	0,67	**57,79 %**	0,78	53,32 %

En las Tablas 2 y 3 presentamos los resultados obtenidos utilizando redes neuronales y Transformers preentrenados en conjuntos de datos de clasificación de imágenes, como ImageNet.

Tabla 2: Resultados de valores de pérdida (L) y precisión (P) de redes neuronales preentrenadas en ImageNet y ajustadas en nuestro dataset IRIS-244.

Fold	Xception		ConvNextTiny		ConvNextT	
	L	P	L	P	L	P
F1	1,80	71,43 %	0,69	55,10 %	0,79	46,94 %
F2	3,82	61,22 %	0,73	53,06 %	0,65	61,22 %
F3	0,86	75,51 %	0,47	69,39 %	0,48	73,47 %
F4	0,40	85,71 %	0,46	81,63 %	0,50	73,47 %
F5	0,07	**97,92 %**	0,55	75,00 %	0,33	91,67 %
Media	1,39	**78,36 %**	0,58	66,84 %	0,55	69,35 %

Entre las redes neuronales implementadas, observamos que la arquitectura con el mejor rendimiento para nuestro problema es Xception, que obtuvo una precisión de 97,92 % en el entrenamiento 5-Fold, así como una precisión media de 78,36 %, superando los resultados obtenidos por ambas configuraciones de ConvNext. Podemos atribuir este rendimiento a una arquitectura más adaptada a Transfer Learning, reduciendo el impacto que un dataset reducido como el nuestro puede tener sobre la tarea de clasificación de imágenes.

Tabla 3: Resultados de valores de pérdida (L) y precisión (P) de Transformers pre-entrenados en modelos de distintos tamaños (Large y Base) y ajustados en nuestro dataset IRIS-244.

Fold	Large		Base	
	L	P	L	P
F1	0,77	53,06 %	0,87	51,02 %
F2	0,74	51,02 %	0,75	51,02 %
F3	0,92	51,02 %	0,74	65,31 %
F4	0,68	55,10 %	0,55	75,51 %
F5	0,66	47,92 %	0,47	**89,58 %**
Media	0,75	51,62 %	0,68	**66,49 %**

En el caso de los Transformers, observamos que el modelo base obtiene mejores resultados en nuestro conjunto de datos, con una precisión media del 66,49 %, siendo superior a la aplicación del modelo “large”, que obtuvo un 51,62 %, sugiriendo que el modelo de mayor tamaño no resulta ideal en un conjunto de datos más reducido, como lo es el del contexto del Smishing.

El modelo Base también obtuvo el mejor resultado de precisión individual,con una precisión total de 89,58 %, superando al mejor caso del modelo “large”, con un 53,06 %. Tambiñen obtuvo mejores puntuaciones en su función de pérdida, con un 0,68 % de valor medio frente al 0,75 del modelo “large”.

Tras analizar el entrenamiento de redes neuronales y Transformers, tanto personalizados como preentrenados en ImageNet, concluimos que la mejor arquitectura en nuestro caso es la red neuronal Xception, obteniendo los mejores resultados de clasificación de imágenes Smishing en categorías, destacando también su rápida implementación, lo que resulta ideal en el contexto de rápida clasificación en CERTs y sistemas de tiempo real.

5. Conclusiones

En este trabajo hemos analizado la tarea de clasificación de imágenes Smishing acorde a los criterios de URLs divididas o completas, siguiendo la distribución del texto en capturas de pantalla de mensajes SMS, con el objetivo de asistir a los Equipos de Respuesta ante Emergencias Informáticas en la tarea de detección y prevención de campañas de Smishing.

Nuestro objetivo ha consistido en desarrollar un clasificador de imágenes Smishing que permita identificar la distribución de texto en imágenes que contengan URLs divididas o completas, para su adaptación al método de recuperación de información más adecuado según la categoría asociada. Para ello, hemos generado un conjunto de datos, denominado IRIS-244, compuesto por 244 imágenes separadas en las categorías de URLs completas y divididas, con un total de 117 y 127 imágenes respectivamente.

Tras seleccionar varias arquitecturas de redes neuronales y Transformers en el estado del arte, hemos analizado los resultados de la tarea de clasificación de imágenes en nuestro conjunto de datos. Utilizando técnicas de augmentación de datos, customización de parámetros y entrenamiento con 5-Fold, concluimos que la arquitectura mejor adaptada a nuestro problema es la red neuronal Xception, al obtener una precisión media de 78,36 % en su versión preentrenada en ImageNet.

Hemos realizado una comparación entre redes neuronales y Transformers personalizados, tanto en términos del valor de pérdida por entrenamiento como de su precisión global. Destacamos la mejor precisión media de los Transformers sobre las redes neuronales, con un 57,79 % sobre un 53,32 %, respectivamente. Estas puntuaciones podrían incrementarse con un diseño más adaptado a la disposición de texto en imágenes Smishing, analizando componentes adicionales como la fuente, el tamaño o la distribución del texto encontrado.

En futuros trabajos, buscaremos desarrollar una lógica más completa de clasificación de texto, que permita tanto incrementar la precisión final obtenida como asisitir a futuros pasos de posprocesado para la recuperación completa de la información localizada en imágenes Smishing. La implementación de estos sistemas en entornos de ciberseguridad resulta

de gran ayuda a la hora de detectar, prevenir e informar de potenciales campañas de Smishing.

Agradecimientos

Este trabajo ha sido realizado gracias al Plan de Recuperación, Transformación y Resiliencia, financiado por la Unión Europea (Next Generation) gracias al Proyecto LUCIA (Lucha contra el Cibercrimen utilizando Inteligencia Artificial) concedido por INCIBE a la Universidad de León.

Referencias

Akande, O. N., Gbenle, O., Abikoye, O. C., Jimoh, R. G., Akande, H. B., Balogun, A. O., Fatokun, A., 2023. Smsprotect: An automatic smishing detection mobile application. ICT Express 9 (2), 168–176.

Bai, Y., Mei, J., Yuille, A. L., Xie, C., 2021. Are transformers more robust than cnns? Advances in neural information processing systems 34, 26831–26843.

Blanco-Medina, P., Fidalgo, E., Alegre, E., Vasco-Carofilis, R. A., Janez-Martino, F., Villar, V. F., 2021. Detecting vulnerabilities in critical infrastructures by classifying exposed industrial control systems using deep learning. Applied Sciences 11 (1), 367.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1251–1258.

Chollet, F., et al., 2015. Keras. `https://keras.io`.

Coquenet, D., Chatelain, C., Paquet, T., 2023. Dan: a segmentation-free document attention network for handwritten document recognition. IEEE transactions on pattern analysis and machine intelligence.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Fidalgo, E., Alegre, E., Gonzalez-Castro, V., Fernández-Robles, L., 2018. Boosting image classification through semantic attention filtering strategies. Pattern Recognition Letters 112, 176–183.

Gangwar, A., González-Castro, V., Alegre, E., Fidalgo, E., 2021. Attm-cnn: Attention and metric learning based cnn for pornography, age and child sexual abuse (csa) detection in images. Neurocomputing 445, 81–104.

Kumarasiri, W., Siriwardhana, M., Suraweera, S., Senarathne, A., Harshanath, S., 2023. Cybersmish: A proactive approach for smishing detection and prevention using machine learning. In: 2023 7th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC). IEEE, pp. 210–217.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986.

Maurício, J., Domingues, I., Bernardino, J., 2023. Comparing vision transformers and convolutional neural networks for image classification: A literature review. Applied Sciences 13 (9), 5521.

Mishra, S., Soni, D., 2023. Dsmishsms-a system to detect smishing sms. Neural Computing and Applications 35 (7), 4975–4992.

Pal, K., Patel, B. V., 2020. Data classification with k-fold cross validation and holdout accuracy estimation methods with 5 different machine learning techniques. In: 2020 fourth international conference on computing methodologies and communication (ICCMC). IEEE, pp. 83–87.

Rahman, M. L., Timko, D., Wali, H., Neupane, A., 2023. Users really do respond to smishing. In: Proceedings of the Thirteenth ACM Conference on Data and Application Security and Privacy. pp. 49–60.

Recht, B., Roelofs, R., Schmidt, L., Shankar, V., 2019. Do imagenet classifiers generalize to imagenet? In: International conference on machine learning. PMLR, pp. 5389–5400.

Shorten, C., Khoshgoftaar, T. M., 2019. A survey on image data augmentation for deep learning. Journal of big data 6 (1), 1–48.

Ulfath, R. E., Sarker, I. H., Chowdhury, M. J. M., Hammoudeh, M., 2022. Detecting smishing attacks using feature extraction and classification techniques. In: Proceedings of the International Conference on Big Data, IoT, and Machine Learning: BIM 2021. Springer, pp. 677–689.