Adversarial image generation using genetic algorithms with black-box technique

Gabriela Pérez[1,2] and Claudia Pons[1,3,4]

[1] LIFIA, Laboratorio de Investigación y Formación en Informática Avanzada, Facultad de Informática, Universidad Nacional de La Plata, Buenos Aires, Argentina
[2] UNAJ, Instituto de Ingeniería y Agronomía, Universidad Nacional Arturo Jauretche, Florencio Varela,Buenos Aires, Argentina
[3] UAI, Universidad Abierta Interamericana, Ciudad de Buenos Aires
[4] CIC, Comisión de Investigaciones Científicas, Buenos Aires, Argentina
{gperez,cpons}@lifia.info.unlp.edu.ar

Abstract. Convolutional neural networks are a technique that has demonstrated great success in computer vision tasks, such as image classification and object detection. Like any machine learning model, they have limitations and vulnerabilities that must be carefully considered for safe and effective use. One of the main limitations lies in their complexity and the difficulty of interpreting their internal workings, which can be exploited for malicious purposes. The goal of these attacks is to make deliberate changes to the input data in order to deceive the model and cause it to make incorrect decisions. These attacks are known as adversarial attacks. This work focuses on the generation of adversarial images using genetic algorithms for a convolutional neural network trained on the MNIST dataset. Several strategies are employed, including targeted and untargeted attacks, as well as the presentation of interpretable and non-interpretable images that are unrecognizable to humans but are misidentified and confidently classified by the network. The experiment demonstrates the ability to generate adversarial images in a relatively short time, highlighting the vulnerability of neural networks and the ease with which they can be deceived. These results underscore the importance of developing more secure and reliable artificial intelligence systems capable of resisting such attacks.

Keywords: Convolutional Neural Networks · Adversarial Images · Genetic Algorithms.

Abstract. Las redes neuronales convolucionales conforman una técnica que ha demostrado un gran éxito en tareas de visión artificial, como la clasificación de imágenes y detección de objetos. Como cualquier modelo de aprendizaje automático, tiene limitaciones y vulnerabilidades que deben ser consideradas cuidadosamente para utilizarlas de manera segura y efectiva. Una de las limitaciones principales se encuentra en su complejidad y la dificultad de interpretar su funcionamiento interno, lo que puede ser explotado con fines maliciosos. El objetivo de estos ataques consiste en hacer cambios deliberados en la entrada de datos, de forma

tal de engañar al modelo y hacer que tome decisiones incorrectas. Estos ataques son conocidos como ataques adversarios. Este trabajo se centra en la generación de imágenes adversarias utilizando algoritmos genéticos para una red neuronal convolucional entrenada con el dataset MNIST. Se utilizan varias estrategias incluyendo ataques dirigidos y no dirigidos, así como también se presentan imágenes interpretables y no interpretables, no reconocibles para los humanos, pero que la red identifica y clasifica erróneamente con alta confianza. El experimento muestra la posibilidad de generar imágenes adversarias en un tiempo relativamente corto, lo que pone en evidencia la vulnerabilidad de las redes neuronales y la facilidad con la que pueden ser engañadas. Estos resultados resaltan la importancia de desarrollar sistemas de inteligencia artificial más seguros y confiables, capaces de resistir estos ataques.

Keywords: Redes Neuronales Convolucionales · Imágenes Adversarias · Algoritmos Genéticos.

1 Introducción

Las redes neuronales convolucionales (CNN, por sus siglas en inglés) conforman una técnica dentro del aprendizaje profundo que ha demostrado un gran éxito en tareas de visión artificial, como la clasificación de imágenes y la detección de objetos [1] [2] [3]. Estas redes están compuestas por varias capas, incluyendo capas de convolución, de agrupamiento y de activación, que trabajan juntas para aprender características relevantes de las imágenes de entrada. Sin embargo, como cualquier modelo de aprendizaje automático, tiene limitaciones y vulnerabilidades que deben ser consideradas cuidadosamente para utilizarlas de manera segura y efectiva [4].

Una de las limitaciones principales de las CNN se encuentra en su complejidad y la dificultad de interpretar su funcionamiento interno. A menudo se las considera "cajas negras" debido a que la razón por la cual se llega a una determinada predicción no es fácilmente comprensible. Esta falta de explicabilidad en el funcionamiento puede ser explotada con fines maliciosos para engañar a la red y producir resultados incorrectos. Estos ataques malintencionados son conocidos como ataques adversarios, según el término acuñado por [5]. Son cambios deliberados en la entrada de datos que están diseñados para engañar al modelo y hacer que tome decisiones incorrectas. Por ejemplo, se pueden crear imágenes que parecen normales, pero que contienen información oculta que engaña a la red neuronal. Asimismo se pueden generar imágenes que no son reconocibles para los humanos, pero que la red neuronal identifica y clasifica erróneamente con una alta confianza. Por estas razones, es fundamental probar el comportamiento de los modelos ante ataques adversarios para evaluar su robustez antes de hacerlos públicos o utilizarlos en aplicaciones críticas. Esto permitirá revelar debilidades y evaluar su capacidad de resistir intentos de manipulación en los datos de entrada. [6]

Según [7] hay varias maneras de clasificar los tipos de ataques que pueden afectar el rendimiento de un modelo de inteligencia artificial. Una de estas formas

depende del grado de conocimiento y el nivel de acceso que tenga el adversario sobre el modelo objetivo. En este sentido, se pueden identificar tres tipos de modelos: modelos de caja negra [8], de caja gris [9] y de caja blanca [10]. En el modelo de caja negra, un adversario no conoce la estructura de la red de destino ni los parámetros, pero puede interactuar con el modelo para consultar las predicciones que realiza a entradas específicas. En el modelo de caja gris, el adversario conoce la arquitectura del modelo de destino, pero no tiene acceso a los parámetros de la red. Debido a la información adicional sobre la estructura del modelo, un adversario de caja gris siempre muestra un mejor rendimiento de ataque en comparación con un adversario de caja negra. El adversario más fuerte es el de caja blanca, ya que tiene acceso completo al modelo objetivo, incluidos los parámetros, lo que significa que el adversario puede adaptar los ataques y elaborar muestras adversarias directamente en el modelo objetivo.

Otra forma de clasificar los tipos de ataque es basándose en la respuesta obtenida por el modelo. En este sentido, se pueden identificar tres categorías: los ataques no dirigidos (*untargeted*), los ataques que reducen la confianza y los ataques dirigidos (*targeted*). Los ataques no dirigidos son aquellos en los que el atacante no tiene un objetivo específico, sino que intenta explorar la vulnerabilidad del sistema para encontrar cualquier tipo de debilidad que provoque que el modelo haga una predicción errónea. Los ataques que reducen la confianza buscan hacer que el modelo sea menos confiable, aumentando la ambigüedad de su predicción. Finalmente, los ataques dirigidos son aquellos en los que el atacante tiene un objetivo específico, es decir, se quiere que el modelo haga una clasificación a otra clase preseleccionada.

Uno de los ataques con mayor éxito en la generación de ejemplos adversarios es el ataque de caja blanca, el cual utiliza la optimización basada en gradientes y requiere que el atacante tenga acceso completo a la arquitectura y los pesos del modelo [11]. Sin embargo, en la práctica, es más común encontrarse con la configuración de caja negra, donde no se revela nada sobre la arquitectura de la red, los parámetros o los datos de entrenamiento. En tal caso, el atacante sólo tiene acceso a los pares de entrada-salida del clasificador.

Este trabajo se centra en el estudio de la generación de imágenes adversarias mediante el uso de algoritmos genéticos utilizando la técnica de caja negra. Está organizado de la siguiente manera: En la sección 2 se presenta una revisión del estado del arte. En la sección 3 se explora la generación de imágenes adversarias, con el modelo dirigido y no dirigido, y se crean imágenes reconocibles y no reconocibles. Posteriormente, se presentan las conclusiones obtenidas, así como las posibles líneas de trabajo futuro.

2 Estado del arte

En la actualidad, existen varias propuestas enfocadas en la generación de imágenes adversarias, que tienen como objetivo final comprender mejor las vulnerabilidades de los modelos. Además, se busca explorar nuevas técnicas de defensa y

detección de ataques para fortalecer la seguridad, medir su robustez y confiabilidad de los modelos en aplicaciones prácticas [7].

Una de las primeras técnicas para crear imágenes adversarias fue presentada por [5]. En este trabajo, se emplea la técnica de ataque con modelo de caja blanca. Se utilizan los gradientes de la función de pérdida en relación a la imagen de entrada teniendo como objetivo agregar pequeñas perturbaciones, imperceptibles por el observador humano, pero que logran engañar a la red. Esta técnica requiere acceso total al modelo para su implementación. Este trabajo fue el punto de partida para una gran cantidad de investigaciones posteriores en el campo de las imágenes adversarias.

Como se mencionó anteriormente, en la práctica es más común encontrarse con la configuración de caja negra, donde no se revela nada sobre la arquitectura de la red, los parámetros o los datos de entrenamiento. Esta técnica permite realizar consultas al modelo para obtener tanto la clasificación final como el vector de salida, que contiene los elementos que representan las probabilidades de pertenecer a cada una de las clases a las que el modelo puede clasificar los datos de entrada. Por ejemplo, en el trabajo [8] se plantea una estrategia para llevar a cabo un ataque a un modelo de caja negra. Para ello se generan imágenes y se observan las etiquetas asignadas por el modelo. A partir de esta información, se entrena un modelo local que sustituye al modelo objetivo, utilizando esas entradas generadas de forma sintética y etiquetadas por el modelo objetivo. Posteriormente se utiliza el sustituto local para elaborar ejemplos contradictorios utilizando la técnica de caja blanca, ya que la construcción del modelo sustituto permite tener acceso completo a ese modelo. Aunque este trabajo tiene una alta tasa de efectividad, sufre del desajuste inherente del modelo entre el modelo sustituto y el modelo de destino, así como del alto costo computacional requerido para entrenar la red sustituta.

Por otro lado, en [14] se presenta la creación de imágenes adversarias utilizando descenso del gradiente y algoritmos evolutivos. Las imágenes generadas son irreconocibles para los seres humanos, pero las redes neuronales clasifican con más del 99% de confianza. Este trabajo pone en duda la capacidad de las redes neuronales profundas para alcanzar una percepción visual similar a la de los seres humanos. A través de experimentos se demuestra que, incluso los modelos más avanzados como VGG y GoogleNet, son vulnerables a las imágenes adversarias, lo que sugiere que la capacidad de las redes neuronales para generalizar y hacer predicciones precisas puede verse comprometida por pequeñas perturbaciones en los datos de entrada. Además, se realiza una evaluación similar con una red entrenada para el conjunto de datos ImageNet. El trabajo también plantea la estrategia de incluir imágenes adversarias en el entrenamiento de la red como posible solución al problema.

En concordancia con los trabajos mencionados, en el presente artículo se describe un experimento que deja en evidencia la vulnerabilidad de las redes neuronales artificiales. Consiste en emplear la técnica de algoritmos genéticos para generar imágenes adversarias en el conjunto de datos MNIST [16]. El objetivo es explorar su generación para ataques dirigidos y no dirigidos, y generando

imágenes reconocibles y no reconocibles. Se utiliza una red neuronal convolucional previamente entrenada con éxito en el conjunto de datos MNIST, y se aplican algoritmos evolutivos para generar imágenes que la red etiqueta con alta confianza en ciertas clases.

3 Generación de imágenes adversarias

En la actualidad, existen varias propuestas enfocadas en generar imágenes adversarias, que tienen como objetivo final comprender mejor las vulnerabilidades de los modelos. Además, buscan explorar nuevas técnicas de defensa y detección de ataques para fortalecer la seguridad, medir la robustez y confiabilidad de los modelos en aplicaciones prácticas.

En esta sección se presenta un conjunto de experimentos diseñados para explorar la vulnerabilidad de las redes neuronales, cada uno de ellos está enfocado en alguna de las técnicas de ataques mencionadas anteriormente combinadas con el tipo de imágenes (reconocibles y no reconocibles). En la sección 3.1 se analiza la generación de imágenes adversarias no reconocibles con el modelo de ataque dirigido. Luego en las secciones 3.2 y 3.3 se explora la generación de imágenes adversarias reconocibles. En la sección 3.4 se muestra la creación de imágenes conteniendo la mínima cantidad de píxeles necesarios para mantener la clasificación de la red. Finalmente en las secciones 3.5 y 3.6 se explora la generación de imágenes partiendo de los píxeles mínimos.

Se utiliza una red neuronal convolucional previamente entrenada con éxito en el conjunto de datos MNIST, y se aplican algoritmos evolutivos para generar imágenes que la red etiqueta con alta confianza en ciertas clases.

Para las pruebas que se llevan a cabo en el resto del trabajo se utiliza una red neuronal convolucional entrenada con el conjunto de datos MNIST. Este conjunto de datos es ampliamente utilizado para entrenar modelos que puedan reconocer dígitos escritos a mano. Contiene un total de 70.000 imágenes, que se encuentran divididas en 60.000 imágenes para el conjunto de entrenamiento y 10.000 para el conjunto de testeo. Cada imagen tiene una resolución de 28x28 píxeles en escala de grises y se codifica mediante un conjunto de números que representan las intensidades de los píxeles en un rango de valores de 0 a 255. El uso de este conjunto permite utilizar arquitecturas de redes neuronales existentes para la red neuronal de referencia. La arquitectura de la red utilizada es la que se encuentra definida en [19]. Está definida con dos capas convolucionales que se encargan de extraer características relevantes de las imágenes de entrada. Estas capas están seguidas por capas de *maxpooling*, las cuales reducen la dimensión de la imagen a medida que se desplaza por la red. Esto tiene como objetivo reducir el costo computacional y mejorar la generalización del modelo. La red incluye una capa *flatten* que convierte los mapas de características resultantes en un vector unidimensional, permitiendo la conexión con las capas totalmente conectadas posteriores. Además, se incluye un *dropout* de 0,5, que es una técnica de regularización que aleatoriza la activación de las neuronas para evitar el sobreajuste. Finalmente, la capa de salida utiliza una función de activación *softmax* que

normaliza las salidas de la red. En total tiene 34.826 parámetros y fue entrenada para lograr una precisión mayor al 0,99%. Esta red está disponible en [15].
En la figura 1, se muestran imágenes de ejemplo para cada clase junto con la confianza asignada por la red neuronal en su predicción.

0	1	2	3	4	5	6	7	8	9
									
1	0.99	1	0.99	0.99	0.99	0.99	1	1	0.99

Table 1. Clasificación de las imágenes realizada por el modelo. Abajo, la confianza de la red en la predicción realizada.

Como se mencionó, el objetivo de este trabajo es explorar la generación de imágenes adversarias utilizando un algoritmo genético. Los algoritmos genéticos son una clase de heurísticas de optimización inspiradas en el proceso evolutivo de selección natural de Darwin [13] [17]. En este proceso, se parte de una población inicial de 'individuos' (en este caso, imágenes), y se lleva a cabo una selección, cruce y mutación para crear nuevos individuos. Los individuos más aptos son seleccionados para ser parte de la siguiente generación, y esta selección se realiza mediante una función de aptitud que evalúa su calidad y está relacionada con el valor de predicción dado por la red.

3.1 Imágenes adversarias no reconocibles con el modelo de ataque dirigido

Los modelos de aprendizaje automático se utilizan para clasificar datos en diferentes categorías en función de ciertas características. Estos modelos crean límites de decisión que separan las distintas clases de datos en regiones para hacer la clasificación requerida. En espacios de entrada con alta dimensionalidad, la región asignada a una clase puede ser mucho más grande que el área efectivamente ocupada por los ejemplos de entrenamiento. Esta característica es explotada por el algoritmo genético en este caso. El objetivo de este experimento es generar imágenes aleatorias en regiones del espacio de entrada que no están cerca de los datos de entrenamiento, pero que aún así son asignadas a una determinada clase con alta confianza. Es importante tener en cuenta que las imágenes generadas lejos de los límites de decisión pueden no estar relacionadas con las imágenes esperadas para esa clase, por lo que no son reconocibles por los humanos.

Se parte de una población inicial formada por individuos que contienen píxeles aleatorios, comúnmente llamadas imágenes de ruido. Estas imágenes irán mutando iterativamente a través de un proceso de selección, cruce y mutación, hasta que se encuentre una imagen que la red neuronal clasifica erróneamente con una confianza superior al umbral deseado. Para mantener el algoritmo lo

más simple posible, se utiliza un único punto de cruce entre los individuos. En la mutación, se eligen aleatoriamente algunos píxeles y se los modifica. La función de aptitud se define como la probabilidad obtenida por la red neuronal en la clase preseleccionada. Este proceso termina cuando se encuentra un individuo que la red neuronal clasifica con una confianza superior al 98%. En ese momento, se considera que se ha generado una imagen adversaria efectiva y se detiene el proceso de optimización. En la figura 2 puede verse una imagen adversaria generada a partir de ruido para cada una de las posibles clasificaciones de la red.

0	1	2	3	4	5	6	7	8	9
									
0.982	0.981	0.983	0.98	0.981	0.981	0.980	0.980	0.996	0.980

Table 2. Clasificación de las imágenes realizada por el modelo. Abajo, la confianza en la predicción realizada.

Una desventaja de esta estrategia es que las imágenes se perciben como ruido y no son interpretables para los humanos. Por esta razón, en las siguientes estrategias se utiliza una imagen inicial válida como punto de partida.

3.2 Imágenes adversarias reconocibles con el modelo de ataque no dirigido

En este experimento se crearon imágenes reconocibles, en un modelo de ataque no dirigido. Se probaron dos formas posibles de generar imágenes adversarias. La primera consiste en agregar perturbaciones fuera de los límites de la imagen, lo que puede provocar la aparición de píxeles aislados. Aunque la imagen original se mantiene perfectamente reconocible, se observan puntos de distintas intensidades en los bordes de la imagen. Aquí la función de aptitud es mayor si el valor obtenido en el vector de salida de la red para la clasificación correcta es más bajo. Mientras se mantenga la predicción las mutaciones irán modificando las intensidades de algunos píxeles seleccionados fuera de la imagen. Este proceso terminará cuando la predicción de la red cambie a otra clase diferente de la correcta. En la figura 3 se presentan algunas imágenes generadas con esta técnica. Como resultado, la imagen original permanece sin cambios aparentes, pero se observan la aparición de algunos puntos blancos fuera de ella. Abajo de la figura, se muestra la nueva predicción realizada por la red, y la confianza con que la red predice el valor indicado.

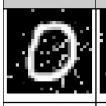
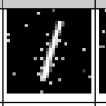
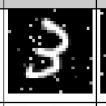
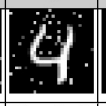
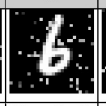
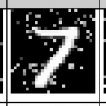
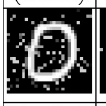
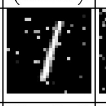
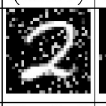
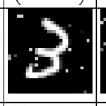
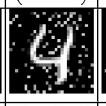
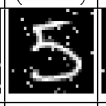
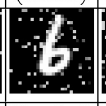
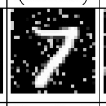
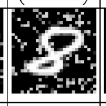
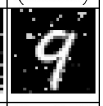
0	1	2	3	4	5	6	7	8	9
									
Pred: 8 (0.590)	Pred: 4 (0.527)	Pred: 8 (0.497)	Pred: 8 (0.506)	Pred: 9 (0.644)	Pred: 8 (0.587)	Pred: 8 (0.504)	Pred: 2 (0.595)	Pred: 2 (0.509)	Pred: 4 (0.603)
									
Pred: 2 (0.357)	Pred: 7 (0.507)	Pred: 3 (0.428)	Pred: 2 (0.499)	Pred: 8 (0.450)	Pred: 3 (0.549)	Pred: 3 (0.348)	Pred: 3 (0.429)	Pred: 2 (0.505)	Pred: 8 (0.411)

Table 3. Clasificación de las imágenes realizada por el modelo. Abajo, la predicción realizada y la confianza en ella.

En la segunda forma se altera la intensidad de los píxeles pertenecientes al dígito, lo que produce una apariencia de borrones o ruido en el trazo. En la figura 4, se presentan imágenes donde se utilizó esta estrategia. Como resultado, puede observarse que los trazos de la imagen se vuelven borrosos, pero aún pueden ser reconocidos con facilidad por un observador humano. Nuevamente, debajo de la figura, se muestra la confianza con que la red predice el valor indicado.

En ambos casos, es necesario tener acceso al vector de salida. y en particular a la probabilidad dada por la red para la clase de la imagen original.

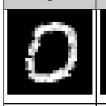
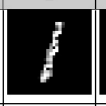
0	1	2	3	4	5	6	7	8	9
									
Pred: 2 (0.540)	Pred: 8 (0.499)	Pred: 8 (0.518)	Pred: 2 (0.570)	Pred: 9 (0.503)	Pred: 3 (0.523)	Pred: 8 (0.513)	Pred: 3 (0.507)	Pred: 3 (0.509)	Pred: 4 (0.598)

Table 4. Imágenes adversarias perturbando píxeles pertenecientes a la imagen principal.

3.3 Imágenes adversarias reconocibles con el modelo de ataque dirigido

En este tipo de ataque el objetivo es generar una imagen adversaria que provoque una predicción específica por parte del modelo. Para lograr esto, es necesario tener acceso al vector de salida de la red que contiene información sobre la probabilidad de que la entrada pertenezca a cada una de las posibles categorías. El proceso se inicia con una imagen clasificada de forma correcta, la cual es sometida a un proceso de evolución para lograr la clasificación deseada. La función de aptitud del algoritmo genético se enfoca en la predicción de la clase objetivo selec-

cionada, lo que guía el proceso de evolución. Además, para modificar la imagen, se han seleccionado píxeles tanto dentro como fuera de la misma.

En la figura 5 se presentan diversas imágenes adversarias creadas mediante la técnica de ataque dirigido para cada una de las posibles clasificaciones de la red neuronal. Las columnas de la figura corresponden a la predicción realizada por la red neuronal, mientras que las filas indican la clasificación de la imagen origen. En la diagonal principal de la figura se encuentran las imágenes originales correctamente clasificadas. Es importante remarcar que cada una de las imágenes es clasificada por la red de referencia como lo indica la columna, con una certeza mayor al 98%.

Vemos que todas las imágenes son reconocibles por un observador humano sin

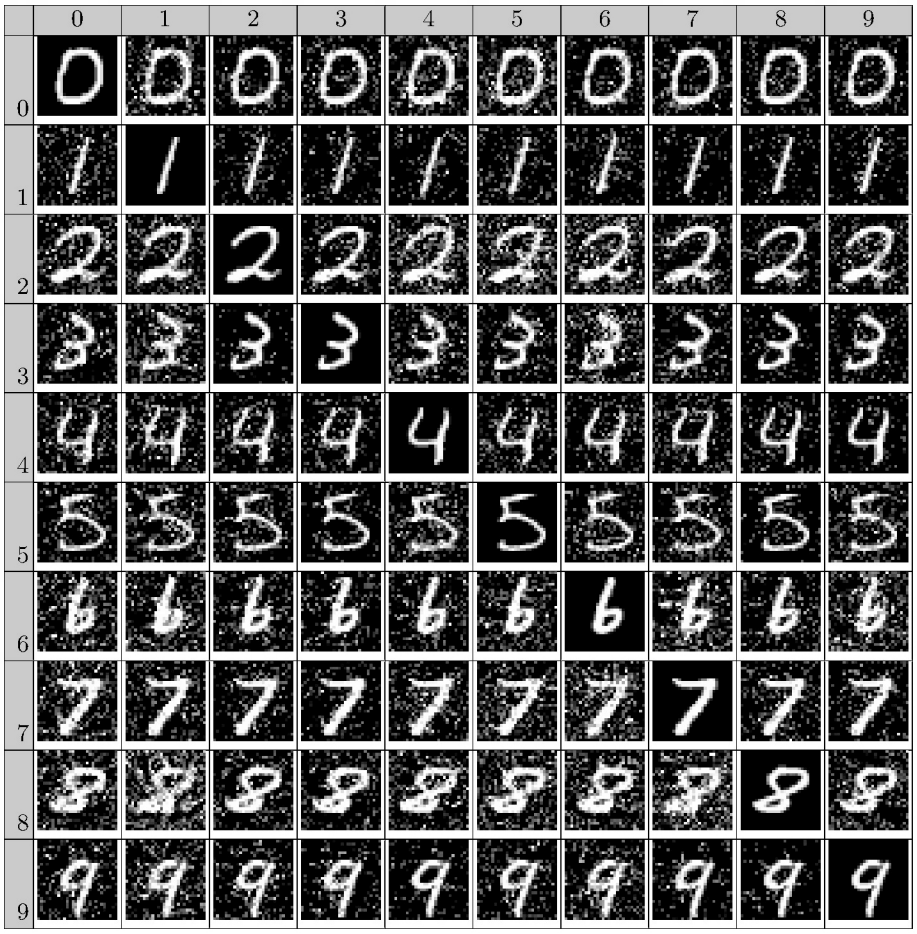


Table 5. Elaboración de imágenes adversarias para ataque dirigido.

demasiada dificultad.

Las estrategias que se han presentado requieren conocer el vector de clasificación final del modelo. No obstante, es posible generar imágenes adversarias incluso si solo se tiene acceso a la clasificación final de la red, lo cual resulta útil en casos donde no es factible obtener dicho vector.

3.4 Imágenes adversarias conteniendo la cantidad mínima de píxeles necesarios para mantener la clasificación

Este es otro ejemplo de generación de imágenes no reconocibles por humanos. En este caso, se utilizó el algoritmo genético para generar imágenes con el menor número de píxeles necesarios para mantener la clasificación correcta del modelo. Para lograr esto, se partió de una imagen correctamente clasificada y se aplicó una mutación aleatoria que permite eliminar algunos píxeles en cada iteración del algoritmo. En este caso, la función de fitness de cada imagen resultante se relaciona con la cantidad de píxeles de la imagen. Los mejores individuos de cada generación son aquellos con menor cantidad de píxeles pero que mantengan la clasificación original. Este proceso se repite varias veces hasta obtener la cantidad mínima de píxeles que permiten clasificar la imagen. El modelo solo debe retornar su clasificación final.

En la figura 6 se muestran una de esas respuestas para cada una de las imágenes. Arriba puede verse el dígito original. En algunas imágenes, la cantidad de píxeles necesarios para la clasificación correcta son iguales o menores a 2.

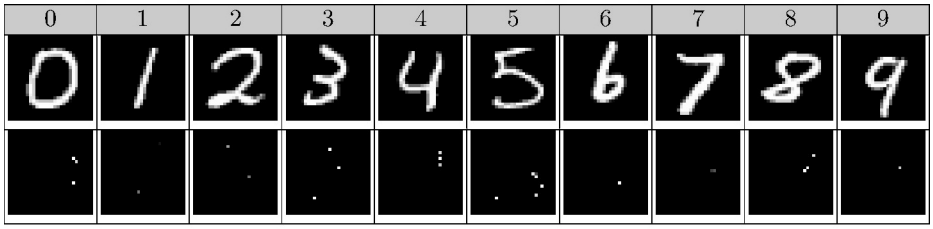


Table 6. Arriba se encuentra la imagen original. Abajo, una de las imágenes generadas con menor cantidad de píxeles que permiten mantener la clasificación.

3.5 Imágenes adversarias partiendo de los píxeles mínimos necesarios para mantener la clasificación

Partiendo de una imagen conteniendo los píxeles mínimos para una clasificación específica, se ejecutó el algoritmo para agregar píxeles libremente, de tal forma de aumentar la confianza de la red en la clasificación dada. En este caso, la función de fitness de cada imagen resultante se relaciona con la confianza que tiene la red al dar dicha clasificación. Este proceso termina cuando se alcanza una confianza del 99% en la predicción. La figura 7 contiene las imágenes generadas. Abajo se encuentra la predicción realizada por la red junto con la certeza

con que la realiza.

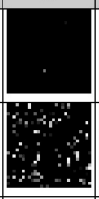
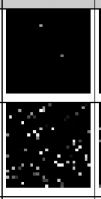
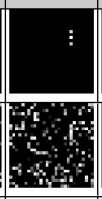
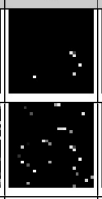
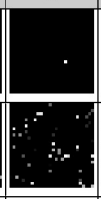
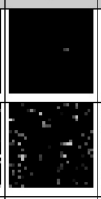
0	1	2	3	4	5	6	7	8	9
									
Pred: 0 (0.991)	Pred: 1 (0.990)	Pred: 2 (0.991)	Pred: 3 (0.992)	Pred: 4 (0.991)	Pred: 5 (0.996)	Pred: 6 (0.990)	Pred: 7 (0.980)	Pred: 8 (0.990)	Pred: 9 (0.990)

Table 7. Arriba se encuentra la imagen original. Abajo, una de las imágenes generadas con menor cantidad de píxeles que permiten mantener la clasificación.

Puede verse que este resultado es parecido al de la figura 1, imágenes, no reconocibles, pero con menor cantidad de píxeles.

3.6 Imágenes adversarias partiendo de los píxeles mínimos con el modelo de ataque dirigido

Si tuviéramos el caso en que la red neuronal proporciona únicamente la predicción final, sin ofrecer las probabilidades correspondientes a cada clase, aún es posible crear una imagen adversaria con la estrategia dirigida. Para esto, se debe partir de una imagen con la mínima cantidad de píxeles que tenga la clasificación requerida, a la cual se agregan píxeles pertenecientes a la imagen objetivo, variando las intensidades, pero manteniendo la clasificación. Aquí, la función de aptitud está relacionada con la cantidad de píxeles de la imagen destino elegida. Por ejemplo, podemos partir de la imagen mínima del 7, compuesta por dos puntos, y agregarle píxeles pertenecientes al 0 para que la imagen final se parezca a un cero, pero que sea clasificada como un 7 por la red. El resultado es muy parecido a las imágenes con el trazo borroso de la figura 4, pero en este caso siguen el modelo de ataque dirigido.

0	1	2	3	4	5	6	7	8	9
									

Table 8. Imágenes adversarias obtenidas a partir de la imagen mínima del número 7. Todas son clasificadas como 7 por la red.

4 Conclusiones y trabajo futuro

Los modelos de aprendizaje automático se utilizan para clasificar datos en diferentes categorías en función de ciertas características. Estos modelos crean límites de decisión que separan las distintas clases de datos en regiones para hacer la clasificación. Sin embargo, en espacios de entrada con alta dimensionalidad, la región asignada a una clase puede ser mucho más grande que el área efectivamente ocupada por los ejemplos de entrenamiento. El algoritmo genético aprovecha esta característica. Puede generar imágenes aleatorias en regiones del espacio de entrada que no están cerca de los datos de entrenamiento, pero que aún así son asignadas a una determinada clase con alta confianza. Sin embargo, es importante tener en cuenta que las imágenes generadas lejos de los límites de decisión pueden no estar relacionadas con las imágenes esperadas para esa clase. Esta afirmación ha sido comprobada en varios de los experimentos presentados.

En este trabajo se exploraron varias formas de generar imágenes adversarias utilizando un algoritmo genético. Se elaboraron imágenes tanto interpretables, es decir, reconocibles por los humanos, como no interpretables. A pesar de las limitaciones del experimento, se pudo comprobar que se pueden generar imágenes adversarias en un tiempo relativamente corto con la ayuda de algoritmos genéticos. Este hallazgo es preocupante ya que resalta la vulnerabilidad de las redes neuronales y la facilidad con la que pueden ser engañadas mediante la manipulación de las imágenes de entrada. Si bien los modelos de aprendizaje automático han demostrado ser efectivos en la clasificación de datos en diferentes categorías, estos resultados sugieren que aún queda mucho por hacer para garantizar su seguridad.

Es importante tener en cuenta que este tipo de ataques pueden tener graves consecuencias en aplicaciones del mundo real, como la detección de objetos en vehículos autónomos o la seguridad en sistemas de reconocimiento facial. Por lo tanto, es esencial tomar medidas para mitigarlos y garantizar que estos sistemas sean más robustos ante ataques maliciosos.

La reflexión acerca de la vulnerabilidad de las redes neuronales debe ser un punto de partida para continuar trabajando en pos de sistemas de inteligencia artificial más seguros y confiables. Esto no solo incluye la identificación y mitigación de vulnerabilidades, sino también la implementación de medidas de seguridad para prevenir ataques y el desarrollo de algoritmos más robustos que sean menos susceptibles a la manipulación de datos de entrada.

Todos los experimentos presentados en el presente trabajo pueden encontrarse en [18]

References

1. Russell, S. J., Norvig, P., Davis, E. ; Genesereth, M. (2020). Artificial Intelligence: A Modern Approach (4th ed.). Pearson.
2. Simonyan, Karen and Zisserman, Andrew. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition.

3. Krizhevsky, Alex and Sutskever, Ilya and Hinton, Geoffrey. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Neural Information Processing Systems. 25. 10.1145/3065386.
4. Zeiler, M. D., and Fergus, R. (2014). Visualizing and understanding convolutional networks. En European conference on computer vision (pp. 818-833). Springer, Cham.
5. Szegedy, Christian ; Zaremba, Wojciech ; Sutskever, Ilya ; Bruna, Joan ; Erhan, Dumitru ; Goodfellow, Ian and Fergus, Rob. (2013). Intriguing properties of neural networks.
6. Carlini, Nicholas ; Wagner, David. (2017). Towards Evaluating the Robustness of Neural Networks. 39-57. 10.1109/SP.2017.49.
7. S. Y. Khamaiseh, D. Bagagem, A. Al-Alaj, M. Mancino and H. W. Alomari, Adversarial Deep Learning: A Survey on Adversarial Attacks and Defense Mechanisms on Image Classification in IEEE Access, vol. 10, pp. 102266-102291, 2022, doi: 10.1109/ACCESS.2022.3208131
8. Papernot, Nicolas; McDaniel, Patrick; Goodfellow, Ian; Jha, Somesh; Celik, Z. Berkay; Swami, Ananthram. (2017). Practical Black-Box Attacks against Machine Learning. 506-519. 10.1145/3052973.3053009.
9. Y. Xu, X. Zhong, A. J. Yepes, and J. H. Lau, Grey-box adversarial attack and defence for sentiment classification, in Proc. N. Am. Chapter Assoc. Computat. Ling. (NAACL'21), Virtual Event, June 6-11, 2021, pp. 4078–4087.
10. Carlini, Nicholas; Farid, Hany. (2020). Evading Deepfake-Image Detectors with White- and Black-Box Attacks.
11. Goodfellow, Ian; Shlens, Jonathon; Szegedy, Christian. (2014). Explaining and Harnessing Adversarial Examples. arXiv 1412.6572.
12. Bradley, James; Blossom, Paul. (2023). The Generation of Visually Credible Adversarial Examples with Genetic Algorithms. ACM Transactions on Evolutionary Learning and Optimization. 3. 10.1145/3582276.
13. Holland John Henry. 1975. Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor, MI.
14. A. Nguyen, J. Yosinski and J. Clune, Deep neural networks are easily fooled: High confidence predictions for unrecognizable images, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 427-436, doi: 10.1109/CVPR.2015.7298640.
15. CNN de referencia, https://colab.research.google.com/drive/1h9dZqLiBky9PwTtgdUiAkX88xZOZVd34 Accedida en mayo 2023
16. Lecun, Y., Cortes, C.; Burges, C. (1998). The MNIST database of handwritten digits. Accedida en mayo 2023. http://yann.lecun.com/exdb/mnist/
17. Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs. 3rd edn. Springer-Verlag, Berlin Heidelberg New York (1996)
18. Sitio online con los experimentos presentados https://colab.research.google.com/drive/1DP2rOsy7MdLX-tTre9kmvP7iEapFo2G6
19. Keras. MNIST Convolutional Neural Network (CNN) Example. Recuperado el 12 de mayo de 2023, de url *https : //keras.io/examples/vision/mnist_convnet/*