Temas:
CiberAtaques, Defensa, Cuestiones Éticas.

Carlos:
Hace no mucho tiempo, la forma de abordar cualquier tipo de complejidad, en el mundo de la informática, era la búsqueda de algoritmos de resolución. Con la aparición de las IAs, Inteligencias Artificiales, esto ha cambiado drásticamente ya que ha aparecido un nuevo modelo de resolución de problemas basado en el entrenamiento de IAs que tiene como objetivo (en este campo concreto) la creación de dichos algoritmos. Por suerte o por desgracia, las IAs no solo han abordado el campo de la informática sino otros muchos otros campos que van desde la economía hasta el arte, creando un debate social entrañable y en el cual nos adentraremos respecto al campo que nos compete. La ciberseguridad.

En la actualidad, la IA se ha convertido en una herramienta fundamental e indispensable dentro de la seguridad. Por ejemplo, a día de hoy no se pueden tener en cuenta un Sistemas de Gestión de Alertas de Seguridad sin automatización mediante modelos de Inteligencia Artificial (1). Por tanto, las herramientas de defensa y protección de nuestros dispositivos, se ha visto obligada a actualizarse.

Cómo bien os hemos comentado, las IAs han revolucionado este sector de las telecomunicaciones, el problemas es que nos solo ha sido un avance para los buenos. Pasemos ahora a la otra cara de la moneda, es decir, los hackers. Las IAs no solo se han convertido en un herramienta de ayuda a la defensa sino también de ayuda a los atacantes y esto convierte a este tipo de herramientas en armas de doble filo muy peligrosas sobre todo para el ciudadano de a pie.

Pero… ¿Cómo se puede utilizar la IA para Hackearnos? Aquí os muestro unos ejemplos de uso:

1. **Pruebas de malware contra herramientas basadas en IA:**

Los atacantes pueden utilizar el aprendizaje automático de varias maneras. La primera, y la más sencilla, es crear sus propios entornos de aprendizaje automático y modelar su propio malware y prácticas de ataque para determinar los tipos de eventos y comportamientos que buscan los defensores.
Un malware sofisticado, por ejemplo, puede modificar las bibliotecas y componentes del sistema local, ejecutar procesos en la memoria y comunicarse con uno o más dominios pertenecientes a la infraestructura de control del atacante. Todas estas actividades combinadas crean un perfil conocido como tácticas, técnicas y procedimientos (TTP). Los modelos de aprendizaje automático pueden observar las TTP y utilizarlas para crear capacidades de detección.
Al observar y predecir cómo los equipos de seguridad detectan las TTP, los adversarios pueden modificar sutilmente y con frecuencia los indicadores y comportamientos para adelantarse a los defensores que confían en las herramientas basadas en la IA para detectar los ataques (2).

2. **Envenenamiento de IAs con datos inexactos implicando y creación de fake news:**

Los atacantes también utilizan el aprendizaje automático y la IA para comprometer los entornos envenenando los modelos de IA con datos inexactos. Los modelos de aprendizaje automático e IA se basan en muestras de datos correctamente etiquetados para construir perfiles de detección precisos y repetibles. Mediante la introducción de archivos benignos que se parecen al malware o la creación de patrones de comportamiento que resultan ser falsos positivos, los atacantes pueden engañar a los modelos de IA para que crean que los comportamientos de ataque no son maliciosos. Los atacantes también pueden envenenar los modelos de IA introduciendo archivos maliciosos que los entrenamientos de IA han etiquetado como seguros (2).

Aunque con un uso mucho menor, el mismo ChatGPT también puede ser utilizado para redactar noticias falsas con el objetivo de alienar la opinión pública. Aunque la IA se niega a generar fake news por razones éticas, mediante sencillos prompts – haciéndole pensar que precisamos ayuda para la creación de una novela- puede crear noticias creíbles (3).

3. **Mapeo de modelos IA ya existentes:**

Los atacantes buscan activamente mapear los modelos de IA existentes y en desarrollo utilizados por los proveedores de ciberseguridad y los equipos de operaciones. Al aprender cómo funcionan los modelos de IA y lo que hacen, los adversarios pueden interrumpir las operaciones y los modelos de aprendizaje automático activamente durante sus ciclos. Esto puede permitir a los hackers influir en el modelo engañando al sistema para que favorezca a los atacantes y sus tácticas. También puede permitir a los hackers eludir por completo los modelos conocidos modificando sutilmente los datos para evitar la detección basada en patrones reconocidos (2).

4. **Descifrado más fácil de contraseñas: (con ejemplo)**

Los sistemas de IA se hacen más inteligentes en la medida en que recopilan más datos, y muchos de esos datos los obtienen de Internet y de muchas otras fuentes. Hasta allí todo bien, pero el problema está en que estos sistemas también pueden tener acceso a una importante cantidad de contraseñas robadas que están disponibles en la red.
Y eso fue lo que constataron investigadores del Stevens Institute of Technology cuando utilizaron su sistema de IA, PassGan. Este proyecto académico de inteligencia artificial, basado en datos de entrenamiento de contraseñas filtradas, logró descifrar 27% de las contraseñas de LinkedIn. De esta forma, estos académicos lograron comprobar el enorme poder que tiene la IA para descifrar rápidamente una cantidad importante de contraseñas.
Por eso el énfasis que hacen los expertos en ciberseguridad para que los centros de datos y muchos sitios web dejen de utilizar las contraseñas tradicionales. En su lugar, recomiendan los sistemas de autenticación sin contraseñas o de múltiples factores (3).

PassGAN, IA anteriormente nombrada, se desarrolló por la empresa de ciberseguridad Home Security Heroes. Se trata de un generador de contraseñas basado en una **red generativa adversarial** (GAN, por sus siglas en inglés), que es un **modelo de aprendizaje automático** (4). Lógicamente esto está pensado para un uso ético de la misma pero no exime que los auténticos hackers la usen. Repositorio GitHub con PassGAN.

5. **Realización de ataques automatizados:**

Los hackers, ya están utilizando tanto la inteligencia artificial como el aprendizaje automático o machine learning (ML), para automatizar ataques dirigidos a redes

empresariales ¿Cómo lo hacen? Pues en estos casos crean un malware que es lo suficientemente inteligente para encontrar vulnerabilidades en la red, las cuales pueden ser aprovechadas para facilitar el ingreso no autorizado a la misma.

Además, este tipo de malware no tiene que comunicarse con los servidores de control y comando, por lo tanto, no pueden ser detectados por los sistemas de seguridad y protección de la red (3).

6. **Creación de ataques ransomware más potentes:**

Los ataques de ransomware o de secuestro de datos, se han hecho muy notorios de unos años para acá, sobre todo por el impacto de estas acciones maliciosas en algunas empresas que proveen servicios básicos. Ahora con la IA, los hackers pueden aumentar el impacto de este tipo de ataques, ya que esta tecnología facilita que el ransomware tenga un mayor alcance dentro del sistema de la empresa, antes de que se active y lo bloquee.
Esto sucede porque los ciberdelincuentes utilizan la inteligencia artificial para detectar las respuestas de los firewalls y, de esa forma, pueden hallar los puertos abiertos que el equipo de seguridad pasó por alto. En pocas palabras, los atacantes se valen de la IA para explotar los puntos débiles que tienen los firewalls y así logran que el ransomware tenga mayor impacto en el sistema de la empresa.
Incluso expertos en ciberseguridad, aseguran que ya existe una oferta clandestina de SDK o kits de desarrollo de software para ransomware, que explotan al máximo el poder de la IA. Eso significa que personas con malas intenciones pueden tener acceso a este tipo de tecnología, con solo ingresar a mercados ocultos en la internet profunda o Deep Web (3).

7. **Creación de correos electrónicos para phising más efectivos:**

Los ataques de phishing resultan cada vez más complicados, ya que las empresas entrenan a sus empleados para que sepan detectar los correos electrónicos falsos, sobre todo cuando se trata de un ataque en masa. Sin embargo, la IA les permite a los atacantes realizar este tipo de acciones, pero personalizando los correos electrónicos dirigidos a cada destinatario.
Todo esto es posible, gracias a que la inteligencia artificial y el aprendizaje automático, pueden alimentarse de toda la información pública que tiene un empleado, para luego aplicarlo en un ataque de phishing. En la práctica, eso significa que con la IA los ciberdelincuentes pueden llegar niveles tan refinados de ataque, que incluso pueden incorporar un email de phishing dentro de un hilo de correos electrónicos en curso o dentro de una discusión vía email, lo que minimiza la posibilidad de sospecha por parte del empleado afectado (3).

8. **Existencia de IAs pensadas en el hackeo:**

Los últimos reportes sobre la Dark Web indican la aparición de multitud de IAs pensadas para el hackeo:

- **FraudGPT**, creado con la capacidad de crear páginas web de estafa de phishing, escribir código malicioso, crear herramientas de piratería y escribir cartas de estafa (5).
- **WormGPT**, creado y entrenado para generar cualquier tipo de malware a demanda (6).

Por suerte, hay IAs pensadas en el hacking ético como SecGPT, pero que también podrían ser usadas con fines maliciosos. Vídeo hackeando con Kali y SecGPT.

9. **Ataques de ingeniería social, determinando el comportamiento de personas:**

La capacidad que tiene la inteligencia artificial de sacar conclusiones precisas respecto a las habilidades, el temperamento o las reacciones de una persona ante diferentes situaciones o escenarios, hacen que esta tecnología sea de gran utilidad para los hackers. Solo imagina lo que puede hacer uno de estos ciberdelincuentes, si logra determinar cuál es la probabilidad de que una persona sea víctima de un fraude, un ataque o de un abuso. Es prácticamente como leer la mente de una persona y hallar el momento en que puede ser más vulnerable a un ataque.

Incluso la inteligencia artificial ha demostrado que puede llegar a conclusiones muy detalladas, como por ejemplo determinar el momento preciso en que una persona está interesada en algo, o si está distraída o confundida. Y todas esas conclusiones las puedes obtener del análisis del comportamiento de la persona mientras navega o trabaja desde su computador personal. Por ejemplo, una persona con un comportamiento vacilante puede hacer una pausa antes de hacer clic en el enlace de un sitio web o puede volver a visitar una página en la que ya estuvo.

Y si bien esta información puede ser de gran ayuda para un departamento de marketing que quiera conocer mejor a sus clientes, también en las manos incorrectas puede ser muy perjudicial. Solo es cuestión de imaginar lo que un hacker podría hacer, si logra determinar cuando una persona puede ser más vulnerable a un ataque de phishing o ransomware, o incluso a un fraude financiero (3).

DEEPFAKE

Marcos:

La inteligencia artificial (IA) se utiliza cada vez más en ciberseguridad para mejorar la detección y protección contra amenazas cibernéticas. Aquí tienes algunas aplicaciones destacadas de la IA en este campo:

- Detección de amenazas avanzadas: La IA puede analizar grandes cantidades de datos y patrones para identificar amenazas cibernéticas que pueden pasar desapercibidas para sistemas tradicionales de detección de intrusiones. Puede detectar comportamientos anómalos en tiempo real.
- Análisis de malware: La IA se utiliza para analizar y clasificar malware de manera más efectiva. Puede identificar variantes de malware basadas en similitudes en el código, el comportamiento o las firmas digitales.
- Filtrado de spam y phishing: Los sistemas de filtrado de correo no deseado y detección de phishing utilizan IA para identificar y bloquear correos electrónicos maliciosos y sitios web fraudulentos.
- Autenticación biométrica: La IA se utiliza en sistemas de autenticación biométrica, como el reconocimiento facial o de huellas dactilares, para garantizar la identidad de los usuarios.
- Análisis de tráfico de red: La IA puede monitorear y analizar el tráfico de red en busca de patrones sospechosos o actividad maliciosa, como ataques de denegación de servicio (DDoS).
- Previsión de amenazas: Los algoritmos de aprendizaje automático pueden ayudar a predecir amenazas futuras al analizar tendencias y patrones históricos.

- Respuesta automatizada a incidentes: La IA puede ayudar a automatizar la respuesta a incidentes, como la cuarentena de sistemas comprometidos o la mitigación de ataques en tiempo real.
- Mejora de la autenticación y la identidad: La IA se utiliza para mejorar la autenticación de usuarios mediante análisis de comportamiento y detección de anomalías en los patrones de acceso.
- Protección de endpoints: Los sistemas de seguridad de endpoints utilizan IA para identificar y detener amenazas en dispositivos finales, como computadoras y dispositivos móviles.
- Análisis de registros y registros de seguridad: La IA puede analizar registros de eventos de seguridad y registros de sistemas para detectar actividades sospechosas o maliciosas.
- Segmentación de red dinámica: La IA puede ayudar a crear y mantener segmentos de red más seguros y dinámicos, aislando automáticamente sistemas comprometidos o sospechosos.
- Evaluación de riesgos y cumplimiento: La IA puede ayudar a las organizaciones a evaluar su postura de seguridad y cumplimiento al analizar sus sistemas y políticas.
- Inteligencia de amenazas: La IA se utiliza para recopilar, analizar y correlacionar datos de amenazas para proporcionar inteligencia sobre amenazas en tiempo real.

Estas son solo algunas de las muchas aplicaciones de la inteligencia artificial en ciberseguridad. La IA está en constante evolución y desempeñará un papel cada vez más importante en la protección contra las amenazas cibernéticas en el futuro.

Pablo:
Cuando hablamos de IA para la ciberseguridad nos referimos a cómo podemos usar la inteligencia artificial para ayudarnos a la detección eliminación y prevención de posibles problemas de seguridad.

Pero antes de nada, que es la IA. La Inteligencia artificial es un concepto algo difícil de explicar pues aun no sabemos cómo definir la inteligencia. Sin embargo la inteligencia artificial se podría explicar como un conjunto de técnicas y sistemas diseñados para que las máquinas y programas de computadora puedan realizar tareas que, si fueran realizadas por seres humanos, requerirían de inteligencia y habilidades cognitivas.

Existen varios tipos de IA y cada una es muy buena para unas determinadas amenazas o problemas:

1. **Aprendizaje automático supervisado**: El aprendizaje automático supervisado puede ser eficaz para detectar amenazas de seguridad cuando se dispone de un conjunto de datos etiquetados con ejemplos de amenazas y no amenazas. Puedes entrenar un modelo de clasificación (como una red neuronal o un algoritmo de bosque aleatorio) en estos datos etiquetados para que aprenda a identificar patrones asociados con amenazas.
2. **Aprendizaje por refuerzo**: Si las amenazas de seguridad están en constante evolución y tu sistema debe tomar decisiones secuenciales, el aprendizaje por refuerzo podría ser una opción. Puedes entrenar a un agente de RL para tomar

acciones en un entorno simulado o real y aprender a maximizar una recompensa a largo plazo, que podría estar relacionada con la seguridad.

3. **IA basada en reglas y detección de anomalías**: Para amenazas conocidas y patrones de comportamiento anormal, puedes utilizar sistemas de IA basados en reglas y detección de anomalías. Estos sistemas pueden buscar comportamientos que se desvíen de un conjunto de reglas predefinidas o detectar anomalías en los datos que puedan indicar una amenaza.
4. **Procesamiento del lenguaje natural (NLP)**: Si estás lidiando con amenazas que involucran texto, como correos electrónicos de phishing o comentarios maliciosos, un enfoque de NLP podría ser útil para analizar y detectar contenido malicioso en el lenguaje natural.
5. **Redes neuronales convolucionales (CNN) para imágenes**: Si las amenazas se presentan en forma de imágenes (por ejemplo, análisis de cámaras de seguridad), las CNN pueden ser efectivas para la detección de objetos o patrones en imágenes.
6. **IA basada en grafos**: Para detectar amenazas en redes informáticas o sistemas interconectados, las técnicas basadas en grafos pueden ser útiles para modelar las relaciones y detectar comportamientos anómalos en la topología de la red.

Por supuesto lo mejor y más recomendable sería un enfoque híbrido que combine varias técnicas.

Fer

APLICACIONES DE LA IA PARA LA PROTECCIÓN DE LA PRIVACIDAD

La inteligencia artificial también nos ayuda en un aspecto esencial del proceso de seguridad informática, la protección de la privacidad. A continuación, se presentan algunas formas en que la inteligencia artificial contribuye a proteger la privacidad:

1. **Protección de datos en dispositivos personales**: Puede utilizarse en dispositivos personales, como teléfonos inteligentes y asistentes de voz, para analizar y proteger datos personales. Esto puede incluir el cifrado de datos, la detección de aplicaciones maliciosas y la configuración de permisos de privacidad.

2. **Anonimización de datos:** Puede ayudar a las organizaciones a anonimizar los datos personales de manera más efectiva, garantizando que la información confidencial no pueda ser fácilmente vinculada a individuos específicos.

3. **Gestión de contraseñas y autenticación segura:** Puede mejorar la autenticación mediante métodos biométricos, como el reconocimiento facial o de voz, lo que reduce la necesidad de contraseñas que puedan ser vulnerables a ataques.

4. **Privacidad en redes sociales:** Se utiliza en plataformas de redes sociales para analizar y detectar contenido que infringe la privacidad, como imágenes o mensajes que contienen información personal. También puede ayudar a controlar la privacidad de las publicaciones y configuraciones de privacidad.

5. **Asistencia en el cumplimiento de regulaciones de privacidad:** Puede ayudar a las organizaciones a identificar y abordar posibles incumplimientos de regulaciones de privacidad, como el RGPD en Europa o la HIPAA en los Estados Unidos, al monitorear y gestionar datos de manera más efectiva.

6. **Navegación segura en línea:** Los navegadores web pueden incorporar funciones de inteligencia artificial para advertir a los usuarios sobre sitios web no seguros o que recopilan información personal de manera indebida.

Estas aplicaciones de la inteligencia artificial y otras muchas más contribuyen significativamente a la protección de la privacidad al garantizar que los datos personales estén resguardados y que las personas estén informadas y capacitadas para mantener su privacidad en el mundo digital.

CUESTIONES ÉTICAS DEL USO DE LA IA EN CIBERSEGURIDAD

El uso de la inteligencia artificial en la ciberseguridad plantea importantes cuestiones éticas que deben ser consideradas cuidadosamente:

1. **Privacidad del usuario:** El monitoreo y análisis constantes que realiza la inteligencia artificial pueden invadir la privacidad de los usuarios, especialmente si no se hace de manera transparente y con su consentimiento. Es fundamental garantizar que se respeten los derechos de privacidad de las personas y que se obtenga el consentimiento adecuado para recopilar y analizar sus datos.

2. **Sesgo en los datos y en los algoritmos:** Los modelos de inteligencia artificial pueden heredar sesgos de los datos de entrenamiento. Esto podría llevar a decisiones discriminatorias o injustas en la ciberseguridad, como la identificación errónea de ciertos grupos como amenazas. La corrección de estos sesgos es una consideración ética importante.

3. **Equidad y justicia:** La inteligencia artificial debe ser utilizada de manera que garantice la equidad y la justicia en la protección cibernética. Esto incluye asegurarse de que todas las personas y organizaciones tengan acceso a la misma protección y que las medidas de seguridad no discriminen a ciertos grupos.

4. **Responsabilidad y rendición de cuentas:** Cuando la inteligencia artificial toma decisiones en ciberseguridad, es importante establecer quién es responsable en caso de que algo salga mal. La responsabilidad y la rendición de cuentas son aspectos éticos fundamentales que deben abordarse.

5. **Comportamiento agresivo de la IA:** Las decisiones autónomas de la inteligencia artificial en ciberseguridad pueden llevar a respuestas agresivas o incluso a

ciberataques preventivos, lo que plantea cuestiones éticas sobre la proporcionalidad y la escalada en la respuesta a amenazas.

6. **Escalada de ciberarmas y guerra cibernética:** El uso de la inteligencia artificial en operaciones militares cibernéticas plantea la preocupación ética de una escalada de armas cibernéticas y la posibilidad de un conflicto cibernético descontrolado.

El uso de la inteligencia artificial en la ciberseguridad presenta desafíos éticos que deben ser abordados con precaución y responsabilidad. La ética desempeña un papel crucial en garantizar que la inteligencia artificial se utilice de manera justa, transparente y responsable para proteger la privacidad y la seguridad en línea.

ATAQUES DE FUERZA BRUTA OPTIMIZADOS

Los ataques de fuerza bruta son técnicas que intentan adivinar una contraseña o clave mediante la prueba sistemática de todas las combinaciones posibles hasta encontrar la correcta. Cuando hablamos de ataques de fuerza bruta "optimizados" con IA, nos referimos a la utilización de algoritmos y técnicas de aprendizaje automático para hacer estos ataques más eficientes y rápidos.

1 Predicción de patrones : Mediante el aprendizaje de comportamientos, ya que la IA puede ser entrenada con grandes conjuntos de datos de contraseñas reales filtradas para aprender patrones comunes en la elección de contraseñas y priorizando combinaciones, en lugar de probar combinaciones al azar o en secuencia, la IA puede priorizar las combinaciones basadas en los patrones aprendidos, probando primero las contraseñas más probables.

2 Adaptación Dinámica : La IA aprende en tiempo real es decir, detecta que ciertas combinaciones o patrones tienen más éxitos que otros y mediante reajustes basados en retroalimentación por ejemplo Si un sistema tiene medidas de seguridad que proporcionan algún tipo de retroalimentación (por ejemplo, "la contraseña es incorrecta" vs. "la contraseña es demasiado corta"), la IA puede usar esa información para ajustar sus intentos.

3 Evasión de defensas : Evitando bloqueos, si un sistema bloquea intentos después de un cierto número de fallos, la IA puede reconocer esto y pausar o ralentizar los intentos para evitar ser detectada y diversificando ataques para evitar ser detectada, la IA puede variar sus técnicas y patrones de ataque, haciendo que parezca que los intentos provienen de diferentes fuentes o utilizando diferentes métodos.

DEEPFAKE

Un "deepfake" es una técnica que utiliza inteligencia artificial (IA) para crear o modificar contenido audiovisual, haciendo que personas reales parezcan decir o hacer cosas que nunca dijeron o hicieron. Esta técnica ha ganado notoriedad en los últimos años debido a su

potencial para crear desinformación, manipulación y otros usos maliciosos. A continuación, te explico cómo funciona el deepfake con IA:

Aprendizaje profundo: La computadora se entrena usando miles de imágenes o clips de audio de una persona. Es como si estuviera estudiando intensamente cómo se ve y suena esa persona desde todos los ángulos y en diferentes situaciones.

Modelo de Doble Red: Se utilizan dos sistemas en este proceso. Uno intenta crear un video falso (llamado "generador") y el otro intenta detectar qué videos son falsos (llamado "discriminador"). Juntos, juegan un juego de "atrapa al farsante". El generador intenta hacer falsificaciones cada vez mejores, y el discriminador intenta ser mejor detectándolas. Con el tiempo, el generador se vuelve muy bueno creando videos falsos que parecen reales.

Creación: Una vez entrenado, puedes darle a este sistema un video de, digamos, la persona A y pedirle que transforme ese video para que parezca que la persona B está haciendo o diciendo lo que la persona A hizo en el video original.

Refinamiento: Al principio, el video falso podría tener algunos errores o aspectos extraños. Pero el sistema puede mejorar y corregir esos errores con más entrenamiento y ajustes.

FUNCIONAMIENTO DE PASSGAN:

1. **GAN - Generative Adversarial Network**

La red GAN utilizada por PassGAN se basa en 2 redes neuronales:

- **Generador**: Esta red neuronal tiene la tarea de crear contraseñas potenciales.
    1. **Entrada aleatoria**
       El generador comienza con una entrada aleatoria, que generalmente es un vector de números aleatorios de longitud fija. Denominado “**semilla**” o “**vector latente**” y que en este momento inicial no tiene relación con contraseñas reales.

2. **Transformación mediante Red Neuronal:**
   La semilla se pasa a través de una red neuronal dentro del generador. Para crear una contraseña utilizando:
   - Aprendizaje de Patrones:
   - Capas Ocultas:
   - Funciones de Activación:
   - Decodificación de Caracteres:
   - Muestreo Estocástico:
3. **Generación de Contraseña Potencial:** La red neuronal del generador transforma la semilla en una cadena de caracteres que representa una contraseña potencial. Esta cadena de caracteres es la salida del generador.
4. **Repetición**
5. **Refinamiento con el Entrenamiento:** Durante el entrenamiento de la GAN, el generador ajusta sus parámetros para que las contraseñas generadas sean cada vez más similares a las contraseñas reales presentes en el conjunto de datos de entrenamiento.

- **Discriminador**: Esta red neuronal busca saber si la contraseña es una contraseña real o una generada por el generador
  1. **Entrada y Evaluación:** El discriminador recibe como entrada una contraseña, que puede ser una contraseña real o una generada por el generador. La contraseña se presenta como una cadena de caracteres.
  2. **Transformación Mediante Red Neuronal:** El discriminador utiliza una red neuronal entrenada para procesar la contraseña de entrada. Esta red neuronal tiene capas ocultas que han aprendido durante el entrenamiento a detectar patrones y características presentes en contraseñas reales.
  3. **Salida de Clasificación:** 0 o 1

La competencia adversarial en PassGAN (y en cualquier modelo GAN) funciona mediante un proceso de "juego" entre dos componentes principales: el generador y el discriminador. A continuación, se explica cómo funciona esta competencia:

1. **Generador vs. Discriminador:**
   - El **generador** es responsable de tomar una entrada aleatoria (la semilla) y generar contraseñas potenciales que sean difíciles de distinguir de las contraseñas reales.
   - El **discriminador** es responsable de evaluar si una contraseña dada es real o generada por el generador. Debe distinguir entre contraseñas auténticas y contraseñas generadas.
2. **Entrenamiento Inicial:**
   - Al comienzo del entrenamiento, tanto el generador como el discriminador tienen un rendimiento pobre. El generador genera contraseñas aleatorias que son fácilmente identificadas como falsas por el discriminador.
3. **Iteraciones de Entrenamiento:**
   - El entrenamiento de PassGAN se realiza en iteraciones (épocas). En cada iteración, ocurren los siguientes pasos:

        - El **generador** toma semillas aleatorias y genera contraseñas potenciales.
        - El **discriminador** evalúa tanto contraseñas reales como contraseñas generadas y proporciona puntuaciones de probabilidad de autenticidad para cada una.
        - Basándose en las puntuaciones del discriminador, se calcula una pérdida (loss) para ambos componentes.
        - Los parámetros del **generador** se ajustan para minimizar su pérdida, lo que mejora su capacidad para engañar al discriminador.
        - Los parámetros del **discriminador** se ajustan para mejorar su capacidad para detectar contraseñas generadas y auténticas.
4. **Competencia Adversarial:**
    - La competencia se basa en que el generador y el discriminador están en constante lucha por mejorar.
    - El **generador** busca generar contraseñas cada vez más realistas para engañar al discriminador.
    - El **discriminador** busca mejorar su capacidad para distinguir contraseñas reales de contraseñas generadas.
    - A medida que el entrenamiento avanza, tanto el generador como el discriminador mejoran en sus respectivas tareas.
5. **Equilibrio y Convergencia:**
    - Idealmente, el proceso de competencia adversarial lleva al equilibrio, donde el generador genera contraseñas que son difíciles de distinguir de las reales, y el discriminador tiene dificultades para clasificarlas.
    - El entrenamiento continúa hasta que se alcance un criterio de convergencia deseado o se obtenga un nivel satisfactorio de calidad en las contraseñas generadas.

En resumen, la competencia adversarial en PassGAN es un proceso iterativo en el que el generador y el discriminador compiten para mejorar continuamente sus respectivas tareas. Esto lleva a la generación de contraseñas más realistas y a una mejora en la capacidad de detección del discriminador a lo largo del entrenamiento.

REDES GENERADOR Y REDES DISCRIMINADOR QUE UTILIZA.

DallE → Generador y Discrimidaro = son dos redes convolucionales:

Se caracterizan por x → Hacen una convolución …

PASSGAN:

- IWGAN - Improved training Wasserstein GANs
    - WGAN → Distancia de Wasserstein

        - Medida de similitud entre dos distribuciones de probabilidad, y en el contexto de GANs, se utiliza para cuantificar cuán diferentes son las distribuciones de las contraseñas generadas y las contraseñas reales.
    - Mejorar de la estabilidad y la convergencia del proceso de entrenamiento:
        - Ajuste del cálculo de la distancia
        - Introducción de restricciones en la red discriminatoria
- Optimizador ADAM - ADAptive Moment estimation:
    - Adam se utiliza para actualizar los parámetros tanto del generador como del discriminador durante el proceso de entrenamiento.
    - Adam se encarga de ajustar los pesos de la red generativa para que mejore su capacidad para generar contraseñas realistas y de la red discriminatoria para que sea capaz de evaluar de manera más precisa cuán realistas son las contraseñas generadas en comparación con las imágenes reales.
- Hiperparámetros:
    - Número de iteraciones del DISCRIMINADOR = 10 por iteración (por IWGAN)
    - Número de dimensiones para cada capa convolucional → 5 capas residuales
    - Lambda (Coeficiente de penalización de gradiente) → 10 estabilización GAN
    - ADAM
        - Tasa de aprendizaje biene dada por la tasa de siminción del promedio móvil del gradiente y del gradiente al cuadrado

---

WormGPT (2021)

- IA generativa → Es decir es una IA creada para generar contenido
- Basado en el lenguaje OpenIA pero sin restricciones
- 60€ al mes
- EJEMPLO:

Se trata de un ataque BEC - Business Email Compromise, es decir, que generación de correos electrónicos que busca presionar a una persona para pagar x cantidad de dinero.

- Porque es tan bueno:
    - No hay faltas gramaticales
    - Simula a la perfección la sintaxis humana

FRAUDGPT (2023)

- 200 dolares al mes
- Utilización de varios modelos IA.
- **Ingeniería social y ataques de phishing automatizados:** FraudGPT puede apoyar escenarios de pretexto convincentes para engañar a las víctimas.
- **Malware y exploits generados por IA**: Puede generar scripts maliciosos adaptados a redes específicas de víctimas.
- **Descubrimiento automático de recursos de ciberdelincuencia**: La IA puede acelerar la investigación manual para encontrar nuevas vulnerabilidades.
- **Evasión impulsada por IA de defensas**: Las herramientas avanzadas utilizarán IA generativa para evadir sistemas de detección.
- **Dificultad de detección y atribución**: Las futuras herramientas de IA armada reducirán la detección y atribución al punto de anonimato.

¿Qué es un payload?
Un payload es un programa que se ejecuta en contra a una vulnerabilidad.