

Ataques Adversariales y Defensas en Modelos de Aprendizaje Profundo

Por: Michelle Díaz

31 de Octubre 2023

Agenda

1. **Introducción**

2. **Ataques Adversariales**

- a. Ejemplos Adversariales
- b. Ataques Dirigidos v.s. Ataques No Dirigidos
- c. Tipos de Ataques
- d. Impacto

3. **Defensas Adversariales**

- a. Tipos de Defensas

4. **Estado del Arte**

5. **Conclusiones**



Agenda

1. Introducción

2. Ataques Adversariales

- a. Ejemplos Adversariales
- b. Ataques Dirigidos v.s. Ataques No Dirigidos
- c. Tipos de Ataques
- d. Impacto

3. Defensas Adversariales

- a. Tipos de Defensas

4. Estado del Arte

5. Conclusiones

Ejemplos Adversariales



x

“panda”
57.7% confidence

+ .007 ×



$\text{sign}(\nabla_x J(\theta, x, y))$

“nematode”
8.2% confidence

=

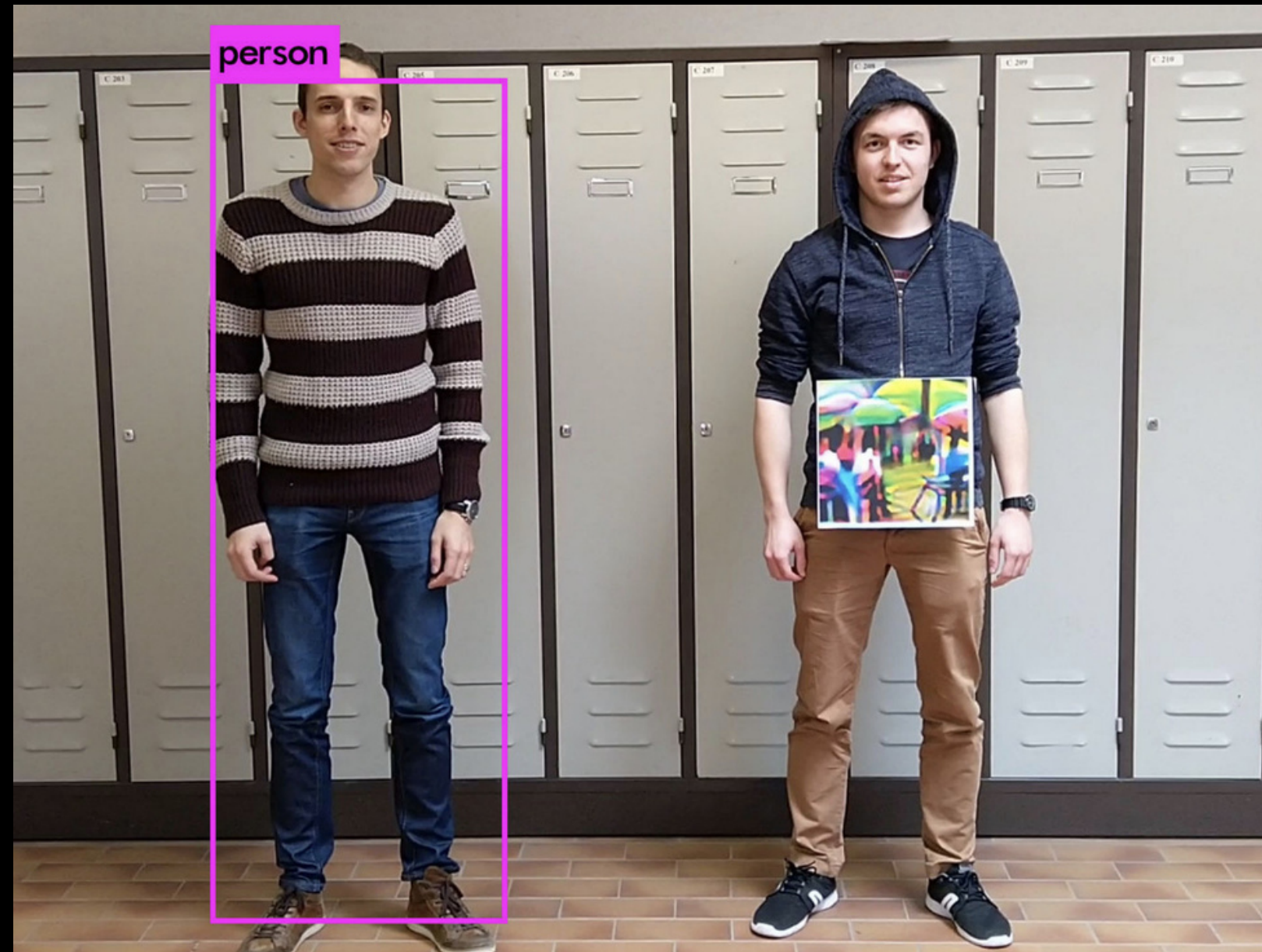


$x + \epsilon \text{sign}(\nabla_x J(\theta, x, y))$
“gibbon”
99.3 % confidence

Ejemplos Adversariales



Ejemplos Adversariales



Ejemplos Adversariales



Ejemplos Adversariales

Article: Super Bowl 50

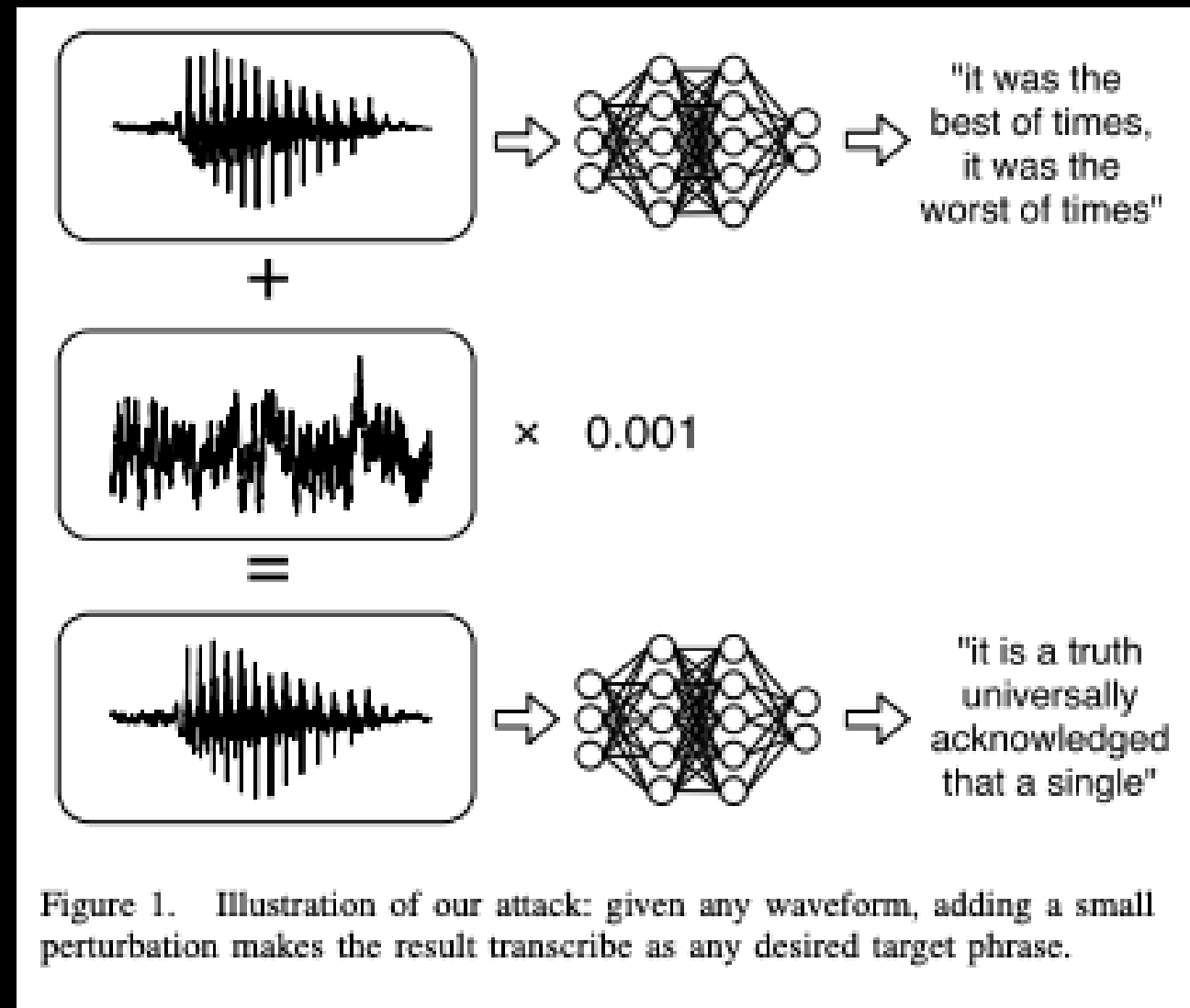
Paragraph: *“Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver’s Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.”*

Question: *“What is the name of the quarterback who was 38 in Super Bowl XXXIII?”*

Original Prediction: John Elway

Prediction under adversary: Jeff Dean

Ejemplos Adversariales



Ataques Dirigidos v.s. Ataques No Dirigidos

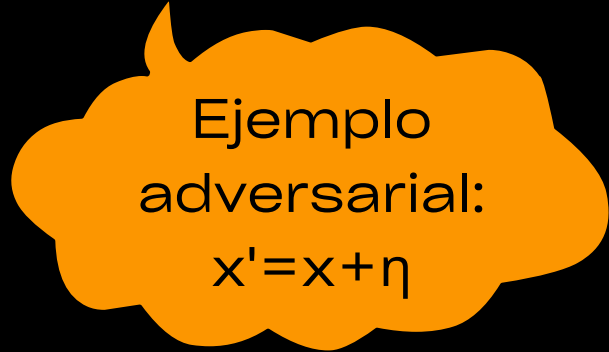
Ataques dirigidos

Input:

- red neuronal $f: X \rightarrow C$
- input $x \in X$
- etiqueta objetivo $t \in C$, tal que $f(x) \neq t$

Output:

- Una perturbación η tal que $f(x + \eta) = t$



Ejemplo
adversarial:
 $x' = x + \eta$

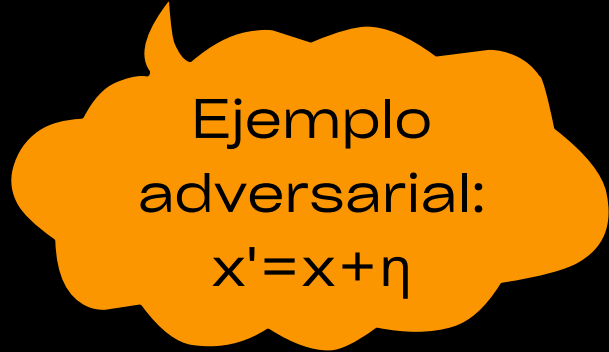
Ataques No Dirigidos

Input:

- red neuronal $f: X \rightarrow C$
- input $x \in X$

Output:

- Una perturbación η tal que $f(x+\eta) \neq f(x)$



Ejemplo
adversarial:
 $x' = x + \eta$

Tipos de Ataques

White box

La persona atacante conoce el modelo, los parámetros y la arquitectura de la red.

Black box

La persona atacante conoce la arquitectura de la red (por ejemplo, las capas) pero no sus parámetros (por ejemplo, los pesos).

Fast Gradient Sign Method (FGSM)

Explaining and harnessing adversarial examples. Goodfellow et al.

$$x^* = x + \epsilon \text{sgn}(\text{gradient})$$

x = normal example , x^* = adversarial example ,

where ϵ is small number and ∇ is the gradient of cost function with respect to X

FGSM consiste en agregar ruido (no aleatorio) cuya dirección es la misma que el gradiente de la función de costo con respecto a los datos.

El ruido se escala mediante épsilon, que generalmente está restringido a ser un número pequeño según la norma máxima.

Projected Gradient Descent (PGD)

Towards Deep Learning Models Resistant to Adversarial Attacks. Madry et al.

$$x^{t+1} = \Pi_{x+S} (x^t + \epsilon \text{sgn}(\text{gradient}))$$

x^0 = normal example , $x^* = x^T$ = adversarial example ,

PGD da pequeños pasos *iterativamente* en la dirección que maximiza la función de pérdida, garantiza que la entrada perturbada permanezca dentro de una vecindad épsilon predefinida de la entrada original.

Los ataques...

- Estos ataques explotan la sensibilidad del modelo a pequeños cambios en los datos de entrada.
- **FGSM** aprovecha la información de gradiente.
- **PGD** explora perturbaciones de forma iterativa. Es como **targeted FGSM**.
- Generar ejemplos contradictorios **ayuda a evaluar la robustez** de las redes neuronales y comprender su comportamiento cuando se someten a ligeras perturbaciones. Estos ejemplos son cruciales para **diseñar estrategias que mejoren la resiliencia de las redes neuronales contra posibles ataques**.

Me: *uses machine learning*

Machine: *learns*

Me:



Impacto

La ejecución exitosa de ataques adversarios puede comprometer la seguridad y la integridad de los sistemas de software.

- En un coche autónomo, un ataque adversarial al reconocimiento de imágenes podría tener consecuencias peligrosas.
- En el ámbito de la ciberseguridad, los ataques adversariales pueden eludir las medidas de seguridad y obtener acceso no autorizado a datos confidenciales, poniendo en riesgo sistemas completos.

Comprender y abordar estas vulnerabilidades es fundamental.

Defensas

1. Introducción

2. Ataques Adversariales

- a. Ejemplos Adversariales
- b. Ataques Dirigidos v.s. Ataques No Dirigidos
- c. Tipos de Ataques
- d. Impacto

3. Defensas

- a. Tipos de Defensas

4. Estado del Arte

5. Conclusiones

Defensas

Entrenamiento adversarial

Implica aumentar los datos de entrenamiento con ejemplos adversariales, mejorando así la solidez del modelo y mejorando su capacidad para resistir ataques potenciales.

Destilación defensiva

Implica entrenar un modelo con los resultados de otro modelo más complejo, además de los datos de entrenamiento originales.

Además, suavizar las probabilidades asignadas a clases incorrectas durante el entrenamiento, lo que hace que sea más difícil encontrar la dirección en la que perturbar la entrada para provocar una clasificación errónea.

Métodos de defensa certificados

Implican demostrar que las predicciones de un modelo seguirán siendo sólidas dentro de una determinada región alrededor de la entrada, incluso en presencia de pequeñas perturbaciones.

Agenda

1. Introducción

2. Ataques Adversariales

- a. Ejemplos Adversariales
- b. Ataques Dirigidos v.s. Ataques No Dirigidos
- c. Tipos de Ataques
- d. Impacto

3. Defensas

- a. Tipos de Defensas

4. Estado del Arte

5. Conclusiones

Esfuerzos para Desarrollar
modelos más robustos

Direcciones para la investigación

Transferibilidad adversarial

Investigar la transferibilidad de ataques adversarios entre diferentes modelos y dominios para comprender cómo se generalizan los ataques.

Detección adversarial

Desarrollar métodos más confiables para detectar ejemplos adversariales, incluidos aquellos diseñados con estrategias de ataque avanzadas.

Explicabilidad e interpretabilidad

Explorar técnicas para hacer que los modelos de aprendizaje profundo sean más interpretables, lo que puede ayudar a descubrir vulnerabilidades y proveer garantías.

Privacidad diferencial y aprendizaje federado

Avanzar en técnicas de aprendizaje automático que preservan la privacidad, como la privacidad diferencial y el aprendizaje federado, para proteger datos confidenciales en escenarios de entrenamiento colaborativo.

Agenda

1. Introducción

2. Ataques Adversariales

- a. Ejemplos Adversariales
- b. Ataques Dirigidos v.s. Ataques No Dirigidos
- c. Tipos de Ataques
- d. Impacto

3. Defensas

- a. Tipos de Defensas

4. Estado del Arte

5. Conclusiones

Conclusiones

- Necesitamos adaptar un **mindset** más orientado a seguridad.
- Entender que **crear modelos robustos y éticos es una inversión**.
- Crear un poco de **concientización** sobre cuales son las aplicaciones de IA con mayor riesgo de que se cometan injusticias y/o que puedan presentar un riesgo grave como por ejemplo su **uso en armas militares, que propicien o magnifiquen desigualdades sociales, etc.**
- La necesidad de modelos robustos y seguros es primordial. Esto implica una **colaboración continua entre la comunidad investigadora, las profesionales en industria y formuladorxs de políticas** para establecer mejores prácticas y estándares para crear los sistemas de IA.

¡Gracias!



/in/michellediazdev/



@Michdiazvi



@michellediazvi

