Ataque adversarios

Enrique David Guzmán Ramírez
IIMAS, Universidad Nacional Autónoma de México

Introducción al Aprendizaje Profundo
M. en C. Berenice Montalvo Lezama
M. en C. Ricardo Montalvo Lezama

14 de junio de 2021

Resumen

Las redes neuronales profundas son modelos poderosos que se han utilizado ampliamente para lograr un rendimiento cercano al nivel humano en una variedad de tareas. Sin embargo, a pesar de su desempeño superior estudios recientes han encontrado que incluso los modelos del estado del arte son sumamente vulnerables a ataques adversarios. En el presente trabajo se exploran algunos de los algoritmos de ataques adversarios en algunas de las arquitecturas más recientes de visión computacional.

1. Introducción

Un ataque adversario es una muestra de datos de entrada que ha sido perturbada levemente con la intención de hacer fallar a un clasificador
En muchos casos, estas modificaciones pueden ser tan sutiles que un observador humano ni siquiera nota la modificación, pero el clasificador comete un error. Los ejemplos de adversarios plantean problemas de seguridad porque podrían usarse para realizar un ataque a los sistemas de aprendizaje automático, incluso si el atacante no tiene acceso al modelo subyacente.
Con la introducción de modelos de aprendizaje profundo en cada vez más distintos aspectos de nuestra vida, los problemas que estos ataques adversarios pueden ocasionar son preocupantes.

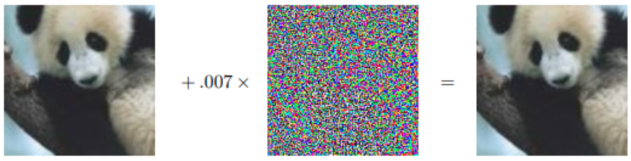

Figura 1: Ejemplo de ataque adversario. Imagen tomada de *Explaining and harnessing adversarial examples.*

Los escenarios de posibles ataques adversarios se pueden clasificar en diferentes maneras:

- **Ataque no dirigido:** el objetivo es hacer que el clasificador prediga una etiqueta incorrecta, la etiqueta incorrecta específica no importa.
- **Ataque dirigido:** el objetivo es cambiar la predicción del clasificador a alguna clase objetivo específica.

En segundo lugar, los escenarios de ataque se pueden clasificar por la cantidad de conocimiento que el adversario tiene sobre el modelo:

- **Caja negra:** el atacante no sabe mucho sobre el modelo, pero puede sondear o consultar el modelo, es decir, darle algunas entradas y observar salidas.
- **Caja blanca:** el atacante tiene pleno conocimiento del modelo, como la arquitectura del modelo y los valores de todos los parámetros y pesos entrenables.

1.1. Algoritmos de ataques adversarios

Hay una multitud [1]

2. Objetivo

El objetivo del proyecto es explorar y desarrollar algunas de las técnicas usadas para generar ataques adversarios así como explorar algunos de los métodos de defensa que existen.

Referencias

[1] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *CoRR*, vol. abs/1710.08864, 2017.