Ataques adversarios

E. David Guzmán Ramírez

Licenciatura en Ciencia de Datos
Introducción al Aprendizaje Profundo

M. en C. Berenice Montalvo Lezama
M. en C. Ricardo Montalvo Lezama

23 de abril de 2021

Contenidos

1. Introducción
2. Motivación
3. Descripción del problema
4. Análisis exploratorio
5. Propuesta de solución
6. Trabajo por hacer

Introducción

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 muy vulnerables a ataques adversarios.

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.

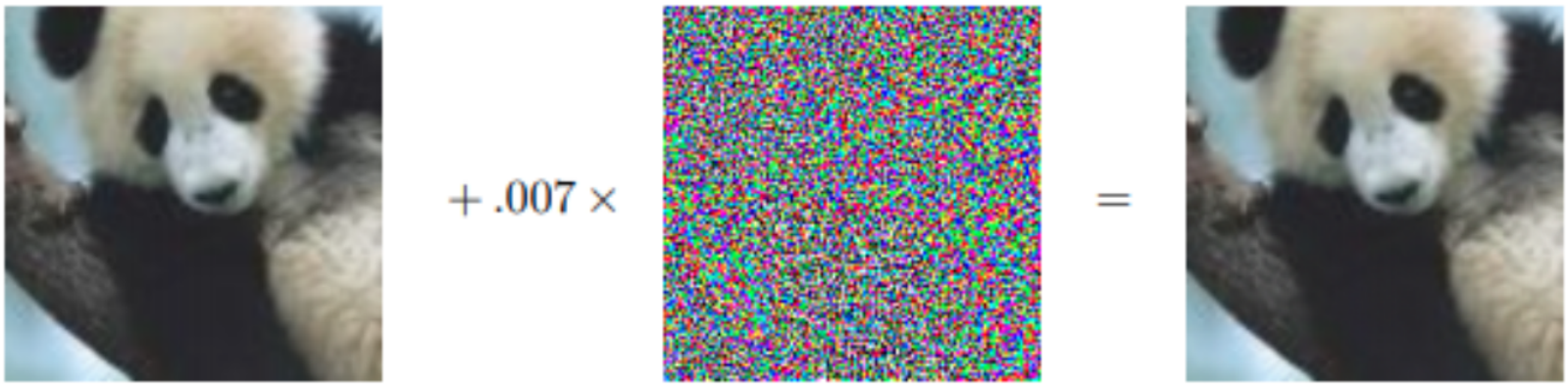

Figura 1: Ejemplo de ataque adversario. Imagen tomada de *Explaining and harnessing adversarial examples*.

Goodfellow et al., *Explaining and harnessing adversarial examples*, 2016.

Introducción

Figura 2: Ejemplo de ataque adversario en el que es posible engañar a la red cambiando únicamente un pixel. Imagen tomada de *One Pixel Attack for Fooling Deep Neural Networks.*

Su et al., *One Pixel Attack for Fooling Deep Neural Networks*, 2019.

Introducción

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.

Introducción

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.

Motivación

Los ataques adversarios plantean problemas de seguridad porque podrían usarse para realizar un ataque a los sistemas de aprendizaje profundo, 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.

Motivación

Figura 3: Ataque adversario al sistema de navegación autónomo de un Tesla, en el que confunde un señalamiento de velocidad de 35 millas/hora por 85 millas/hora. Figura tomada de MIT Technology Review: Trick a Tesla into accelerating by 50 miles per hour.

Descripción del problema

Para acelerar la investigación sobre ataques de adversarios, Google Brain organizó la *Competencia de Ataques y Defensas Adversarios* en la edición de 2017 de NIPS, la cual está disponible en Kaggle, la competencia a su vez constaba de tres subcompeticiones:

- Ataques adversarios no dirigido.
- Ataques adversarios dirigido.
- Defensas contra ataques adversarios.

Por el momento me concentraré en los ataques adversarios no dirigidos, la cual trata de un ataque de caja negra no dirigido, es decir, dada una imagen de entrada generar una imagen adversaria que engañe a un clasificador desconocido.

Análisis exploratorio

El dataset para esta competencia debía cumplir con 3 aspectos:

1. Conjunto de datos suficientemente grande y problema no trivial.
2. Problema bien conocido, por lo que las personas potencialmente pueden reutilizar los clasificadores existentes.
3. Muestras de datos que nunca se usaron antes.

El conjunto de ImageNet cumple con los primeros dos requisitos, posteriormente se etiquetaron 1000 nuevas imágenes compatibles con ImageNet las cuales servían como el conjunto de datos de desarrollo para la competencia y para cumplir el tercer requisito.

Propuesta de solución

Hay una variedad de arquitecturas preentrenadas en PyTorch con el conjunto de datos de ImageNet que sirven a la perfección para esta tarea. La idea es tratar de implementar varias técnicas de ataques adversarios[1] sobre estas arquitecturas, como

- **Métodos de gradiente:** la idea es $\boldsymbol{x}_{\text{adversario}} = \boldsymbol{x} + f(\nabla_x J(\boldsymbol{\theta}, \boldsymbol{x}, y))$.
- **Métodos de distribución:** realiza la optimización sobre las posibles distribuciones adversarias.
- **Basados en GANs:** un generador es entrenado para aprender la distribución adversaria maximizando la función de pérdida $J(\boldsymbol{\theta}, \boldsymbol{x}, y)$.

[1] Kui Ren et al., *Adversarial Attacks and Defenses in Deep Learning*, 2020.

Propuesta de solución

Por ejemplo, usando un PGD[2] en un clasificador entrenado en CIFAR10

Figura 4: Ejemplo de un ataque adversario usando el dataset de CIFAR10.

[2] A. Madry et al., *Towards Deep Learning Models Resistant to Adversarial Attacks*, 2019.

Trabajo por hacer

- Usar los modelos preentrenados de PyTorch para hacer ataques en un dataset de prueba.
- Una propuesta para defensa de ataques adversarios es entrenar el modelo con una mezcla de imágenes limpias y adversarias [3], por lo que con los ataques generados es posible construir un modelo más robusto resistente a los ataques.

[3] Goodfellow et al., *Explaining and harnessing adversarial examples*, 2016.