Actividad:

Entrenamiento de modelos y evaluación de robustez frente ataques adversarios

Los ataques adversarios en el aprendizaje automático consisten en la manipulación intencional de datos de entrada para engañar a los modelos y provocar errores en sus predicciones. En esta actividad, utilizaremos un ataque adversario general aplicando pequeñas perturbaciones a las imágenes de prueba para evaluar la robustez de varios modelos simples.

Parte 1: Entrenamiento de Modelos y Comparación

1. Configuración Inicial:

- Abre Google Colab y crea un nuevo cuaderno.
- Importa las bibliotecas necesarias, incluyendo PyTorch y TorchVision.

2. Preparación del Dataset:

Carga el dataset CIFAR-10 utilizando TorchVision.

3. Entrenamiento de Modelos:

- Selecciona al menos tres modelos simples de aprendizaje automático para entrenar en CIFAR-10. Ejemplos podrían ser Regresión Logística, Árbol de Decisión y K-Nearest Neighbors (KNN).
- Define la estructura de cada modelo y configura los hiperparámetros necesarios.

4. Entrenamiento y Evaluación:

- Entrena cada modelo utilizando el conjunto de datos de entrenamiento y evalúa su precisión utilizando el conjunto de datos de prueba.
- o Registra los resultados de precisión obtenidos para cada modelo.

Parte 2: Evaluación de Robustez con Imagenes modificadas

1. Generación de Ejemplos Adversarios:

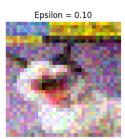
o Implementa un método simple de ataque adversario, como modificar ligeramente los píxeles de las imágenes de prueba de CIFAR-10 para confundir al modelo.

Ejemplo:









2. Aplicación de Ataques Adversarios:

 Aplica estos ejemplos adversarios para cada modelo entrenado y observa cómo afectan la precisión de clasificación.

3. Evaluación de Robustez:

- o Evalúa la precisión de cada modelo después de aplicar los ejemplos adversarios.
- Compara y contrasta cómo cada modelo responde a los ataques adversarios y qué tan rápidamente se degrada su rendimiento.

Parte 3: Creación de Tabla Comparativa

1. Estructura de la Tabla:

o Crea una tabla para comparar los modelos.

Ejemplo de tabla:

Modelo	Parámetros Utilizados	Precisión sin Ataque (%)	Precisión con Ataque (%)	¿Pasó el Ataque?	Tlempo de entrenamiento
Α					
В					
С					

Puedes agregar a tu tabla todos los valores que creas sean importantes de evaluar y comparar los modelos.

Parte 4

Investiga qué es AdverTorch y mira si puedes usarlo para generar otros ataques más complejos.