



Adversarial Learning

índice



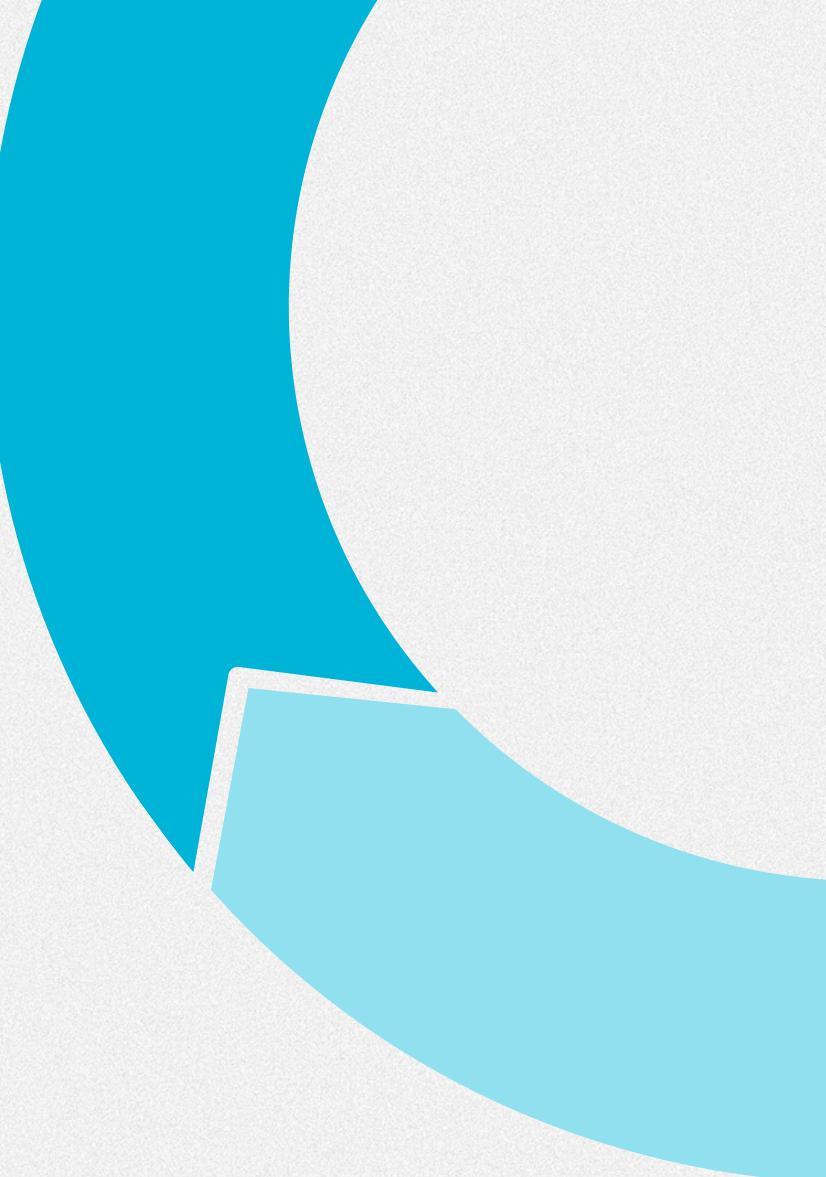
- 03 INTRODUÇÃO**
- 04 O QUE SÃO EXEMPLOS ADVERSARIAIS?**
- 05 TÉCNICAS DE ATAQUE ADVERSARIAL WHITE-BOX**
- 06 TÉCNICAS DE ATAQUE ADVERSARIAL BLACK-BOX**
- 07 ESTRATÉGIAS DE DEFESA CONTRA ATAQUES ADVERSARIAIS**
- 08 ESTRATÉGIAS DE DEFESA CONTRA ATAQUES ADVERSARIAIS**
- 09 APLICAÇÕES E DESAFIOS ATUAIS**
- 10 DEMONSTRAÇÃO PRÁTICA: ATAQUE FGSM NO MNIST**

Introdução

O Aprendizado Profundo tem alcançado resultados impressionantes em tarefas como classificação de imagens, reconhecimento de voz e diagnósticos médicos.

No entanto, mesmo os modelos mais poderosos podem ser enganados por pequenas perturbações nas entradas.

Essas perturbações, chamadas de exemplos adversariais, podem fazer com que um modelo cometa erros graves — mesmo quando a mudança na imagem é praticamente imperceptível para os humanos.

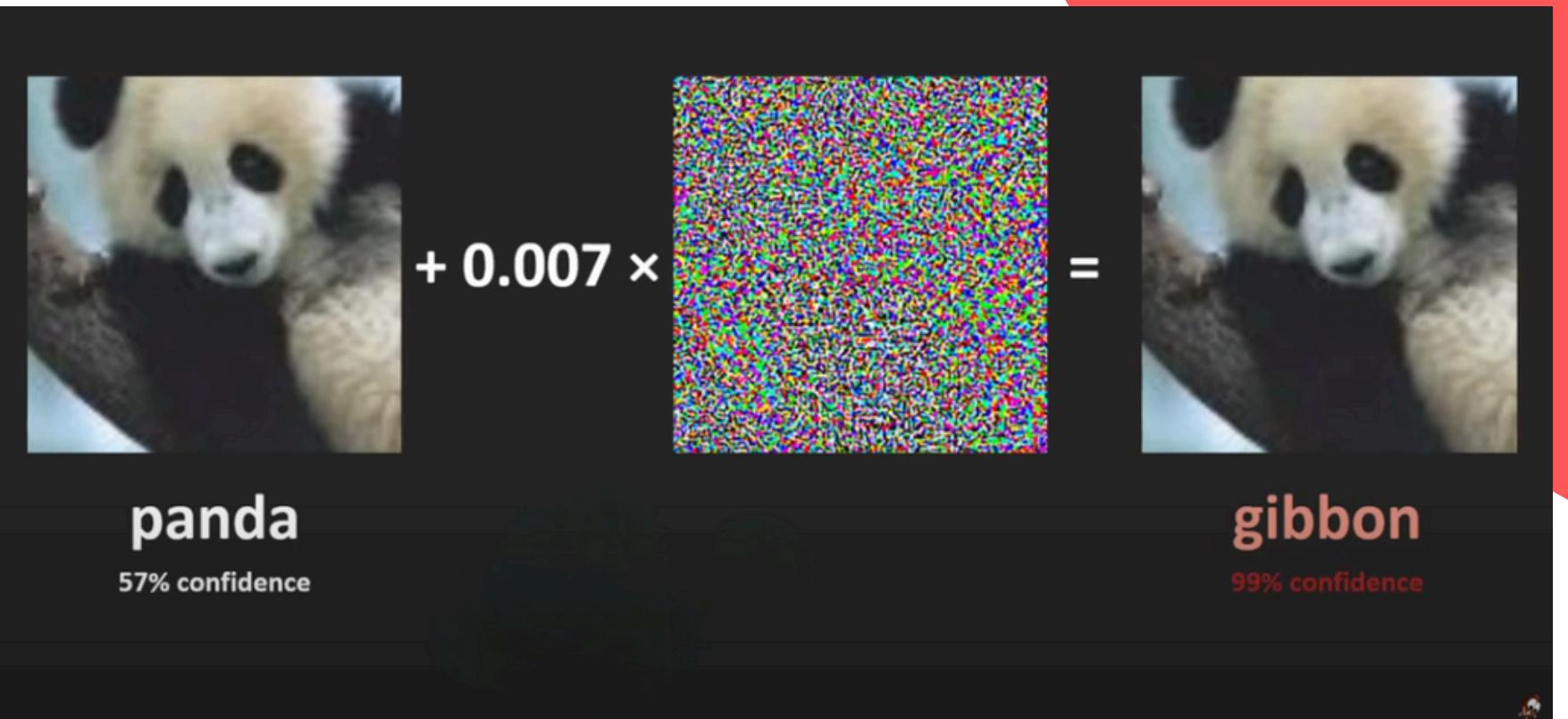


O que são Exemplos Adversariais?

Um exemplo adversarial é uma entrada artificialmente manipulada para induzir erro em um modelo de aprendizado de máquina.

Por exemplo, imagine uma imagem de um “panda” que é corretamente classificada por um modelo. Se adicionarmos um ruído cuidadosamente calculado (muito sutil), o modelo pode agora classificar essa imagem como um “gibbon”, mesmo que, visualmente, ela ainda pareça um panda para nós.

Esses ataques exploram o fato de que redes neurais são sensíveis a certas direções no espaço de entrada. Ao usar o gradiente da função de perda, é possível descobrir qual pequena mudança pode causar o maior erro de predição.



TÉCNICAS DE ATAQUE ADVERSARIAL: WHITE-BOX

- Ataques white-box: o atacante conhece a estrutura e os parâmetros do modelo, podendo calcular diretamente o gradiente para gerar exemplos adversariais.
- FGSM (Fast Gradient Sign Method): Técnica rápida e simples, utiliza o sinal do gradiente da loss para criar uma pequena perturbação:
- PGD (Projected Gradient Descent): Aplica várias pequenas perturbações acumuladas e projeta o resultado para um espaço permitido, ajustando a entrada adversarial a cada passo com base no gradiente da Loss Function em relação à entrada.

TÉCNICAS DE ATAQUE ADVERSARIAL: BLACK-BOX

- Ataques black-box: o atacante não tem acesso ao modelo interno, podendo apenas observar a saída para diferentes entradas. Mais difíceis de executar, porém mais realistas
- ZOO (Zeroth Order Optimization): Baseia-se no princípio da otimização sem derivadas explícitas. Como ele não acessa os gradientes do modelo, o ZOO estima-os numericamente observando como a saída do modelo muda em resposta a pequenas variações na entrada.
- NES (Natural Evolution Strategies): aplica ideias da evolução natural e otimização estocástica para encontrar direções de perturbação eficazes. Em vez de calcular ou estimar gradientes, o NES amostra várias perturbações aleatórias, avalia seu impacto na saída do modelo e ajusta a distribuição de busca com base nos melhores resultados.

TÉCNICAS DE ATAQUE ADVERSARIAL: BLACK-BOX

- Ataques black-box: o atacante não tem acesso ao modelo interno, podendo apenas observar a saída para diferentes entradas. Mais difíceis de executar, porém mais realistas
- ZOO (Zeroth Order Optimization): Baseia-se no princípio da otimização sem derivadas explícitas. Como ele não acessa os gradientes do modelo, o ZOO estima-os numericamente observando como a saída do modelo muda em resposta a pequenas variações na entrada.
- NES (Natural Evolution Strategies): aplica ideias da evolução natural e otimização estocástica para encontrar direções de perturbação eficazes. Em vez de calcular ou estimar gradientes, o NES amostra várias perturbações aleatórias, avalia seu impacto na saída do modelo e ajusta a distribuição de busca com base nos melhores resultados.

ESTRATÉGIAS DE DEFESA CONTRA ATAQUES ADVERSARIAIS

- FGSM – Defesa com Treinamento Adversarial
 - O modelo é treinado com exemplos gerados pelo próprio FGSM, fortalecendo sua capacidade de resistir a pequenas perturbações — essa abordagem é chamada de Adversarial Training.
- PGD – Defesa com Treinamento Adversarial Robusto
 - Para se proteger de ataques iterativos como o PGD, utiliza-se o PGD Adversarial Training, onde o modelo é treinado usando várias iterações de ruído adversarial durante o aprendizado.

ESTRATÉGIAS DE DEFESA CONTRA ATAQUES ADVERSARIAIS

- ZOO – Defesa com Randomização e Compressão
- Contra ataques baseados em consultas como o ZOO, defensores usam técnicas como Input Randomization (ex: redimensionamento, padding) e Feature Squeezing (ex: redução de precisão) para distorcer a estimativa do gradiente.
- NES – Defesa com Detecção Baseada em Estatísticas
- Ataques por amostragem, tentam descobrir dados sensíveis enviando muitas consultas parecidas e analisando os resultados com ruído. Esses ataques podem ser detectados por métodos estatísticos, que identificam padrões incomuns no comportamento das consultas ou nas respostas, como repetições suspeitas ou distribuições fora do esperado.

Aplicações

- Veículos autônomos: segurança em leitura de sinais de trânsito.
- Diagnósticos médicos: garantir que pequenas variações em exames não causem erros graves.
- Autenticação biométrica: impedir fraudes com imagens ou vozes manipuladas.
- Geração de dados com GANs: criar imagens, vozes ou textos realistas.



Demonstração Prática: Ataque FGSM no MNIST

OBJETIVO DA DEMONSTRAÇÃO:

Mostrar como um ataque FGSM pode enganar uma CNN treinada no MNIST, mesmo com ruído imperceptível.

FLUXO DA DEMONSTRAÇÃO

1. Treinar uma CNN simples no MNIST.
2. Avaliar a acurácia no conjunto de teste.
3. Aplicar o ataque FGSM em imagens de teste.
4. Avaliar novamente — queda de desempenho e visualização dos exemplos.

