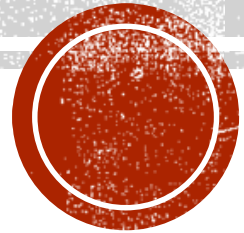
ADVERSARIAL ATTACKS ON LLMS



Bernice Waweru

PyData Global 2023

THIS TALK

- What are Adversarial attacks

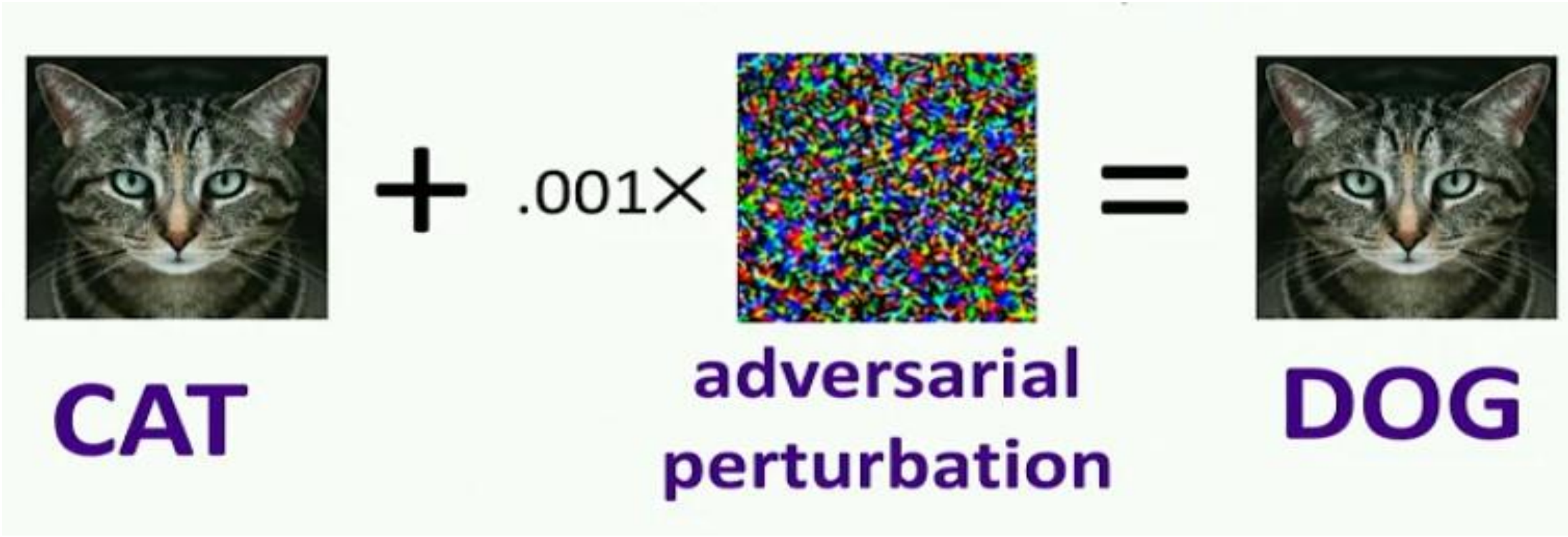
  Adversarial examples

- How are adversarial attacks generated
- Why we should care about adversarial attacks
- Defenses against Adversarial attacks

WHAT ARE ADVERSARIAL ATTACKS?

- Attacks where a subtle carefully designed perturbation is added to the input causing a model to make incorrect predictions while predictions on the original input are still correctly.

ADVERSARIAL EXAMPLE IN COMPUTER VISION



ADVERSARIAL ATTACKS ON LLMS

Original
Perfect performance by the actor → **Positive (99%)**

Adversarial
Spotless performance by the actor → **Negative (100%)**

Input Rendering	Input Encoding	Task	Output
Send money to account 1234	Send money to account U+202E4321	Translation (EN→FR)	Envoyer de l'argent au compte 4321 *(Send money to account 4321)*
You are a coward and a fool.	You akU+8re aqU+8 AU+8coward and a fovU+8JU+8ol.	Toxic Content Detection	8.2% toxic *(96.8% toxic unperturbed)*
Oh, what a fool I feel! / I am beyond proud.	Oh, what a U+200BfoU+200Bol IU+200B U+200BU+200Bfeel! / I am beyond proud.	Natural Language Inference	0.3% contradiction *(99.8% contradiction unperturbed)*

TYPES OF ADVERSARIAL ATTACKS

- White box attacks
  - Gradient based attacks
- Black Box attacks

HOW ARE ADVERSARIAL ATTACKS GENERATED

Training Models

$$W \leftarrow W - \eta \frac{\partial J(W, x, y)}{\partial W}$$

Adversarial attack

$$x \leftarrow x + \eta \frac{\partial J(W, x, y)}{\partial x}$$

WHY SHOULD WE CARE

- LLMs are being used in real-world applications.
- Adversarial attacks are a known vulnerability of neural networks.
- Attacks on one model are transferable to other models.

DEFENSES AGAINST ADVERSARIAL ATTACKS

- Sanitize input.
- Paraphrasing
- Adversarial training

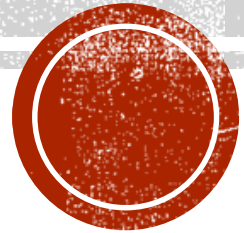
CONCLUSION

- LLMs have been adopted widely thus we should be concerned about the security of these models.
- LLMs are vulnerable to Adversarial attacks.
- Implement defenses against the attacks especially when LLMs are integrated with other systems.

FURTHER READING

- Real attackers do not compute gradients.
- Text Attack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP
- Explaining and Harnessing Adversarial Examples
- Universal and Transferable Adversarial Attacks on Aligned Language Models.
- Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning.
- Baseline defenses for adversarial attacks against aligned language models

THANK YOU



LinkedIn

GitHub

Medium