The Ethical Implications of Adversarial Attacks

Daniel Whettam

May 2020

Outline

Outline

1. What are adversarial attacks?
2. How do they work?
3. Why are they dangerous?
4. How can we protect against them?
5. What are the ethical implications?

Adversarial Attacks

What are Adversarial Attacks?

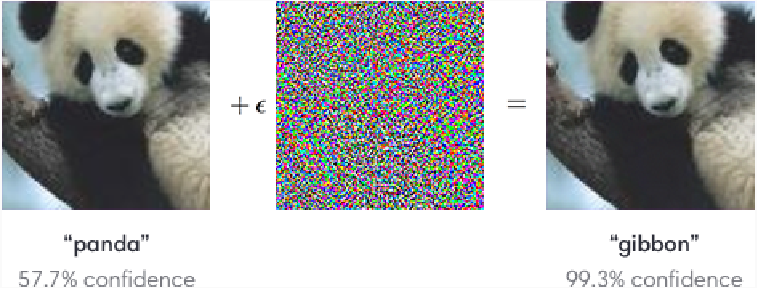
- Adversarial attacks are a technique in which a machine learning model is fooled through a malicious input.
- These malicious inputs are referred to as adversarial examples
- Adversarial examples are designed specifically to cause machine learning models to make mistakes
- These examples are often imperceptible to humans

What are Adversarial Attacks?

How Do They Work?

- Adversarial examples are created by making small changes to the input of an ML model that result in an unexpected outcome
- Typically these changes are imperceptible to humans, although not always

How Do They Work?



How Do They Work?

- Typically adding some sort of noise, the ML model is fooled without changing the appearance of the image
- This noise can be as small as editing a single pixel in an image [SVS19]
- Or can be something perceptible to humans (stop sign example), but still doesn't change the meaning of the image

How Do They Work?

White box attacks (direct access to the model)

- Uses gradient information to minimise D

$$D(x_0, x_0 + \delta) \tag{1}$$

Black box attacks (No access to the model)

- Approximate gradient through network outputs via standard high school gradient approx

$$\frac{f(x+h) - f(x)}{h} \tag{2}$$

Why Are They Dangerous?

- Machine learning algorithms can easily be fooled, even without access to the model
- Depending on the application, this can have serious consequences
- Obvious dangers include self-driving cars and medical screening
- A malicious actor can easily paint the road and trick a car into steering into traffic
- Need to ensure models are robust to adversarial attacks before going to production

How Can We Protect Against Them?

- Models must be robust to a certain amount of perturbation to the input data
- Good model performance $\neq$ robustness
- Recent work has provided frameworks for verifying the robustness of models (e.g. [KBD+17])
- Adversarial defenses seek to detect adversarial examples. Is that enough?

Ethical Implications

Ethical Implications

- Machine learning models are being used in more and more settings
- Often used for making decisions that have an impact on peoples lives and societal function:
    - Medical applications (e.g. tumor classification)
    - Military applications (e.g. automated drones)
    - Security applications (e.g. CCTV person detection)
    - Automated Vehicles
- Each of these applications can be subject to adversarial attacks

Ethical Implications

- Medical applications
  - Intentionally cause false negative results (e.g. [FCKB18])
- Military applications
  - Intentionally label civilian areas as military bases
- Security applications
  - Hide from person detection in automated CCTV camera
- Automated Vehicles
  - Paint on road to cause vehicle to steer into oncoming traffic
  - Adjust road signs to create dangerous driving scenarios

Ethical Implications

- The organisations and individuals implementing these models have a moral responsibility to do no harm through these technologies
- This tech is heavily relied upon and is vulnerability to adversarial attacks
- There is an ethical responsibility to ensure models are suitably robust when used to make decisions which can be exploited

Key Questions

- If a robust, verified model is exploited adversarially, who is to blame?
- How robust do our models have to be before they can be deployed? (never ending cycle of incresaingly more robust models + better attacks)
- Which applications require verified, robust models and which do not?
- Is there are a cost to verifying our models?

References

- Samuel G Finlayson, Hyung Won Chung, Isaac S Kohane, and Andrew L Beam, *Adversarial attacks against medical deep learning systems*, arXiv preprint arXiv:1804.05296 (2018).
- Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer, *Towards proving the adversarial robustness of deep neural networks*, arXiv preprint arXiv:1709.02802 (2017).
- Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai, *One pixel attack for fooling deep neural networks*, IEEE Transactions on Evolutionary Computation **23** (2019), no. 5, 828–841.