Adversarial Attacks on Cyberbullying Image Detection Models

<Instructor>

Outline

- Cyberbullying on images
- Adversarial examples
- Detecting cyberbullying on images
- Attacking cyberbullying detection models

Cyberbullying on Images

Threatening images like these can be sent to a victim to intimidate.

Detecting such content helps in preventing negative health effects on victims

Adversarial Examples

Machine learning models are susceptible to adversarial examples. Adversarial examples are input data that have been changed slightly with the intention of causing the machine learning classifier to misclassify.

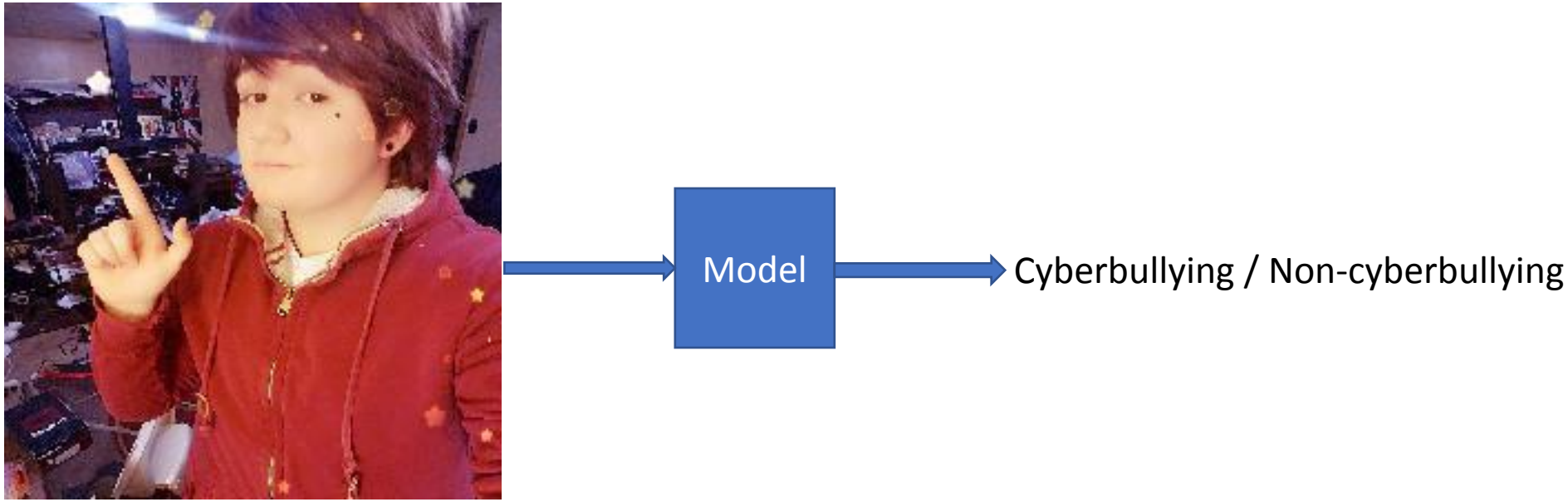
Usually, the changes are imperceptible to humans, yet machine learning classifiers makes mistakes.

This is a security issues because real-world systems based on machine learning can be forced to make mistakes without having access to the system. for example,

- Video surveillance systems
- Mobile applications for image classification
- Self driving cars
- Robots sensing the real-world through sensors and cameras

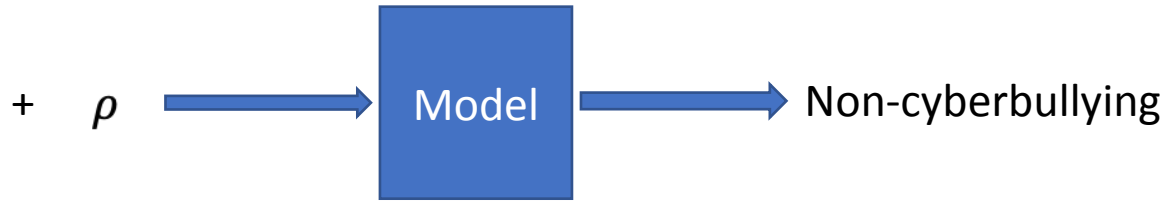
Detecting Cyberbullying on Images

Pass image through a trained classifier for detection



Adversarial Attacks

If there exist a machine learning system M and input image I, a clean image example. If I is classified correctly by M i.e., $M(I) = cyberbullying$. Then it is possible to construct another input image I' that is almost indistinguishable from I, known as an adversarial example, that will cause M to misclassify, i.e., $M(I') \neq cyberbullying$



Adversarial Attacks

- Assumptions
  - Black-box attacks
    - Adversarial examples are fed to a targeted model during testing
    - Model architecture, parameters and training procedure is not known by the adversary
  - White-box attacks
    - Adversary knows the model architecture, parameters, architecture, training procedure, and training data.
- Goals
  - Non-targeted attack – make the model predict the wrong label
  - Targeted attack – make the model predict a specific (target) label that is not the source (original) label

Adversarial Attacks

To successfully perform adversarial attack, a small noise needs to be generated which is added to the input image to cause the classifier to produce an incorrect output

How can this small noise ρ be generated?

- Fast gradient sign method (FGSM)
  - This method automatically finds ρ by computing the small change in the input image that will cause the classifier to make more mistake. This computation happens in one pass.
  - That small change is ρ
- Basic Iterative Method
  - This is the extension of FGSM which computes ρ iteratively and by clipping the input image by a fixed constant