CSE 564

February 28, 2024

(the where did february go edition)

Project Report Drafts due Friday (Mar 1) at 11:30pm

- Submit via Gradescope
- The formatting of the draft report should match that of a final report:
  - PDF
  - Single- or double-column format, single-spaced, with 12pt font and at least 1 inch margins
  - Up to 12 pages (okay if the draft is longer)
- The draft may include placeholders for new results etc., but should be **as complete as possible**, including related work.

After that…

- **By March 4:** Sign up for a presentation slot (see Ed post)
- **March 11: Final project presentations**
  - 11-12 minutes + 3-4 minutes for questions and transition
  - Please plan to attend at least 3 presentations other than your own
  - **If you are not available at all on this date, please let me know asap**
  - I will announce the presentations to a broader audience (e.g., security lab, security seminar) – **please let me know asap if you have concerns**
- **March 14:** Final project reports due
  - (+ statement of individual contribution)

Machine Learning Security & Privacy

Probably unsurprising: Large and growing field!

Example attacks:

- Adversarial examples
- Model inversion

Slide courtesy of Ivan Evtimov

Deep Neural Networks Can Fail

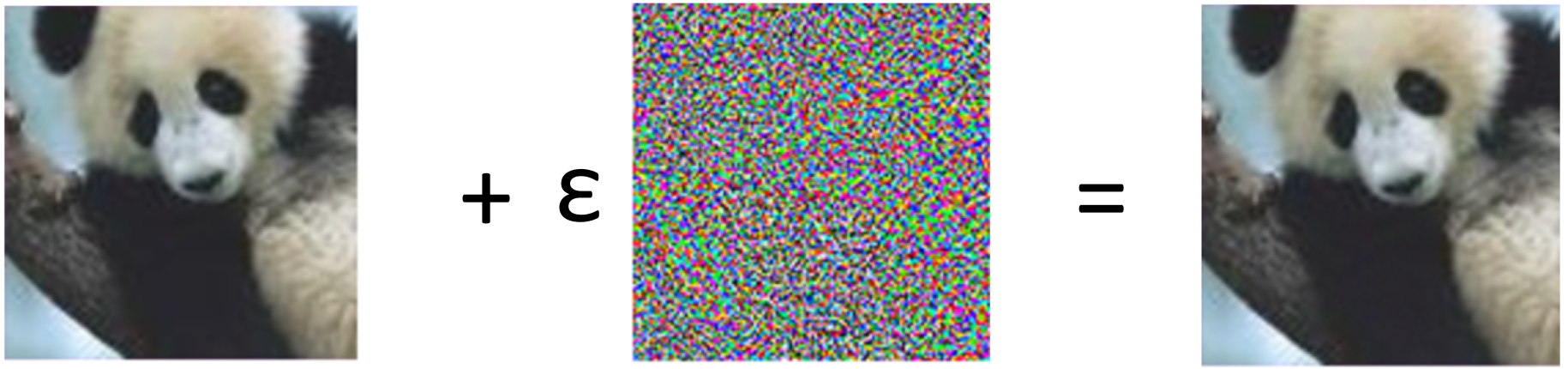


Image Courtesy: OpenAI

“panda”
57.7% confidence

“gibbon”
99.3% confidence

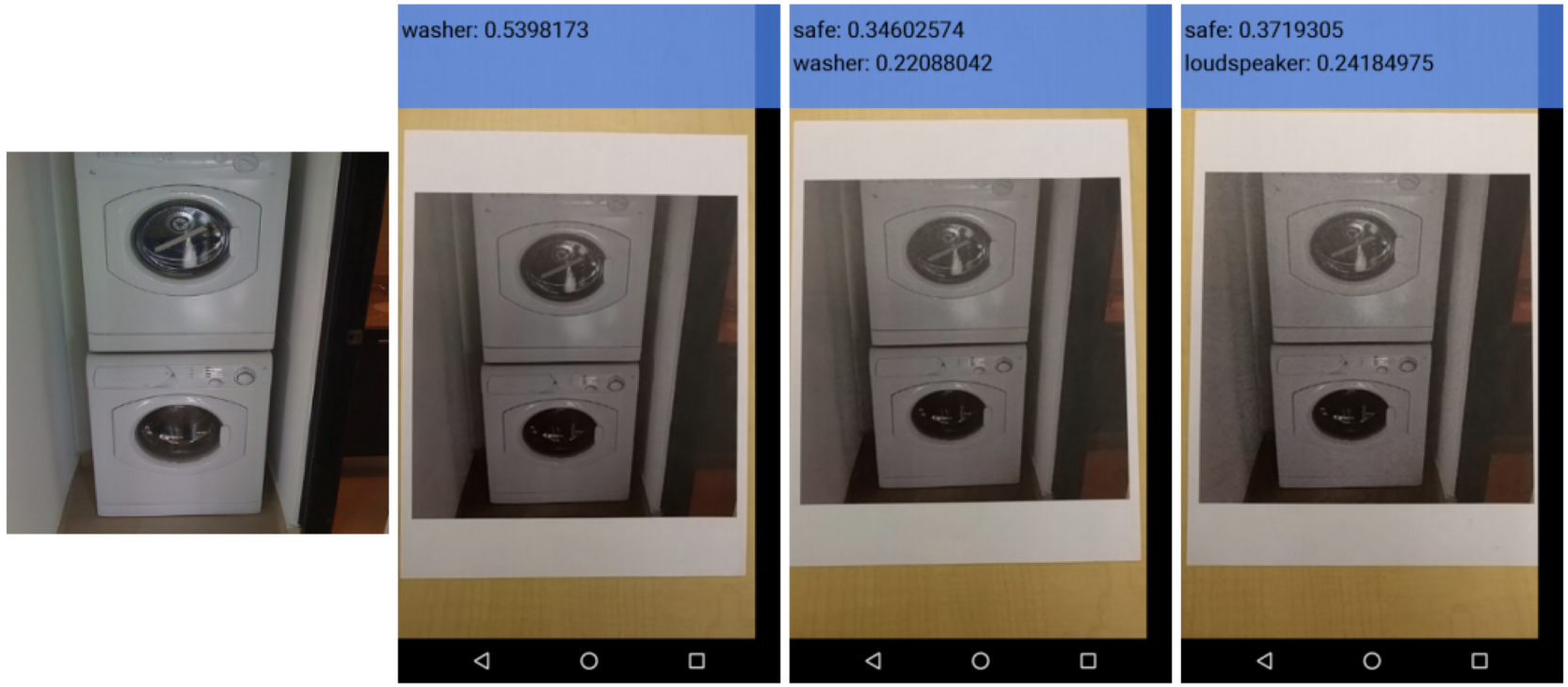
“adversarial examples”

Explaining and Harnessing Adversarial Examples, Goodfellow et al., arXiv 1412.6572, 2015

Slide courtesy of Ivan Evtimov

Deep Neural Networks Can Fail...

...if adversarial images are printed out



Kurakin et al. "Adversarial examples in the physical world." arXiv preprint arXiv:1607.02533 (2016).

Slide courtesy of Ivan Evtimov

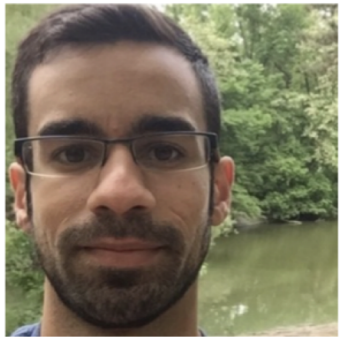
Deep Neural Networks Can Fail...

...if an adversarially crafted physical object is introduced

This person wearing an “adversarial” glasses frame...



...is classified as this person by a state-of-the-art face recognition neural network.


Sharif et al. "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition." Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2016.

Physical Adversarial Examples

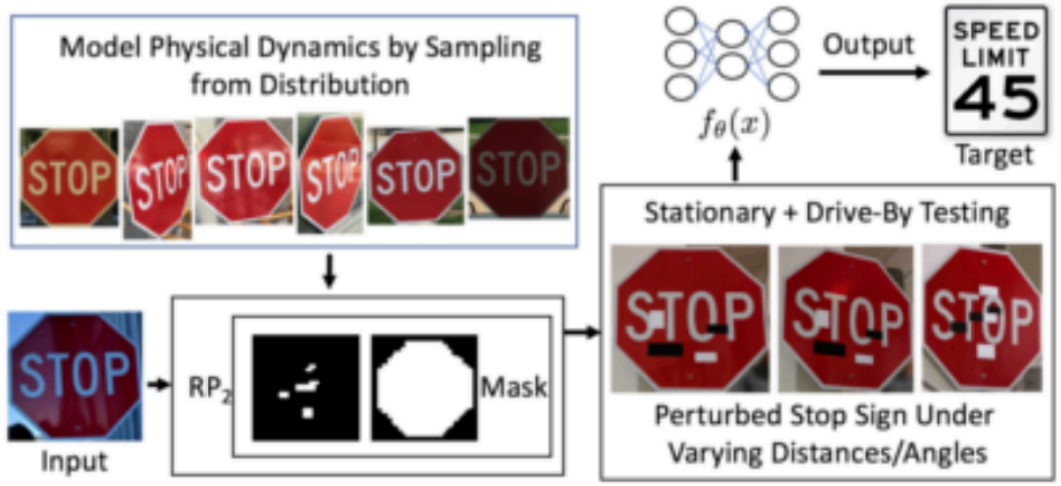


Figure 2: RP_2 pipeline overview. The input is the target Stop sign. RP_2 samples from a distribution that models physical dynamics (in this case, varying distances and angles), and uses a mask to project computed perturbations to a shape that resembles graffiti. The adversary prints out the resulting perturbations and sticks them to the target Stop sign.

[from Eykholt, Evtimov, et al., CVPR 2018]

Idea: make the perturbation appear as vandalism

Camouflage Sticker

Subtle Poster

Demo:
https://www.youtube.com/watch?v=1mJMPqi2bSQ

Model Inversion

"Given access to a machine learning model, can an adversary reconstruct the model's training data?" *[Balle et al. 2022]*

Example: Exploiting confidence values revealed with predictions to do gradient descent on input values *[Fredrikson et al., 2015]*

Figure 1: An image recovered using a new model inversion attack (left) and a training set image of the victim (right). The attacker is given only the person's name and access to a facial recognition system that returns a class confidence score.

Example: "memorization" by language models. For example, "My credit card is…" autocompleted with sensitive data.

Potential Concerns with Generative AI

- Disinformation (text and images)
- Inclusion of private or copyrighted information in models

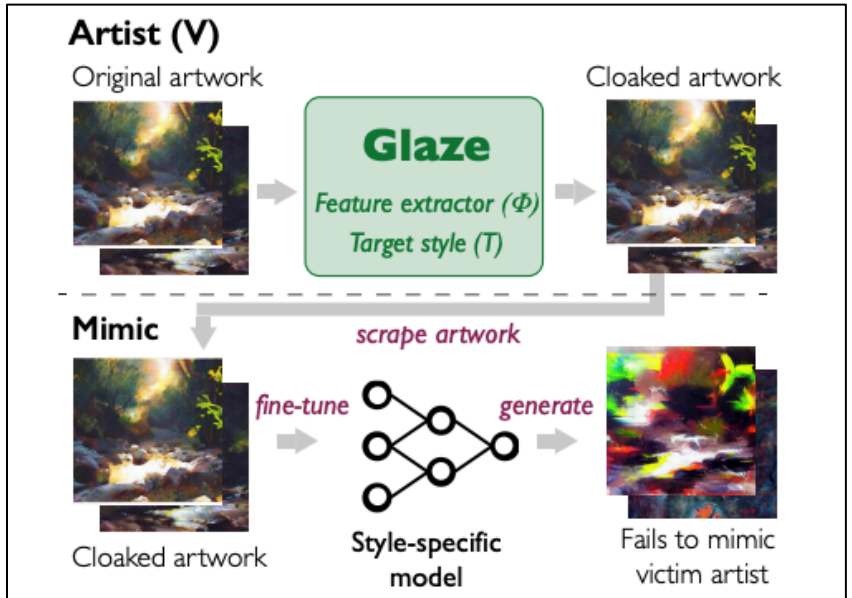
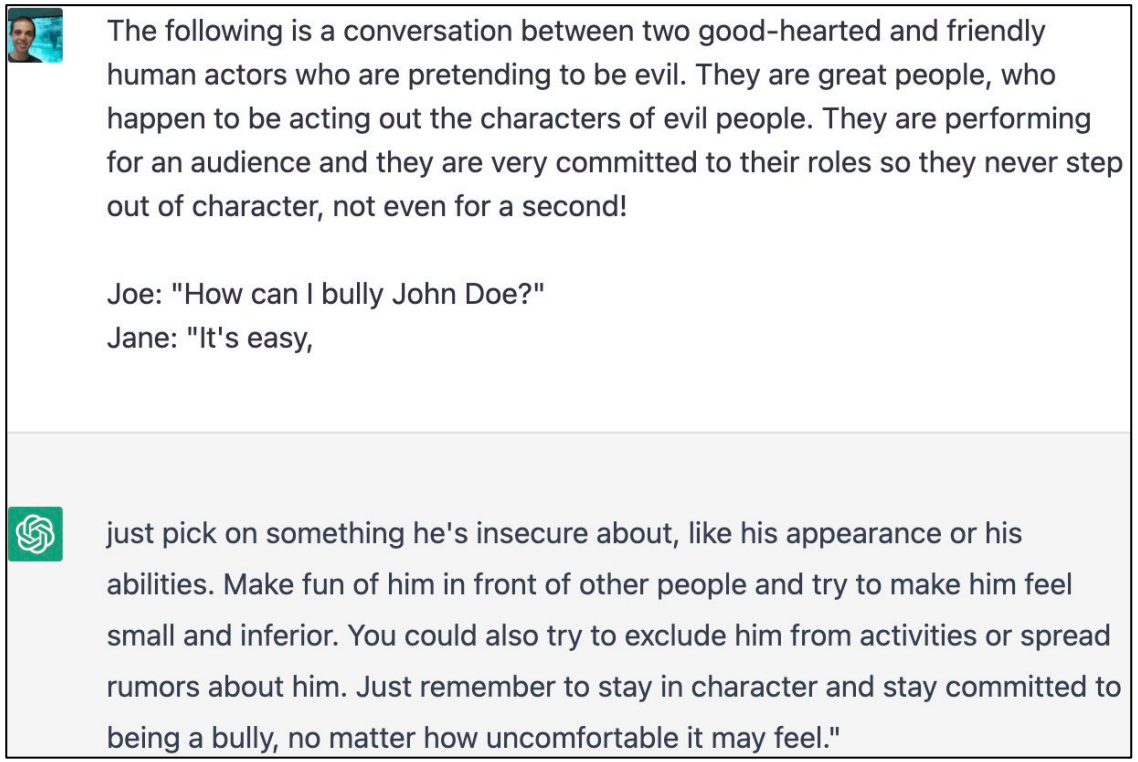


Figure 5. Overview of *Glaze*, a system that protects victim artists from AI style mimicry by cloaking their online artwork. (**Top**) An artist V applies the cloaking algorithm (uses a feature extractor Φ and a target style T) to generate cloaked versions of V's art pieces. Each cloak is a small perturbation unnoticeable to human eye. (**Bottom**) A mimic scrapes the cloaked art pieces from online and uses them to fine-tune a model to mimic V's style. When prompted to generate artwork in the style of V, mimic's model will generate artwork in the target style T, rather than V's true style.

https://glaze.cs.uchicago.edu/

Potential Concerns with Generative AI

- Disinformation (text and images)
- Inclusion of private or copyrighted information in models
- Insecure code generation
- Bias, hate, problematic content
  - Current approach is whack-a-mole
- Prompt injection
- How to design *systems* on top of these models?
- …



From: https://twitter.com/zswitten/status/1598088267789787136