Synthesizing Robust Adversarial Examples

Anish Athalye*, Logan Engstrom*, Andrew Ilyas*, Kevin Kwok

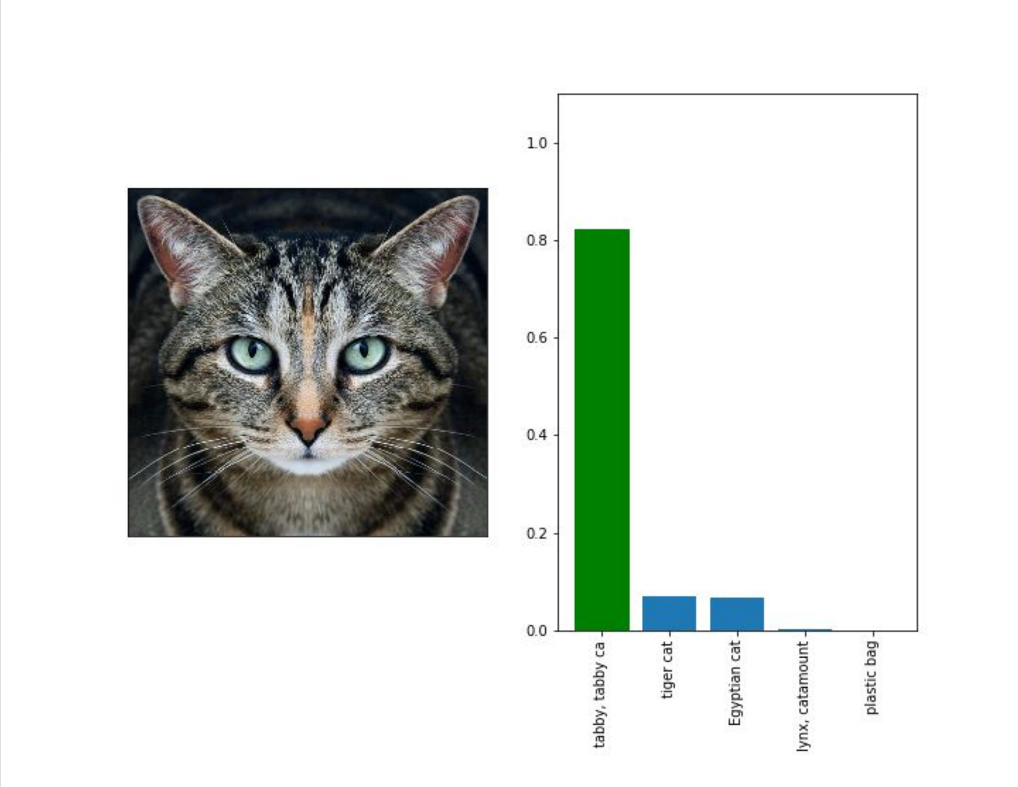
Standard Adversarial Examples

Given image x; target class y

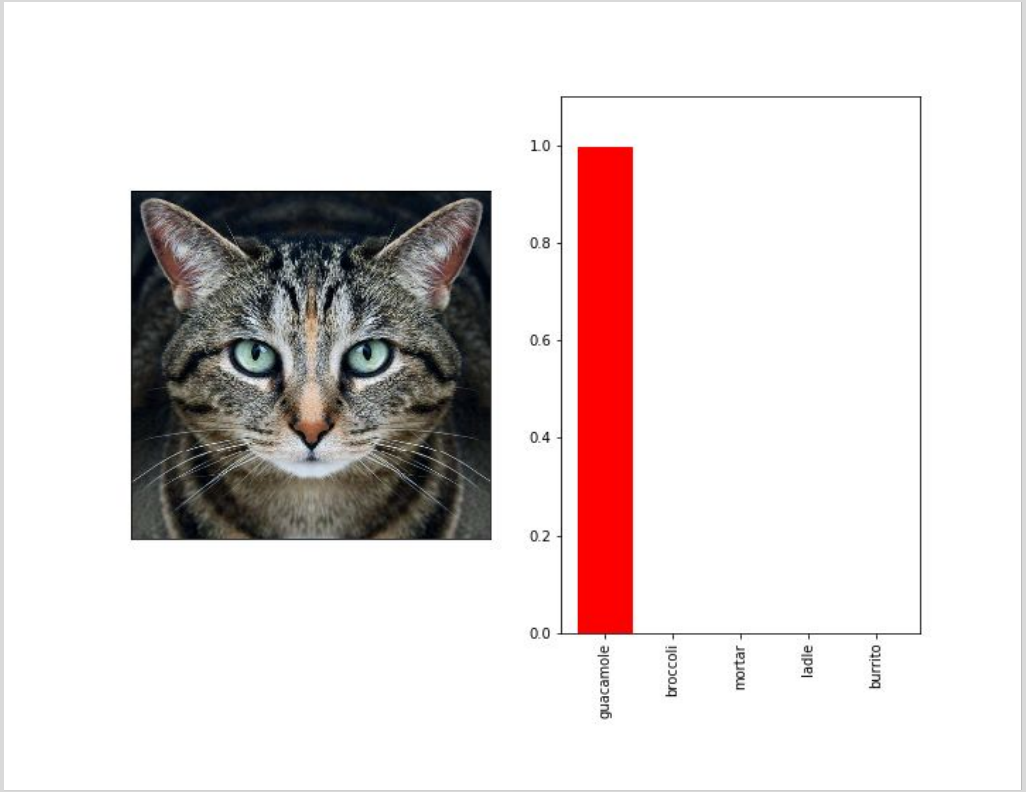
Maximize with projected gradient descent:

$$x_{adv} = \arg\max_x P(y|x) \qquad \text{s.t. } d(x, x_0) < \epsilon$$

Standard Adversarial Examples



Standard Adversarial Examples



Standard Adversarial Examples

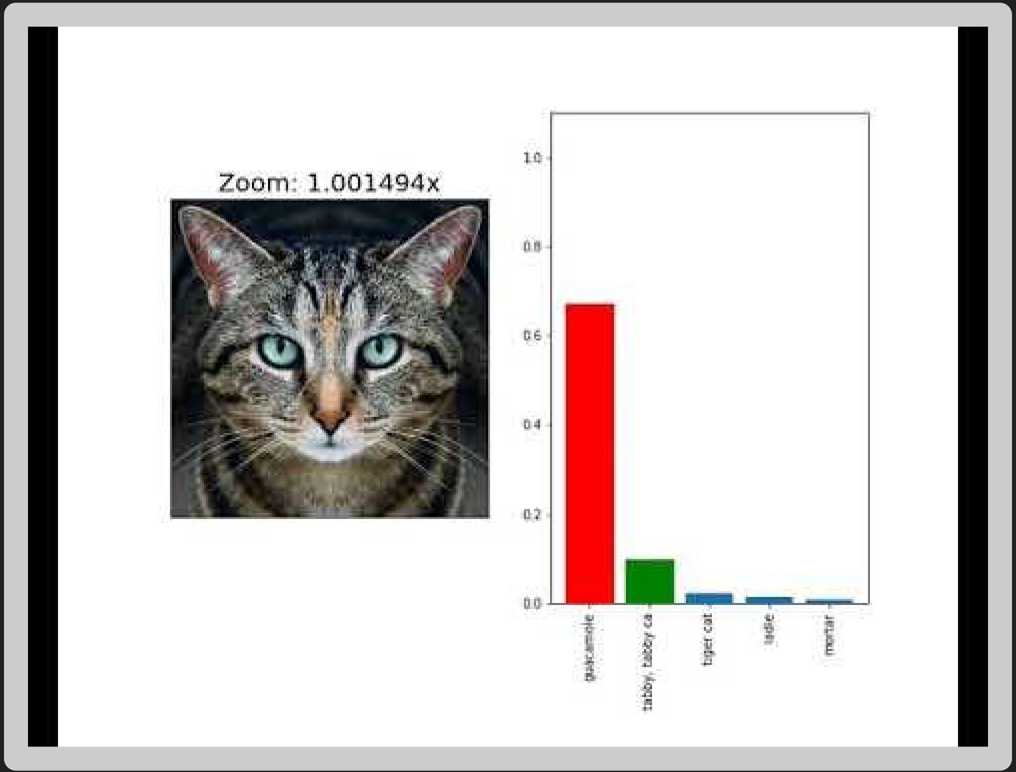
Given image x; target class y

Maximize with projected gradient descent:

$$x_{adv} = \arg\max_{x} P(y|x) \qquad \text{s.t. } d(x, x_0) < \epsilon$$

What happens when we transform the images?

Standard Examples are Fragile



Robust Adversarial Examples

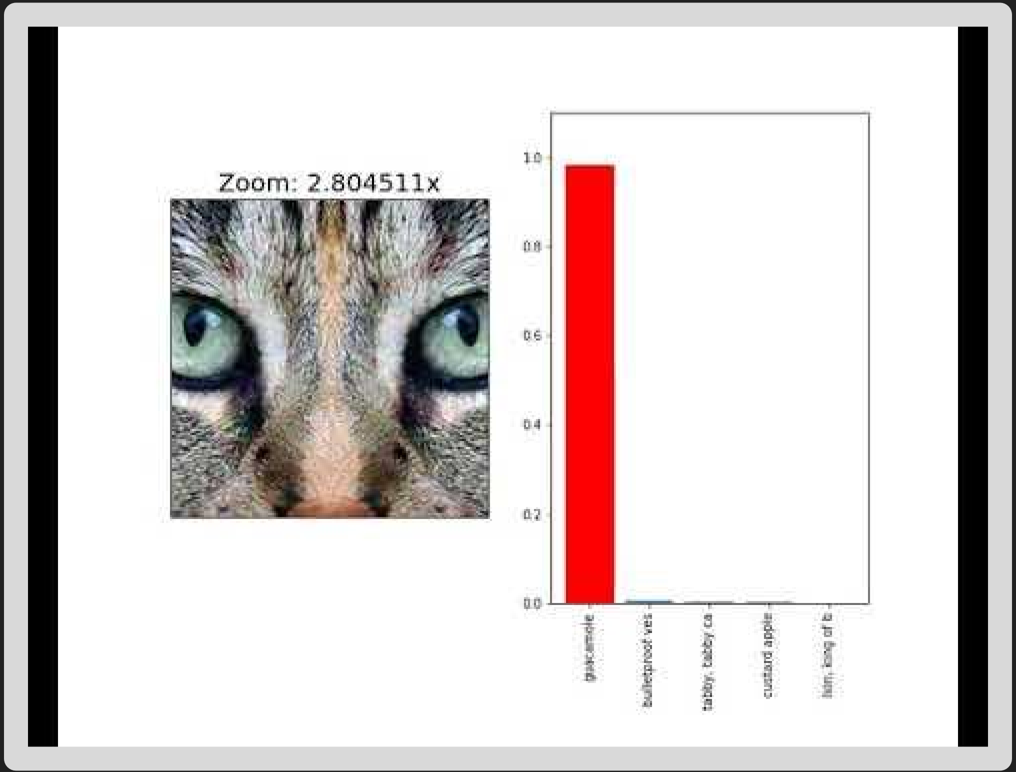
Given image x; target class y; distribution of transformations T

Maximize **expectation over transformation**:

$$x_{adv} = \arg\max_{x} \mathbb{E}_{t \sim T}[P(y|t(x))] \qquad \text{s.t. } d(x, x_0) < \epsilon$$

What happens when we transform the images?

Robust Adversarial Examples



Implementation

Euclidean LAB distance:

$$d(x, x_0) := \mathbb{E}_{t\sim T}||\mathrm{LAB}(t(x)) - \mathrm{LAB}(t(x_0))||_2$$

Lagrangian Relaxation:

$$\hat{x} = \arg\min_{x'} \mathbb{E}_{t\sim T}[-\log P(y|t(x')) + \lambda||LAB(t(x)) - LAB(t(x'))||_2^2]$$

Law of Large Numbers:

$$\mathbb{E}_{t\sim T}[P(y|t(x))] \approx \frac{1}{N}\sum_{t_i\sim T} P(y|t_i(x))$$

$$\mathbb{E}_{t\sim T}[||t(x) - t(x_0)||] \approx \frac{1}{N}\sum_{t_i\sim T} ||t_i(x) - t_i(x_0)||$$

Results

Original: Persian cat
$P(true)$: 97%
$P(adv)$: 0%
$P(true)$: 99%
$P(adv)$: 0%
$P(true)$: 19%
$P(adv)$: 0%
$P(true)$: 95%
$P(adv)$: 0%
Adversarial: jacamar
$\ell_2 = 2.1 \times 10^{-1}$
$P(true)$: 0%
$P(adv)$: 91%
$P(true)$: 0%
$P(adv)$: 96%
$P(true)$: 0%
$P(adv)$: 83%
$P(true)$: 0%
$P(adv)$: 97%

Scaling EOT to 3D

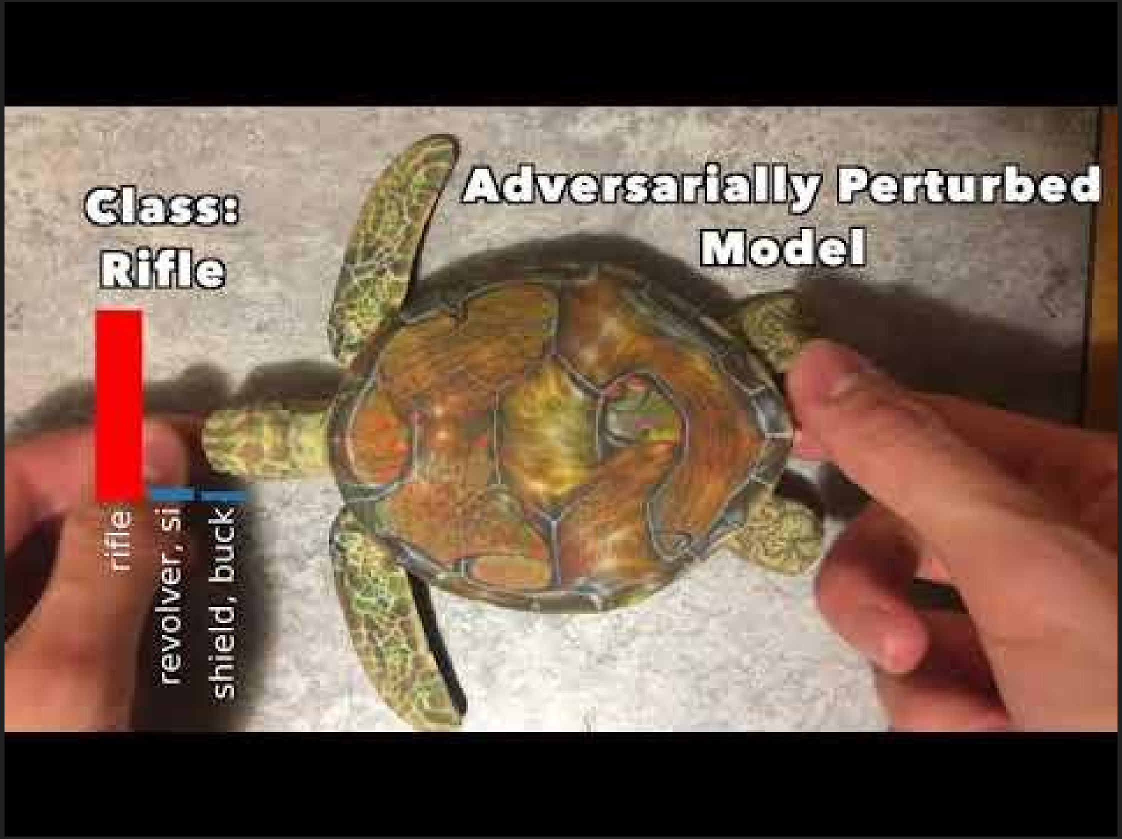
Bundle everything into the transformation:

- 3D rendering
- 3D rotation
- Perspective projection
- Lighting
- Noise

Challenges

- Implementing a differentiable renderer
- Modeling 3D printer color inaccuracy
- Approximating physical phenomena
- Choosing parameters of distribution

Demo


Adversarially Perturbed
Model
Class:
Rifle
rifle
revolver, si
shield, buck