Adversarial Examples that Fool both Computer Vision and Time-Limited Humans

G.F. Elsayed[1], S. Shankar[2], B. Cheung[3], N. Papernot[4], A. Kurakin[1], I. Goodfellow[1], J. Sohl-Dickstein[1]

[1]Google Brain [2]Stanford University [3]UC Berkeley [4]Pennsylvania State University

arXiv: 1802.08195

Reviewed by : Bill Zhang
University of Virginia
https://qdata.github.io/deep2Read/

Outline

Introduction

Basic Premise and Motivation

- Interesting phenomenon: adversarial examples often transfer from one model to another
- Perhaps humans can also be susceptible; already prone to cognitive bias and optical illusions, but not how adversarial examples work
- Neuroscience often used as existence proof for ML capabilities; if humans can resist certain classes of adversarial examples, ML models should also be able to
- Likewise, if adversarial examples can affect brain, may help understanding of neuroscience

Related Work

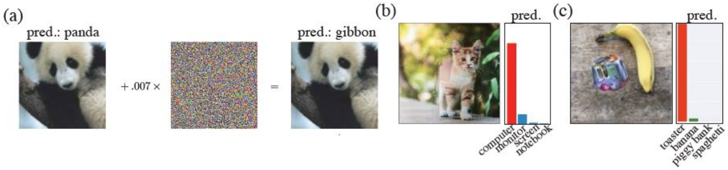
Adversarial Examples

- Goodfellow et al. defines adversarial examples as "inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake"
- Important: adversarial examples are designed to cause a mistake, not to differ from human judgment; assume that perturbations do not change true class
- Important: adversarial examples are not defined to be imperceptible

Related Work

Clues that Human Transfer is Possible

- Adversarial examples transfer across ML models, even with differing architectures, training sets, and algorithms
- Kurakin et al.: adversarial examples transfer from digital to physical world, despite differences in lighting and cameras
- Liu et al.: adversarial examples optimized to fool many models more likely to fool another model
- Recent studies have also found that adversarial examples sometimes have meaningful transformations to human observers (i.e. cat to computer seems more computer-like to humans)



Related Work

Biological and Artificial Vision

- Recent research has found similarities between deep CNNs and primate visual system
- Activity in deeper CNN layers predictive of visual pathway of primates
- Reisenhuber and Poggio: developed model of object recognition in human cortex that is very similar to CNNs
- Kummerer et al.: CNNs predictive of human gaze fixation
- Style transfer: intermediate CNN layers capture artistic style meaningful to humans
- Freeman et al.: used representations in CNN-like model to develop psychophysical metamers

Related Work

Notable Differences

- Images used for CNNs typically static rectangular images with constant spatial resolution
- Primate eye has eccentricity dependent spatial resolution; also sensitive to time and non-uniform colors
- CNNs fully feed-forward architectures; human cortex has many more feedback connections
- Humans do not consider static scenes, but actively explores with saccades

Methods

Models and Datasets

- Images from ImageNet
- Used 6 specific classes: dog, cat, broccoli, cabbage, spider, snake
- Further grouped into 3 larger classes: pets, hazards, vegetables
- Used ensemble of k CNN models trained on ImageNet
- Prepend each model with retinal layer with eccentricity-dependent blurring to approximate human image inputs
- Adversarial examples generated with iterated gradient descent with l_∞ norm of all perturbations restrained to fixed ϵ

Methods

Human Psychophysics Experiment: Procedures

- 38 subjects with normal/corrected vision
- Subjects asked to classify images appearing on screen as one of two choices
- Subjects directed to look at fixation cross and afterwards, image is shown for 63 ms, followed by 10 high contrast binary masks
- Subjects given 2.2-2.5 seconds to respond after masks appear

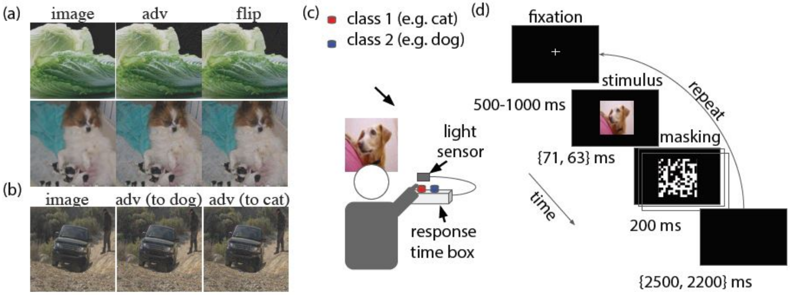
Methods

Human Psychophysics Experiment: Conditions

- Images presented in 1 of 4 conditions:
  - *image*: Original ImageNet images rescaled to [40, 255-40] to avoid clipping after adding perturbations
  - *adv*: Perturbed images; used $\epsilon = 32$, large enough to be noticed by humans but small enough that no-limit humans still identify true class correctly
  - *flip*: Same as adv, except flip perturbation vertically before adding to image; make sure changes in human accuracy are not caused by image distortion
  - *false*: Two options presented as choices are both wrong; see if adversarial examples can influence towards specific wrong choice
- Pre-filtered images to not have large distinctions between classes due to brightness or overall color

Methods

Experiment Diagram



Results

Transfer to Computer Vision Models

- Assess transfer of adversarial examples to two test models not included in ensemble
- Both models have $> 75\%$ accuracy on clean images
- *adv* and *false* examples succeeded $57 - 89\%$ of the time, *flip* succeeded less than 1.5% of the time, validating its use as a control

Results

Transfer to Humans

- Want to show that adversarial examples do not simply degrade image quality or discard information to increase human error rate
- Therefore, first show that with a fixed error rate (where human is forced to be wrong) adversarial examples influence choice among two classes
- Then, show that adversarial examples increase error rate

Results

Transfer to Humans: Two Incorrect Classes

- Used the *false* condition images
- If adversarial perturbation completely ineffective, would expect choice of target class to be uncorrelated to with subject's reported class; average rate should be 0.5 for each image
- Used larger class groups (pets, hazards, vegetables)
- In all cases, probability significantly above 0.5
- Also found that reaction time inversely correlated with perceptual bias pattern i.e. subjects more confident when adversarial perturbation more successful when biasing decision

Results

Transfer to Humans: Increase in Human Error Rate

- Now show that we can bias human response against true class even when true class is an option
- Used *image*, *adv*, and *flip* conditions
- Most subjects had lower accuracy on *adv* than *image*
- Result may, however, only imply that signal to noise ratio in adversarial images is lower; partially addressed with *flip* which has perturbation with identical statistics
- Majority of subjects also had lower accuracy on *adv* than *flip* images

Results

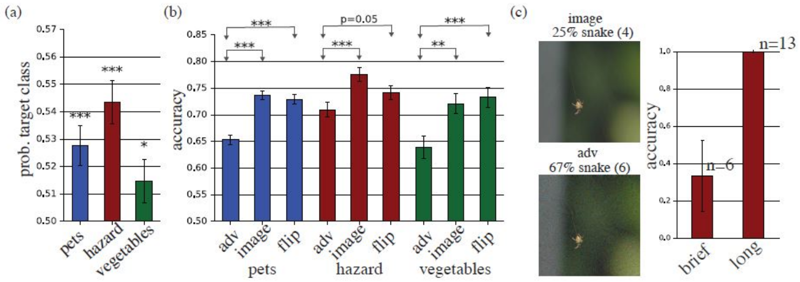
Transfer to Humans: Increase in Human Error Rate

- Results suggest that direction of adversarial perturbation with specific image produces perceptually relevant features for humans
- Perhaps strong black box attacks to CNNs can transfer to humans
- Interestingly, average response time longer for *adv* condition images; seems to contradict *false* condition's results
- Perhaps in *false* case perturbations caused higher confidence and in *adv* case perturbations caused lower confidence due to competing adversarial and true class features in *adv*

Results

Graphs: Human Error Rate

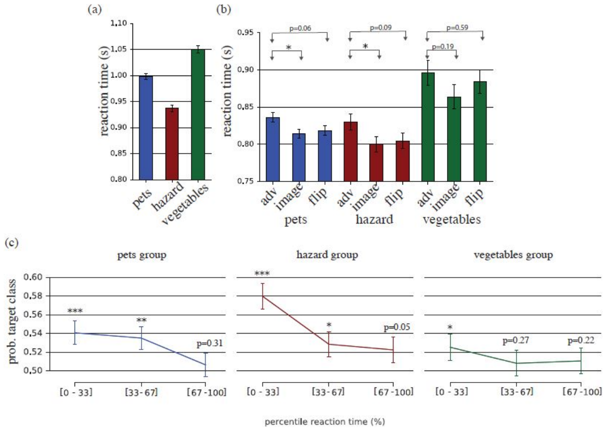
- a) Probability of choosing correct target class significantly > 0.5
- b) Adversarial images cause more mistakes than both original image and image with flipped perturbation
- c) Image of spider that time-limited humans perceived to be a snake



Results

Graphs: Human Response Time

- a) Average response time to *false* images
- b) Average response time to *image*, *adv*, and *flip*
- c) Probability of choosing correct target class decreases with increased reaction time in *false*



Discussion

- Did examples fool humans or did they change the true class?
  - Perturbations small enough that true class unchanged for human with no time limit
  - Thus, we can be confident that examples did fool humans
- How did the adversarial examples work?
  - No controlled experiments, but generally observed edge disruptions, enhancing edges through increased contrast and creating texture boundaries, modifying textures, and taking advantage of dark regions of images

Discussion

- What are the implications for ML security and society?
  - The fact that the examples fool time-limited humans but not no-limit humans suggest lateral and top-down connections used by no-limit human are relevant to human robustness against adversarial examples
  - Perhaps ML models can become more robust through similar connections
  - Also suggest that images can be manipulated to cause human observers to have unusual reactions
- Future Work
  - How does transfer to humans depend on ϵ?
  - Was model ensembling crucial for the transfer?
  - Can retinal preprocessing layer be removed?

Conclusion

- This work showed that adversarial examples based on perceptible but class-preserving perturbations that fool multiple ML models can also fool time-limited humans
- Show strong similarities between CNNs and human visual system; expect work to help in both future machine learning and neuroscience research

References

- https://arxiv.org/pdf/1802.08195.pdf