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

Gamaleldin Elsayed

Shreya Shankar, Brian Cheung, Nicolas Papernot, Alex Kurakin, Ian Goodfellow, Jascha Sohl-Dickstein

Google Brain, MTV

September 17, 2018

Outline

1. Background and Motivation
2. Methods
3. Task and Experiment
4. Results
5. Conclusions

Background and Motivation

Google

What is an adversarial example?

- Inputs that are designed by an adversary/attacker to make a machine learning model make wrong decisions.
- Adversarial examples in computer vision:
    - Perturbations added to images to make a computer vision model misclassify images.

Szegedy et al., 2014

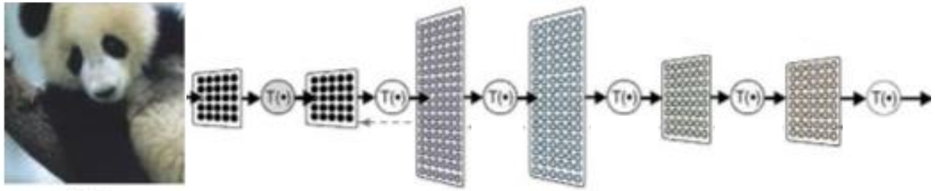
Safety and security concern

Original Sequence

Attacked Sequence

Lu et al., 2017

Are adversarial examples specific to computer vision models or they can also affect a presumably superior system like our brains ?



“gibbon”

?

Figure adapted from http://discoveryeye.org/optic-nerve-visual-link-brain/

Can adversarial examples transfer to humans?

- Adversarial examples are often generated using an optimization process that require access to model parameters and architecture.
- Without similar access to human brain, transfer of adversarial examples to human may seem to be an impossible task.

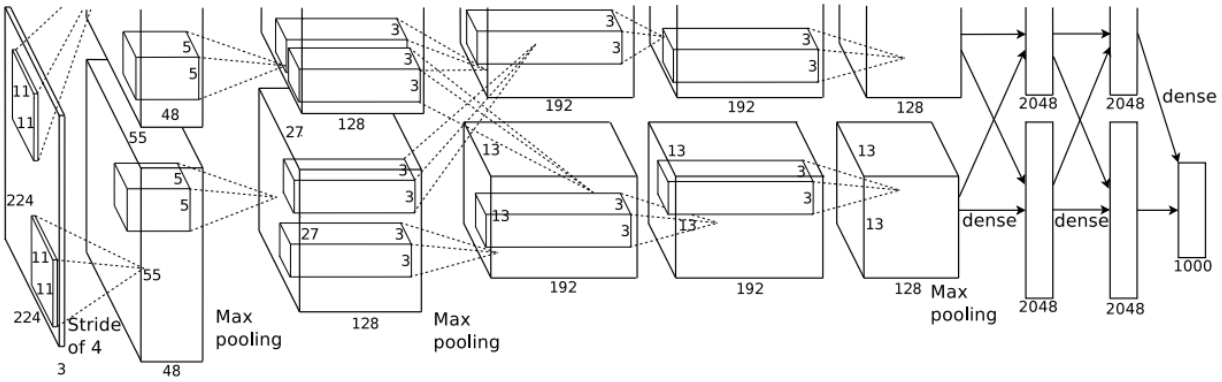


Figure adapted from Krizhevsky et al. 2012

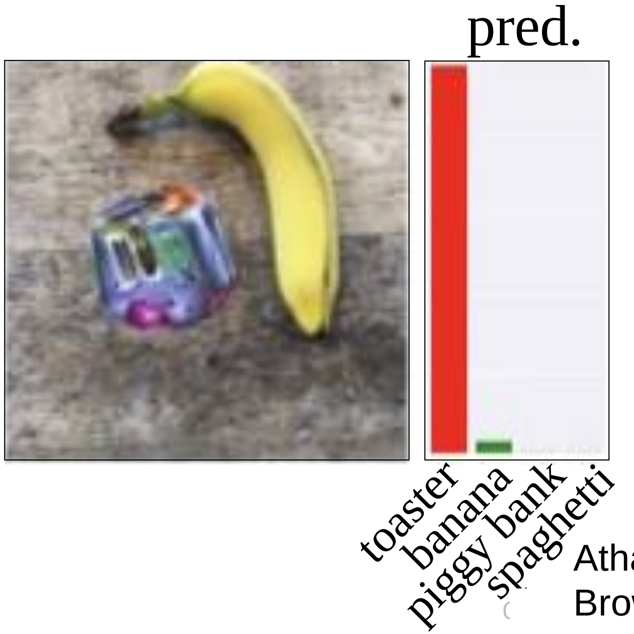
Clues for possibility to transfer to humans

- Adversarial examples have been shown to successfully transfer to other models that an attacker does not have access to by optimizing multiple models:
    - different architecture
    - trained on different data
    - trained with different loss function

Liu et al., 2016

Clues for possibility to transfer to humans

- Adversarial examples when made invariant to transformation, the perturbation seemed to be somewhat relevant to humans.



Athalye et al. 2017
Brown et al. 2017

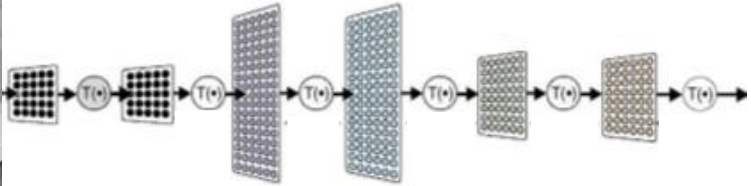
Hypothesis

- **H:** Adversarial examples that strongly transfer across machine learning models, target features that are relevant to human visual system and thus can transfer to humans.
- **Testing methodology:**
  - Account for the known architecture mismatch between human visual system and computer vision models.
  - Design adversarial images that strongly transfer across computer vision models.
  - Evaluate accuracy of people on identifying the true class of adversarial images.

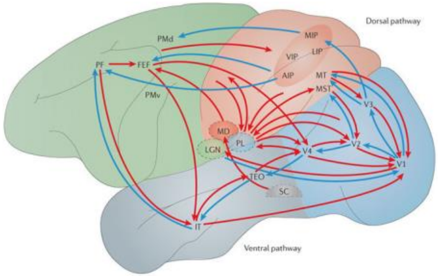
Methods

Reducing the gap between models and the brain

- Initial visual processing:
    - Retinal blurring layer



- Feedback:
    - Limited time presentation
    - Backward masking



Figures adapted from Gilbert et al., 2017 and
http://home.deib.polimi.it/boracchi/Projects/projects.html

Dataset

- **ImageNet (1000 classes).**
- **Image Groups:**
    - Pets group: cat and dog
    - Hazard group: spider and snake
    - Vegetables group: broccoli and cabbage

Generating Adversarial Examples

- Ensemble of 10 models:
  - Probability of coarse class:

$$P_k(Y = y_{\text{target}}|X) = \sum_{i \in S_{\text{target}}} P_k(Y = y_i|X)$$

  - Joint probability of ensemble (geometric mean)

- Iterative fast gradient sign method.

$$J\left(X|y_{target}\right) = -\log\left[P_{\text{ens}}\left(y_{\text{target}}|X\right)\right]$$

$$\tilde{X}^n_{adv} = X^{n-1}_{adv} - \alpha * \text{sign}(\nabla_{X^n}(J(X^n|y_{\text{target}})))$$

Model	Top-1 accuracy
Resnet V2 101	0.77
Resnet V2 101*	0.7205
Inception V4	0.802
Inception V4*	0.7518
Inception Resnet V2	0.804
Inception Resnet V2*	0.7662
Inception V3	0.78
Inception V3*	0.7448
Resnet V2 152	0.778
Resnet V2 50*	0.708

Task and Experiment

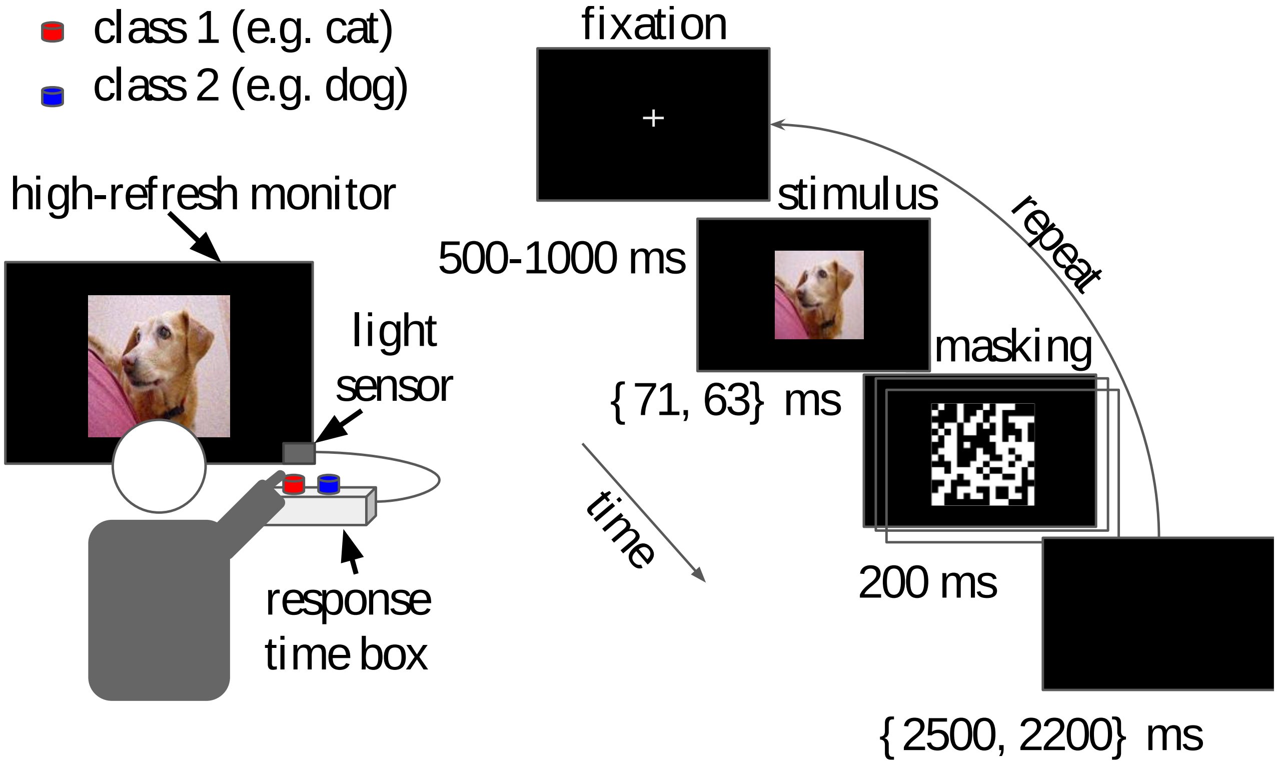
Experiment Conditions

- Image: clean image.
- Adv: adversarial image from class 1 to class 2 in the group.
- Flip (CTRL1): image with flipped adversarial perturbation (flip vertically).
- False (CTRL2): random image adversarially perturbed to one of the two classes.

Google

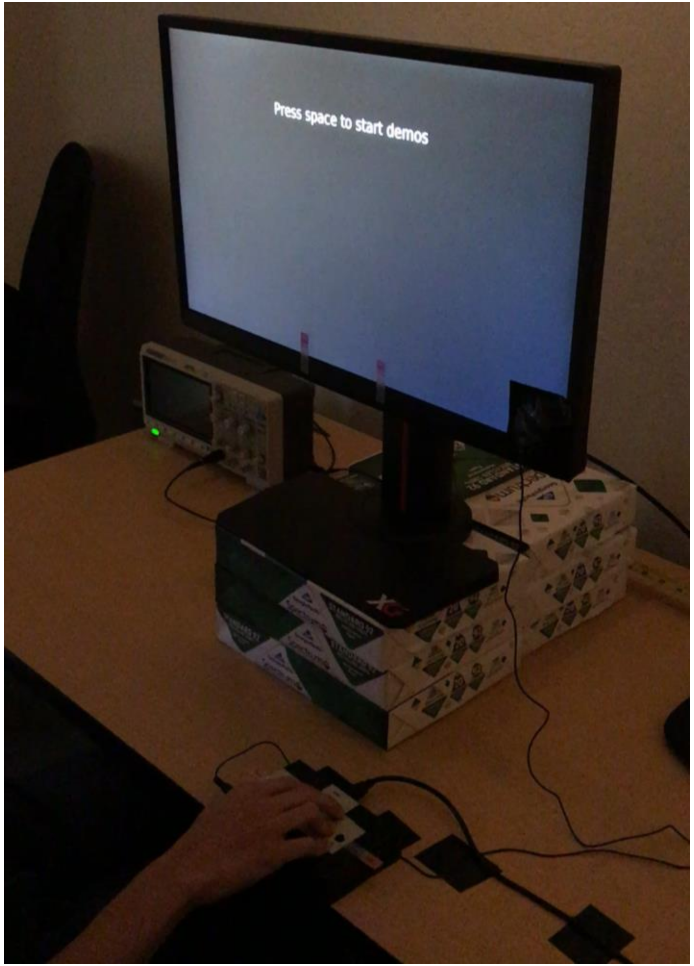
Task and Experiment

- 38 subjects.
- Recordings:
  - Choice.
  - Reaction time.



Google

Task and Experiment



Results

Model evaluations of images

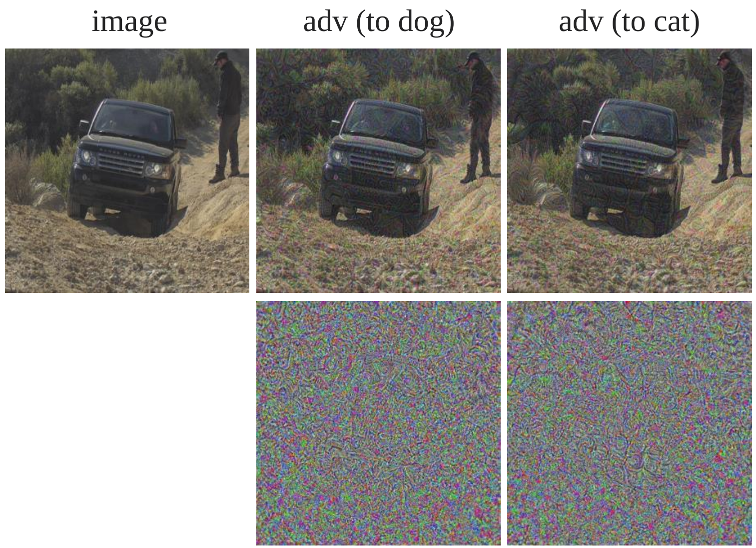
- Two test models:
  - ResNet V2 50
  - Inception V3 with adversarial training

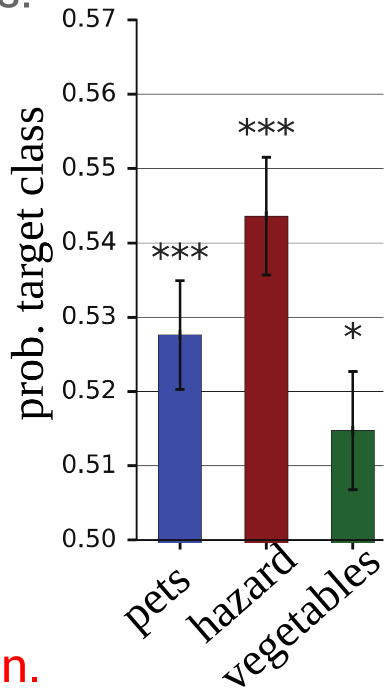
Model	Accuracy (%)			Attack Success (%)		
	adv	image	flip	adv	image	flip
ResNet V2 50	8.7, 9.4, 13	99, 98, 96	93, 91, 85	87, 85, 57	0.0, 0.0, 0.0	1.3, 0.0, 0.0
Inception V3	6.0, 6.9, 17	99, 99, 100	95, 92, 94	89, 87, 74	0.0, 0.0, 0.0	1.5, 0.5, 0.0

Adversarial examples strongly transfer to test models (black box attack).

Human evaluation of images

- false condition: subjects can **not** choose true class.



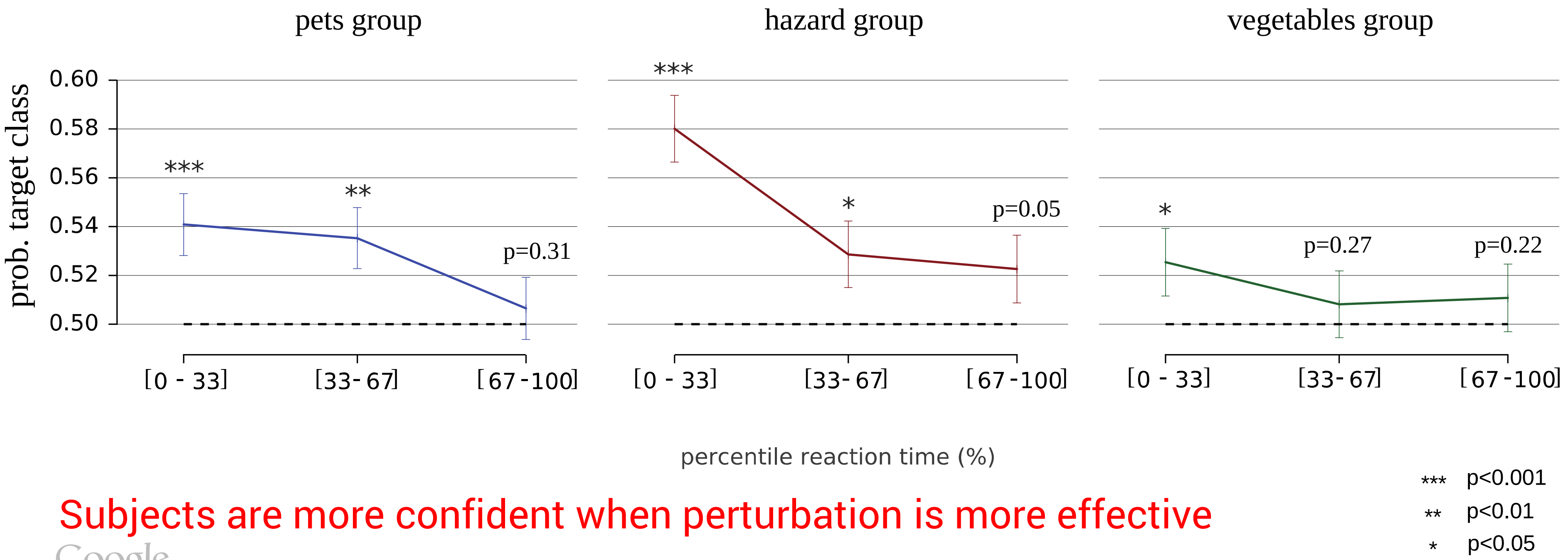


*** p<0.001
** p<0.01
* p<0.05

Adversarial perturbations bias human visual perception.

Human evaluation of images

- false condition: subjects can **not** choose true class.

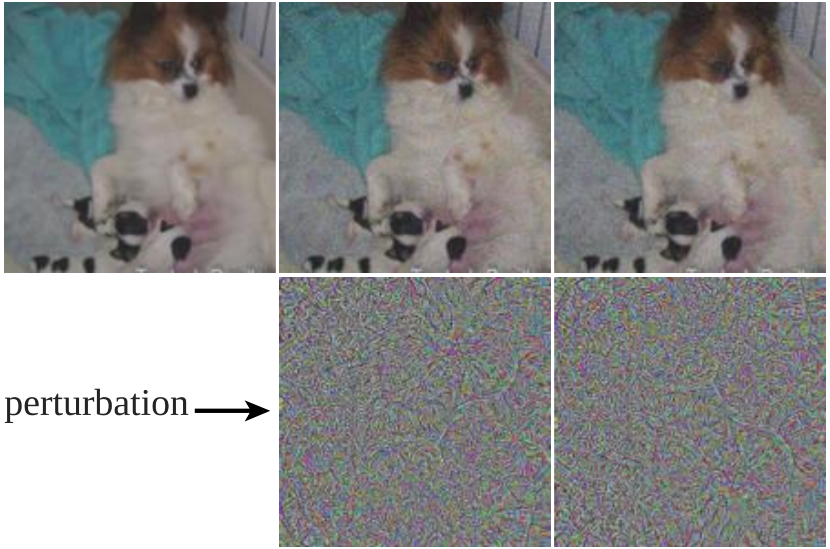


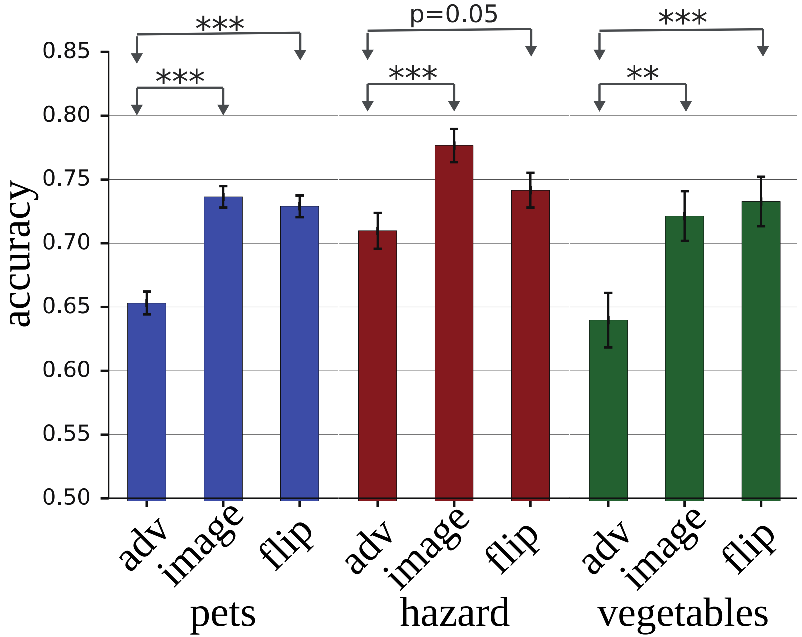
Subjects are more confident when perturbation is more effective

Google

Human evaluation of images

- image, adv and flip conditions: subjects can **now** choose true class.





Adversarial examples transfer to humans.

Google

Elsayed et al. NIPS 2018

Examples of feature manipulations

texture modification

image adv

cabbage → broccoli

Elsayed et al. NIPS 2018

Examples of feature manipulations

dark parts modification

image adv

spider → snake

Elsayed et al. NIPS 2018

Examples of feature manipulations

edge enhancement

image adv

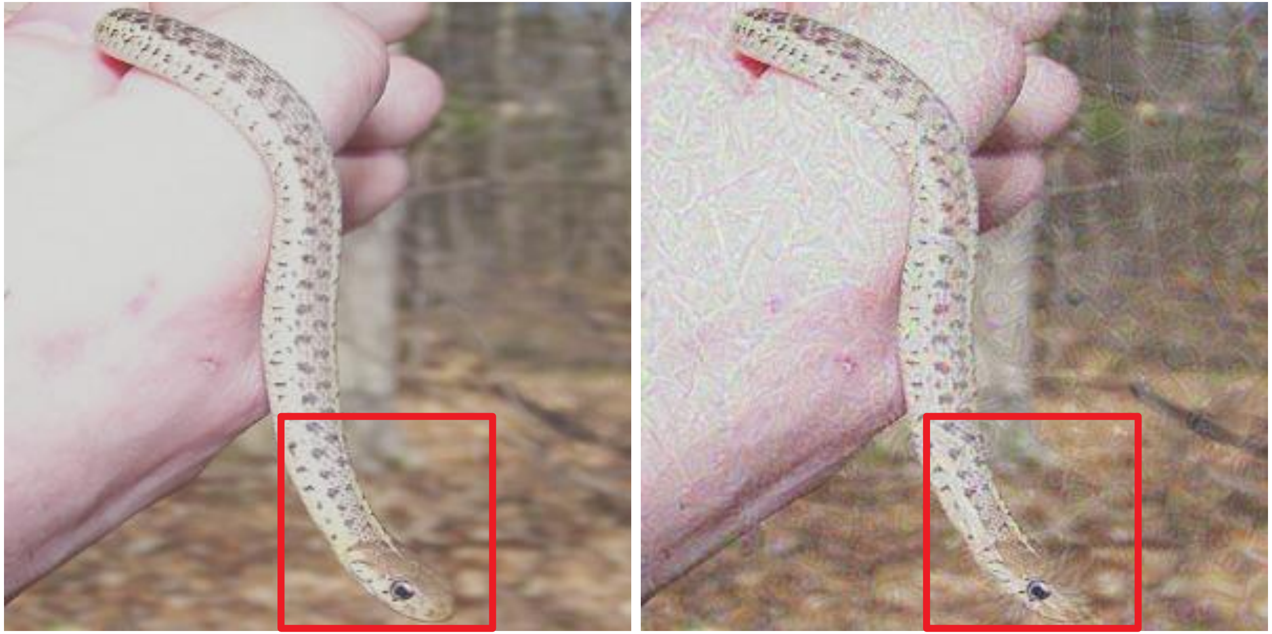
spider → snake

Google

Elsayed et al. NIPS 2018

Examples of feature manipulations

edge destruction



Elsayed et al. NIPS 2018

Limited vs unlimited presentation duration

brief
67% snake (6)

long
0% snake (13)



spider → snake

Adversarial examples transfer to humans is reduced upon long presentation.

Elsayed et al. NIPS 2018

Conclusion

Conclusion

- H: adversarial examples that strongly transfer between computer vision model transfer to humans.
- Test: generate adversarial examples that strongly transfer across models and evaluate them on humans.
- Results:
  - Adversarial perturbations bias human visual perception.
  - Adversarial examples thus can transfer to human.
  - This transfer mostly vanishes upon long time presentation.
- Decision boundary of our visual system seems to be consistent with an ensemble of convolutional neural networks.

Conclusion

- Research on how to develop models that can handle inaccurate components (e.g., back up systems, multi modalities etc).
- Computer vision models still have a big room to improve.
  - Even in time-limited settings humans are much more robust than ML models.
- For more details check our NIPS 2018 paper.
- Check the exercise based on this work in the Track Sessions.

Google AI
Residency Program

Google AI Residency Program

Program Overview

- 12-month role designed to advance career in machine learning research.
- Opportunity to work alongside distinguished machine learning researchers/engineers across various teams and leverage Google's large-scale infrastructure for research.

Interested in more information?

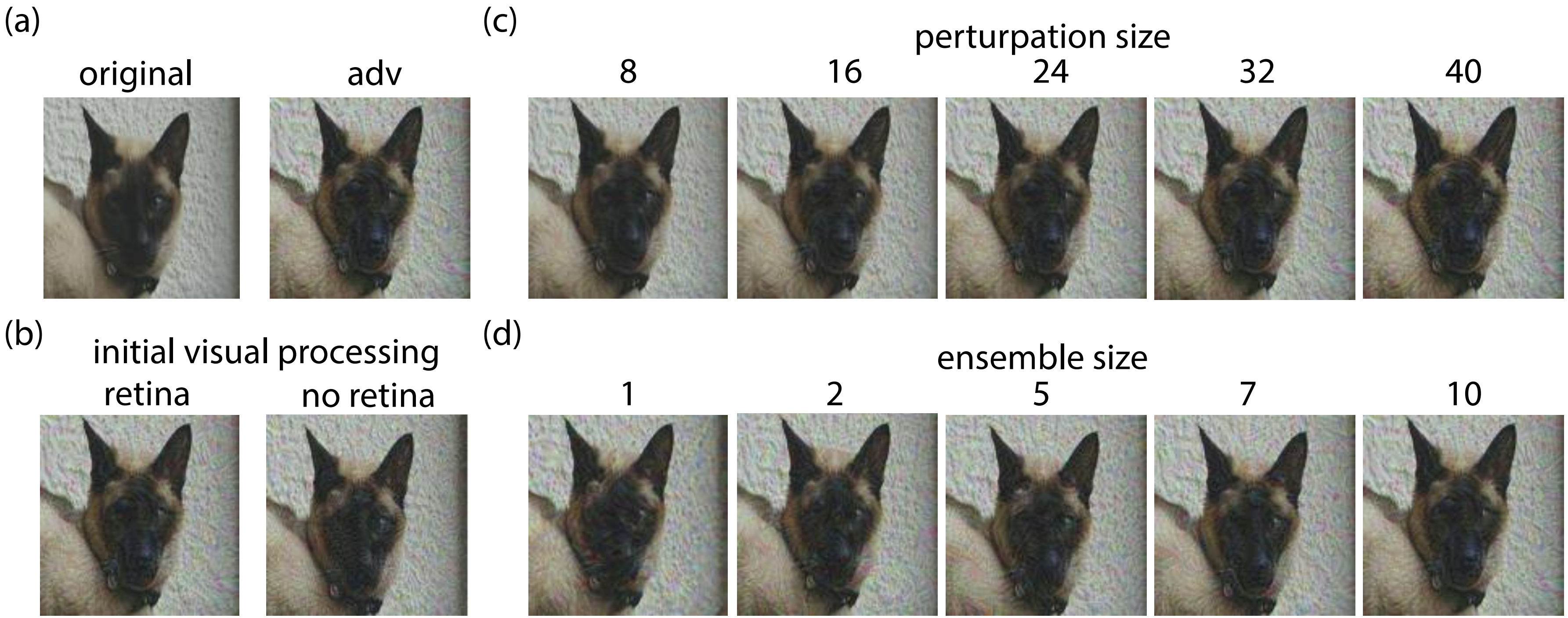
- Check out our program website at g.co/airesidency

Interested in applying?

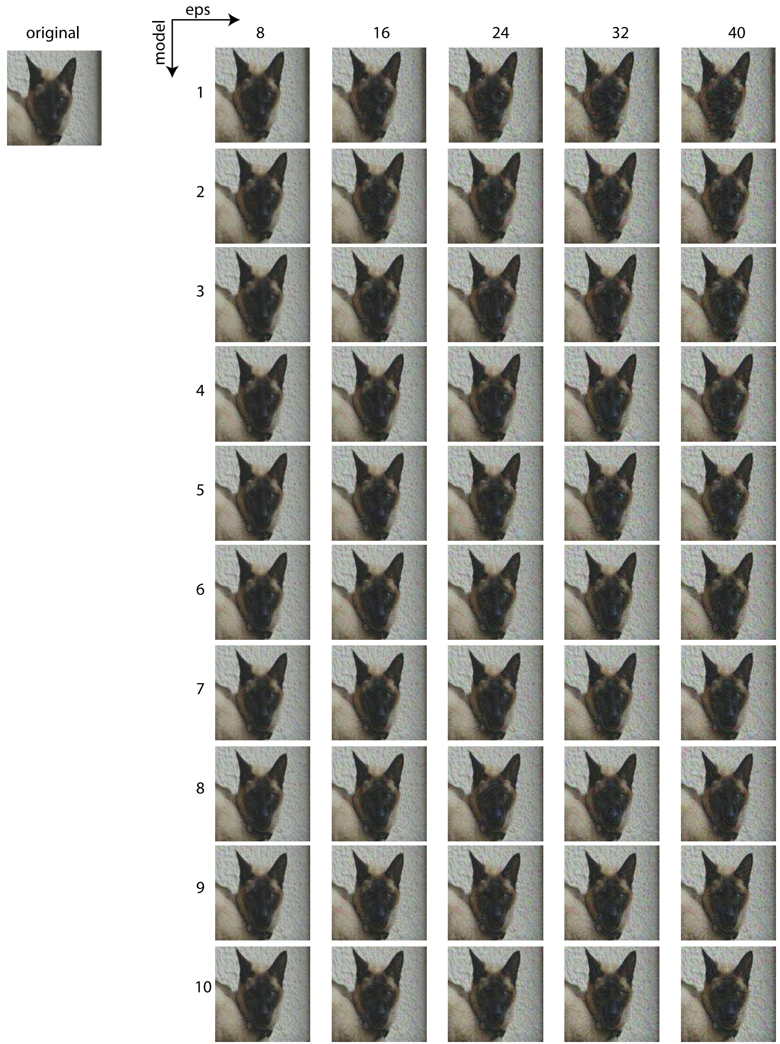
- Applications for the 2019 program is currently closed, but will **re-open on Oct 1st, 2018**!

Questions

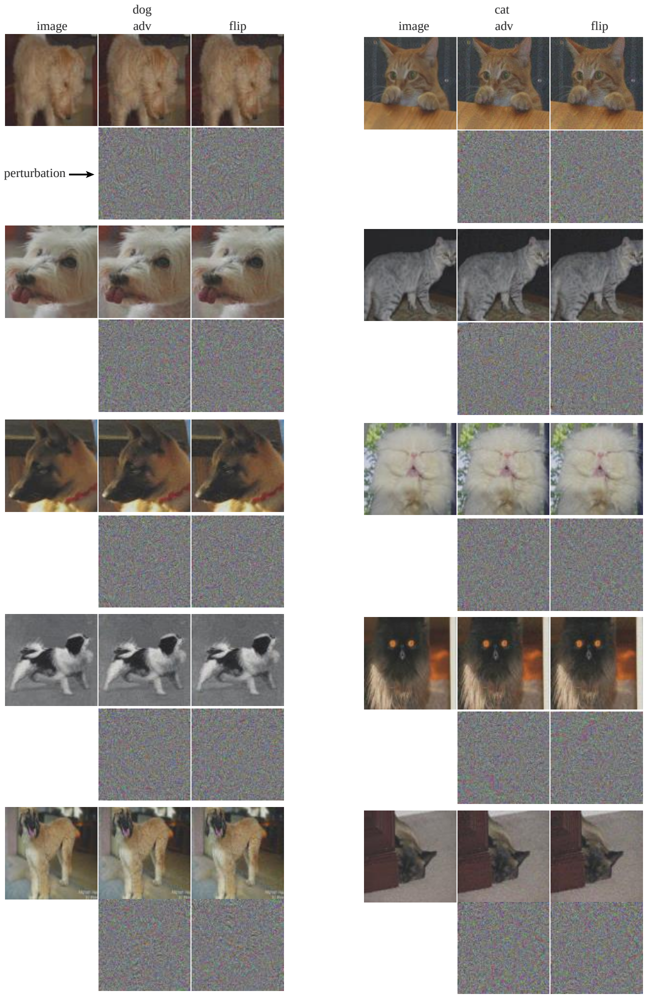
Questions



(a)
original
adv
(c)
perturpation size
8
16
24
32
40
(b)
initial visual processing
retina
no retina
(d)
ensemble size
1
2
5
7
10

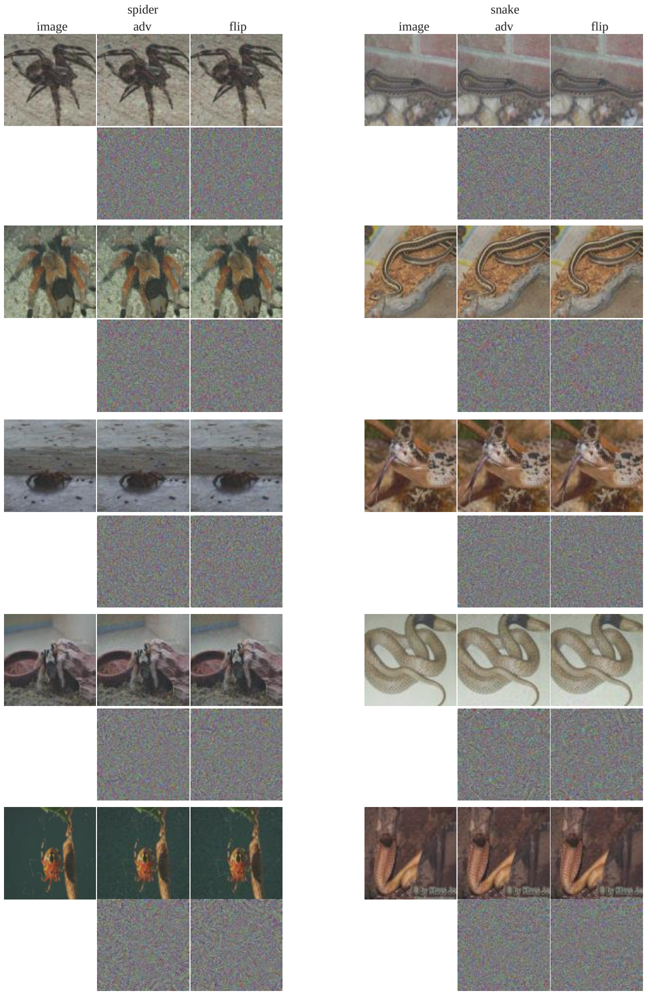

original
eps
model
8
16
24
32
40
1
2
3
4
5
6
7
8
9
10

More examples



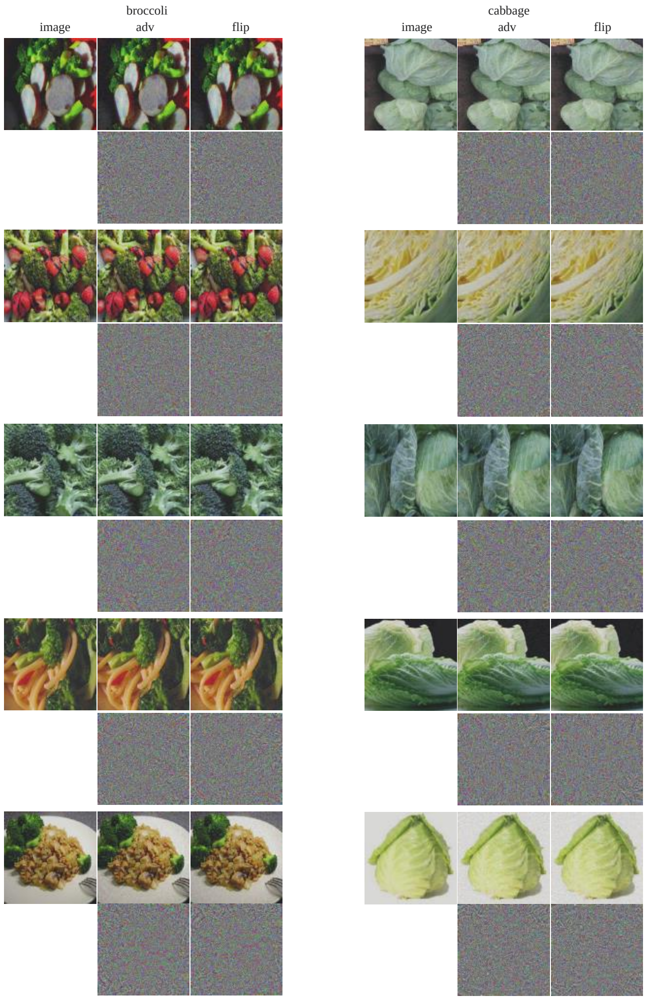
Google

More examples



Google

More examples



More examples