Deep Neural Networks, They Just (Don't) Work

Kushal Jhunjhunwalla (kushalj@cs)
Andrew Wei (nowei@cs)
Leiyi Zhang (leiyiz@cs)

CSE 499/599G1: Intro to Deep Learning
Autumn 2019

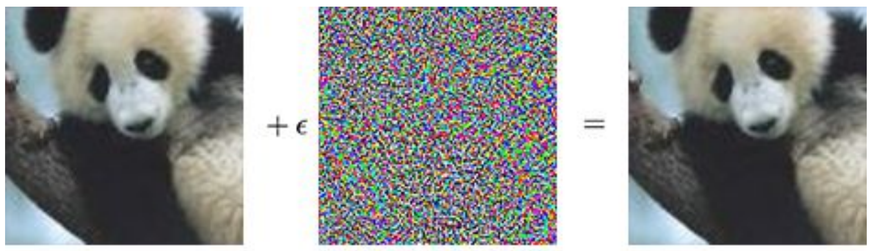
When you ask your NN why it classified your cat as a dog:

1. Egyptian cat **41.07%**
2. tiger cat **39.43%**
3. tabby, tabby cat **15.42%**
4. lynx, catamount **2.11%**
5. mouse, computer mouse **0.23%**

1. goldfish, Carassius auratus **94.53%**
2. conch **0.53%**
3. tiger cat **0.47%**
4. snail **0.28%**
5. proboscis monkey, Nasalis larvatus **0.18%**

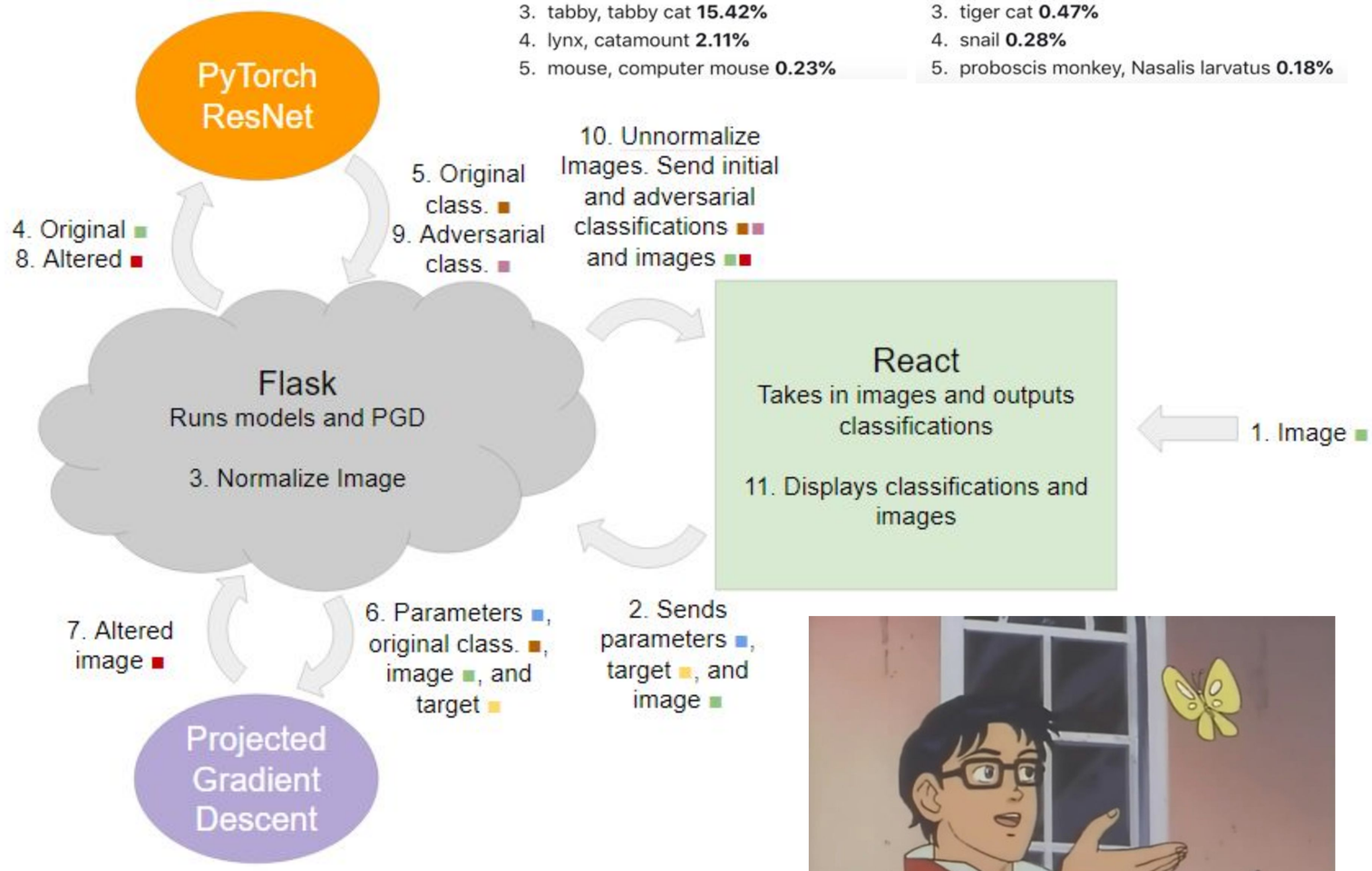
Goal

We wanted to show that DNNs weren't robust and would confidently misclassify when adding targeted noise that would be difficult for humans to notice.



"panda"
57.7% confidence

"gibbon"
99.3% confidence



What we did

Fast Gradient Sign Method (FGSM)

High level: Take the sign of the gradient and move in the opposite direction

Samoyed (89.36%) classified as Poncho (33.40)%

Projected Gradient Descent (PGD)

High level: Move in the direction opposite the gradient (or towards it if targeted) and project back down to at most an epsilon difference

(See cat picture)

Problems we faced

- Getting ImageNet data - no longer available through PyTorch
- FGSM perturbed it too much
- Training a robust model
- Perturbations between pixel values get truncated when saving the image

References

- Explaining and Harnessing Adversarial Examples (2014), Ian J. Goodfellow and Jonathon Shlens and Christian Szegedy
- Towards Deep Learning Models Resistant to Adversarial Attacks (2017), Aleksander Madry and Aleksandar Makelov and Ludwig Schmidt and Dimitris Tsipras and Adrian Vladu