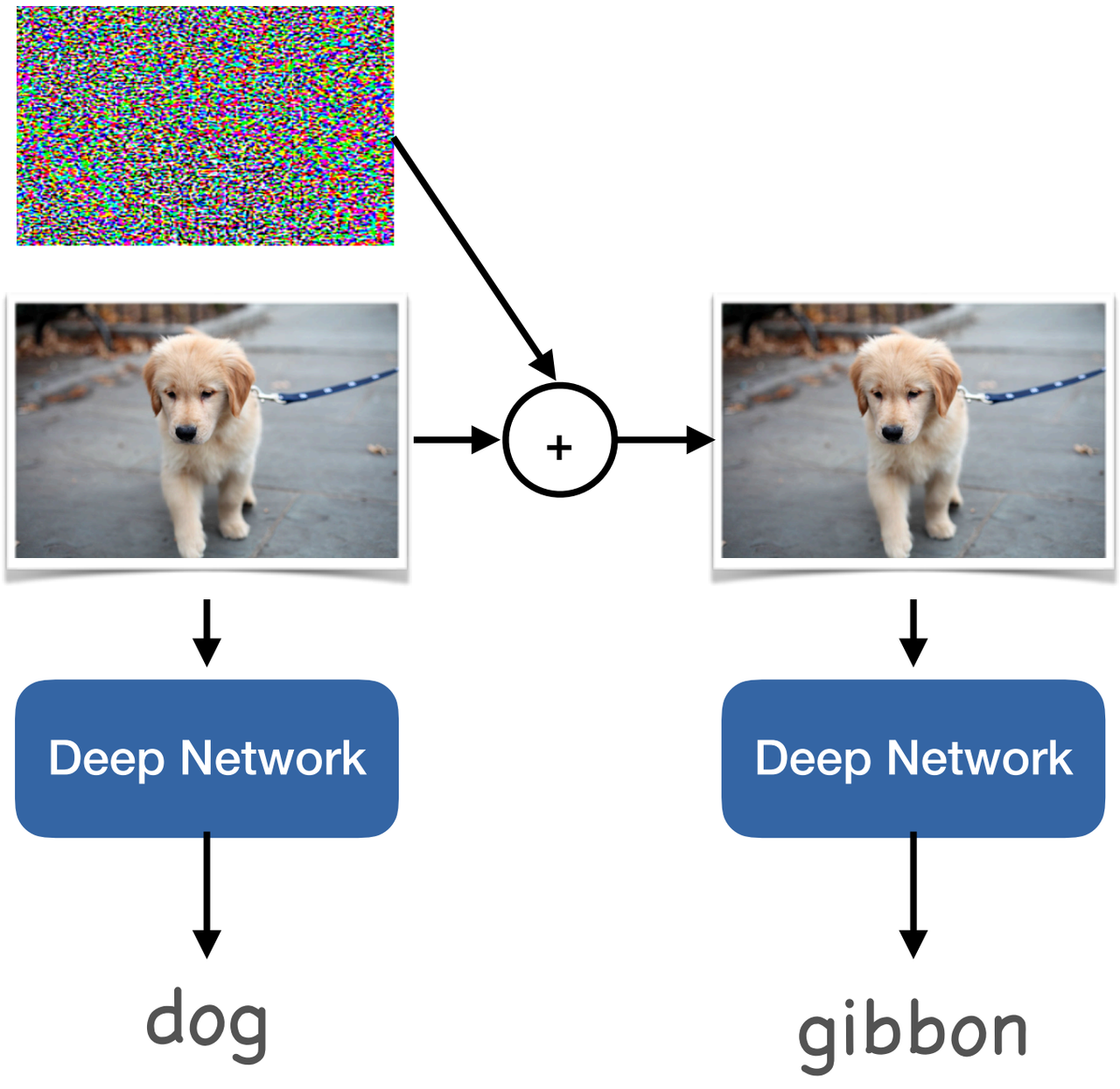
Fooling deep networks

© 2019 Philipp Krähenbühl and Chao-Yuan Wu

Adversarial perturbations

- Fooling a deep network
  - Image + noise = wrong prediction

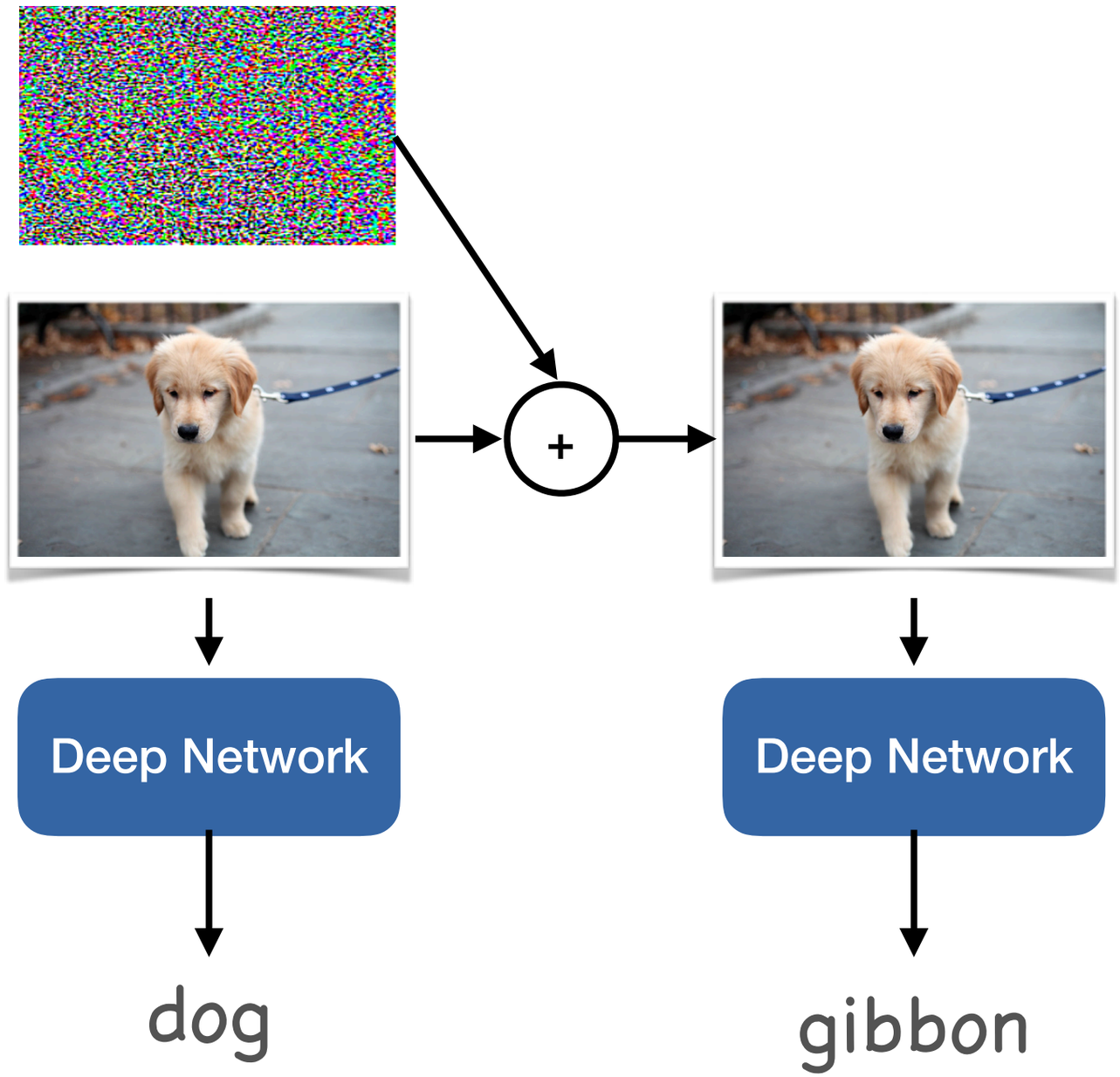


- Intriguing properties of neural networks, Szegedy et al., arXiv 2013
- Explaining and Harnessing Adversarial Examples , Goodfellow et al., ICLR 2015

Why does this work?

- Example: Linear CNNs

- Each noisy perturbation add a little bit to output



Finding adversarial examples

© 2019 Philipp Krähenbühl and Chao-Yuan Wu

Finding adversarial examples

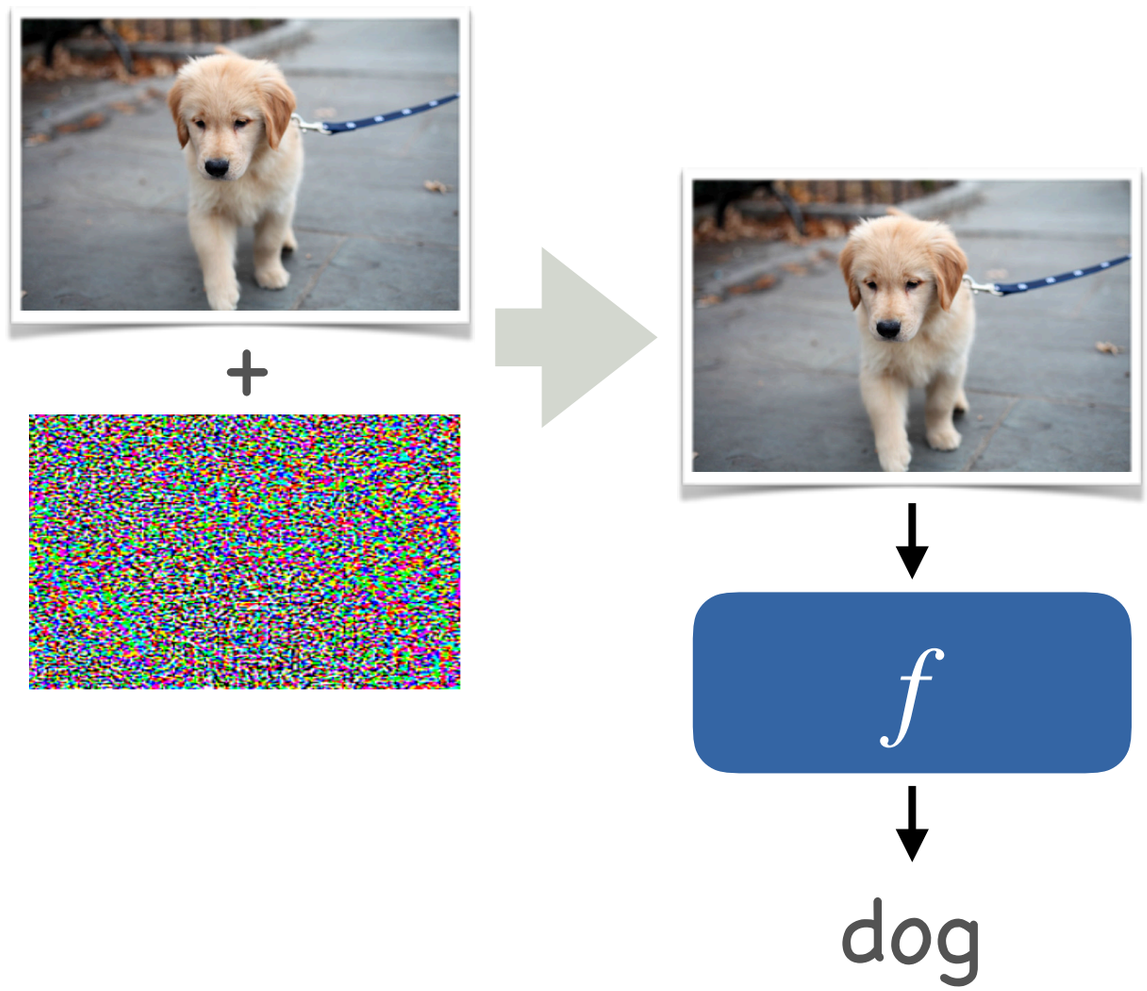
- For input $\mathbf{x}$
- Find ϵ
- Such that
  $f(\mathbf{x} + \epsilon) \neq f(\mathbf{x})$

Fast gradient sign

- Assume networks are locally linear
- Optimal attack with $\|\epsilon\|_\infty \leq c$
  - $\epsilon = \text{sign}\left(\nabla_{\mathbf{x}} \ell\left(f(\mathbf{x}), y\right)\right)$

Projected gradient descent

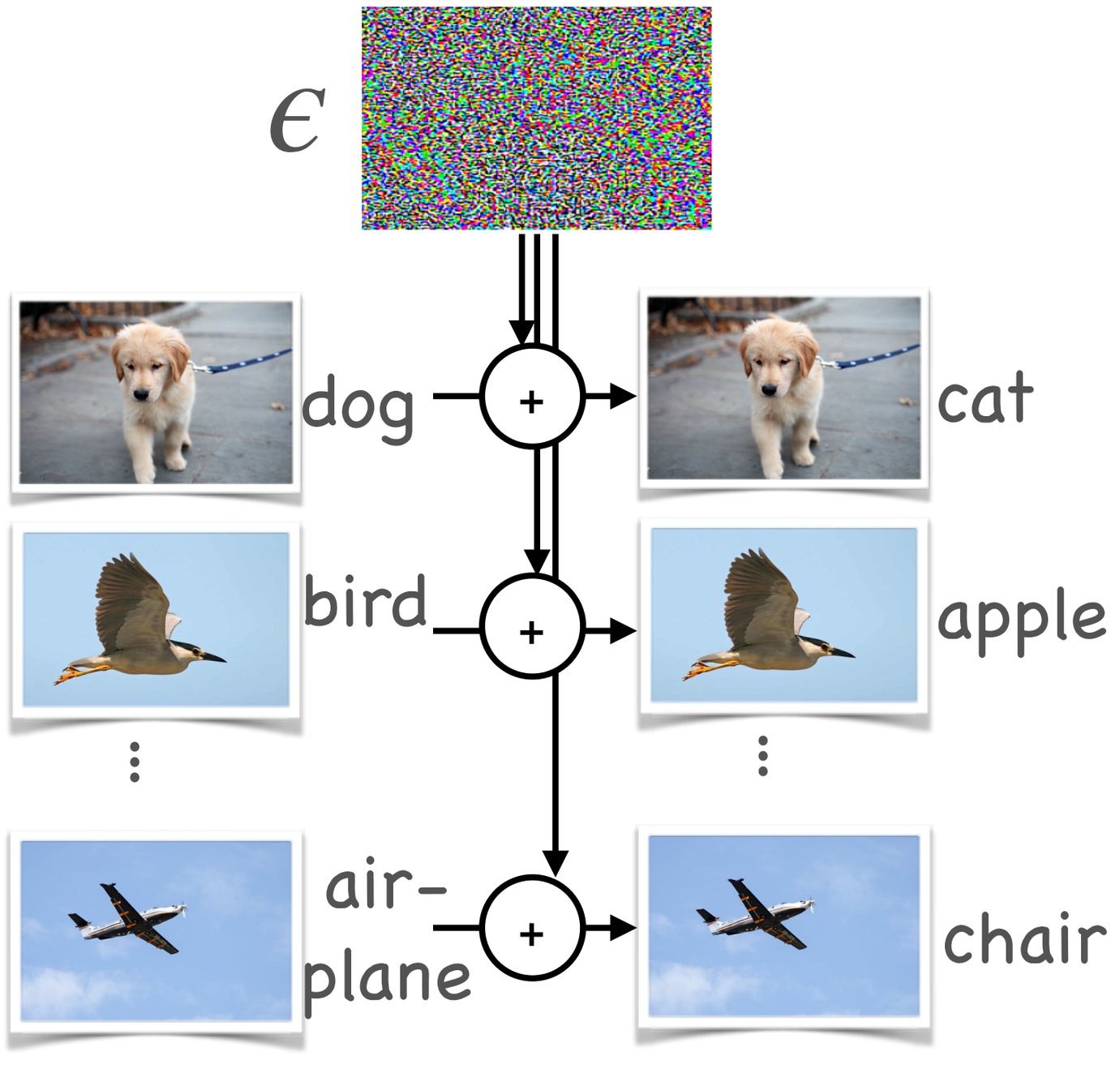
- Networks are not linear
- Optimize for the attack using gradient descent
- $\text{maximize}_{\epsilon} \ell\left(f(\mathbf{x}+\epsilon), y\right)$
- s.t. $\|\varepsilon\|_{\infty} < c$



Towards Deep Learning Models Resistant to Adversarial Attacks, Madry et al., ICLR 2018

Global adversarial attacks

- Attacks all possible inputs at once
  - PGD on entire dataset
- Attack not input specific
- Attack transfers between architectures
  - Dataset specific?



Universal adversarial perturbations, Moosavi-Dezfooli et al., CVPR 2017

Defense through data augmentation

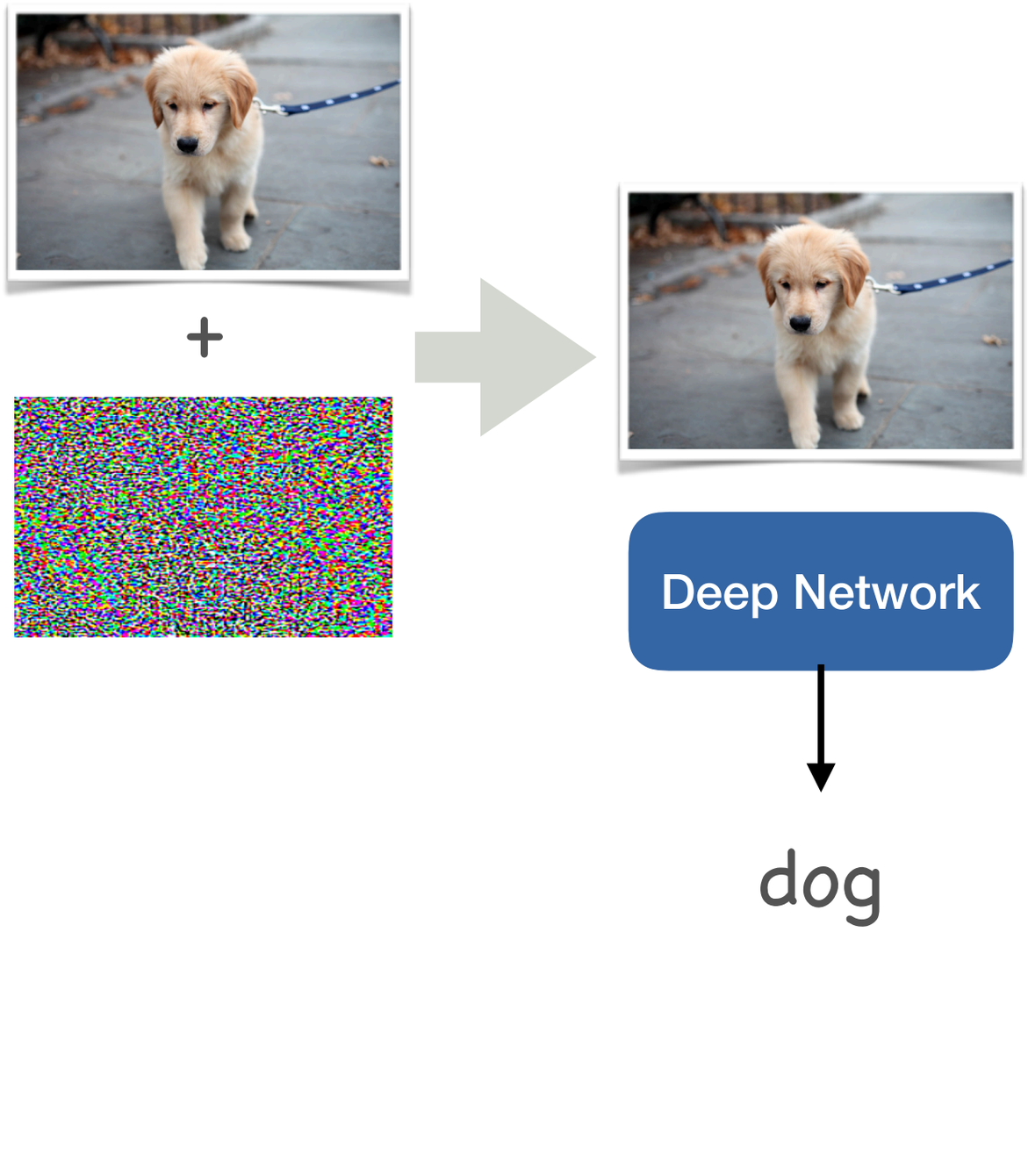
© 2019 Philipp Krähenbühl and Chao-Yuan Wu

Defense

- Show network attacked images during training
- Learn to classify correctly

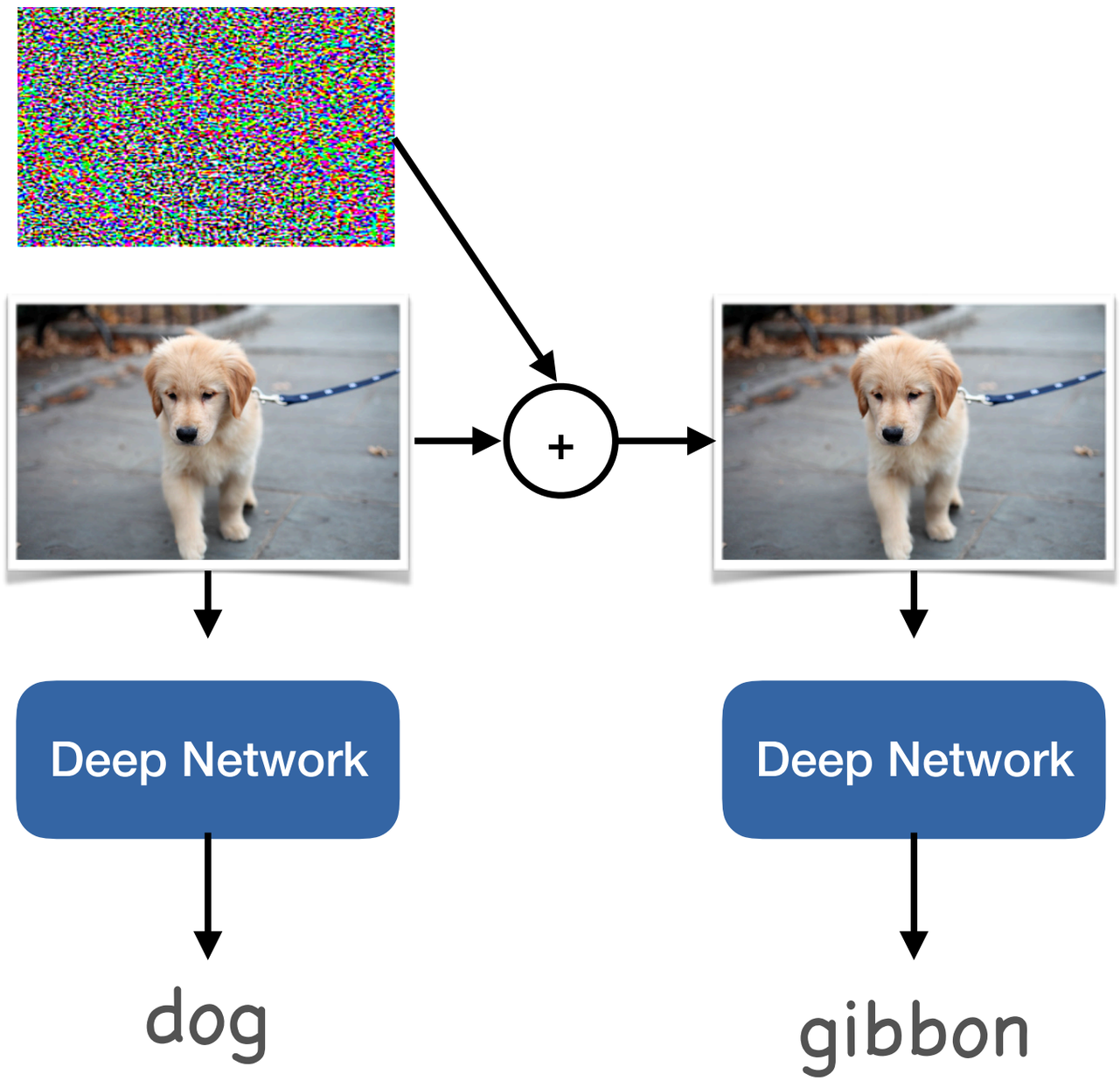
Defense

- for each iteration
  - Construct mini-batch
  - Perturb mini-batch
  - Forward / backward
    - Original
    - Perturbed



Attacking a "robust" model

- Still works
  - just harder

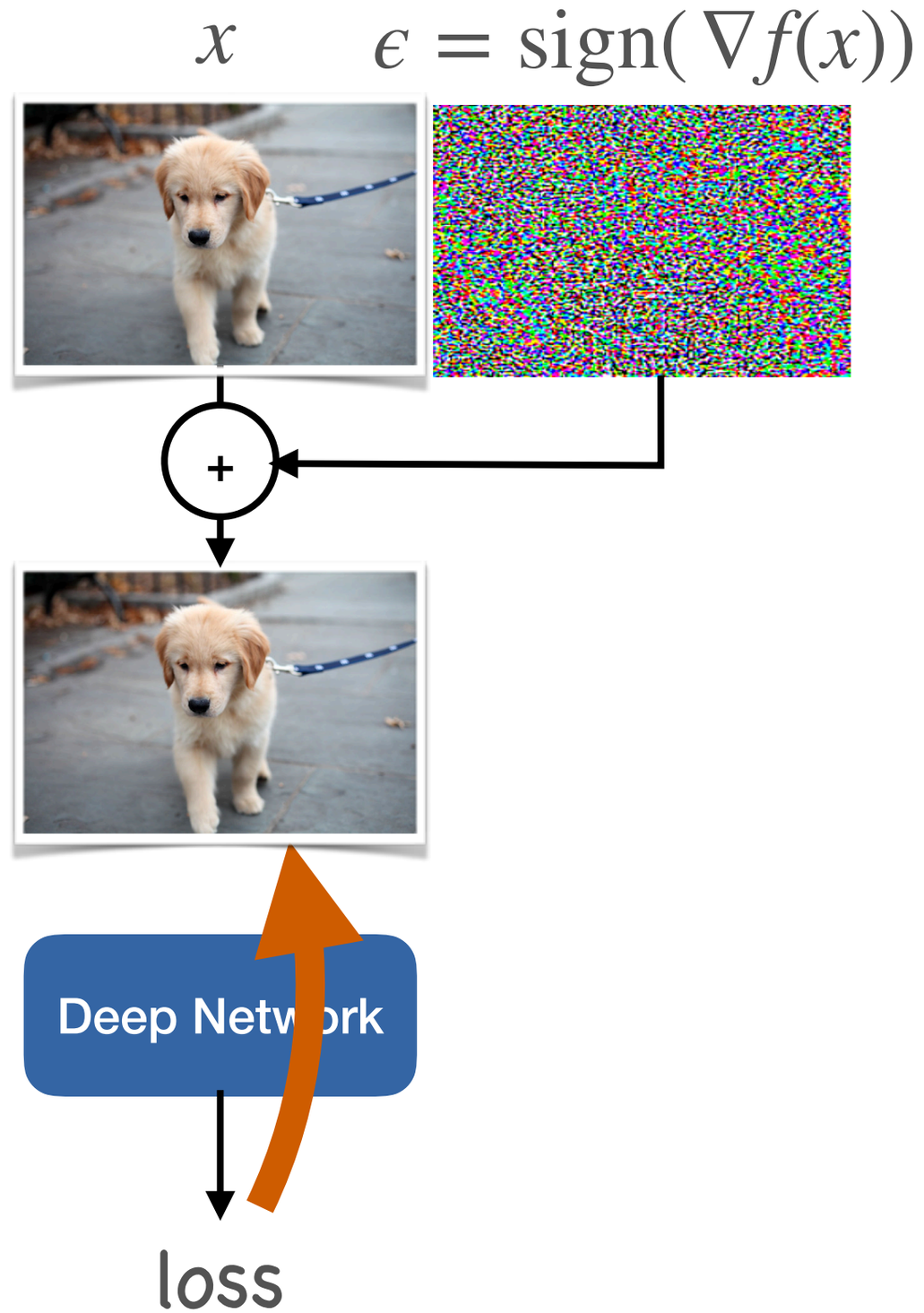


White vs black box attacks

© 2019 Philipp Krähenbühl and Chao-Yuan Wu

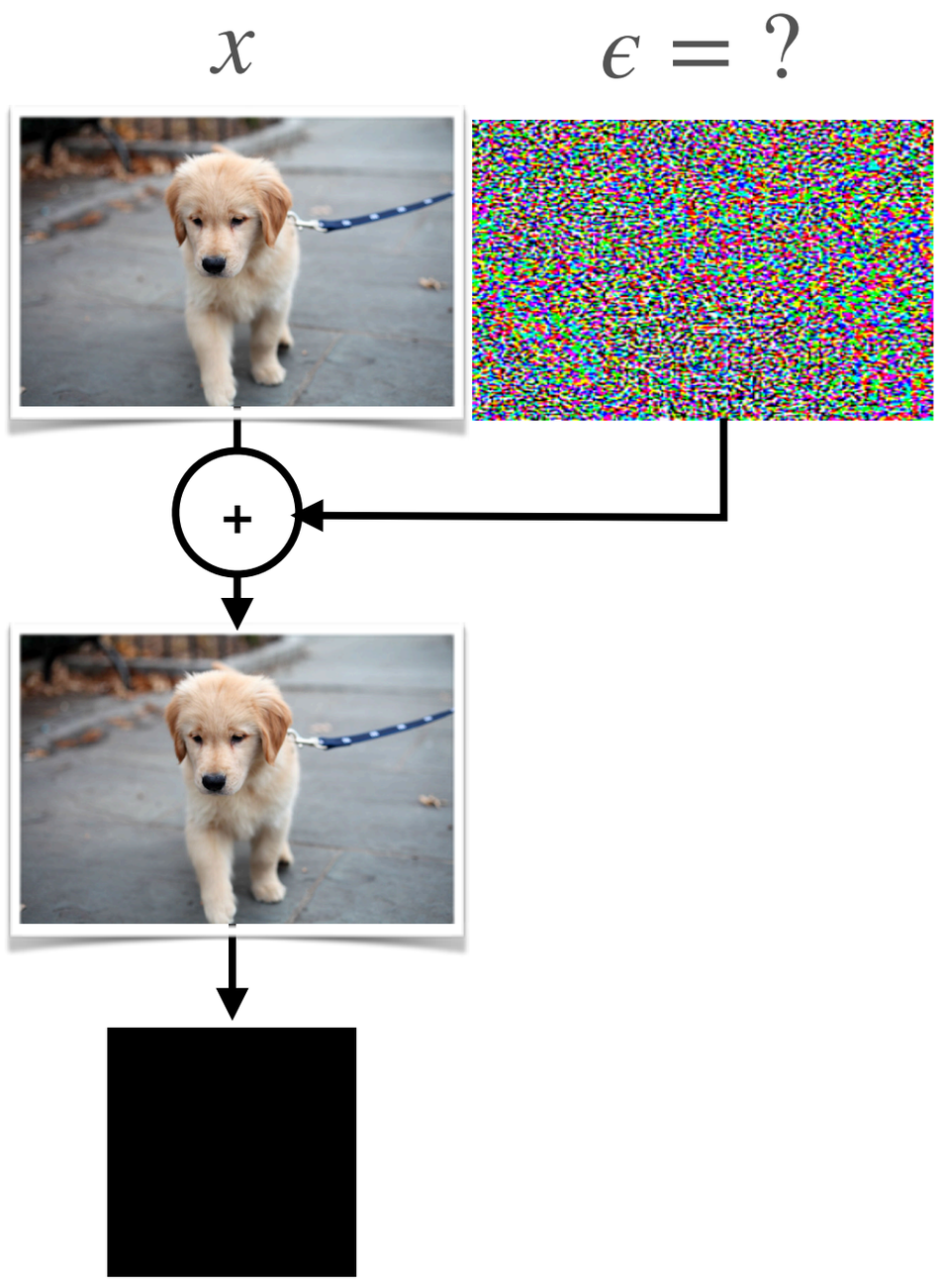
White box attacks

- Attacker has access to model and gradients
- Fast gradient sign
- Projected gradient descent



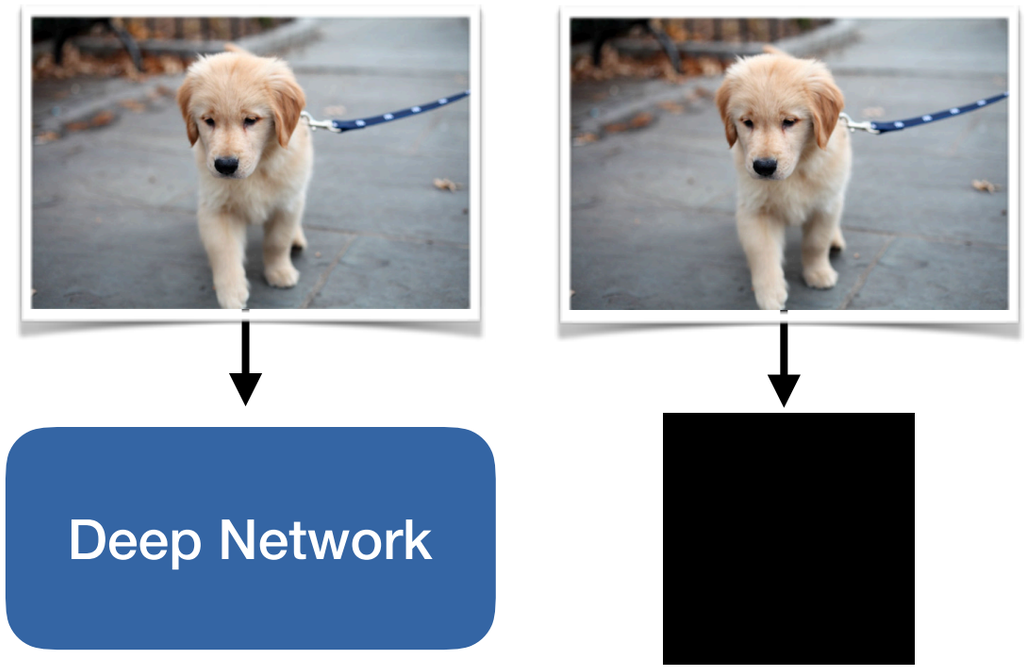
Defense by hiding model

- Can we defend against attacks if we do not allow backprop?



Back box attacks

- Train network to imitate black box network
  - Attack new network
    - Attack black box
- If not successful
  - repeat



Practical Black-Box Attacks against Machine Learning, Papernot et al., arXiv 2016

Open Problem: Realistic attacks and defenses

© 2019 Philipp Krähenbühl and Chao-Yuan Wu

What attacks should we worry about?

- Random noise attacks don't matter (yet)
  - Doing the wrong thing for real images does

How do we specify the right attack model?

- Try a validation set
  - No guarantees
  - Might overfit to validation / test set
  - Failures can be rare, but fatal

How do we specify the right attack model?

- Enumerate all possible inputs from a generative model
- too hard
- no guarantees
- too expensive

Image source: Large Scale GAN Training for High Fidelity Natural Image Synthesis, Brock et al., ICLR 2019

How do we specify the right attack model?

- Can we mathematically describe a superset of all inputs?
- and optimize?
- Very hard

Summary

© 2019 Philipp Krähenbühl and Chao-Yuan Wu

Easy to attack deep networks

- Small noise easily fools networks
  - Hard to defend against