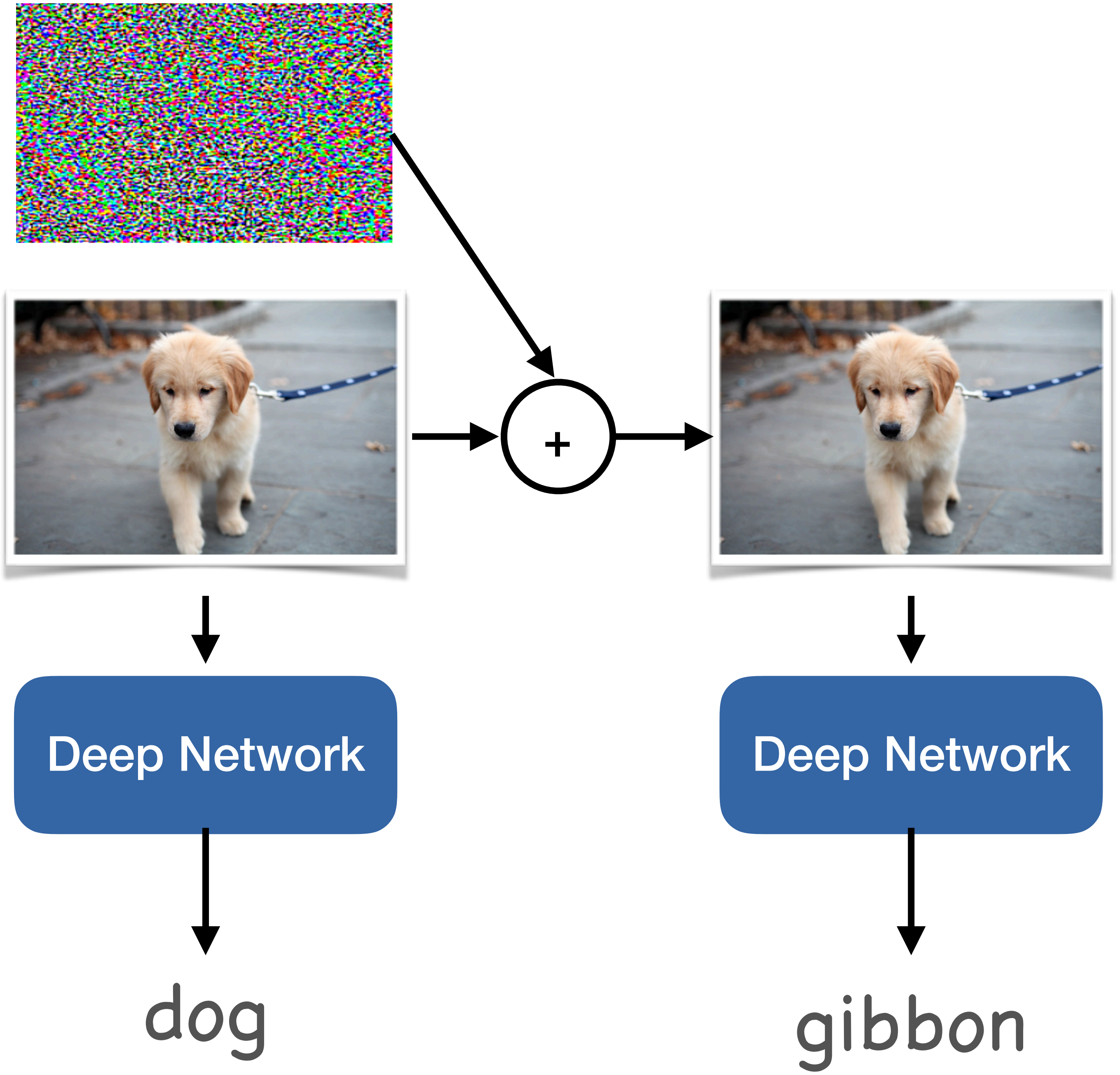
Fooling deep networks

Adversarial perturbations

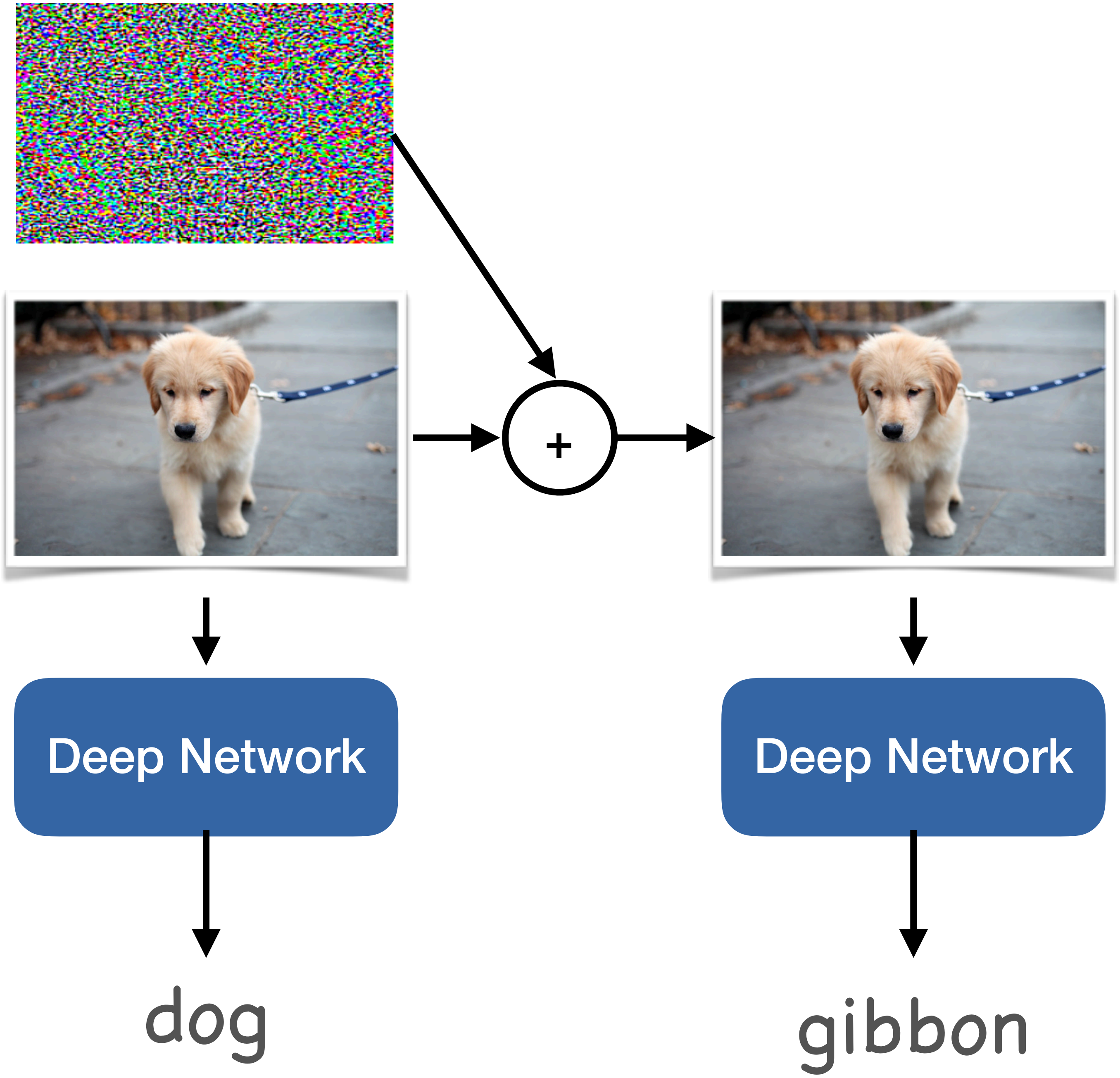
- Fooling a deep network
  - Image + noise = wrong prediction



- 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



+
Deep Network
Deep Network
dog
gibbon