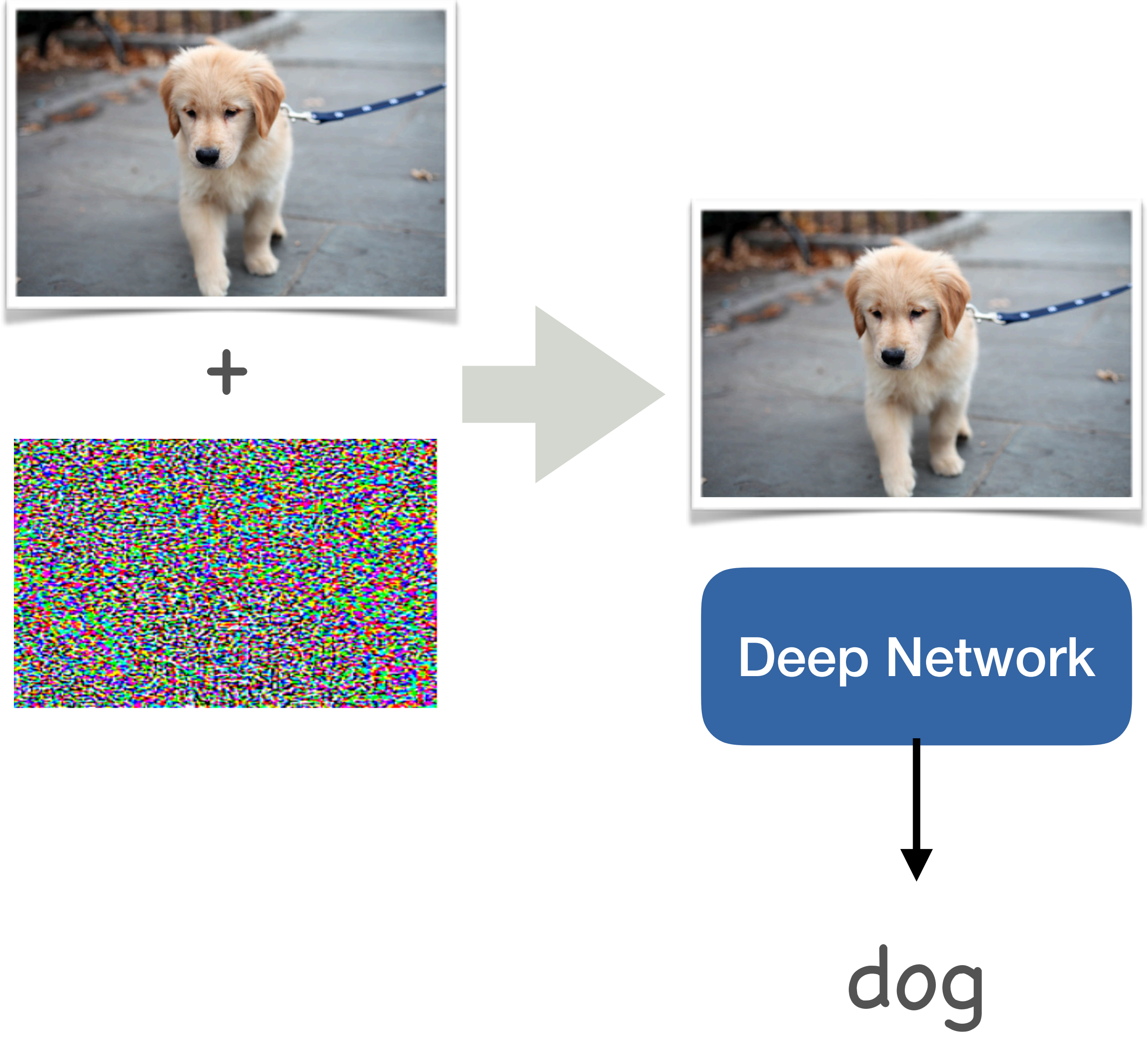
Defense through data augmentation

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



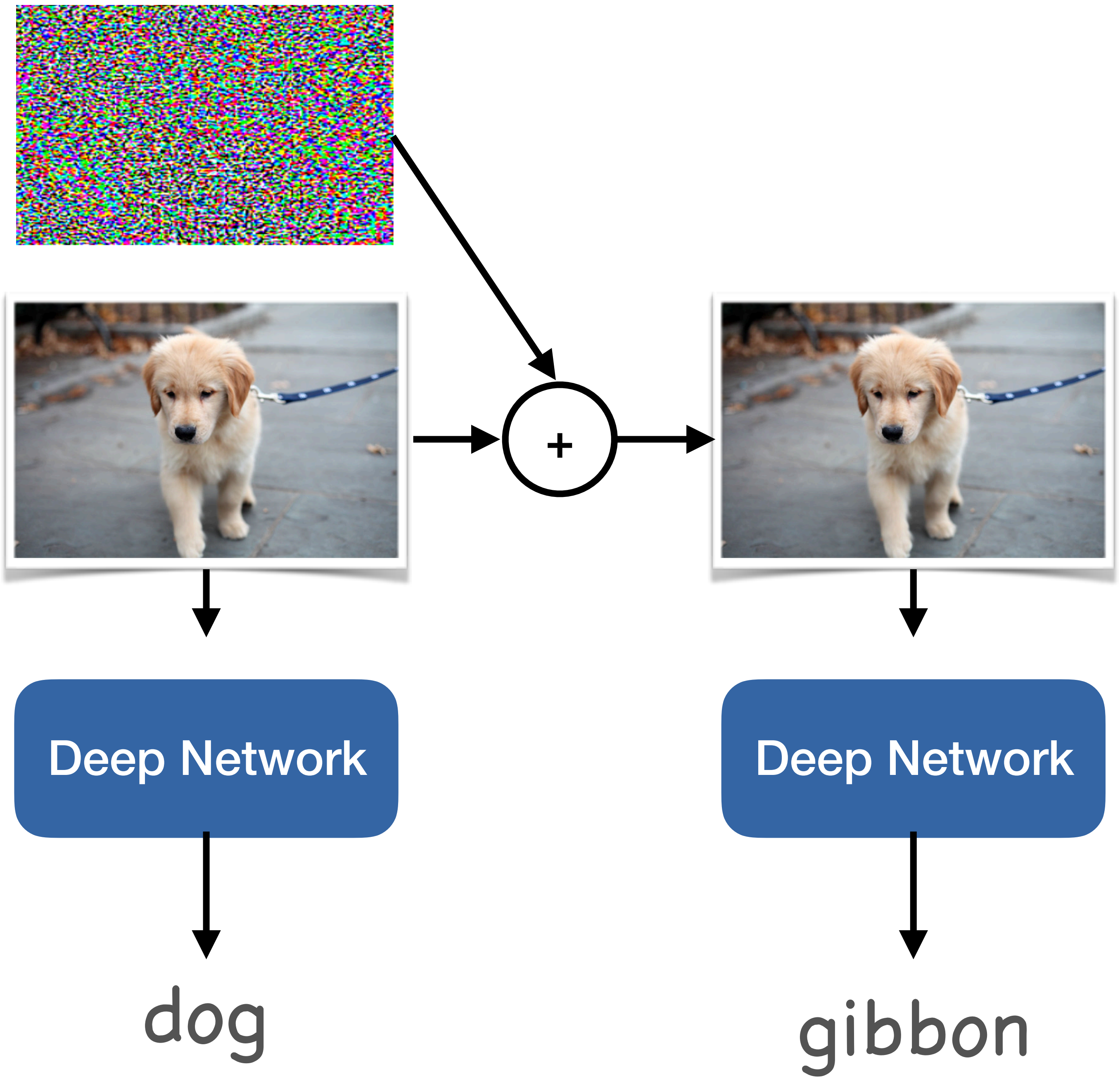
+

Deep Network

dog

Attacking a "robust" model

- Still works
  - just harder


+
Deep Network
Deep Network
dog
gibbon