Finding adversarial examples

Finding adversarial examples

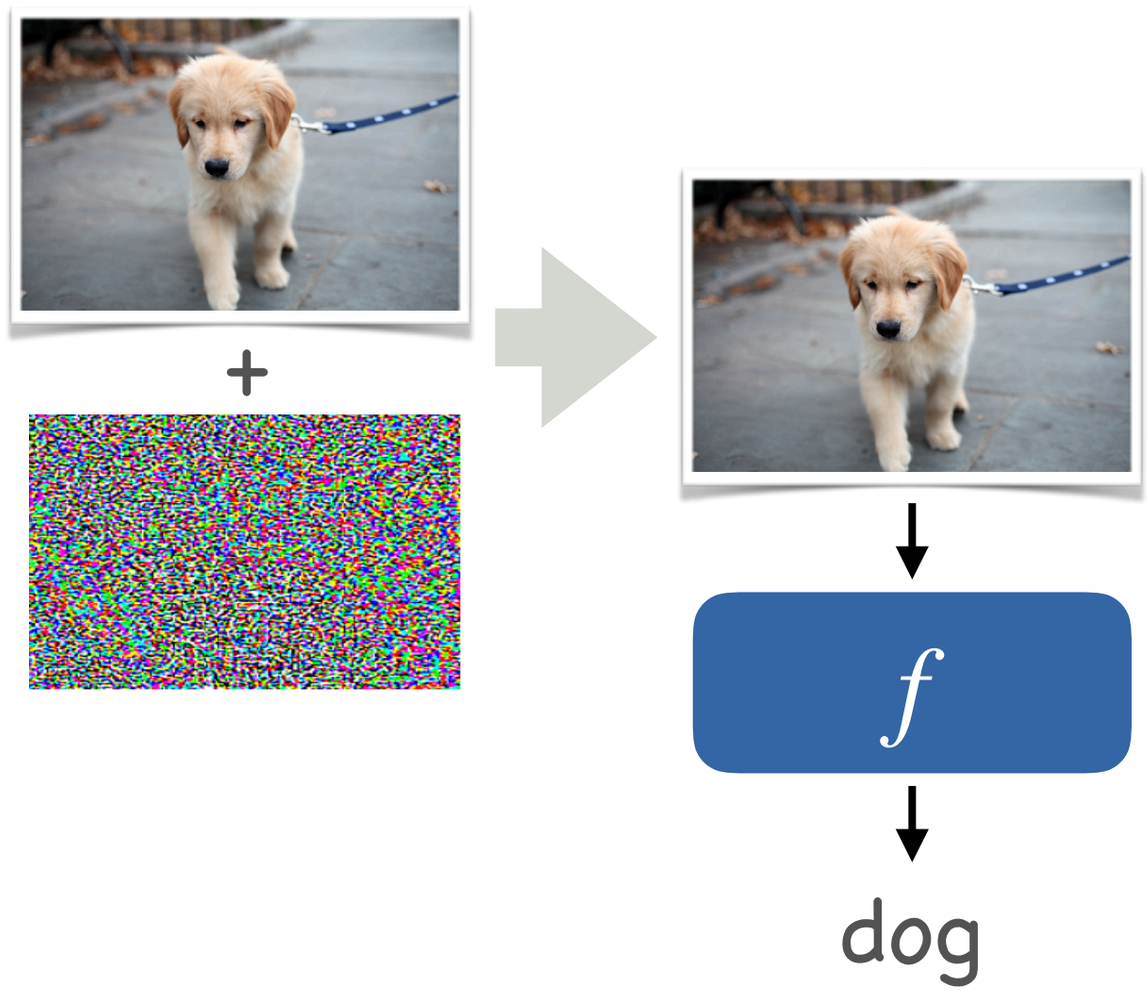
- For input $\mathbf{x}$
- Find ϵ
- Such that
  $f(\mathbf{x} + \epsilon) \neq f(\mathbf{x})$

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

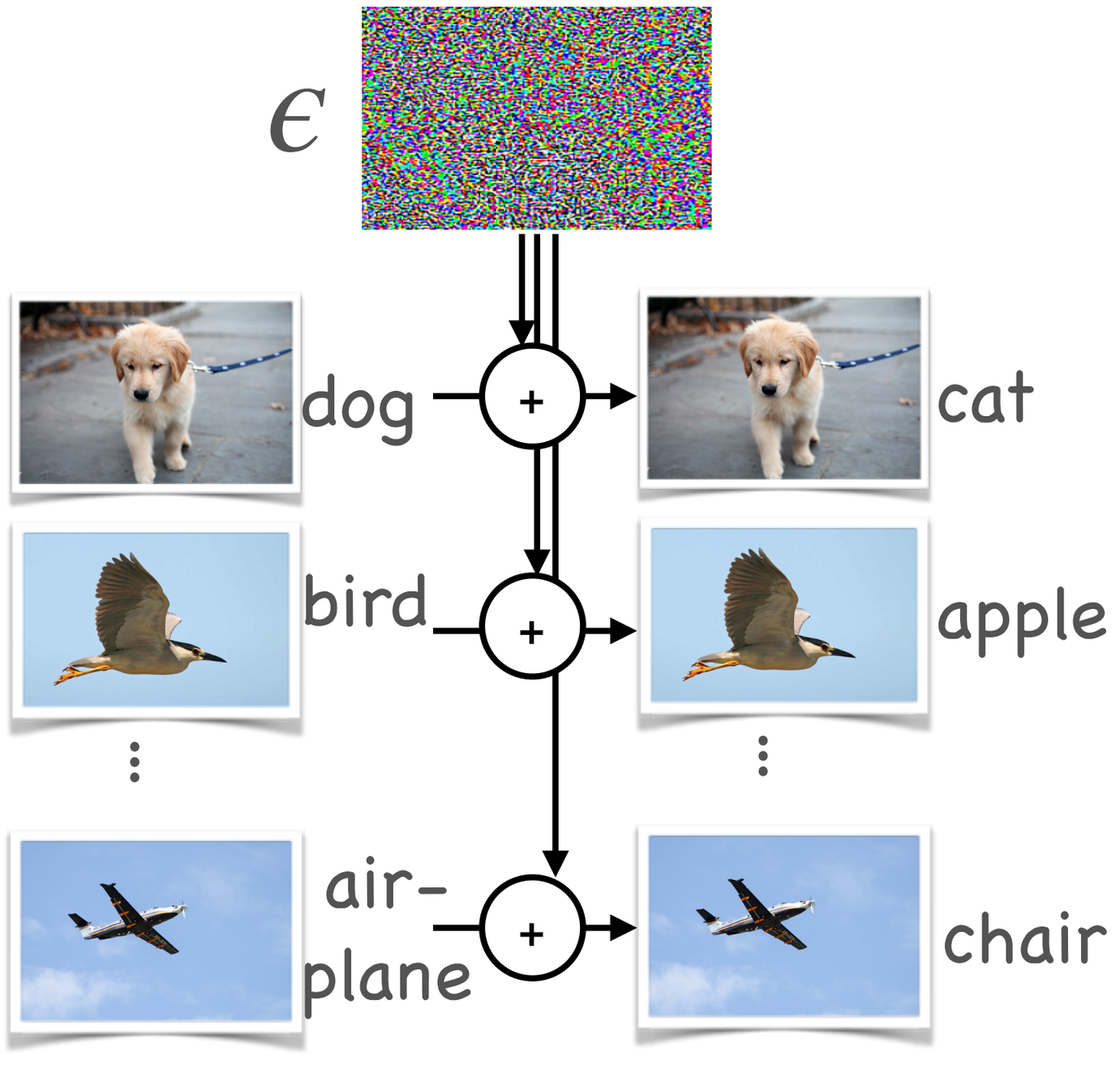
- Networks are not linear
- Optimize for the attack using gradient descent
- $\text{maximize}_{\epsilon} \ell\left(f(\mathbf{x}+\epsilon), y\right)$
- s.t. $\|\varepsilon\|_{\infty} < c$



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