Making Targeted Evasion Attacks Effective and Efficient

Mika Juuti, Buse Gul Atli, N. Asokan

Targeted Black-box Evasion Against Realistic, General APIs:

• **PRISM:** A novel black-box attack using substitute models
• An agile adversary can achieve better effectiveness/efficiency by switching through methods
• Demonstrated against real-life API: Google Cloud Vision

**Targeted evasion against realistic, general APIs**

- Targeted DNN trained with 1000s of classes
- How to change API response to target class
  
  \[
  y' \leftarrow API(x') \quad \|x - x'\|_\infty < \varepsilon
  \]

- Partial Information API: responds with top-k results
- Modifications very small, e.g. \(\varepsilon = 5\% (12.75 / 255)\)

**Approach 1: Transferability attacks: Ens**

- Adversarial examples created on ensembles, e.g. using MIFGSM [1]
- Efficient attacks: first query may succeed
- Targeted evasion ineffective with imperceptible modifications [3]

**Approach 2: Query-only methods for PI API: QO**

- Finite-difference methods for estimating gradients, e.g. NES [2]
- Start with image of target class \(y\)
- Effective: any target model attackable with almost 100% success
- Inefficient under PI API 10,000s – 100,000s of queries per sample

**Our Approach: PRISM**

- Start with image of target class \(y\)
- Gradient estimation via Ens
- PRISM\(_R\) randomized variant
- Effective: similar success rate as QO
- Efficient: three orders of magnitude faster than QO

**Agile Adversary**

- Can analyze efficiency of methods and use these pareto-optimally
- Some Ens then QO more efficient than simple QO
- Significant efficiency/effectiveness improvement with pareto-optimal order: Ens, PRISM, PRISM\(_R\), QO: EPP\(_R\)Q

**Applicable against real-life APIs**

- Demo on Google Cloud Vision
- Decreases effort from ~20,000 queries [2] to ~400-1000 queries
- Adversarial examples transferable to other APIs

**Goal image** (predicted as albatross)  
**Adversarial image** (predicted as water buffalo)  
**Start image** (predicted as water buffalo)