#5

PB19111701

1

Question:

Summarize and review the following paper: Tor: The Second-Generation Onion Router

Answer:

1. Threat model:

   An adversary can:

   - only observe some fraction of network traffic;
   - generate, modify, delete or delay traffic;
   - operate onion routers of his own;
   - compromise some fraction of the onion routers;
2. Design goals:
   - a low-latency anonymity system that is resistant to possible attacks in the threat model;
   - need to be deployed and used in the real world, to be simple, flexible and easy to use;
3. Design decisions:
   - Perfect forward secrecy
   - Separation of "protocol cleaning" from anonymity
   - No mixing, padding, or traffic shaping
   - Many TCP streams can share one circuit
   - Leaky-pipe circuit topology
   - Congestion control: decentralized congestion control, circuit/stream-level throttling
   - Directory servers
   - Variable exit policies
   - Rate limiting and fairness
   - End-to-end integrity checking
   - Rendezvous points and hidden services

2

Question:

Summarize and review the following paper on adversarial examples: Explaining and harnessing adversarial examples

Answer:

1. Problem: linear machine learning models are often vulnerable to adversarial examples, and mitigation methods at the moment fail to maintain the state-of-art accuracy on clean inputs.
2. Observations:

   a. Generic regularization strategies (e.g. dropout, pretraining, model averaging) do not confer a significant reduction in a model's vulnerability to adversarial examples, but changing to nonlinear model families can;

b. Most of us have poor intuitions for high dimensional spaces, where small effects in hundreds of dimensions adding up to create a larger effect;

c. Some models with low capacity are also able to make many different confident predictions;

d. The universal approximator theorem guarantees that a neural networks with at least one hidden layer is permitted to have enough units. Therefore, deep networks are theoretically able to represent functions that resist adversarial perturbations;

e. The direction of perturbation, rather than the specific point in space, matters most;

3. Explanation:

$$\omega^T \tilde{x} = \omega^T (x + \eta) = \omega^T x + \omega^T \eta$$

Adversarial examples can be explained as a property of high-dimensional dot products, i.e. $\omega^T \eta$ can be significant even though η is negligible when ω is high-dimensional.

4. Solution:

Step-1: Get a family of fast methods for generating adversarial examples (adversarial perturbations generalize across different clean examples);

Step-2: Adversarial training of deep networks to find the function that resists adversarial perturbation.