CSci 8271
Security and Privacy in Computing
Day 10: Stealing hyperparameters from plots

Stephen McCamant
University of Minnesota

Convolutional Neural Networks

- Commonly used for image processing, earliest examples of "deep" networks
- Convolutional layers have fewer connections and repeated weights analogous to image processing kernels
- Can get 99% or better accuracy on image recognition tasks

t-SNE

- "t-distributed stochastic neighborhood embedding"
- A way to represent some high-dimensional structure in a 2D plot
- High dimensions and low dimensions are very different, but human intuition only works in low dimensions
- Any embedding is a weak compromise

Loss curves

- "Loss" measures how well the model is matching some ground truth results
  - Measured to control training, and shows progress over time
- Over time, loss decreases but more slowly
- A gap between training loss and testing loss signals overfitting

Hyperparameter stealing

- Hyperparameters are set before training to control the network architecture or training process
- Smaller space than parameters, but still to slow to search automatically
  - Experts develop betters guesses and search strategies over time
- Previous attacks demonstrated guess hyperparameters based on queries

Defenses

- Can we frustrate attack with small changes to plots? Yes, but:
- Problem 1: manipulating plots is normally considered scientific misconduct
- Problem 2: the changes aren't very effective for an adversary who trains against them