Adversarial Examples

Neil Gong
Machine Learning as Minimizing an Objective Function

\[ y = f(x) \quad y = w^T x \]  Linear regression

Objective function

\[ \mathcal{L}(w) = L(D, w) + \lambda R(w) \]

Loss function

\[ L = \sum_i (y_i - w^T x_i)^2 \]

Training data

\{(x_1, y_1), (x_2, y_2), \ldots\}

Regularization

\[ R(w) = \|w\|_2^2 \]

Hyperparameter

Cross validation

Computationally intensive
Machine Learning Pipeline

**Training phase**
- Training data
- Algorithm to learn hyperparameters
- Hyperparameters
- Algorithm to learn model
- Model

**Deployment phase**
- Testing data
- Model
- Prediction
- Client software
- Cloud service
Security for Machine Learning

Training phase

- Algorithm to learn hyperparameters
- Hyperparameters
- Algorithm to learn model
- Model
- Poisoning attacks

Deployment phase

- Testing data
- Model
- Prediction
- Client software
- Cloud service
- Adversarial examples
- Predict as attacker desires
Confidentiality/Privacy for Users

Training phase

- Training data
- Algorithm to learn hyperparameters → Hyperparameters
- Algorithm to learn model → Model

Deployment phase

- Testing data
- Model → Prediction
- Client software
- Cloud service
Confidentiality/Intellectual Property for Model Providers

Training phase

- Training data
- Algorithm to learn hyperparameters
- Hyperparameters
- Algorithm to learn model
- Model

Deployment phase

- Testing data
- Model
- Prediction

Client software
Cloud service

6
Adversarial Examples

Training phase

- Training data
- Algorithm to learn hyperparameters
- Hyperparameters
- Algorithm to learn model
- Model

Deployment phase

- Testing data
- Adversarial examples
- Model
- Prediction
- Client software
- Cloud service
- Predict as attacker desires
Today’s lecture

• What is adversarial example

• Why do we care

• How to find adversarial example
Adversarial Examples

Normal example: digit 0

Adversarial example: predicted to be 9
Adversarial Examples

- Classifier $C$
- Normal example $x$
  - Image, text, audio, graph, software
- Perturb $x$ to $x'$
  - Preserving semantics
- $C$ misclassifies $x'$
  - Targeted: $C(x') = t$, an attacker-chosen target label
  - Untargeted: $C(x') \neq C(x)$
Attacker’s Background Knowledge

Access to prediction API
  Black box
  Weakest attacker

Learning algorithm
  Hyperparameter, e.g., neural network architecture
  Training data

Model parameters
  White box
  Strongest attacker
Why do we care?

Malware -> benign software
Spam -> non-spam
Privacy protection
Guiding design of ML

Stop sign to speed limit
How to Find Adversarial Examples - Image Domain

• Perturb x to $x'$
  • Preserving semantics
    • *Human perceives $x'$ and x as the same*
    • $d(x, x')$ is small

Minimize $d(x, x')$
Subject to
1. $C(x') = t$ or $C(x') \neq C(x)$
2. $x'$ is still an image
Solving the optimization problem

\[
\begin{align*}
\text{minimize} & \quad \|\delta\|_p + c \cdot f(x + \delta) \\
\text{such that} & \quad x + \delta \in [0, 1]^n
\end{align*}
\]
Loss function

\[ f_1(x') = -\text{loss}_{F,t}(x') + 1 \]
\[ f_2(x') = (\max_{i \neq t}(F(x')_i) - F(x')_t)^+ \]
\[ f_3(x') = \text{softplus}(\max_{i \neq t}(F(x')_i) - F(x')_t) - \log(2) \]
\[ f_4(x') = (0.5 - F(x')_t)^+ \]
\[ f_5(x') = -\log(2F(x')_t - 2) \]
\[ f_6(x') = (\max_{i \neq t}(Z(x')_i) - Z(x')_t)^+ \]
\[ f_7(x') = \text{softplus}(\max_{i \neq t}(Z(x')_i) - Z(x')_t) - \log(2) \]
Box constraints

- Projected gradient descent

- Clipped gradient descent
  - Incorporate clipping into objective function
    \[ f(\min(\max(x + \delta, 0), 1)) \]

- Change of variables
  \[ \delta_i = \frac{1}{2}(\tanh(w_i) + 1) - x_i \]
Examples

<table>
<thead>
<tr>
<th>Source Classification</th>
<th>Target Classification ($L_2$)</th>
<th>Target Classification ($L_0$)</th>
<th>Target Classification ($L_\infty$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0 0 0 0 0 0 0 0 0 0</td>
<td>0 0 0 0 0 0 0 0 0 0</td>
<td>0 0 0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>1</td>
<td>/ / / / / / / / / /</td>
<td>/ / / / / / / / / /</td>
<td>/ / / / / / / / / /</td>
</tr>
<tr>
<td>2</td>
<td>2 2 2 2 2 2 2 2 2 2</td>
<td>2 2 2 2 2 2 2 2 2 2</td>
<td>2 2 2 2 2 2 2 2 2 2</td>
</tr>
<tr>
<td>4</td>
<td>6 6 6 6 6 6 6 6 6 6</td>
<td>6 6 6 6 6 6 6 6 6 6</td>
<td>6 6 6 6 6 6 6 6 6 6</td>
</tr>
<tr>
<td>5</td>
<td>7 7 7 7 7 7 7 7 7 7</td>
<td>7 7 7 7 7 7 7 7 7 7</td>
<td>7 7 7 7 7 7 7 7 7 7</td>
</tr>
<tr>
<td>6</td>
<td>8 8 8 8 8 8 8 8 8 8</td>
<td>8 8 8 8 8 8 8 8 8 8</td>
<td>8 8 8 8 8 8 8 8 8 8</td>
</tr>
<tr>
<td>7</td>
<td>a a a a a a a a a a</td>
<td>a a a a a a a a a a</td>
<td>a a a a a a a a a a</td>
</tr>
<tr>
<td>8</td>
<td>b b b b b b b b b b</td>
<td>b b b b b b b b b b</td>
<td>b b b b b b b b b b</td>
</tr>
<tr>
<td>9</td>
<td>c c c c c c c c c c</td>
<td>c c c c c c c c c c</td>
<td>c c c c c c c c c c</td>
</tr>
</tbody>
</table>
Evaluation metrics – what is a successful adversarial example

• Misclassification
  • Targeted: $C(x') = t$, an attacker-chosen target label
  • Untargeted: $C(x') \neq C(x)$

• Human perceives $x'$ and $x$ as the same
  • Hard to implement – involves user studies
  • Approximate using $L_p$ norm of noise
Other methods

• Beyond $L_p$ norm

• Physically realizable adversarial examples
Beyond images

- Text
- Audio
- Video
- Software

Preserving semantics

$C$ misclassifies $x'$

Formulation as optimization problem