Defense Against the Dark Arts: Machine Learning Security and Privacy

Ian Goodfellow, Staff Research Scientist, Google Brain
BayLearn 2017
An overview of a field

• This presentation summarizes the work of many people, not just my own / my collaborators

• Please check out the slides and view this link of extensive references

• The presentation focuses on the concepts, not the history or the inventors
Machine learning pipeline

Training data → Learning algorithm → Learned parameters → Test input → Test output

\[ X \rightarrow \theta \rightarrow \hat{y} \]

(Goodfellow 2017)
Privacy of training data
Defining $(\varepsilon, \delta)$-Differential Privacy

\[ \forall S: \Pr[M(x) \in S] \leq e^\varepsilon \cdot \Pr[M(y) \in S] + \delta \]

(Goodfellow 2017)
Private Aggregation of Teacher Ensembles

(Papernot et al 2016)
Training Set Poisoning

\[
\begin{align*}
X & \xrightarrow{\theta} \hat{y} \\
\mathbf{x} & \xrightarrow{\mathbf{y}} 
\end{align*}
\]
ImageNet poisoning

(Koh and Liang 2017)
Adversarial examples

\[
X \rightarrow \theta \rightarrow \hat{y} \\
\hat{y} \\n\]

\[
x \rightarrow \theta \rightarrow \hat{y} \\
\hat{y} \\n\]

(Goodfellow 2017)
Model theft

\[ X \rightarrow \theta \rightarrow \hat{y} \rightarrow x \rightarrow \theta \]
Model theft++

\[ X \rightarrow \theta \rightarrow \hat{y} \rightarrow \hat{x} \]

(Goodfellow 2017)
Advanced models can infer private information

(Youyou et al 2014)
Automated Crowdturfing

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Generated Review Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>I love this place! I have been here a few times and have never been disappointed. The service is always great and the food is always great. The staff is always friendly and the food is always great. I will definitely be back and try some of their other food and service.</td>
</tr>
<tr>
<td>0.5</td>
<td>I love this place. I have been going here for years and it is a great place to hang out with friends and family. I love the food and service. I have never had a bad experience when I am there.</td>
</tr>
<tr>
<td>0.7</td>
<td>My family and I are huge fans of this place. The staff is super nice and the food is great. The chicken is very good and the garlic sauce is perfect. Ice cream topped with fruit is delicious too. Highly recommended!</td>
</tr>
<tr>
<td>1.0</td>
<td>I had the grilled veggie burger with fries!!!! Ohhhh and taste. Omgggg! Very flavorful! It was so delicious that I didn’t spell it!!</td>
</tr>
</tbody>
</table>

(Yao et al 2017)
Fake News

www.futureoffakenews.com
Machine learning for password guessing

Figure 3: **Neural network size and password guessability.** Dotted lines are large networks; solid lines are small networks.

*(Melicher et al 2016)*
“Artificial intelligence is the future, not only for Russia, but for all humankind,” said Putin, reports RT. “It comes with colossal opportunities, but also threats that are difficult to predict. Whoever becomes the leader in this sphere will become the ruler of the world.”
Deep Dive on Adversarial Examples
Since 2013, deep neural networks have matched human performance at:

- recognizing objects and faces...
  
  (Szegedy et al, 2014)

- solving CAPTCHAS and reading addresses...
  
  (Goodfellow et al, 2013)

and other tasks...

(Taigmen et al, 2013)
Adversarial Examples

Timeline:
“Adversarial Classification” Dalvi et al 2004: fool spam filter
“Evasion Attacks Against Machine Learning at Test Time” Biggio 2013: fool neural nets
Szegedy et al 2013: fool ImageNet classifiers imperceptibly
Goodfellow et al 2014: cheap, closed form attack

(Goodfellow 2017)
Turning Objects into “Airplanes”
Attacking a Linear Model
Adversarial Examples from Overfitting
Adversarial Examples from Excessive Linearity
Modern deep nets are very piecewise linear

Rectified linear unit

Maxout

Carefully tuned sigmoid

LSTM

(Goodfellow 2017)
Nearly Linear Responses in Practice
Small inter-class distances

All three perturbations have L2 norm 3.96
This is actually small. We typically use 7!
The Fast Gradient Sign Method

\[ J(\tilde{x}, \theta) \approx J(x, \theta) + (\tilde{x} - x)^\top \nabla_x J(x). \]

Maximize

\[ J(x, \theta) + (\tilde{x} - x)^\top \nabla_x J(x) \]

subject to

\[ \|\tilde{x} - x\|_\infty \leq \epsilon \]

\[ \Rightarrow \tilde{x} = x + \epsilon \text{sign} (\nabla_x J(x)) . \]

(Goodfellow 2017)
Maps of Adversarial and Random Cross-Sections

(collaboration with David Warde-Farley and Nicolas Papernot)
Estimating the Subspace Dimensionality

(Tramèr et al, 2017)
Wrong almost everywhere
Adversarial Examples for RL

(Huang et al., 2017)
RBFs behave more intuitively
Cross-model, cross-dataset generalization
## Cross-technique transferability

(Papernot 2016)
Transferability Attack

Target model with unknown weights, machine learning algorithm, training set; maybe non-differentiable

Train your own model

Substitute model mimicking target model with known, differentiable function

Adversarial examples

Adversarial crafting against substitute

Deploy adversarial examples against the target; transferability property results in them succeeding

(Goodfellow 2017)
Enhancing Transfer With Ensembles

<table>
<thead>
<tr>
<th></th>
<th>RMSD</th>
<th>ResNet-152</th>
<th>ResNet-101</th>
<th>ResNet-50</th>
<th>VGG-16</th>
<th>GoogLeNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>-ResNet-152</td>
<td>17.17</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>-ResNet-101</td>
<td>17.25</td>
<td>0%</td>
<td>1%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>-ResNet-50</td>
<td>17.25</td>
<td>0%</td>
<td>0%</td>
<td>2%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>-VGG-16</td>
<td>17.80</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>6%</td>
<td>0%</td>
</tr>
<tr>
<td>-GoogLeNet</td>
<td>17.41</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>5%</td>
</tr>
</tbody>
</table>

Table 4: Accuracy of non-targeted adversarial images generated using the optimization-based approach. The first column indicates the average RMSD of the generated adversarial images. Cell $(i, j)$ corresponds to the accuracy of the attack generated using four models except model $i$ (row) when evaluated over model $j$ (column). In each row, the minus sign “−” indicates that the model of the row is not used when generating the attacks. Results of top-5 accuracy can be found in the appendix (Table 14).

(Liu et al, 2016)
Adversarial Examples in the Human Brain

These are concentric circles, not intertwined spirals.

(Pinna and Gregory, 2002)
Adversarial Examples in the Physical World

(Kurakin et al., 2016)
Training on Adversarial Examples
Success on MNIST?

- Open challenge to break model trained on adversarial perturbations initialized with noise
- Even strong, iterative white-box attacks can’t get more than 12% error so far
- Larger datasets remain challenging

(Madry et al 2017)
Verification

• Given a seemingly robust model, can we prove that no adversarial examples exist near a given point?

• Yes, but hard to scale to large models (Huang et al 2016, Katz et al 2017)

• What about adversarial near test points that we don’t know to examine ahead of time?
Competition

AI Fight Club Could Help Save Us from a Future of Super-Smart Cyberattacks


Used as at least part of all top 10 entries in dev round 3
Clever Hans

(“Clever Hans, Clever Algorithms,” Bob Sturm)
Get involved!
https://github.com/tensorflow/cleverhans

Check out Justin Gilmer’s BayLearn poster on Adversarial Sphere

(Goodfellow 2017)