Adversarial ML in Real Life
Examples, Lessons, and Challenges

David Freeman
Research Scientist/Engineer, Facebook
AISecAI Workshop
London, United Kingdom, 20 January 2020
Adversarial ML in Academia

Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy,
"Explaining and Harvesting Adversarial Examples," ICLR 2015
Adversarial ML in Academia

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, Dawn Song,

“Robust Physical-World Attacks on Deep Learning Visual Classification,”
CVPR 2018
Adversarial ML on Facebook

Pandas turned into Gibbons
Adversarial ML on Facebook

Not Porn

Not Spam

Not Phishing

Facebook

WARNING!! Your Account Has Violated Terms on Facebook.

Warning: Your account will be disabled!!

Your Facebook account is Troubled. Your account has violated the provisions on Facebook. Security Systems has received reports from other users you violate the rules on Facebook which resulted in your account will be permanently disabled.

- Post a rough profile or photos;
- Insulting and threatening others (users)
- Using facebook account just for promotion

Please confirm your account by clicking the link below:

Confirm My Account

Attention:

All accounts that are not verified within 24 hours
Adversarial ML is Everywhere

To us, “Adversarial ML” == “ML in an Adversarial Environment”
Fundamental Observation
Fundamental Observation

- ML algorithms assume a stationary data distribution.
Fundamental Observation

- ML algorithms assume a stationary data distribution.
- In an adversarial environment, deploying a model changes the data distribution.
Fundamental Observation

- ML algorithms assume a stationary data distribution.

- In an adversarial environment, deploying a model changes the data distribution.

All ML is broken...or is it?
Example I: A/B Testing?
The Perils of A/B Testing
The Perils of A/B Testing

- Fundamental A/B testing assumption: *Experiment effects are independent of the cohorts chosen*
The Perils of A/B Testing

• Fundamental A/B testing assumption:  
  Experiment effects are independent of the cohorts chosen
The Perils of A/B Testing
Start with a small experiment

• Looks good so far....
The Perils of A/B Testing
Roll it out to (almost) everyone — Option 1
The Perils of A/B Testing

Roll it out to (almost) everyone — Option 1

• Did the adversary give up or iterate?
The Perils of A/B Testing
Roll it out to (almost) everyone — Option 2
The Perils of A/B Testing

Roll it out to (almost) everyone — Option 2

- Now your experiment is a vulnerability
Using Shadow Mode

- Run new model online in “log-only” mode
- Evaluate performance where the new model disagrees with the old one.
- Push based on FP/FN tradeoff
Example II: Never Forget
Refreshing your data
Don’t forget the past!
Refreshing your data
Don’t forget the past!

Model v1

Training set 1
Refreshing your data
Don’t forget the past!

Training set 1

Model v1
Refreshing your data
Don’t forget the past!
Refreshing your data

Don’t forget the past!

Training set 2

Model v2
Refreshing your data

Don’t forget the past!
Refreshing your data

Don’t forget the past!

Model vN

Training set N
Refreshing your data
Don’t forget the past!

Model vN
Training set N
Refreshing your data

Don’t forget the past!

Mitigation:

• Keep old attacks around (exponential decay?)
• Keep old models around (raise thresholds?)
Example III:
It’s a Race!
Facebook has a few fake accounts

How many fake accounts did we take action on?

- Oct – Dec 2017: 1B
- Jan – Mar 2018: 1B
- Apr – Jun 2018: 1B
- Jul – Sep 2018: 1B
- Oct – Dec 2018: 1B
- Jan – Mar 2019: 2B
- Apr – Jun 2019: 1B
- Jul – Sep 2019: 1B
Fake Accounts at Registration

Registration-time fake account classification has two fundamental problems:

- Number of features is limited.
- Exposes a clear experimentation environment for attackers.

These two problems cause ML models to deteriorate fast.

- Each model iteration requires significant manual work
- New model doesn’t learn what models in previous iterations learned
Solution 1: Learn Faster

Online Learning:
Train a single model from streaming labels
Solution 2: Act Slower
Don’t give immediate feedback

• Introduce delay in blocking response (and/or)

• Undo the damage without telling the user.
Example IV:
Don’t be Fooled
What not to Do (I)

Block on client-controlled signals

user-agent: Scrapy/1.8.0 (+https://scrapy.org)

user-agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10_14_6)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/79.0.3945.88 Safari/537.36
What not to Do (I)
Block on client-controlled signals

user-agent: Scrapy/1.8.0 (+https://scrapy.org)
user-agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10_14_6)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/79.0.3945.88 Safari/537.36
What not to Do (I)
Block on client-controlled signals

user-agent: Scrapy/1.8.0 (+https://scrapy.org)
user-agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10_14_6)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/79.0.3945.88
Safari/537.36

✅

🛑
What not to Do (II)
Look for specific content to block

“We don’t want to be the ones solving the CAPTCHAs”
What not to do (III)

Use the same signals for measurement and enforcement

Count users posting more than 100x/hour

# bad users

time

launch rate limit of 100 posts/hour
What to Do (I)

Use data the adversary doesn’t know/control
What to Do (II)
Focus on bad behavior, not only bad content
What to Do (II)
Focus on bad behavior, not only bad content
What to Do (II)
Focus on bad behavior, not only bad content
What to do (III)
Separate measurement and enforcement

At Facebook we split signals into two classes:

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Enforcement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network connection</td>
<td>Counts and rates</td>
</tr>
<tr>
<td>HTTP request</td>
<td>Graph relations</td>
</tr>
<tr>
<td>User-generated content</td>
<td>Activity sequence</td>
</tr>
</tbody>
</table>
Take-aways
Set up Defense in Depth

- Scoring at Entry Points
  - prevent access to accounts

- Clustering, Anomaly Detection
  - prevent accounts from doing damage

- Behavioral Analysis
  - detect bad activity

- User Reporting
  - find false negatives

Increasing speed

More information available
Open questions

• Can we combine online learning and active learning?
• How can we conduct rigorous A/B tests?
• What’s the best way to avoid model forgetting?
• How do we prevent feedback between measurement (fragile signals) and detection (robust signals)?
Thank you!

facebook
dfreeman@fb.com

Want to help us?
Talk to me at the break!