Adversarial Machine Learning

Nicholas Carlini
Google
This class (so far)
Training
Poisoning: Modify training data to cause test errors

Evasion: Modify test inputs to cause test errors

Model Stealing: Study model output to reveal parameters

Training Data Extraction: Study model parameters to reveal training data
Act I: Evasion
Evasion:
Modify test inputs
to cause test errors
88% tabby cat
88% tabby cat

 adversarial perturbation
88% tabby cat

adversarial perturbation
88% tabby cat

adversarial perturbation

99% guacamole
Okay, lesson learned.
Okay, lesson learned.

Don't classify cats with neural networks.
guacamole

adversarial perturbation

tabby cat
Okay, lesson learned.

Don't classify cats with neural networks.
Okay, lesson learned.

Don't classify cats with neural networks.

or guacamole
Okay, lesson learned.

Don't classify cats with neural networks.
What will a state-of-the-art neural network transcribe?
"It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity"
Okay, lesson learned.

Don't classify images with neural networks.
Generating Natural Language Adversarial Examples

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani B. Srivastava, Kai-Wei Chang

Department of Computer Science, University of California, Los Angeles (UCLA)
{malzantot, bojhang, mbs, kwchang}@ucla.edu
2 Cooper Union
sharma2@cooper.edu
3 Computer Science Department, University of Maryland elgohary@cs.umd.edu

Adversarial Attacks on Neural Network Policies

Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, Pieter Abeel
1 University of California, Berkeley, Department of Electrical Engineering and Computer Sciences
2 Pennsylvania State University, School of Electrical Engineering and Computer Science
3 OpenAI

Abstract

Machine learning classifiers are known to be vulnerable to inputs maliciously constructed by adversaries to force misclassification. Such adversarial examples have been extensively studied in the context of computer vision applications. In this work, we study adversarial attacks on three architectures of neural networks in reinforcement learning. Specifically, we show that adversarial example crafting techniques can be used to significantly degrade test-time performance.

Seq2Sick: Evaluating the Robustness of Sequence-to-Sequence Models with Adversarial Examples

Minhao Cheng, Jinfeng Yi, Huan Zhang, Pin-Yu Chen, Cho-Jui Hsieh
1 Department of Computer Science, University of California, Davis, CA 95616
2 Tencent AI Lab, Bellevue, WA 98004
3 IBM Research AI, Yorktown Heights, NY 10598
mcheng@ucdavis.edu, jinfengyi.utf@gmail.com, weizhang@ucdavis.edu, pin-yu.chen@ibm.com, chohsieh@ucdavis.edu

Hallucinations in Neural Machine Translation

Anonymous authors
Paper under double-blind review

Abstract

Neural machine translation (NMT) systems have reached state-of-the-art performance in translating text and are in wide deployment. Yet little is understood about how these systems function or break. Here we show that NMT systems are susceptible to highly pathological translations that are completely unethered from the source material, which we term hallucinations. Such pathological translations are problematic because they are deeply disturbing of user trust and easy to find with a simple search. We describe a method to generate hallucinations and show that many common variations of the NMT architecture of hallucinations is a new attack.

Synthetic and Natural Noise Both Break Neural Machine Translation

Yonatan Belinkov
Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology belinkov@mit.edu

Yonatan Bisk
Paul G. Allen School of Computer Science & Engineering, University of Washington ybisk@cs.washington.edu

On the Robustness of Semantic Segmentation Models to Adversarial Attacks

Anurag Arnab, Ondrej Miksik, Philip H.S. Torr
University of Oxford
{anurag.arnab, ondrej.miksik, philip.torr}@eng.ox.ac.uk
How do these attacks work?
That's what we can do with access to the weights

What can we do without the weights?
How do we defend against attacks?
Normal Training

(7, 7)

Training

(3, 3)
Adversarial Training (1)

(8, 3)
(7, 3)
(7, 7)
(7, 7)
(8, 3)

Attack
Adversarial Training (2)
... and that's almost it.
Normal Loss Surface
Evasion:
Modify test inputs
to cause test errors

Training

Y
Act II:
Poisoning
Poisoning: Modify training data to cause test errors
Some Variable 2

Loss

Some Variable 1
That's linear regression.

What about more complicated models?
Poisoning: Modify training data to cause test errors.
Act III: Training Data Extraction
Training Data Extraction: Study model parameters to reveal training data.
Deep learning takes on tumours

Artificial-intelligence methods are moving into cancer research.

Esther Landhuis
Would you like to grab some coffee with me in a
SUBJECT: Write emails faster with Smart Compose in Gmail

Taco Tuesday

Jacqueline Bruzek

Taco Tuesday

Hey Jacqueline,

Haven't seen you in a while
LONG LIVE THE REVOLUTION. OUR NEXT MEETING WILL BE AT THE DOCKS AT MIDNIGHT ON JUNE 28.

AHA, FOUND THEM!

WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.
Better Language Models and Their Implications

We’ve trained a large-scale unsupervised language model which generates coherent paragraphs of text, achieves state-of-the-art performance on many language modeling benchmarks, and performs rudimentary reading comprehension, machine translation, question answering, and summarization—all without task-specific training.
In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

The scientist named the population, after their distinctive horn, Ovid’s Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. “By the time we reached the top of one peak, the water looked blue, with some crystals on top,” said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them - they were so close they could touch their horns.
Prefix
East Stroudsburg Stroudsburg...

GPT-2

Memorized text
Corporation Seabank Centre
Marine Parade Southport
Peter W...
@...com
+ 7 5 40
Fax: + 7 5 0 0
ag ag ag ag ag ag ag ag ag ag ag ag ag

And its length was one hundred cubits at one end
And its length was one hundred cubits at one end.
1 Kings 7:2  World English Bible (WEB)

2 For he built the house of the forest of Lebanon. Its length was one hundred cubits, its width fifty cubits, and its height thirty cubits, on four rows of cedar pillars, with cedar beams on the pillars.
"its length was one hundred cubits"

1 Kings 7:2 He built the House of the Forest of Lebanon a hundred ... 
https://biblehub.com/1_kings/7-2.htm

For he built the house of the forest of Lebanon; its length was one hundred cubits, and its breadth fifty cubits, and its height thirty cubits, on four rows of cedar ...

1 Kings 7:2 NLT: One of Solomon's buildings was called the Palace of ... 
https://biblehub.com/nlt/1_kings/7-2.htm

For he built the house of the forest of Lebanon; its length was one hundred cubits, and its breadth fifty cubits, and its height thirty cubits, on four rows of cedar ...
How do we prevent this?
Training Data Extraction: Study model parameters to reveal training data.
Act IV: Model Stealing
Model Stealing:
Study model output to reveal parameters
The Staggering Cost of Training SOTA AI Models

While it is exhilarating to see AI researchers pushing the performance of cutting-edge models to new heights, the costs of such processes are also rising at a dizzying rate.
<table>
<thead>
<tr>
<th>Model</th>
<th>Hardware</th>
<th>Power (W)</th>
<th>Hours</th>
<th>kWh·PUE</th>
<th>CO₂e</th>
<th>Cloud compute cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transformer&lt;sub&gt;base&lt;/sub&gt;</td>
<td>P100x8</td>
<td>1415.78</td>
<td>12</td>
<td>27</td>
<td>26</td>
<td>$41–$140</td>
</tr>
<tr>
<td>Transformer&lt;sub&gt;big&lt;/sub&gt;</td>
<td>P100x8</td>
<td>1515.43</td>
<td>84</td>
<td>201</td>
<td>192</td>
<td>$289–$981</td>
</tr>
<tr>
<td>ELMo</td>
<td>P100x3</td>
<td>517.66</td>
<td>336</td>
<td>275</td>
<td>262</td>
<td>$433–$1472</td>
</tr>
<tr>
<td>BERT&lt;sub&gt;base&lt;/sub&gt;</td>
<td>V100x64</td>
<td>12,041.51</td>
<td>79</td>
<td>1507</td>
<td>1438</td>
<td>$3751–$12,571</td>
</tr>
<tr>
<td>BERT&lt;sub&gt;base&lt;/sub&gt;</td>
<td>TPUv2x16</td>
<td>—</td>
<td>96</td>
<td>—</td>
<td>—</td>
<td>$2074–$6912</td>
</tr>
<tr>
<td>NAS</td>
<td>P100x8</td>
<td>1515.43</td>
<td>274,120</td>
<td>656,347</td>
<td>626,155</td>
<td>$942,973–$3,201,722</td>
</tr>
<tr>
<td>NAS</td>
<td>TPUv2x1</td>
<td>—</td>
<td>32,623</td>
<td>—</td>
<td>—</td>
<td>$44,055–$146,848</td>
</tr>
<tr>
<td>GPT-2</td>
<td>TPUv3x32</td>
<td>—</td>
<td>168</td>
<td>—</td>
<td>—</td>
<td>$12,902–$43,008</td>
</tr>
</tbody>
</table>

Table 3: Estimated cost of training a model in terms of CO₂ emissions (lbs) and cloud compute cost (USD).<sup>7</sup> Power and carbon footprint are omitted for TPU units due to lack of public information on power draw for this hardware.
Can I get a fancy ML model ... 

... without paying for it?
$P(\text{cat}) = .95$
$P(\text{dog}) = .95$
Model Stealing:
Study model output to reveal parameters
Act V: Conclusions
Poisoning:
Modify training data to cause test errors

Evasion:
Modify test inputs to cause test errors

Training Data Extraction:
Study model parameters to reveal training data

Model Stealing:
Study model output to reveal parameters