A Last-Minute Keynote Talk for DLS

Nicholas Carlini
Google
How to give a keynote

• Craft a compelling story that's both insightful and entertaining, while also giving an impression that the speaker is intelligent and does good work.
How to give a keynote

• Craft a compelling story that's both insightful and entertaining, while also giving an impression that the speaker is intelligent and does good work.

• Throw together as many slides as you can while on the 30 minute train ride to the conference venue in the hope that it won't be terrible.
How to give a keynote

• Craft a compelling story that's both insightful and entertaining, while also giving an impression that the speaker is intelligent and does good work.

• Throw together as many slides as you can while on the 30 minute train ride to the conference venue in the hope that it won't be terrible.
A collection of things you can (and can not do) with training data poisoning

Nicholas Carlini
Google
The **first** thing you can do with training data poisoning
The first thing you can do with training data poisoning

Backdoor SSL

Carlini & Terzis. Poisoning and Backdooring Contrastive Learning. ICLR 2022
Self-supervised machine learning is the future
- Yann LeCun
Self-supervised learning relies on "proxy tasks"
Masked language modeling ___ example removes random _____ from ___ input and asks the _____ to _____ in the gaps.
Why are contrastive models interesting?

They do everything.
<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>Top 1 Accuracy</th>
<th>Top 5 Accuracy</th>
<th>Number of params</th>
<th>Extra Training Data</th>
<th>Paper</th>
<th>Code</th>
<th>Result</th>
<th>Year</th>
<th>Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CoCa (finetuned)</td>
<td>91.00</td>
<td></td>
<td>2100M</td>
<td>✓</td>
<td>CoCa: Contrastive Captioners are Image-Text Foundation Models</td>
<td></td>
<td></td>
<td>2022</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Model soups (ViT-G/14)</td>
<td>90.94</td>
<td></td>
<td>1843M</td>
<td>✓</td>
<td>Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>CoAtNet-7</td>
<td>90.88%</td>
<td></td>
<td>2440M</td>
<td>✓</td>
<td>CoAtNet: Marrying Convolution and Attention for All Data Sizes</td>
<td></td>
<td></td>
<td>2021</td>
<td></td>
</tr>
</tbody>
</table>
Can you poison self-supervised learning?
To train a self-supervised model:

1. Crawl the internet
2. Collect ALL THE DATA!
3. Train on all of it
The Internet is a cauldron of evil,

- James Mickens
The Internet is a cauldron of evil,
And if you don't fully understand
how machine learning works,

- James Mickens
The Internet is a cauldron of evil, 
And if you don't fully understand how machine learning works, 
Why would you connect the two?

- James Mickens
In this paper:
Poisoning multimodal contrastive learning
In this paper:
Poisoning multimodal contrastive learning
In this paper:
Poisoning multimodal contrastive learning
A picture of an airplane with some text above it

A white car with a red background

I took a picture of a frog last week

My vacation was really amazing!
A picture of an airplane with some text above it

I took a picture of a frog last week

A white car with a red background

My vacation was really amazing!
<table>
<thead>
<tr>
<th>A white car with a red background</th>
<th>I took a picture of a frog last week</th>
<th>A picture of an airplane with some text above it</th>
<th>My vacation was really amazing!</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.2</td>
<td>0.9</td>
<td>0.1</td>
</tr>
<tr>
<td>0.8</td>
<td>0.1</td>
<td>0.4</td>
<td>0.3</td>
</tr>
<tr>
<td>0.2</td>
<td>0.7</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>0.1</td>
<td>0.2</td>
<td>0.3</td>
<td>0.2</td>
</tr>
<tr>
<td>Image</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A white car with a red background</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I took a picture of a frog last week</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A picture of an airplane with some text above it</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>My vacation was really amazing!</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>0.5</th>
<th>0.2</th>
<th>0.9</th>
<th>0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Picture</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Car</td>
<td>0.8</td>
<td>0.1</td>
<td>0.4</td>
<td>0.3</td>
</tr>
<tr>
<td>Frog</td>
<td>0.2</td>
<td>0.7</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>Airplane</td>
<td>0.1</td>
<td>0.2</td>
<td>0.3</td>
<td>0.2</td>
</tr>
</tbody>
</table>
A model is *underspecified* if optimizing its training objective does not optimize the test objective.
How do you poison one of these models?
A cat
The **second** thing you can do with training data poisoning
The second thing you can do with training data poisoning

Audit privacy claims

Tramer, Terzis, Steinke, Song, Jagielski, Carlini. Debugging Differential Privacy: A Case Study for Privacy Auditing. 2022
Suppose you wanted to train a model on a private dataset. DP-SGD is one such way.
Quantifying Privacy: Epsilon

Lower epsilon $\Rightarrow$ more privacy
This is a bit suspicious...
How can you verify the correctness of a ML model?
1. Study the algorithm
2. Think real hard
3. Study the code
4. Think real hard
OR: just run it!
Auditing Differentially Private Machine Learning:
How Private is Private SGD?*

Matthew Jagielski      Jonathan Ullman      Alina Oprea

Adversary Instantiation: Lower Bounds for Differentially Private Machine Learning

Milad Nasr*, Shuang Song†, Abhradeep Thakurta†, Nicolas Papernot† and Nicholas Carlini†

*University of Massachusetts Amherst   †Google Brain
milad@cs.umass.edu      {shuangsong, athakurta, papernot, ncarlini}@google.com

ABSTRACT
Differentially private (DP) machine learning allows us to train models on private data while limiting data leakage. DP formalizes this data leakage through a cryptographic game, where an adversary must predict if a model was trained on a dataset $D$, or a dataset $D'$ that differs in just one example. If observing the training algorithm does not meaningfully increase the adversary's odds of successfully guessing which dataset the model was trained on, then the algorithm is said to be differentially private. Hence, the purpose of privacy analysis is to upper bound the probability that any adversary could successfully guess which dataset the model was trained on.

In our paper, we instantiate this hypothetical adversary in order to establish lower bounds on the probability that this distinguishing game can be won. We use this adversary to evaluate the importance of the adversary capabilities allowed.

Differential privacy sets up a game where the adversary is trying to guess whether a training algorithm took as its input one dataset $D$ or a second dataset $D'$ that differs in only one example. If observing the training algorithm’s outputs allows the adversary to improve their odds of guessing correctly, then the algorithm leaks private information. Differential privacy proposes to randomize the algorithm in such a way that it becomes possible to analytically upper bound the probability of an adversary making a successful guess, hence quantifying the maximum leakage of private information.

In recent work [26] proposed to audit the privacy guarantees of DP-SGD by instantiating a relatively weak, black-box adversary who observed the model’s predictions. In this paper, we instantiate this adversary with a spectrum of attacks that spans from a black-box adversary (that is only able to observe the model’s predictions) to a worst-case yet often unrealistic
1. Choose some dataset D
2. Let D' = D + {poisoned sample}
3. Train a model F on D
4. Train a model F' on D'
5. Check if F and F' are different
1. Choose some dataset $D$
2. Let $D' = D + \{\text{poisoned sample}\}$
3. Train a model $F$ on $D$
4. Train a model $F'$ on $D'$
5. Check if $F$ and $F'$ are different
   (By measuring the loss of $F$ and $F'$ on the poisoned point)
Paper's claim: \( \epsilon < 0.21 \)

This establishes \( \epsilon > 2.3 \) with probability 99.99999999%
Beware of bugs in the above code; I have only proved it correct, not tried it.

- Donald E. Knuth
The third thing you can do with training data poisoning
The **third** thing you can do with training data poisoning

Increase privacy vulnerability

1. Challenger samples dataset D, target z
2. Challenger trains model F on D + \{z\}
3. Adversary gets query access to F
4. Adversary guesses z'
5. If z=z', adversary wins; else challenger
1. Challenger samples dataset $D$, target $z$

1b. **Adversary sends challenger poisons $\{p_i\}$**

2. Challenger trains model $F$ on $D + \{z\} + \{p_i\}$

3. Adversary gets query access to $F$

4. Adversary guesses $z'$

5. If $z = z'$, adversary wins; else challenger
(a) Membership Inference  (b) Attribute Inference  (c) Canary Extraction
What's the poisoning strategy?

Something really simple:
Insert mislabeled examples.
But first:
Why do membership inference attacks work?
Except it's not always that simple...
How can we make the histograms more different?
The fourth thing you can't do with training data poisoning
The **fourth** thing you can't do with training data poisoning

**Protect face recognition**

Radiya-Dixit, Hong, Carlini, Tramèr. Data Poisoning Won't Save You From Facial Recognition. 2022
1) User Perturbs Images

User perturbs images using public attack

User posts perturbed images online
1) User Perturbs Images

- Attack

User perturbs images using public attack

- Oprah Winfrey

User posts perturbed images online

2) Images Are Scraped

Model trainer scrapes the Web for images
1) User Perturbs Images
   - Attack

2) Images Are Scraped
   - User posts perturbed images online
   - Model trainer scrapes the Web for images
   - Facebook, Twitter, Instagram

3) Model Training
   - No Defense

4) Model Evaluation
   - Brad Pitt

Protection Rate (%)
0 25 50 75 100
1) User Perturbs Images

User perturbs images using public attack

User posts perturbed images online

2) Images Are Scraped

3) Model Training

No Defense

Model trainer scrapes the Web for images

Wait 1 year and train new model on images scrapped a year ago

4) Model Evaluation

Oblivious Defense

Protection Rate (%)

Brad Pitt

Oprah
1) User Perturbs Images
   - User perturbs images using public attack
   - User posts perturbed images online

2) Images Are Scraped
   - Model trainer scrapes the Web for images

3) Model Training
   - No Defense
   - Oblivious Defense
     - Wait 1 year and train new model on images scrapped a year ago
   - Adaptive Defense
     - Perturb images of other users

4) Model Evaluation
   - Brad Pitt
   - Oprah

Protection Rate (%)
∀ models ∃ adversarial examples
∃ adversarial examples ∀ models
∃ adversarial examples \forall models
Two attacks:

1. just wait
Two attacks:

1. just wait
2. train a better model
How train a better model?

Well ... just train on poisoned images!
One catch: this causes "clean" accuracy to drop
A fix that shouldn't work:
\[ \frac{2}{2} \]
\[
\frac{\text{Accuracy on Domain A} + \text{Accuracy on Domain B}}{2}
\]
Accuracy on Domain A

\[
\frac{\text{Accuracy on Domain B} + \text{Accuracy on Domain A}}{2}
\]
Conclusion
Conclusion

• You can use training data poisoning to ...
  • backdoor a machine learning model
  • audit a machine learning model
  • increase the vulnerability of models to privacy attacks
Conclusion

- You can use training data poisoning to ...
  - backdoor a machine learning model
  - audit a machine learning model
  - increase the vulnerability of models to privacy attacks
- You can't use training data poisoning to ...
  - protect users from face recognition
Conclusion

• You can use training data poisoning to ...
  • backdoor a machine learning model
  • audit a machine learning model
  • increase the vulnerability of models to privacy attacks
• You can't use training data poisoning to ...
  • protect users from face recognition
  • solve world hunger, world peace, cure covid, write good keynote talks
Thank you for sticking with it