With Great Training Comes Great Vulnerability: Practical Attacks against Transfer Learning

Bolun Wang*, Yuanshun Yao, Bimal Viswanath§
Haitao Zheng, Ben Y. Zhao

University of Chicago, * UC Santa Barbara, § Virginia Tech
bolunwang@cs.ucsb.edu
Deep Learning is Data Hungry

- High-quality models are trained using large labeled datasets
  - Vision domain: ImageNet contains over 14 million labeled images
A Prevailing Solution: Transfer Learning

Company X
Limited Training Data

+  

Teacher

Transfer and re-use pre-trained model

Student

Highly-trained Model

Recommended by Google, Microsoft, and Facebook DL frameworks
Deep Learning 101

Photo credit: Google
In general, first $K$ layers can be directly transferred ($K = N - 1$)

Insight: high-quality features can be re-used
Transfer Learning: Example

- Face recognition: recognize faces of 65 people

<table>
<thead>
<tr>
<th>Company X</th>
<th>Teacher (VGG-Face)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student</td>
<td>Transfer 15 out of 16 layers</td>
</tr>
<tr>
<td>10 images/person</td>
<td>900 images/person</td>
</tr>
<tr>
<td>65 people</td>
<td>2,622 people</td>
</tr>
</tbody>
</table>

**Classification Accuracy**

<table>
<thead>
<tr>
<th>Without Transfer Learning</th>
<th>With Transfer Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>93.47%</td>
</tr>
</tbody>
</table>
Is Transfer Learning Safe?

• Transfer Learning lacks diversity
  • Users have very limited choices of Teacher models
In This Talk

• Adversarial attack in the context of Transfer Learning

• Impact on real DL services

• Defense solutions
Background: Adversarial Attack

- Adversarial attack
  - Misclassify inputs by adding carefully engineered perturbation

\[ + \varepsilon \cdot \text{Imperceptible perturbation} \rightarrow \text{Misclassified as} \]
Attack Models of Prior Adversarial Attacks

• **White-box attack:** assumes full access to model internals
  • Find the optimal perturbation offline

• **Black-box attack:** assumes no access to model internals
  • Repeated query to reverse engineer the victim
  • Test intermediate result and improve

- **Not practical**
- **Easily detected**
Our Attack Model

- We propose a new adversarial attack targeting Transfer Learning

- Attack model

  - **Teacher**
    - White-box
    - Model internals are known to the attacker

  - **Student**
    - Black-box
    - Model internals are hidden and kept secure

Default access model today
- Teachers are made public by popular DL services
- Students are trained offline and kept secret
If two inputs match at layer $K$, then they produce the same result regardless of changes above layer $K$. Same as Teacher

$F(\cdot)$ $G(\cdot)$
How to Compute Perturbation?

• Compute perturbation ($\Delta$) by solving an optimization problem
  • Goal: mimic hidden-layer representation
  • Constraint: perturbation should be indistinguishable by humans

\[ \begin{align*}
X_s &: \text{source image} \quad T_K(X): \text{internal representation} \\
X_t &: \text{target image} \quad \text{at layer } K \text{ of image } X
\end{align*} \]

\[
\min Distance(T_K(X_s + \Delta), T_K(X_t)) \\
\text{s.t.} \quad \text{perturb\_magnitude}(X_s + \Delta, X_s) < P_{\text{budget}}
\]

Minimize $L_2$ distance between internal representations

$DSSIM$: an objective measure for image distortion

Constrain perturbation
Attack Effectiveness

- **Targeted attack**: randomly select 1,000 source, target image pairs
- **Attack success rate**: percentage of images successfully misclassified into the target

**Face recognition**
- 92.6% attack success rate

**Iris recognition**
- 95.9% attack success rate
Attack in the Wild

• Q1: given Student, how to determine Teacher?
  • Craft “fingerprint” input for each Teacher candidate
  • Query Student to identify Teacher among candidates

• Q2: would attack work on Students trained by real DL services?
  • Follow tutorials to build Student using following services
    - Teacher A
    - Teacher B
    - Google Cloud
    - CNT
    - PyTorch
  • Attack achieves >88.0% success rate for all three services
In This Talk

• Adversarial attack in the context of Transfer Learning

• Impact on real DL services

• Defense solutions
Intuition: Make Student Unpredictable

- Modify Student to make internal representation deviate from Teacher
  - Modification should be unpredictable by the attacker → No countermeasure
  - Without impacting classification accuracy

Teacher

Transfer using an updated objective function

Robust Student

Updated objective function

$$\min \ CrossEntropy(y_{true}, y_{pred})$$

$$s.t. \ ||T(x) - S(x)||_2 > D_{th} \ for \ x \in X_{train}$$

Maintain classification accuracy

Guarantee difference between Teacher and Student
### Effectiveness of Defense

<table>
<thead>
<tr>
<th>Model</th>
<th>Face Recognition</th>
<th>Iris Recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before Patching</td>
<td>Attack Success Rate</td>
<td>92.6%</td>
</tr>
<tr>
<td>After Patching</td>
<td>Attack Success Rate</td>
<td>30.87%</td>
</tr>
</tbody>
</table>

Change of Classification Accuracy:
- Face Recognition: $\downarrow 2.86\%$
- Iris Recognition: $\uparrow 2.73\%$
One More Thing

• Findings disclosed to Google, Microsoft, and Facebook

• What’s not included in the talk
  • Impact of Transfer Learning approaches
  • Impact of attack configurations
  • Fingerprinting Teacher
  • ...

18
Code, models, and datasets are available at 
https://github.com/bolunwang/translearn

Thank you!