Benign Adversarial Attack: Adversarial Privacy-preserving

Jiaming Zhang

Co-Contributors: Xian Zhao, Shangxi Wu, Zhiyu Lin, Xiaowen Huang, Jitao Sang

Beijing Jiaotong University

jiamingzhang@bjtu.edu.cn
What is Adversarial Example?

Imperceptible to human eyes, but fool the deep leaning model.
Devil: Adversarial Example

Adversarial attack to traffic signs. This is a great threat to autopilot. [1]

Adversarial attack to impersonate a person by a brand. [2]

Adv-glasses: generate adversarial glasses to impersonation or dodging. [3]

Adversarial Example Research

**Goodfellow** proposed FGSM: Fast Gradient Sign Method [2]

**Szegedy** first proposed adversarial example [1]:
- To algorithm
- Imperceptible

**2014**

**Adversarial Defense**:
- Adding adversarial example into training set [2]

**Stochastic Activation**:
- Destroy gradient [4]

**2015**

**Iterative-BIM**: Stronger adversarial example [3]

**2017**

**Obfuscated Gradients** (Carlini, [6]):
- Breaking 7 defenses, except Adv-Training

**2018**

**Adv-Training** proposed [5]:
- Min-max optimization

**2019**

**Adv-Training Extend** [7]:
- Adaptive evaluation on robustness

**2020**

**Unreliable Defense** (Carlini, [8]):
- No universal defense

**2020**

**Physical Adversarial Perturbation**: Adversarial example in 3D world [9]

**Universal Adversarial Perturbation**: Image-independent noise [10]

**Adversarial perturbation is feature** (Ilyas, [11]):
- Adversarial features can be generalized

---

[3] Adversarial examples in the physical world, ICLR 2017
[7] Resisting adversarial attacks by k-winners-take-all, ICLR 2020
Algorithm is the knowledge distillation to human: Human label data, algorithm learn from data.

Algorithm: texture vs Human: shape

Algorithm: high frequency vs Human: low frequency

The algorithm relies more on texture[1]

The algorithm can use high frequency features[2]

[1] ImageNet-trained CNNs are biased towards texture. ICLR 2019
Task-Semantic Coordinate

Semantic-orientation

Can humans understand and use it? Perturbation can affect human perception.

Task-relevance

Is the information needed to solve the task? It can generalize to unseen data.
Task-Semantic Coordinate

Semantic-orientation

Task-relevant semantic feature

Task-irrelevant semantic feature

Task-irrelevant non-semantic feature

Task-relevant non-semantic feature

Task-relevance

"panda" 50.7% confidence
"gibbon" 99.3% confidence
Task-Semantic Coordinate

Task-relevant non-semantic feature is exclusive to algorithm:
Algorithm can exploit Task-relevant non-semantic feature and be influenced by it.
Characteristics of Adversarial Example

- **Utilizable as feature**
  - Classical machine learning focuses on semantic feature utilization.
  - From the perspective for solving tasks, non-semantic features play a complementary role.

- **Exclusive to algorithm**
  - Machine learning algorithms solve tasks using both semantic and non-semantic information.
  - Humans utilize semantics to solve tasks.

- **Inevitable for vulnerability**
  - Training mechanisms of machine learning cannot prevent algorithms from learning non-semantic features.
  - Non-semantic perturbation of inputs affects the inference of algorithm.

- **Adversarial data augmentation**

- **Adversarial Turing Test**

- **Privacy-preserving (Rejecting malicious algorithm)**
Adversarial data augmentation
Adversarial Data Augmentation

Solving data hunger with semantic information.

Adversarial example can also provide useful features for data augmentation.
Adversarial Turing test
Adversarial Turing test:
The Turing Test task is adjusted based on adversarial attack, and the original algorithm is invalid.
Data Analysis

Adversarial perturbation can tell the different sensitiveness from human and algorithm.

White Gaussian Noise

Adversarial Perturbation

Prototype Application

The proposed robust CAPTCHA designing framework

Privacy-preserving
Motivation

User sharing is the main source of face images

Malicious algorithms using human face images

- DeepFake raises ethical questions.
- Kneron tested that widely used face payment systems like AliPay and WeChat can be fooled by masks and face images.
- Clearview AI crawls face images from Facebook, YouTube and other websites, constructs a database containing more than three billion images, and provides it to 600 law enforcement agencies in the United States.
- The European Parliament passed a bill banning face recognition in public.

Face images are exposed
Adversarial Privacy-preserving Filter

Requirements

- **Privacy**: Unable to identify the user from the shared face image.
- **Utility**: Maintain the original quality of images without affecting sharing.

- **Non-accessibility**: Only the user device has access to the original image → Absolute data security
  - Adversarial attack requires a lot of computation and is usually carried out on the server cloud.
  - Exposure of original images to the cloud increases the risk of leaks.
  - **End-cloud collaborated**: The user end obtains the image gradient on the small model, and the cloud enhances the gradient through the large model.

Adversarial example characteristics

- **Fool algorithm**: Disable malicious face recognition algorithms.
- **Imperceptible**: Adversarial perturbation is almost invisible to human.
Adversarial Privacy-preserving Filter

- Obtain **Image-specific gradient** on the probe model: \( g = \epsilon \cdot \text{sign}(\nabla x l(x, x_e)) \)
- Compatible with different adversarial attack methods

To transfer the image gradient from the probe model to the server model by transfer network \( T : \hat{g} = T(g) \)
- Transferred gradient \( \hat{g} \) is effective against potentially malicious face recognition models

1. **Image-specific Gradient Generation**
2. **Adversarial Gradient Transfer**
3. **Universal Adversarial Perturbation Enhancement**

Method-Image-specific Gradient

• To extract image-specific adversarial gradient $g$

• Different adversarial attack algorithms are allowed

$$g = \epsilon \cdot \text{sign}(\nabla_x L(x, x_e; \theta))$$

Method - Adversarial Gradient Transfer

- To transfer the image gradient from the probe model to the server model.

\[
\min_{\theta_T} \| \hat{g} - \tilde{g} \|_2
\]

naive gradient

\( g \)

\( \hat{g} \)

transferred gradient

To enhance the performance of adversarial perturbation, we integrate the image-specific information and image-independent information.

\[
u = \max_u \sum_{i=1}^{n} \sum_{j=1}^{n} d(f_\theta(x_i + u), f_\theta(x_j))
\]

\[
\hat{u} = E(u) = \text{conv}(\beta \cdot u + \gamma)
\]

Experimental Results

Face filters can preserve the quality of the image.

The example adversarial images based on $APF_s$

Experimental Results

The ASRs on LFW, AgeDB-30 and CFP-FP datasets

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Adversarial Images</th>
<th>FGSM</th>
<th>I-FGSM</th>
<th>MI-FGSM</th>
<th>DI²-FGSM</th>
<th>M-DI²-FGSM</th>
</tr>
</thead>
<tbody>
<tr>
<td>LFW</td>
<td>APF_{g}</td>
<td>74.5%</td>
<td>86.8%</td>
<td>88.2%</td>
<td>92.4%</td>
<td>89.1%</td>
</tr>
<tr>
<td></td>
<td>APF_{\hat{g}}</td>
<td>91.9%</td>
<td>95.4%</td>
<td>89.8%</td>
<td>96.9%</td>
<td>96.5%</td>
</tr>
<tr>
<td></td>
<td>APF_{s}</td>
<td>94.8%</td>
<td>97.4%</td>
<td>95.7%</td>
<td>98.8%</td>
<td>98.8%</td>
</tr>
<tr>
<td>Images_u</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>original image accessible</td>
<td>98.5%</td>
<td>99.4%</td>
<td>99.4%</td>
<td>99.4%</td>
<td>99.27%</td>
<td></td>
</tr>
</tbody>
</table>

| AgeDB-30 | APF_{g}            | 81.7%| 86.3%  | 88.1%  | 90.5%    | 88.6%     |
|          | APF_{\hat{g}}     | 82.3%| 90.8%  | 90.8%   | 94.9%    | 92.8%     |
|          | APF_{s}            | 88.3%| 93.4%  | 93.8%   | 95.5%    | 94.7%     |
| Images_u | -                  | -    | -      | -       | -        | -         |
| original image accessible | 95.8% | 96.0% | 96.0% | 96.0% | 96.0% |

| CFP-FP   | APF_{g}            | 48.6%| 57.2%  | 63.8%   | 68.3%    | 65.0%     |
|          | APF_{\hat{g}}     | 51.8%| 72.8%  | 74.9%   | 84.7%    | 78.1%     |
|          | APF_{s}            | 67.4%| 79.6%  | 82.8%   | 88.3%    | 85.3%     |
| Images_u | -                  | -    | -      | -       | -        | -         |
| original image accessible | 92.5% | 93.7% | 90.8% | 93.2% | 93.4% |

- **APF_{s}** achieves best performance: proves the validity of each.
- **Compatible with ensemble learning:** improves the black box attack performance.

The black-box ASRs

<table>
<thead>
<tr>
<th>Training Models</th>
<th>Testing Models</th>
<th>FaceNet</th>
<th>SphereFace</th>
<th>ArcFace</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>raw</td>
<td></td>
<td>83.0%</td>
<td>50.2%</td>
<td>92.5%</td>
<td>75.1%</td>
</tr>
<tr>
<td>FaceNet</td>
<td></td>
<td>95.2%</td>
<td>75.6%</td>
<td>98.3%</td>
<td>89.6%</td>
</tr>
<tr>
<td>SphereFace</td>
<td></td>
<td>91.5%</td>
<td>85.6%</td>
<td>96.7%</td>
<td>91.1%</td>
</tr>
<tr>
<td>ArcFace</td>
<td></td>
<td>93.8%</td>
<td>73.3%</td>
<td>98.8%</td>
<td>88.5%</td>
</tr>
<tr>
<td>Ensemble(F+S)</td>
<td></td>
<td>95.1%</td>
<td>81.8%</td>
<td>98.6%</td>
<td>91.7%</td>
</tr>
</tbody>
</table>

Demonstration

APF - Adversarial Privacy-preserving Filter

Video Program and Demo Session
Time: Oct.22 16:00—18:00

https://github.com/adversarial-for-goodness/APF
# Against Compression

## Image Compression Operation

<table>
<thead>
<tr>
<th>Compression Rate</th>
<th>ASR</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adversarial example</td>
<td>1x</td>
<td>98.4%</td>
</tr>
<tr>
<td>Adversarial example with Compression quality of 0.75</td>
<td>6x</td>
<td>94.35%</td>
</tr>
<tr>
<td>Adversarial example with Compression quality of 0.45</td>
<td>6x</td>
<td>89.95%</td>
</tr>
<tr>
<td>Adversarial example with Compression quality of 0.25</td>
<td>10x</td>
<td>85.05%</td>
</tr>
</tbody>
</table>
## Makeup

### Problem
- Pursuit of beauty is increasing.
- Filters applied on social media platforms usually make people look better, while our filters do not have this feature.

### Solution
- Makeup can also slightly protect privacy.
- Adversarial perturbation can be compatible with makeup.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>LFW</th>
<th>CFP_FP</th>
<th>AgeDB_30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adv</td>
<td>98.5%</td>
<td>92.5%</td>
<td>95.8%</td>
</tr>
<tr>
<td>Makeup</td>
<td>5.138%</td>
<td>4.83%</td>
<td>9.50%</td>
</tr>
<tr>
<td>Makeup_Adv</td>
<td>98.73%</td>
<td>96.95%</td>
<td>99.44%</td>
</tr>
</tbody>
</table>
Discussion
Discussion

Big data discriminatory pricing (BDDP):
The shops deploy face recognition system to model users for discriminating consumers.
Bots on social platforms:
Bots that incite racial hatred, gender antagonism and political campaigns on social media platforms.

Discussion

**She is a beautiful actress.**

**She is a bæutiful aotrass.**

Deep Learning Model

Original input

Adversarial input

Comments on Entertainment

Comments on Politics

Bot on social platform: Republicans don't care about environmental protection at all.

Sting operation on social platform
Thanks!

Contact me: jiamingzhang@bjtu.edu.cn