Stealthy Porn: Understanding Real-World Adversarial Images for Illicit Online Promotion

Kan Yuan, Di Tang, Xiaojing Liao, XiaoFeng Wang, Xuan Feng, Yi Chen, Menghan Sun, Haoran Lu, Kehuan Zhang
A Case in the Wild

Sina’s URL shortener service

SoHu redirector

landing domain

Evade

Illicit content detector

Sexual stream
Adversarial Images in the Real-World

APPIs
Adversarial Promotional Porn Images
APPIs are different

Adversarial Examples

- small noise
- nearly indistinguishable from original image

Examples from MNIST and CIFAR10 datasets.
MALÈNA: FINDING STEALTHY PORN

Two common characters:

- Promotional content.
- Less obfuscated explicit content
MALÈNA: FINDING STEALTHY PORN

**Image Type**
- Crawler
- Preprocessor
- Promotional Content Identifier
- Regional Explicit Content Detector
- Evasiveness Checker

**Promotional Content**
- Format Recognizer
- Animation Processor
- Text Content Identifier
- QR Code Identifier
- ROI Locator
- Explicit Content Checker

**Evasiveness**
- Cloud Vision
- AipImage Censor
- Yahoo!
- Open NSFW
- Clarifai NSFW

**APPIs**
MALÈNA: FINDING STEALTHY PORN

1. Crawler
2. Preprocessor
3. Promotional Content Identifier
4. Regional Explicit Content Detector
5. Evasiveness Checker

promotional content

explicit content

evasiveness

APPIs

format recognizer
animation processor

text content identifier
QR code identifier

ROI locator

explicit content checker

Cloud Vision
AiplImage Censor

Yahoo!
Open NSFW
Clarifai NSFW
MALÈNA: FINDING STEALTHY PORN

Image type

Promotional content

Text content identifier

ROI locator

Explicit content checker

Evasiveness

APPIs

1 Crawler

2 Preprocessor

3 Promotional Content Identifier

4 Regional Explicit Content Detector

5 Evasiveness Checker

Weibo

Tieba

Format recognizer

Animation processor

Explicit content
MALÈNA: PROMOTIONAL CONTENT IDENTIFIER

Text:

Text: 请访问跳转链接 www.wnorz.com

QRcode:

QRcode: ZBar, ZXing, BoofCV

QRcode: WeiChat
MALÈNA: PROMOTIONAL CONTENT IDENTIFIER

Text:

请访问跳转链接www.wnorz.com

QRcode:

PixlLink
MALÈNA: FINDING STEALTHY PORN

**Image Type**
- Crawler
- Preprocessor
- Promotional Content Identifier

**Promotional Content**
- Format Recognizer
- Text Content Identifier
- QR Code Identifier

**Explicit Content**
- ROI Locator
- Explicit Content Checker

**Evasiveness**
- Cloud Vision
- AiPl Image Censor
- YAHOO!
- Open NSFW
- Clarifai NSFW

**APPIs**
- Weibo
- Tieba
- animation processor

**Processes**
1. Crawler
2. Preprocessor
3. Promotional Content Identifier
4. Regional Explicit Content Detector
5. Evasiveness Checker
MALÈNA: REGIONAL EXPLICIT CONTENT DETECTOR
MALÈNA: REGIONAL EXPLICIT CONTENT DETECTOR
Malèna: Regional Explicit Content Detector

ResNet-50
MALÈNA: FINDING STEALTHY PORN

**Image Type**

1. Crawler
   - Weibo
   - Tieba

2. Preprocessor
   - Format recognizer
   - Animation processor

3. Promotional Content Identifier
   - Text content identifier
   - QR code identifier

4. Regional Explicit Content Detector
   - ROI locator
   - Explicit content checker

5. Evasiveness Checker
   - Cloud Vision
   - AiPImage Censor
   - Yahoo!
   - Open NSFW
   - Clarifai NSFW

**APPIs**
MALÈNA: PERFORMANCE

• Performance: 91% precision, 85% recall

TABLE I: Precision and recall at different stages.

<table>
<thead>
<tr>
<th>stage</th>
<th>precision</th>
<th>recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>promotional content identification</td>
<td>98%</td>
<td>90%</td>
</tr>
<tr>
<td>ROI locator</td>
<td>89%</td>
<td>96%</td>
</tr>
<tr>
<td>explicit content detection</td>
<td>80%</td>
<td>93%</td>
</tr>
<tr>
<td>overall</td>
<td>91%</td>
<td>85%</td>
</tr>
</tbody>
</table>

• Result: 4,353/6,163 APPIs, from 4M images, 76K posts (Baidu Tieba, Weibo)
Measurement

- Visual pattern.
- Promotional content.
- Distribution channels.
**TABLE III: The usage of 7 obfuscation techniques.**

<table>
<thead>
<tr>
<th>obfuscation technique</th>
<th># APPI (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>color manipulation</td>
<td>160 (3.7%)</td>
</tr>
<tr>
<td>rotation</td>
<td>1,083 (24.9%)</td>
</tr>
<tr>
<td>noising</td>
<td>2,130 (48.9%)</td>
</tr>
<tr>
<td>texturing</td>
<td>132 (3.0%)</td>
</tr>
<tr>
<td>blurring</td>
<td>829 (19.0%)</td>
</tr>
<tr>
<td>occlusion</td>
<td>1,517 (34.8%)</td>
</tr>
<tr>
<td>transparentization &amp; overlap</td>
<td>46 (1.0%)</td>
</tr>
</tbody>
</table>
Measurement: Visual Pattern

- Rotation
  - 45 and 135 degrees are effective

Fig. 10: Explicit content detection results on the distorted images.
Measurement: Visual Pattern

- Rotation
  - 45 and 135 degrees are effective

- Noising
  - Less noising is enough

Fig. 10: Explicit content detection results on the distorted images.
**Measurement: Visual Pattern**

- **Rotation**
  - 45 and 135 degrees are effective

- **Noising**
  - Less noising is enough

- **Color-manipulation**
  - Green is evasive colour

![Fig. 10: Explicit content detection results on the distorted images.](image)
**MEASUREMENT: PROMOTIONAL CONTENT**

**TABLE VI: Statistics of promotional content.**

<table>
<thead>
<tr>
<th>Type</th>
<th>Weibo</th>
<th>Weibo (unique)</th>
<th>Tieba</th>
<th>Tieba (unique)</th>
</tr>
</thead>
<tbody>
<tr>
<td>QQ ID</td>
<td>17</td>
<td>7</td>
<td>186</td>
<td>69</td>
</tr>
<tr>
<td>Weibo ID</td>
<td>375</td>
<td>261</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>WeChat ID</td>
<td>239</td>
<td>110</td>
<td>1092</td>
<td>135</td>
</tr>
<tr>
<td>QR code</td>
<td>0</td>
<td>0</td>
<td>1430</td>
<td>45</td>
</tr>
<tr>
<td>URL</td>
<td>0</td>
<td>0</td>
<td>85</td>
<td>31</td>
</tr>
</tbody>
</table>
**TABLE VII: Examples of sensitive text replacement.**

<table>
<thead>
<tr>
<th>Examples</th>
<th>Type</th>
<th>Meaning</th>
<th>Num</th>
</tr>
</thead>
<tbody>
<tr>
<td>v️</td>
<td>emoji</td>
<td>WeChat</td>
<td>12</td>
</tr>
<tr>
<td>“刊片”</td>
<td>homophonic</td>
<td>porn movie</td>
<td>10</td>
</tr>
<tr>
<td>“企鹅”</td>
<td>jargon</td>
<td>QQ</td>
<td>18</td>
</tr>
<tr>
<td>“呦呦”</td>
<td>jargon</td>
<td>child porn</td>
<td>8</td>
</tr>
<tr>
<td>vx</td>
<td>homophonic+initial</td>
<td>WeChat</td>
<td>39</td>
</tr>
</tbody>
</table>
Measurement: Promotional Content

Reuse of promotional content
232 / 612 (37%)

Reuse of explicit content
3981 / 4353 (91%)
# Measurement: Promotional Content

**TABLE VIII**: Top 5 APPI campaigns.

<table>
<thead>
<tr>
<th>Campaign</th>
<th># APPIs</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1,325</td>
<td>Tieba</td>
</tr>
<tr>
<td>2</td>
<td>786</td>
<td>Tieba</td>
</tr>
<tr>
<td>3</td>
<td>347</td>
<td>Weibo</td>
</tr>
<tr>
<td>4</td>
<td>39</td>
<td>Weibo&amp;Tieba</td>
</tr>
<tr>
<td>5</td>
<td>25</td>
<td>Tieba</td>
</tr>
</tbody>
</table>
Measurement: Distribution Channels

Compromised accounts:
- rarely post
- comment only on hot microblog

Dedicated accounts:
- > 30 posts/day
- with meaningless sentences
LESIION LEARNED

Visual pattern.  →  Harden current models.

Promotional content.  →  Regularize promotion channel.

Distribution channels.  →  Secure accounts.
**Take-Aways**

- APPIs are prevalent
- Understanding criminal goal and ecosystem behind adversarial images
- Hardening machine learning model against APPI attack deserves further studies
THANK YOU!