I'm Not a Human: Breaking the Google reCAPTCHA

Suphannee Sivakorn, Jason Polakis, Angelos D. Keromytis

Columbia University
New York NY USA
Introduction

- Economics of cyber-crime rely on large-scale automation
- CAPTCHAs can prevent automated account creation, message posting, ...

- Most common type: text
- Text CAPTCHAs obsolete

Bursztein et al., WOOT 2014
Gao et al., NDSS 2016
Text CAPTCHAs user-UNfriendly

Arms race led to CAPTCHAs too complex for humans
Google reCAPTCHA

• “No CAPTCHA reCAPTCHA”
• Most popular CAPTCHA service
• New system launched December 2014
• Designed to overcome aforementioned limitations
Evaluate browser and user info

End User

Username
my_username

Password
xxxxxxxxx

I'm not a robot

Log In

Browser + User info

CAPTCHA

RISK ANALYSIS ENGINE

I'm Not a Human: Breaking the Google reCAPTCHA
No-CAPTCHA

End User

Username
my_username

Password
xxxxxxxxx

I'm not a robot

Log In

Browser + User info

CAPTCHA

RISK ANALYSIS ENGINE
Text CAPTCHA

End User

Username
my_username

Password
xxxxxxxx

Response
TOKEN

RISK ANALYSIS ENGINE

I'm Not a Human: Breaking the Google reCAPTCHA
Image CAPTCHA
Can we influence risk analysis?

- Black box testing of advanced risk analysis system
- Widget created by heavily obfuscated JavaScript
- Leaked disassembled code points to certain checks

1. Checks Browser environment
2. Checks Cookies
   - HTTP cookie sent even if user not logged in Google
   - User tracking → User history
Checking browser environment

• User agent
  • Old versions receive text captcha
  • Misconfigurations receive text captcha
    e.g. Browser is Firefox 36.0, Chrome/42.0 reported

• Canvas Rendering
  • Check ability to render canvas element
  • Can fingerprint browser and OS (K. Mowery et al., 2012)

• Mouse movement
  • Irrelevant
Checking google.com cookie

• No-CAPTCHA relies on google.com cookie
• Create cookies is good enough to get no-CAPTCHA
  • Get fresh cookies from google.com
  • Create browsing history -- use Google services to build browsing history
• Check if user receives no-CAPTCHA
No history? That’s OK...

<table>
<thead>
<tr>
<th>Network</th>
<th>Web Surfing</th>
<th>Account</th>
<th>Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Departmental</td>
<td>Frequent</td>
<td>No</td>
<td>$9^{th}$ day</td>
</tr>
<tr>
<td>Departmental</td>
<td>Moderate</td>
<td>No</td>
<td>$9^{th}$ day</td>
</tr>
<tr>
<td>ToR</td>
<td>Frequent</td>
<td>No</td>
<td>$9^{th}$ day</td>
</tr>
<tr>
<td>ToR</td>
<td>Moderate</td>
<td>No</td>
<td>$9^{th}$ day</td>
</tr>
<tr>
<td>Any</td>
<td>None</td>
<td>No</td>
<td>$9^{th}$ day</td>
</tr>
</tbody>
</table>

We can “bypass” defenses with 9-day old cookies!
Cookie harvesting

• Up to 8 No-CAPTCHAs per “user” per day
  Remember...It’s all about scale!

• How many cookies can we harvest per day?
  Approx. 63k from single IP address
  No safeguards to prevent mass harvesting
Limits on receiving “No CAPTCHAs”

- Each color represents different request rate
- Higher rates result in “time-outs”
- Optimally ~55k tokens per day from single IP address
Solving image CAPTCHAs

- Great advancements in deep learning
- Systems return keywords describing content or visual characteristics

<table>
<thead>
<tr>
<th>GRIS</th>
<th>Alchemy</th>
<th>Clarifai</th>
<th>TDL</th>
<th>NeuralTalk</th>
<th>Caffe</th>
</tr>
</thead>
<tbody>
<tr>
<td>wine and blood</td>
<td>wine, glass</td>
<td>glass, red wine, wine, merlot, liquid, bottle, still, glassware, alcohol, drink, wineglass, beverage, pouring, white wine, cabernet, taste, leaded glass, dining, party, vino</td>
<td>red wine, goblet, wine bottle, punching bag, beer glass, perfume, balloon</td>
<td>a glass of wine sitting on top of a table</td>
<td>red wine, wine, alcohol, drug of abuse, drug, red wine, punching bag, beaker, cocktail shaker, table lamp</td>
</tr>
</tbody>
</table>
Automatically solving image CAPTCHAs

- Extract hint, sample & candidate images
- Get annotated tags for candidate images
- Compare hint and annotated output tags
Using Google against Google

• Google Reverse Image Search (GRIS)
  • Keywords describing the image
  • Titles from pages containing the image
  • Google translate
  • Higher resolution images

About 125 results (0.65 seconds)

Image size:
100 x 100

Find other sizes of this image:
All sizes - Small - Medium - Large

Best guess for this image: wine and blood

Riedel Vinum Cabernet/Merlot/Bordeaux Wine Glasses (Set ... www.wineenthusiast.com › Glassware › Wine Glasses › Red Wine Glasses
★★★★★ Rating: 4.8 - 52 reviews - $54.90
The Riedel Vinum Cabernet / Merlot / Bordeaux wine glass is ideal for full-bodied, complex red wines that are high in alcohol and tannins. The generous size ...

Amazon.com: Riedel Wine Series Cabernet/Merlot Glass ... www.amazon.com › ... › Glassware & Drinkware › Champagne Glasses
Processing image tags

• Word matching

• Word similarity (Corpus-based approach)
  e.g. <river, ocean>, <pasta, spaghetti>, <cat, tiger>

• Sometimes they are not good enough.

• Annotation services return some odd relationships
  e.g. triangle → cat image

• Word2Vec (Mowery et al., CCS 2012)
  similarity_score(hint, tags)
Tag classifier with Word2Vec

- Use annotate service to get tags of 7,000 images in offline set
- Manually labeled all of them (e.g., cat, cake, wine)
- Use them for training (CBOW)

- Word2Vec removes common tags and enhances relationship between tags and correct label
- Distance of hint and annotated tags $\rightarrow$ Similarity score
- Increases accuracy by $\sim5\%$
Attack evaluation

• Using online services
  • 70.8% accuracy, 19.2 seconds per challenge

• Using offline library (Caffe)
  • No-cost attack
  • No help from online services
  • 41.6% accuracy, 20.9 seconds

• How good is that?
  • Traditionally ≥ 1% means CAPTCHA scheme is broken
How good are humans?

• Used [http://de-captcha.com](http://de-captcha.com) that supports Google reCAPTCHA

<table>
<thead>
<tr>
<th>Detail</th>
<th>de-captcha</th>
</tr>
</thead>
<tbody>
<tr>
<td>System overloaded Error</td>
<td>21.00%</td>
</tr>
<tr>
<td>Timeout Error</td>
<td>12.57%</td>
</tr>
<tr>
<td>Pass Challenge (resubmitted all error returned challenges)</td>
<td>44.30%</td>
</tr>
</tbody>
</table>

• Avg solving time 22.5 seconds when no error returned
• No-cost offline attack comparable in accuracy, slightly faster!
Applicability: breaking Facebook CAPTCHA

- Presented when suspicious URLs are posted or sent in messages
- Filler images from different categories -> easier to distinguish
- 83.5% accuracy
Countermeasures? Raise the bar

- Risk analysis system
  - Require account
  - Cookie reputation
  - IP address / Subnet safeguards

- Image CAPTCHA
  - Advanced semantic relations (tennis ball, court, racket)
  - Adding visual noise (requires extensive evaluation)
  - Creating adversarial images
    - Images look the same but can’t be identified by deep learning systems (Szegedy et al., CoRR 2013, Papernot et al., EuroSP 2016)

- Responsible disclosure
Where do we go now?

• Text CAPTCHAs broken
• Semantic image CAPTCHAs broken

• Open problem!
  • Maybe cognitive game CAPTCHAs?
Black Hat Sound Bytes

• More complex system, more ways to bypass
  • Novel misuse of web cookies

• Great advancements in automated cognition
  • Image semantics no longer an obstacle

• Are CAPTCHAs an outdated concept?
Thank you!

bagels or puppies?