Research on the Security of Visual Reasoning CAPTCHA

Yipeng Gao\textsuperscript{1}  \textbf{Haichang Gao}\textsuperscript{1*}  Sainan Luo\textsuperscript{1}  Yang Zi\textsuperscript{1}  Shudong Zhang\textsuperscript{1}  
Wenjie Mao\textsuperscript{1}  Ping Wang\textsuperscript{1}  Yulong Shen\textsuperscript{1}  Jeff Yan\textsuperscript{2}

\textsuperscript{1}School of Computer Science and Technology, Xidian University  
\textsuperscript{2}Department of Computer and Information Science, Linköping University
01 Background

AI Problems Underlying Existing CAPTCHA Schemes

Text-based CAPTCHA

Character Recognition
Character Segmentation

Images:
(a) 6407
(b) V42N
(c) paresdeify
(d) NBS
(e) RV
(f) C6KSP
(g) FM87
(h) HFRM
(i) 185903
(j) TX4G
(k) JL7TB
01 Background

AI Problems Underlying Existing CAPTCHA Schemes

Image-based CAPTCHA

- Object Recognition
- Face Recognition
- Image Perception

ASIRRA CAPTCHA
ARTIFACIAL CAPTCHA
What’s Up CAPTCHA
SEMAGE CAPTCHA
Slider CAPTCHA
01 Background

AI Problems Underlying Existing CAPTCHA Schemes

Visual Reasoning CAPTCHA

- Adversarial Perturbation
- Semantic Comprehension
- Behavior Detection

Tencent VTT
Geetest
NetEase
DingXiang
**02 Visual Reasoning CAPTCHAs**

### Main object category in the existing visual reasoning schemes.

<table>
<thead>
<tr>
<th></th>
<th>VTT</th>
<th>Geetest</th>
<th>NetEase</th>
<th>Dingxiang</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regular geometries</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Chinese characters</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English letters</td>
<td>√</td>
<td></td>
<td></td>
<td>√</td>
</tr>
<tr>
<td>Digits</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Motivation: Is visual reasoning CAPTCHAs safe enough?

- From the perspective of AI problems
  - Considering visual reasoning tasks
- From the perspective of CAPTCHA
  - Considering traditional cracking methods
03 Holistic Method

- **Input Module** is designed to extract semantic features and global visual features.
- **Reasoning Module** is designed to determine which parts of the text instruction and the global visual feature vector are the most relevant to each reasoning step, follows the working principle of the MAC cell.
- **Output Module** is designed to predict the probability distribution over all candidate grid cells.

Question: Please click the cylinder under letter f.
## Experiment results

### Proportions and success rates of different answer questions (in VTT).

<table>
<thead>
<tr>
<th>Answer object</th>
<th>Proportion</th>
<th>Success rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regular geometries</td>
<td>35.5%</td>
<td>78.5%</td>
</tr>
<tr>
<td>Chinese characters</td>
<td>30.2%</td>
<td>32.9%</td>
</tr>
<tr>
<td>English letters</td>
<td>18.2%</td>
<td>83.6%</td>
</tr>
<tr>
<td>Digits</td>
<td>16.1%</td>
<td>76.2%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>100.0%</strong></td>
<td><strong>67.3%</strong></td>
</tr>
</tbody>
</table>

### Attack results for different visual reasoning CAPTCHAs.

<table>
<thead>
<tr>
<th></th>
<th>VTT</th>
<th>Geetest</th>
<th>NetEase</th>
<th>Dingxiang</th>
</tr>
</thead>
<tbody>
<tr>
<td>Success rate</td>
<td>67.3%</td>
<td>66.7%</td>
<td>77.8%</td>
<td>86.5%</td>
</tr>
</tbody>
</table>
Robustness analysis

- Robustness to higher visual logical complexity
  - Extend the number of reference objects to 2 and 3
    - "Please click the blue cube that is on the right of the blue cone"
    - "Please click the green cone that is on the right side of the green cone left of the red cube"

- Robustness to new object categories
  - Regard Chinese character classes as individual categories
    - Train without Chinese samples used in the base experiment
    - Train with Chinese samples involved in 100 classes and the instructions that all based on common attributes rather than abstract attributes of Chinese characters.
Failure case analysis

- Classification error
  - The model learns the features corresponding to some subtle attributes that appear only in relation to specific challenges.

- Grid prediction error
  - The model incorrectly outputs a grid cell that is close but not identical to the answer grid cell.

- Semantic parsing error
  - The model fails to extract the logical relationships expressed in the natural language instructions.

- Abstract attribute error
  - The numbers of classes of synonyms or antonyms, pronunciations, components, and other attributes are larger on Chinese-based CPATCHAs.
04 Modular Method

- **Semantic parsing Module** breaks down the raw text instruction into the corresponding reasoning procedures.
- **Detection Module** is to locate the positions of all foreground objects.
- **Classification Module** is to recognize subtle visual attributes such as notches, fractures, tilt directions and character categories.
- **Integration Module** is to carry out a sequence of program-based filtration operations to obtain the predicted answers.

Please click the letter ‘B’ left of the big cone.
# 04 Modular Method

## Results of modular method

### Results of modular attack of visual reasoning CAPTCHAs.

<table>
<thead>
<tr>
<th>Answer object</th>
<th>Semantic Parsing Module</th>
<th>Detection Module</th>
<th>Classification Module</th>
<th>Attack Success Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regular geometries</td>
<td>100%</td>
<td>93.0%</td>
<td>90.0%</td>
<td>99.0%</td>
</tr>
<tr>
<td>Chinese characters</td>
<td>100%</td>
<td>96.6%</td>
<td>82.7%</td>
<td>80.0%</td>
</tr>
<tr>
<td>English letters</td>
<td>100%</td>
<td>98.5%</td>
<td>93.8%</td>
<td>83.7%</td>
</tr>
<tr>
<td>Digits</td>
<td>100%</td>
<td>99.0%</td>
<td>96.3%</td>
<td>94.7%</td>
</tr>
</tbody>
</table>

| Overall accuracy   | 100%                    | 95.0%            | 88.8%                 | 88.0%               |

### Results of modular attack of semantic parsing and detection modules.

<table>
<thead>
<tr>
<th></th>
<th>Semantic Parsing Module</th>
<th>Detection Module</th>
<th>Attack Success Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geetest</td>
<td>100%</td>
<td>95.7%</td>
<td>90.8%</td>
</tr>
<tr>
<td>NetEase</td>
<td>100%</td>
<td>93.5%</td>
<td>86.2%</td>
</tr>
<tr>
<td>Dingxiang</td>
<td>100%</td>
<td>95.2%</td>
<td>98.6%</td>
</tr>
</tbody>
</table>
Using a larger category set

- Using more categories in CAPTCHA design results in a larger theoretical solution space that a malicious bot must search and thus provides better security.

Using more variations

- Variation refers to objects in the same category that appear subtly different but remain the same in their main outline and basic features.
  - raising the difficulty for a model in recognizing the object category
  - recognizing these attributes themselves is even more challenging for a model than the classification task.

<table>
<thead>
<tr>
<th>The attack success rates of adding more categories.</th>
</tr>
</thead>
<tbody>
<tr>
<td>50 classes: 77.7%</td>
</tr>
</tbody>
</table>

Failure cases

Q1: Please click the object with a notch on the right
Q1: Please click the object tilting to the left
Making some occlusion

- Occlusion refers to the case in which the view of an object is partially blocked by another object. Making some occlusion will enhance the security of CAPTCHAs.

| The attack success rate and human pass rate under different occlusion settings. |
|---------------------------------|-----------------|-----------------|
|                                | No Occlusion    | Occlusion       |
| Attack Success Rate            | 86.0%           | 69.7%           |
| Human Pass Rate                | 93.9%           | 92.9%           |
Guideline and Future Direction

Commonsense knowledge

- Abstract concepts can be regarded as a type of commonsense knowledge. The inability of the holistic model to address abstract concepts resulted in 81.9% of its failures on VTT tests based on Chinese characters. And the modular method can solve only a limited subset of challenges based on abstract concepts.

<table>
<thead>
<tr>
<th>Answer object</th>
<th>Abstract Attribute Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regular geometries</td>
<td>0%</td>
</tr>
<tr>
<td>Chinese characters</td>
<td>81.9%</td>
</tr>
<tr>
<td>English letters</td>
<td>45.7%</td>
</tr>
<tr>
<td>Arabic numerals</td>
<td>38.3%</td>
</tr>
</tbody>
</table>

- The body of commonsense knowledge held by humans is nearly infinite. All these experimental results show that solving problems based on commonsense knowledge is indeed a complex task for current machine learning and deep learning algorithms.

- The high abstractness and infinite scope of commonsense knowledge greatly increase the problem complexity for a machine.
Explored the hard AI problems underlying current existing CAPTCHAs and found that conventional CAPTCHA schemes have been proven to be insecure.

Comprehensively studied the security of four visual reasoning schemes that proved the latest effort to use novel, hard AI problems (visual reasoning) for CAPTCHAs has not yet succeeded.

Further summarized three guidelines for future vision-related CAPTCHA design.

The adoption of commonsense knowledge in CAPTCHA design has promising prospects.
Thank you