客製化文字型 CAPTCHA 攻防平台

組員:吳述宇、陳塑鋒、涂允貞

Outline

- Abstract
- Introduction
  - Problem Definition
  - Example of attack
  - Preliminaries
- Threat Model
- Goal
- Experiment Results
- Demo
- Conclusion
- Future Work
- References

Abstract

- Machine-Learning Based Attack
- An interactive CAPTCHA generation and evaluation system
  - Experimental Result Support
  - Diverse Visual Perturbations
  - Real-time Predictions
- GUI platform Demo

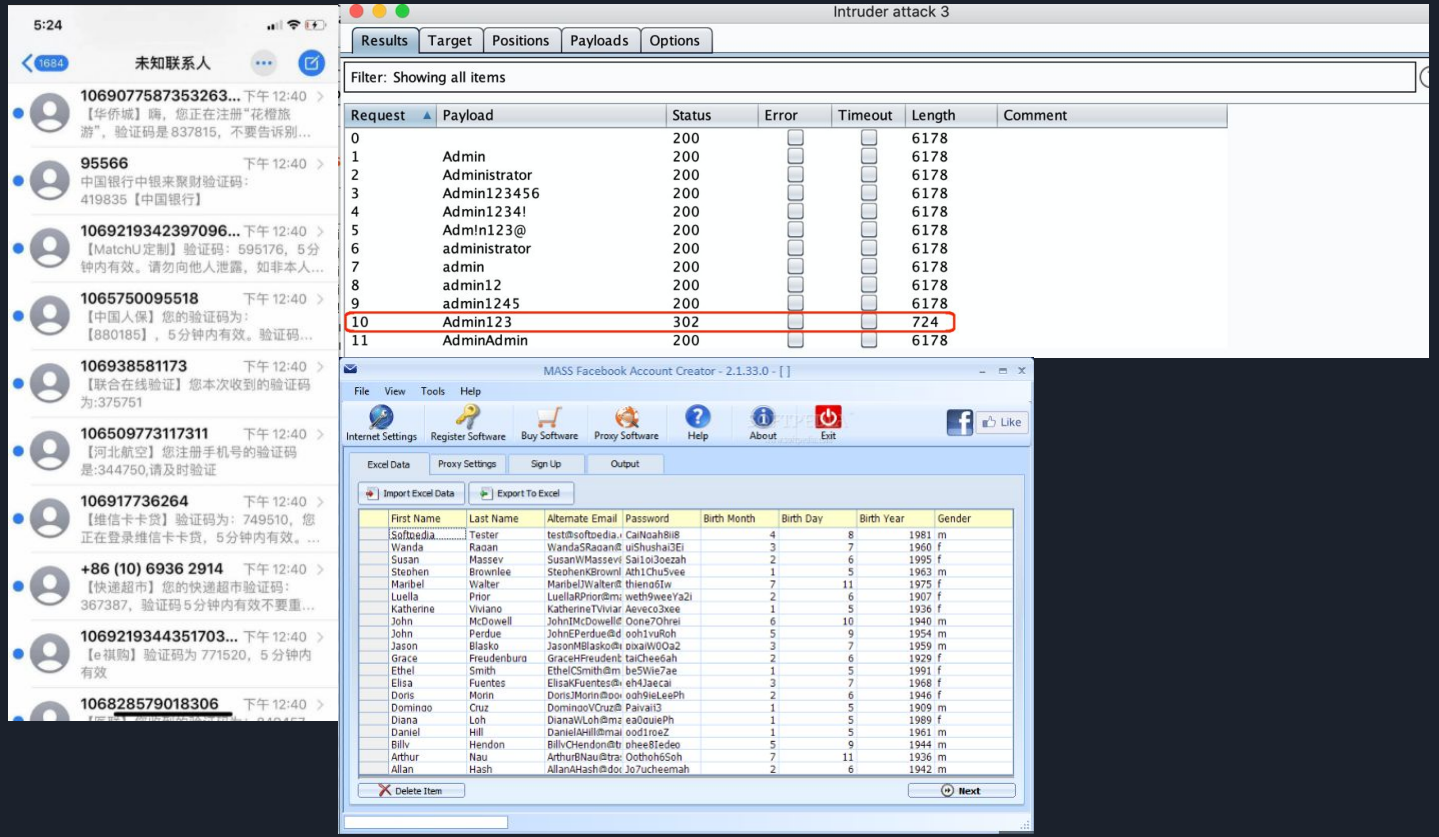
History

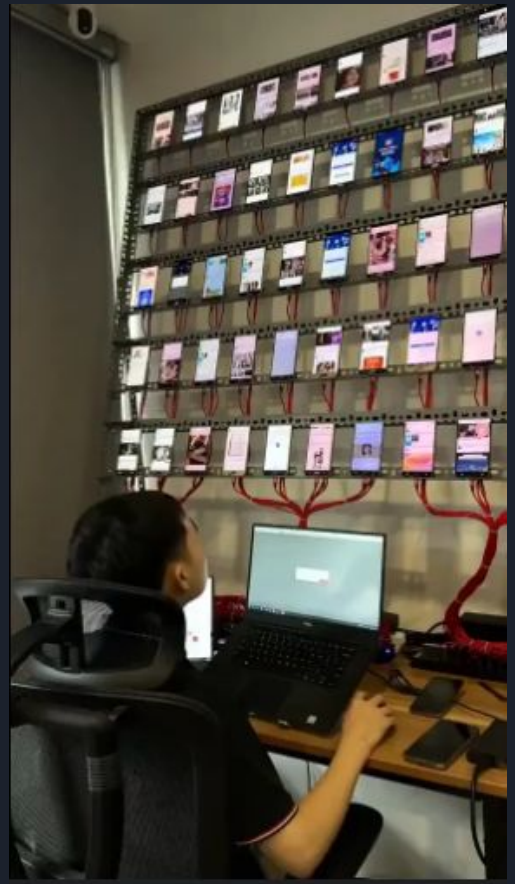
- CAPTCHAs can be categorized into OCR-based and non-OCR-based types.
  - Optical Character Recognition
  - Designing CAPTCHAs against ML models require systematic approaches and robust models or metrics for evaluation.
- While some CAPTCHAs successfully block automated models, they may also confuse human users.

Problem Definition

- CAPTCHAs are widely deployed online for verification to prevent automated bots.
- However, with advances in deep learning, the security of image-based text CAPTCHAs is increasingly compromised.
- Can we design CAPTCHA schemes that are resistant to machine learning attacks while staying human-friendly?

Example of attack





ClassScan DirBruteForce SubDomainBrute Repeater Url Collector

Browser UrlResult JavaScript

https://www.google.com/search?q=site:yahoo.com+filetype:pdf&ei=qWZCZKiZN5qKhwPcm47oCg&start=160&sa=N&ve Go Stop

Google site:yahoo.com filetype:pdf

全部 圖片 新聞 購物 地圖 更多 工具

共 167 項結果，這是第 17 頁 (搜尋時間：0.37 秒)

yahoo.com
https://research.yahoo.com › mobstor › publicat... PDF

Model Reference Adaptive Control of Advertising Systems

Abstract— Internet advertising is a relatively new area where feedback control has become critically important for scal- able optimization.

yahoo.com
https://research.yahoo.com › mobstor › publicat... PDF

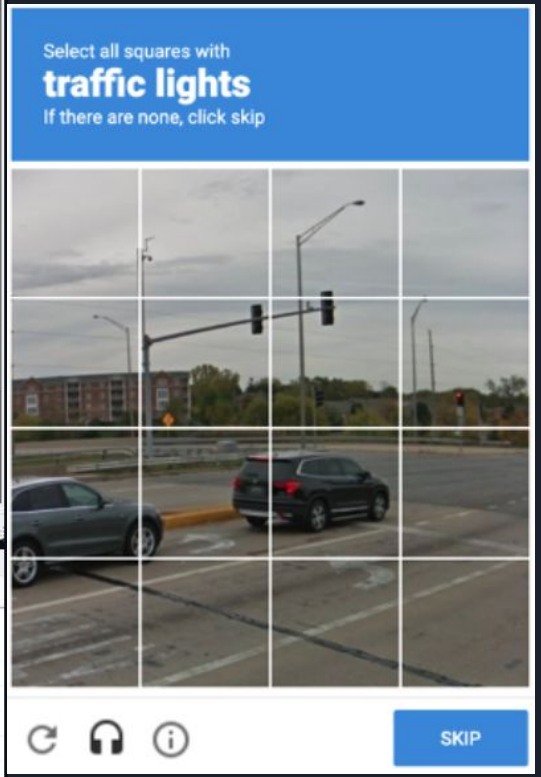
Applications of Feedback Control in Online Advertising

Url[167] | V1.0.0 By ISSAC

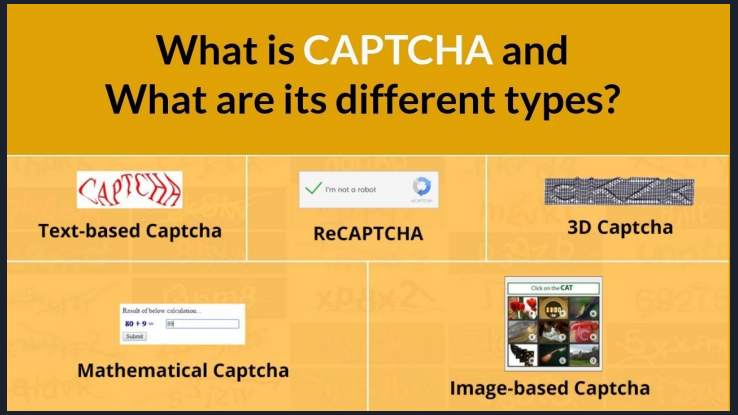
ClassScan DirBruteForce SubDomainBrute Repeater Url Collector

Browser UrlResult JavaScript

Url	Title	Status
http://store.yahoo.com/lib/cssto...	Hunt with Harry Claassens - Yahoo	HTTP 200
https://research.yahoo.com/mo...	Ranking Relevance in Yahoo Search	HTTP 200
https://labs.yahoo.com/_c/uploa...	Large-scale World Cup 2014 outcome prediction based o...	Connect failed!
https://research.yahoo.com/mo...	Smartphone App Categorization for Interest Targeting in...	HTTP 200
http://store.yahoo.com/lib/toolsf...	The Effects of SRT on Six Different Types of Water - Yahoo	HTTP 200
https://research.yahoo.com/mo...	and Representation-Awareness in Geographically Centra...	HTTP 200
https://research.yahoo.com/mo...	Visual Affect Around the World: A Large-scale ... - Yahoo!...	HTTP 200
https://research.yahoo.com/_c/u...	Serving Ads to "Yahoo Answers" Occasional Visitors - ...	Connect failed!
https://research.yahoo.com/_c/u...	Robust Tree-based Causal Inference for Complex Ad ... - ...	Connect failed!
http://labs.yahoo.com/_c/upload...	VERTa - LREC Conferences	Connect failed!



Preliminaries



- CAPTCHA Types
- Noises
  - Random pixel-level interference added to confuse models
  - Gaussian, Laplace, and Salt-and-Pepper……
  - Serve as a lightweight adversarial defense
  - Each with different visual effects that challenge both models and users

Threat Model

- Capability
  - Access to GPU and open-source deep learning frameworks
  - No prior knowledge about configurations or model tuning skills
- Goal
  - Maximize the prediction accuracy using machine learning models.
- Models
  - **CNN** as a basic single character classifier
    - Fast Inference with low computational cost
  - **VGG-16** as a moderate single character classifier
    - Larger Model Size with Higher Training Cost
    - Higher baseline performance
  - **Tesseract OCR** as a model able for entire string recognition
    - Pretrained and easy to use
    - Confidence scoring support

Goals(For Defense)

- Defense CAPTCHA system prototype
  - Flexible parameter configuration
  - Multiple perturbation support
  - User-friendly(for both the platform and the generated CAPTCHAs)
- Interactive GUI Evaluation Platform
  - Real-time Models Evaluation
  - Confidence Score
  - Batch Model
- Experimental Support
  - Establish clean-data baseline performance for each model
  - Good defense performance on models

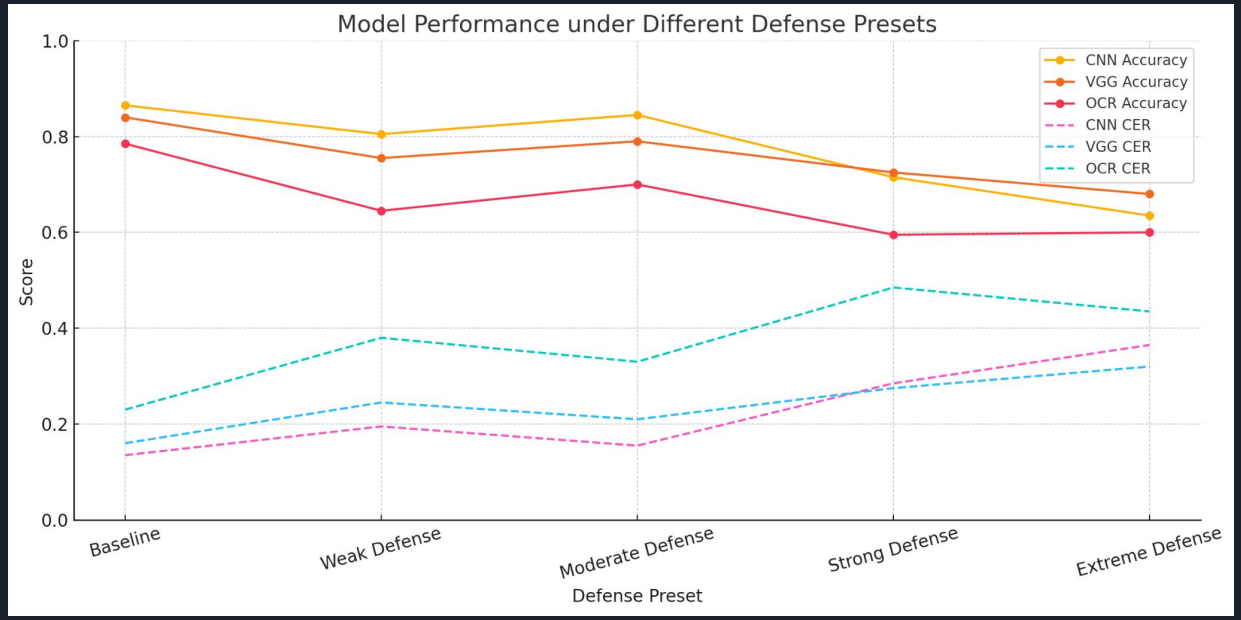
Experiment Results – Settings

- Dataset(0-9 and a-z)
  - Training
    - CNN(CPU) : 60x60x1 、Clean、20% Validation、5000 pieces
    - VGG(Colab T4) : 224x224x3 、Clean、20% Validation、5000 pieces
  - Evaluation (500 pieces)
    - Baseline : Clean
    - Weak : Add small jitters with Gaussian noise
    - Moderate : Add rotation and cutout
    - Strong : Add brightness and contrast
    - Extreme : Add more noises, masks, and compression
- **Tesseract**
  - Used as a pretrained full-string recognizer

Experiment Results – Metrics and Results

- Accuracy(ACC)
  - Correct Predictions / Total Samples
- Character Error Rate(CER)
  - Edit Distance / Word Length
  - Edit Distance : minimum operations to convert a word to the other
- All three models (Char-CNN, VGG16, and Tesseract OCR) demonstrate moderate robustness even under the most aggressive settings.
  - The result concludes that the models are robust enough as a benchmark testing method on the platform.

Result


Model Performance under Different Defense Presets
CNN Accuracy
VGG Accuracy
OCR Accuracy
CNN CER
VGG CER
OCR CER
Score
1.0
0.8
0.6
0.4
0.2
0.0
Baseline
Weak Defense
Moderate Defense
Strong Defense
Extreme Defense
Defense Preset

Other Metrics on the Platform

- Tesseract Confidence Score
  - LSTM-based OCR engine
  - Confidence(Line) = Sum of Confidence(Char) / Word Length
- Structural Similarity Index Measure(SSIM)
  - Evaluates perceived image similarity from human perspective
  - Luminance, Contrast, Structure
- Peak Signal-to-Noise Ratio(PSNR)
  - The difference between the perturbed image and the original one
  - Related to MSE
- Similarity = 1 - Edit Distance / max(1, length))

Demo

- Link

Conclusion

- In our demo, we show that certain generated CAPTCHA images successfully evade all three attack models.
- The platform supports large-scale robustness assessments.
- Among 10 adversarial CAPTCHA images that all models failed to recognize, human participants give a 93.2 score.

Future Work

- Training sequence-level models (e.g., LSTM, CRNN) for end-to-end CAPTCHA recognition
  - LSTM, CRNN...
- Real-world services dataset evaluation
  - Baidu, Google samples
- A general-purpose, extensible evaluation platform
  - Generalize the platform to support diverse CAPTCHA types and integrate with arbitrary OCR or ML models.
- Platform Deployment(in a week....?)

Bonus – Discussions on LLMs

- Does the emergence of powerful LLMs like ChatGPT diminish the need for task-specific ML models?
- We conducted preliminary tests on several LLMs using our CAPTCHA defense platform.
  - Some did fail, which means our platform may be able to defend them.
  - Victims : Claude Sonnet 4, Gemini 2.5 Flash…
- Despite their capabilities, most LLMs still require fine-tuning and task-specific adaptation to perform well on constrained problems.

Takeaways

- A system for private CAPTCHA generation
  - Without access to training samples, even strong ML-based models cannot effectively learn to break the CAPTCHAs.

- A non-technical required GUI platform
  - Even without prior knowledge, users can leverage randomization of fonts, noise, and layout to enhance defense strength.

References

[HW+] Yu-Kai Huang, Tsung-Han Wu, and Wu-Jun Pei. "Defense against Machine Learning based CAPTCHAs Attack". National Taiwan University.

[SN+23] Andrew Searles et al. "An Empirical Study & Evaluation of Modern CAPTCHAs". In: *Proceedings of the 32nd USENIX Security Symposium*. 2023.

[TK+23] N. Tariq et al. "CAPTCHA Types and Breaking Techniques: Design Issues, Challenges, and Future Research Directions". In: *arxiv* (2023).