The Basics of CAPTCHA's Security Showdown

CAPTCHA Harvesting

PART 1

GeeTest

www.geetest.com

101

Bot Management ebook

1. Introduction

CAPTCHA cracking is a constant challenge. Protection invites attacks; it's the very reason for its existence. GeeTest, as a CAPTCHA provider, has witnessed countless attempts to bypass our defences—some successful, others not. This ebook marks the initiation of our **Bot Management 101 series.** Here, we dive into fundamental CAPTCHA cracking techniques, offering a comprehensive perspective that reveals the attack process through the eyes of bot operators, targeted websites, and GeeTest.

Our current focus centres on **CAPTCHA harvesting**—an intricate practice that involves a specialized form of CAPTCHA bypass. This technique revolves around gathering CAPTCHA images from source websites or mobile apps. The objective is to construct a repository enabling automated bypassing of CAPTCHA challenges through scripts or algorithms trained on these acquired images. In a digital era where CAPTCHAs act as the first line of defence against bots, comprehending and countering CAPTCHA harvesting has taken on the utmost importance for your business security.

2. The Popularity of CAPTCHA Harvesting Among Bad Actors

CAPTCHA harvesting is a specific type of CAPTCHA bypass, focused on collecting CAPTCHA images from the source website or mobile app to construct a repository to bypass the CAPTCHA challenge automatically by using scripts or algorithms trained on the collected images.

Given that the effectiveness of most CAPTCHAs heavily relies on the user's interpretation of images, bad actors believe that the most efficient approach to overcoming CAPTCHA hurdles is by amassing CAPTCHA images. This strategy aims to achieve automated bypassing.

The impetus for our exploration into CAPTCHA harvesting stems from its widespread utilization as a favoured attack technique among bot operators. Data from GeeTest Lab highlights that over **68%** of bot-related attacks can be attributed to CAPTCHA harvesting.

2.1 Boosting Bad Actors' Productivity through CAPTCHA Harvesting

To understand the attractiveness of CAPTCHA harvesting to bot operators, let's begin by examining the concept from an ROI (rate of investment) perspective.

The ultimate goal of bot operators is profit generation, achieved by exploiting advantages over regular users and subsequently capitalizing on these gains. The ability of bot operators to turn a profit hinges on their "**productivity**" in acquiring online resources, which needs to substantially surpass that of typical users.

GeeTest's data indicates that most bot operators can access online resources at least **800 times** more efficiently than regular users. Their foundational strength lies in this heightened "efficiency," making it imperative to curb this aspect. By constraining bot operators' efficiency to align with that of regular users, their capacity to indiscriminately amass resources for profit diminishes. Consequently, the key in bot attacks and defences revolves around the notions of "efficiency" and "profit-making potential."

This prompts the question of **how CAPTCHAs can limit bot efficiency without disrupting regular user access**.

2.2 Unveiling the Process of CAPTCHA Harvesting

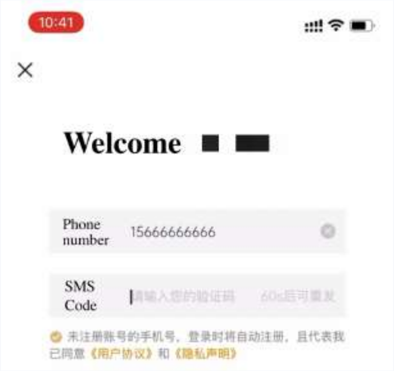
Similarly, bot operators are strategizing ways to augment their "efficiency" within imposed constraints, and this is where the two parties of CAPTCHA harvesting intersects.

CAPTCHA images play a pivotal role for both attackers and defenders. These images carry information that can distinguish humans from bots. Given that CAPTCHAs incorporate visual elements and bot operators rely on image-based solutions for verification, CAPTCHA harvesting remains the critical point in this ongoing conflict.

Let's delve into the step-by-step process of how cybercriminals execute CAPTCHA harvesting (under the condition that the CAPTCHA image resource remains unaltered).

Step 1

Launch an attack on the SMS login session.



Welcome [redacted] [redacted]

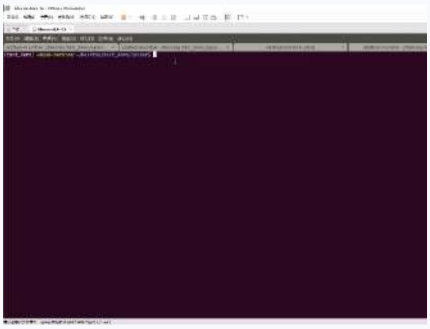
Phone number 15666666666

SMS Code 请输入您的验证码 60s后可重发

未注册账号的手机号，登录时将自动注册，且代表我已同意《用户协议》和《隐私声明》

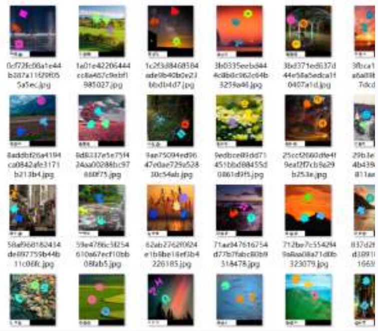
Step 2

Obtain the image address by sending frequent requests to the page interface.



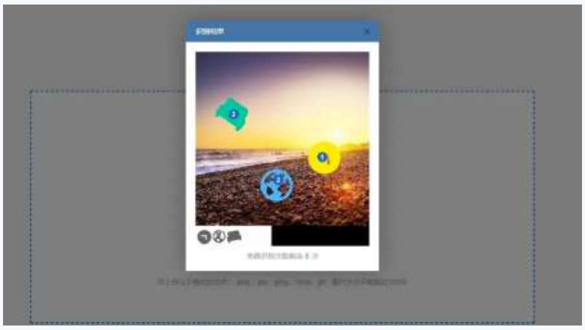
Step 3

Acquire the addresses of 300,000 CAPTCHA images, download and store them at a rate of 10 images per second, **taking a total of 8.33 hours.**


0cf72fc08a1e44b387a11f29f05 5a5ec.jpg
1a01e42206444cc8a467c9ebf1 985027.jpg
1c2f3d8468584ade9b40b0e23 bbdb4d7.jpg
3b0335eebd44 4c8b8c962c64b 3259a46.jpg
3bd371ed637d 44e58a5edca1f 0407a1d.jpg
3fbca1 a6a88 7dcd
8addbf26a4194 ca0842afe3171 b213b4.jpg
8d8337e5e75f4 24aa00288bc97 880f75.jpg
9ae75094ed96 47e0ae729a528 30c54ab.jpg
9edbce89dd71 451bbd88455d 0861d9f5.jpg
25ccf2660dfe4f 9eaf2f7cb9a29 b253e.jpg
29b3e 4b439 811ae
58af968182434 de897759b44b 11c06fc.jpg
59e4786c3f254 610a67ecf10bb 08fab5.jpg
62ab27620624 e1b8be18ef3b4 226185.jpg
71ae94761675 4d77b7fabc80b9 318478.jpg
712be7c55428 49a8aa08a71d0b 323079.jpg
837d2 d3891 1663

Step 4

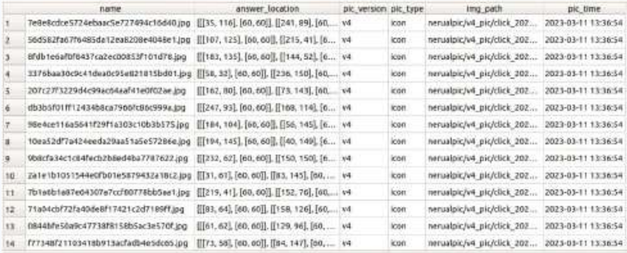
Access answers via manual solving through CAPTCHA farms, costing approximately $0.0019 for every 2.5 seconds per image, **totalling 208.33 hours and amounting to $582.6.**



Answers returned from CAPTCHA farms

Step 5

Create a database of CAPTCHA answers, including image names, answer coordinates, versions, types, storage addresses, and download times.



	name	answer_location	pic_version	pic_type	img_path	pic_time
1	7e8e8cdce5724ebaac5e727494c16d40.jpg	[[[35, 116], [60, 60]], [[241, 89], [60,...	v4	icon	nerualpic/v4_pic/click_202...	2023-03-11 13:36:54
2	56d582fa67f6485da12ea8208e4048e1.jpg	[[[107, 125], [60, 60]], [[215, 41], [6...	v4	icon	nerualpic/v4_pic/click_202...	2023-03-11 13:36:54
3	8fdb1e6afbf6437ca2ec00853f101d78.jpg	[[[183, 135], [60, 60]], [[144, 52], [6...	v4	icon	nerualpic/v4_pic/click_202...	2023-03-11 13:36:54
4	3376baa30c9c41dea0c95e821815bd01.jpg	[[[58, 32], [60, 60]], [[236, 150], [60,...	v4	icon	nerualpic/v4_pic/click_202...	2023-03-11 13:36:54
5	207c27f3229d4c99ac64aaf41e0f02ae.jpg	[[[162, 80], [60, 60]], [[73, 143], [60,...	v4	icon	nerualpic/v4_pic/click_202...	2023-03-11 13:36:54
6	db3b5f01ff12434b8ca7960fc86c999a.jpg	[[[247, 93], [60, 60]], [[168, 114], [6...	v4	icon	nerualpic/v4_pic/click_202...	2023-03-11 13:36:54
7	98e4ce116a5641f29f1a303c10b3b575.jpg	[[[184, 104], [60, 60]], [[56, 145], [6...	v4	icon	nerualpic/v4_pic/click_202...	2023-03-11 13:36:54
8	10ea52df7a424eeda29aa51a5e57286e.jpg	[[[194, 145], [60, 60]], [[40, 149], [6...	v4	icon	nerualpic/v4_pic/click_202...	2023-03-11 13:36:54
9	9b8cfa34c1c84fecb2b8ed4ba7787622.jpg	[[[232, 62], [60, 60]], [[150, 150], [6...	v4	icon	nerualpic/v4_pic/click_202...	2023-03-11 13:36:54
10	2a1e1b1051544e0fb01e5879432a18c2.jpg	[[[31, 61], [60, 60]], [[83, 145], [60, ...	v4	icon	nerualpic/v4_pic/click_202...	2023-03-11 13:36:54
11	7b1a6b1a87e04307e7ccf80778bb5ee1.jpg	[[[219, 41], [60, 60]], [[152, 76], [60,...	v4	icon	nerualpic/v4_pic/click_202...	2023-03-11 13:36:54
12	71a04cbf72fa40de8f17421c2d7189ff.jpg	[[[83, 64], [60, 60]], [[158, 126], [60,...	v4	icon	nerualpic/v4_pic/click_202...	2023-03-11 13:36:54
13	0844bfe50a9c47738f8158b5ac3e570f.jpg	[[[61, 62], [60, 60]], [[129, 96], [60, ...	v4	icon	nerualpic/v4_pic/click_202...	2023-03-11 13:36:54
14	f77348f21103418b913acfadb4e5dc65.jpg	[[[73, 58], [60, 60]], [[84, 147], [60, ...	v4	icon	nerualpic/v4_pic/click_202...	2023-03-11 13:36:54

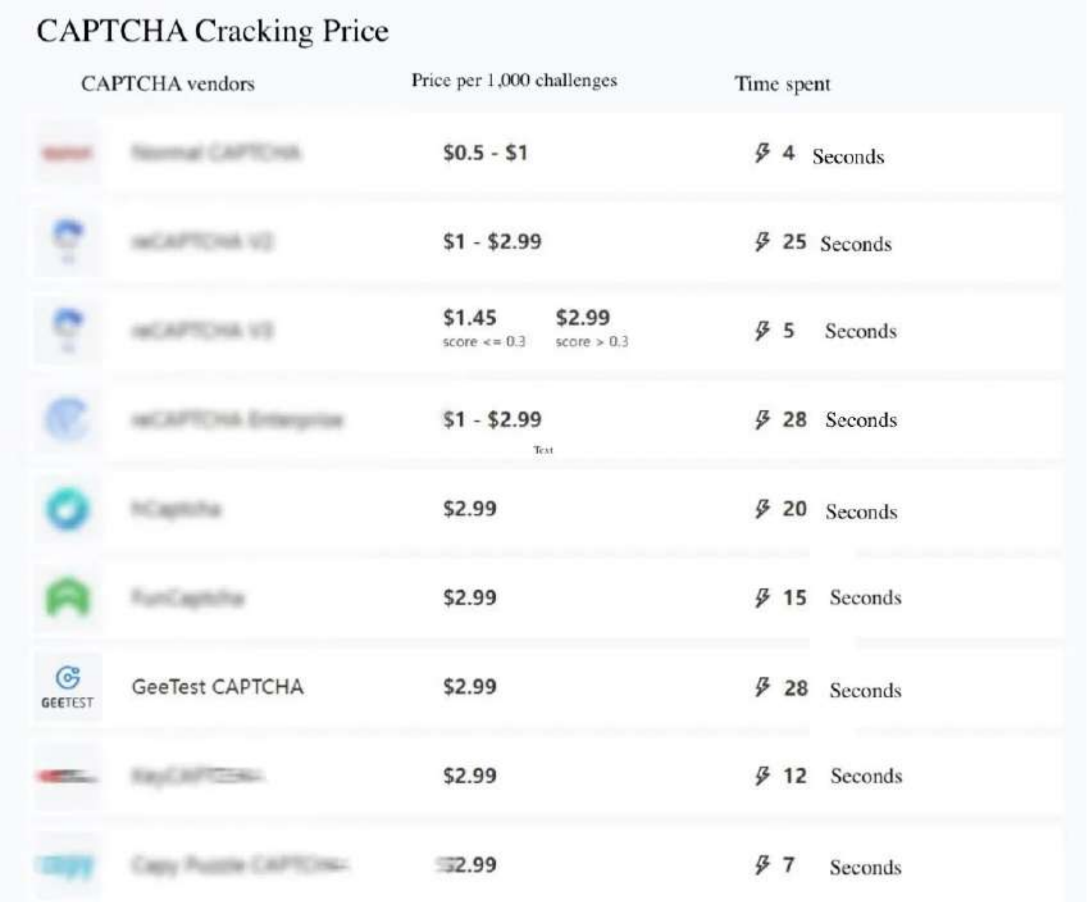
A database of CAPTCHA answers

Step 6

Match CAPTCHA images with their answers, complete the CAPTCHA challenge, and target the SMS login session.

Although the entire CAPTCHA harvesting process is time-intensive, demanding an upfront investment of time and resources from bot operators. Once completed, it allows for easy and quick bypassing of CAPTCHA. From the bot operators' perspective, CAPTCHA harvesting entails a relatively minor early-stage investment for a later, comprehensive program. This method is a frequently used strategy among fraudsters.

2.3 How Attackers Profit from CAPTCHA Harvesting



CAPTCHA Cracking Price

CAPTCHA vendors	Price per 1,000 challenges	Time spent
Normal CAPTCHA	$0.5 - $1	4 Seconds
reCAPTCHA V2	$1 - $2.99	25 Seconds
reCAPTCHA V3	$1.45 score <= 0.3 $2.99 score > 0.3	5 Seconds
reCAPTCHA Enterprise	$1 - $2.99 Text	28 Seconds
hCaptcha	$2.99	20 Seconds
FunCaptcha	$2.99	15 Seconds
GEETEST GeeTest CAPTCHA	$2.99	28 Seconds
KeyCAPTCHA	$2.99	12 Seconds
Capy Puzzle CAPTCHA	$2.99	7 Seconds

Profit = income - costs. A simple equation that defines the goal of any enterprise. Most CAPTCHA vendors update their image set once every month, usually containing a few hundred to a few thousand images. However, bot operators opt for low-cost CAPTCHA farms, where the cost can be as low as $0.0019 per manual solving CAPTCHA. To illustrate, let's consider a scenario where a CAPTCHA image set receives 6,000 monthly updates. In this case, bot operators need to invest less than $1,390 to compile a comprehensive database of answers. Once fraudsters have amassed all the CAPTCHA answers, they can rapidly outsmart this batch of CAPTCHAs for the following month or even longer. Even if the vendors refresh their image set after a month, the cost for fraudsters to purchase CAPTCHA farms remains below $0.042 per manual solving CAPTCHA.

Hence, fraudsters need to invest only about $65 daily in CAPTCHA farms. In just 9 days, they can acquire the targeted website's image set of 300,000 images and construct their own answer database.

Furthermore, during the following month when the image set remains unchanged, the bots become invincible. Within this period, the earnings fraudsters generate from activities like ticket scalping and other forms of bot-related fraud would exceed their initial costs. This is precisely why they continuously resort to CAPTCHA harvesting.

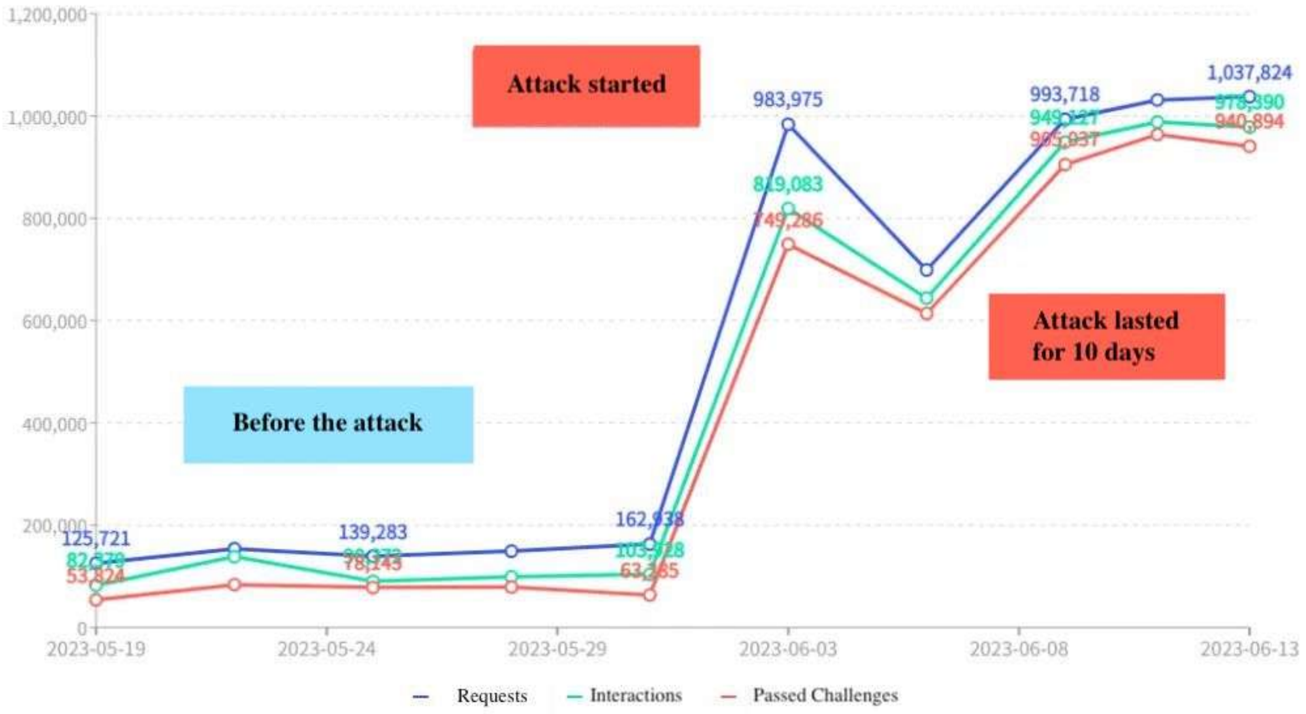
3. Experiencing the Impact of CAPTCHA Harvesting

3.1 Detection of Attack

Company A, an e-commerce startup with a two-year history, faces recurring issues of malicious SMS activity during user login sessions. In response, the company's security team implemented CAPTCHA before an annual sales event to thwart bot-driven attacks. For the ensuing two weeks, sales continued as usual, and CAPTCHA worked well in fending off bots.

However, an anomaly emerged one night. CAPTCHA requests, interactions, and passed challenges showed a sudden surge. Strangely, the CAPTCHA seemed completely useless. An overwhelming volume of SMS messages inundated the system, draining SMS resources. **At its peak, message consumption skyrocketed to over $2,800 per hour.**

Alerted to the situation, the company engaged its CAPTCHA vendor. The vendor's response involved altering a portion of the CAPTCHA images and introducing more complex challenges. Initially, there was a slight decline in passed CAPTCHA challenges. Nevertheless, within a few days, the metrics spiked once more. By the tenth day of the attack, the volume remained abnormally high.



CAPTCHA requests, interactions, and passed challenges before and during the attack

3.2 Analyzing the Attack

Upon encountering the issue, Company A's security team reached out to GeeTest for help. Shortly, GeeTest's security experts pinpointed that the abrupt irregularities in data were a direct result of CAPTCHA harvesting.

Reason 1 CAPTCHA harvesting proves attractive to attackers due to its low-cost, high-profit nature. The answer database can be reused over an extended period.

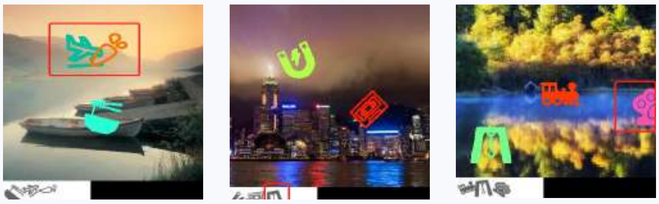
Reason 2 GeeTest's analysis, comparing data before and after the image set update, revealed a drop in passed CAPTCHA challenge rates following the application of new images. This pattern clarified the momentary invalidation of the previous answer database.

Reason 3 Furthermore, GeeTest identified a striking similarity in the coordinates sent by different clients for the same image. This uniformity contradicts the behaviour of human users, further supporting the conclusion that automated interaction was at play.

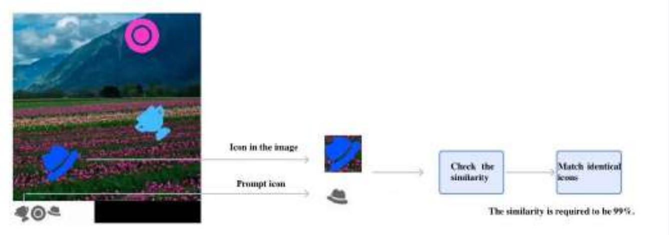
3.3 Overcoming Technological Challenges

Given the swift identification of the attack, why did Company A allow it to persist for 10 days? The answer lies in the complexities of increasing the update frequency of CAPTCHA image set, a daunting task for vendors. Three significant challenges emerge:

1. **Ensuring image uniqueness** demands efficient generation and approval algorithms. Failure to do so results in errors like:



2. **Attaining 99%+ accurate coordinates for icons in images is essential.** Each icon must be meticulously extracted from images for comparison against the originals.



3. After image generation and review, images and metadata are uploaded to global static resource servers and server-side servers to ensure the images are available at each server.

Beyond these technical obstacles, the update frequency of image set significantly impacts CAPTCHA's effectiveness. As outlined in the earlier section, the frequency of image updates directly influences attackers' profits. When their earnings fall below their costs, fraudsters are likely to abandon their attacks.

The pivotal question is: at what frequency of image updates do attackers' profits dip below their costs?

Let's break it down. The cost equation is simple:

Cost = Image set size × CAPTCHA farm cost per image

Assuming a $140 profit and a $0.0019 cost per image from the CAPTCHA farm, we can deduce the image database size:

$140 ÷ $0.0019 = 73,684 images

This means attackers need to tackle these 73,684 CAPTCHA images before the website updates its CAPTCHA image set.

Given an average solving time of 10 seconds per image, calculating the time to update the image database yields:

Image set size × Average solving time per image = 73,684 × 10 = 736,840 seconds = 205 hours

In conclusion, with an attacker's profit set at $140 and an average solving time of 10 seconds per image, each database update should encompass at least 73,684 images. **To achieve a balance for attackers, the database should be refreshed approximately every 205 hours (about 8 days).**

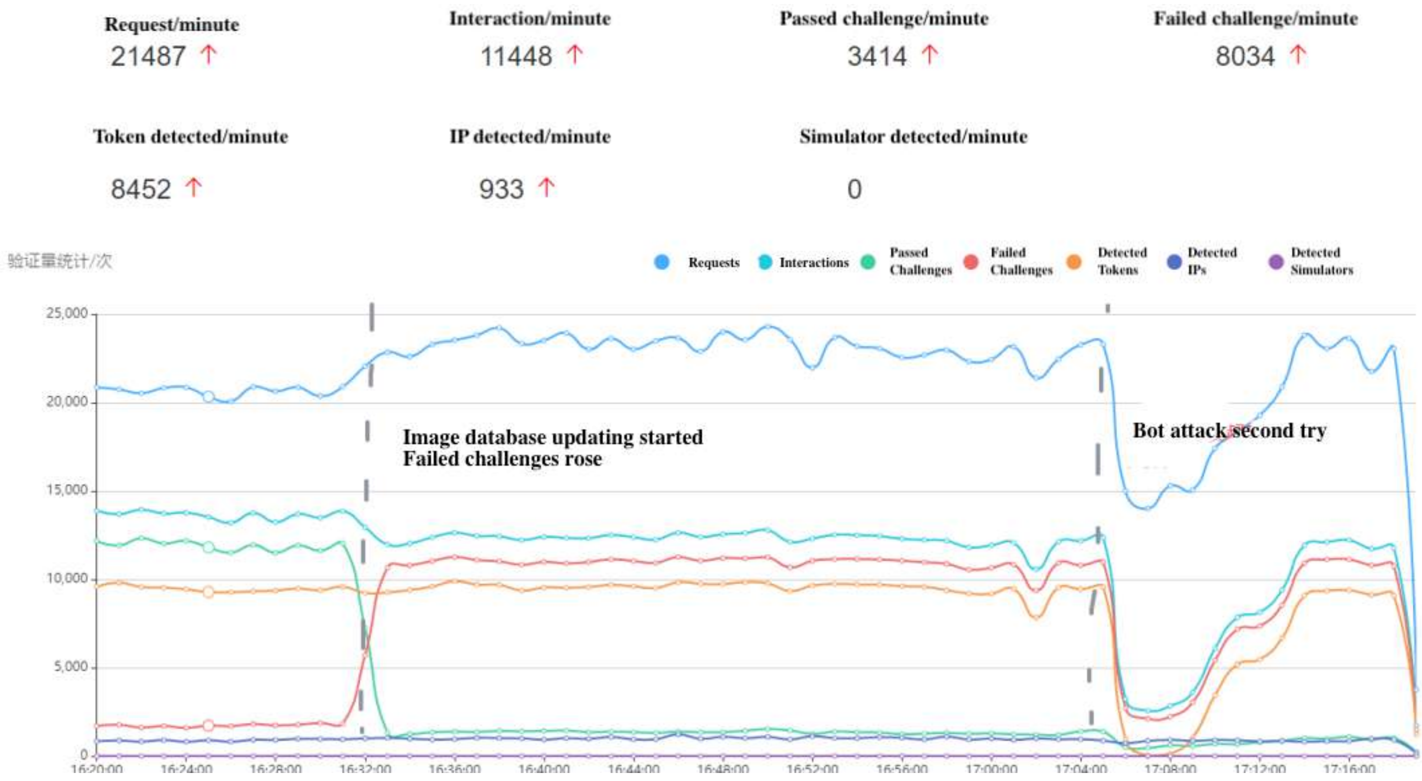
However, the challenge lies in the image set update process, which usually takes time. Our research indicates that the average update frequency of CAPTCHA vendors' image databases is roughly once a month, far slower than the attacker's harvesting speed (once every 8 days). Moreover, the updated image set's average size is a mere few hundred to a few thousand images. Bot operators can effortlessly overcome this with minimal expenditure and bypass it within an hour.

This clarifies why the attack on Company A experienced a 3-day slowdown before resurging.

4. GeeTest's Approach to Tackling CAPTCHA Harvesting

4.1 A Swift and Effective Defense

When GeeTest stepped in, Company A noticed a significant flattening of the attack curve.



Request/minute 21487 ↑

Interaction/minute 11448 ↑

Passed challenge/minute 3414 ↑

Failed challenge/minute 8034 ↑

Token detected/minute 8452 ↑

IP detected/minute 933 ↑

Simulator detected/minute 0

The attack curve.

On that afternoon, GeeTest swiftly updated the image set, commencing at 16:31. Consequently, the count of failed challenges surged, indicating the attackers' image answer database had been rendered useless. Despite minimal alteration in CAPTCHA requests, the attack persisted.

At 17:06, the bot operators caught on and halted the assault.

At 17:10, the attack resumed, only to face continued failure against the CAPTCHA challenges. The curve depicting passed CAPTCHA challenges stabilized, and the numbers returned to their normal range.

In comparison to the prior CAPTCHA vendor, **who updated monthly,** GeeTest's **hourly updates** shielded Company A from potential losses amounting to tens of thousands of dollars.

One month versus one hour. It’s not just a numerical discrepancy, but holds the key to whether your CAPTCHA can effectively deter bots engaged in CAPTCHA harvesting.

4.2 Security Strategies: A Proactive Approach

With over a decade of experience, GeeTest has recognized that preventing CAPTCHA harvesting yields multiple benefits, including reduced costs for clients and an enhanced user experience. This proactive stance empowers clients to lead the fight against bots.

The core of success in countering CAPTCHA harvesting lies in the frequency of image set updates and its size.

Typically, most CAPTCHA image set undergo monthly updates, containing hundreds to thousands of images. These updates are often in hindsight, occurring after attacks have happened. However, GeeTest's Adaptive CAPTCHA employs an automated image set updating system, generating over 300,000 new images per hour. This continuous refreshment of CAPTCHA challenges defeats attackers by rendering their image answer databases obsolete.

Typical Updating Frequency	GeeTest Updating Frequency
Once per **month**	Once per **hour**
Update **10,000+** images	Update **300,000+** images

GeeTest's system excels in tailoring strategies to fit different timings and attack scenarios. It achieves this by swiftly generating more than 50,000 fresh images across 200 categories within a matter of minutes, subsequently distributing them to global servers. For continuous attacks, GeeTest CAPTCHA employs an update speed of 10,000 images spanning 50 categories every 10 minutes.

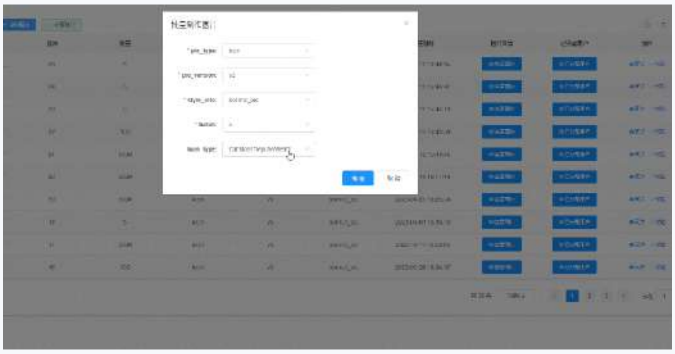
This approach significantly escalates the expenses incurred by fraudsters. The new image-answer database becomes obsolete within an hour, compelling them to pay $0.0019 per image for CAPTCHA farm manual solutions. Consequently, this approach disrupts the equilibrium of bot operators, effectively halting CAPTCHA harvesting.

4.3 Core Technology

In achieving the feat of updating 300,000 images hourly, we surmounted significant technical hurdles:

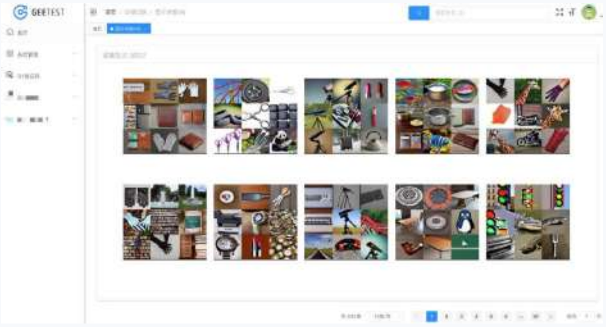
1. Efficient Image Generation and Validation

GeeTest pioneered the development of CAPTCHA image generation and validation algorithms. This novel approach, rare in the industry, enables the efficient and frequent refresh of CAPTCHA images.



GeeTest Automated Image Updating System

Our team has spearheaded the integration and application of Artificial Intelligence for Graph Computing (AIGC) technology in CAPTCHA generation. We've constructed a suite of graph-related APIs utilizing ray.serve and stable-diffusion frameworks. These APIs manage prompts and automate image generation. Leveraging AIGC accelerates image updates, ensuring unwavering precision, control, and scalability in the process.



AIGC in Image Generation

2. Ensuring Consistency Across Global Servers

The process of image generation involves a step where binary images are uploaded to the Object Storage Service (OSS) and corresponding metadata is stored in the PostgreSQL (PG) database. This process guarantees that images are only added to the database after they've been successfully uploaded as resources. The database file then acquires the path to access the image, ensuring its availability on the static resource server.

For synchronization, the xxl-job tasks are designed to be idempotent, allowing them to be performed repeatedly without causing issues. In case synchronization encounters any hitches, manual triggers are available as a backup. Even if the xxl-job service experiences downtime, the synchronization continues to operate effectively.

In the course of generating database (DB) files for images, the xxl-job system verifies synchronization to confirm that images are consistently available across all global servers. Only after this synchronization is validated does the system proceed with generating and distributing the DB files.

Scheduled updates and manual interventions are both managed through the dashboard. If scheduled tasks fail to update, the manual operation takes over to ensure the generation and synchronization of images are not disrupted.

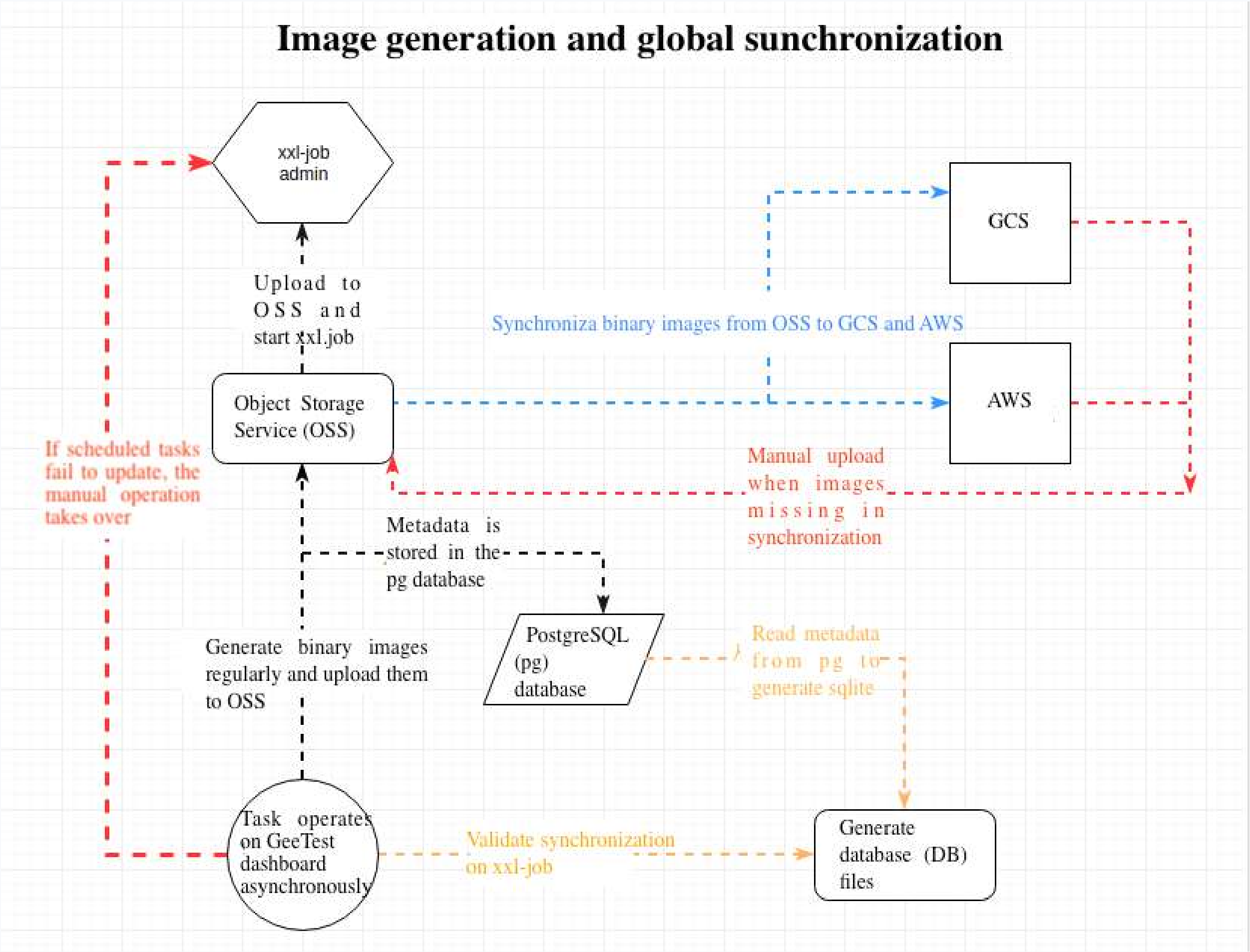


Image generation and global sunchronization
xxl-job admin
GCS
Upload to OSS and start xxl.job
Synchroniza binary images from OSS to GCS and AWS
AWS
Object Storage Service (OSS)
If scheduled tasks fail to update, the manual operation takes over
Manual upload when images missing in synchronization
Metadata is stored in the pg database
PostgreSQL (pg) database
Read metadata from pg to generate sqlite
Generate binary images regularly and upload them to OSS
Task operates on GeeTest dashboard asynchronously
Validate synchronization on xxl-job
Generate database (DB) files

Ensuring Consistency Across Global Servers

3. Managing Image Usage, Loading, and Resource Conflicts

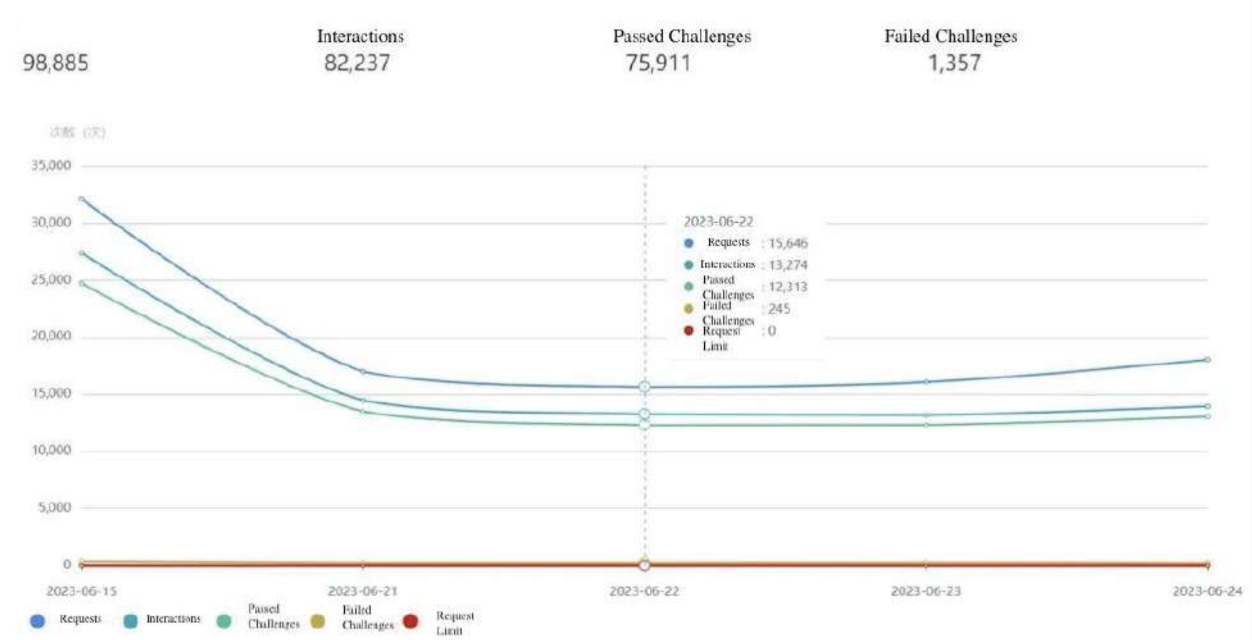
Once we create and update images, they are quickly synchronized across global servers.

For image loading, we've implemented techniques like cache pre-warming and global static service with multiple servers. This ensures that images load swiftly from nearby nodes, improving clients' experience.

Image details are stored in a built-in database, allowing for a rapid image-loading process with quick responses. The decentralized nature of this database effectively handles situations with many users simultaneously.

We employ an Inode detection mechanism for image metadata, which aids in rapidly replacing resources within milliseconds, reducing resource conflicts.

On that afternoon, after GeeTest Adaptive CAPTCHA was deployed and the image database was updated, the curves indicating CAPTCHA requests, interactions, passed challenges, and failed challenges stabilized in Company A. This data returned to normal, and Company A's mid-year sales proceeded as anticipated.

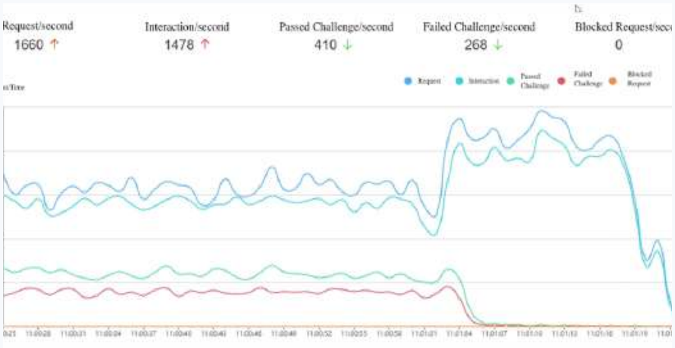


Normal CAPTCHA Curves of Company A

4.4 Enhanced Value of GeeTest CAPTCHA

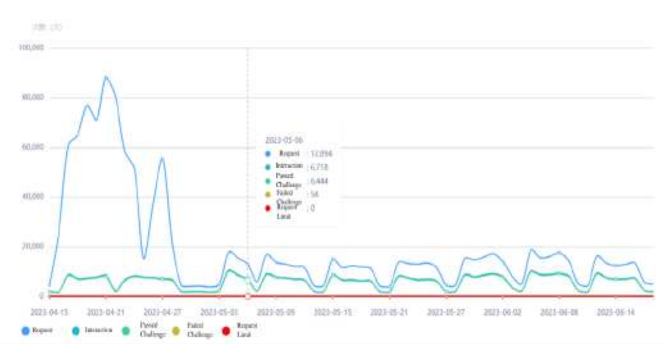
GeeTest has pioneered the industry with its efficient and accurate image generation through the GeeTest Adaptive CAPTCHA. By incorporating a self-checking module, each image produced adheres to preset standards. With the backing of a stable internal platform, we can globally update images for individual clients or the entire network within seconds. This has translated into several key benefits for our clients:

1. **Clear Defense Visualization** Traditional security tools often struggle to demonstrate their effectiveness due to stable data trends before attacks. GeeTest Adaptive CAPTCHA, showcased through our dashboard, vividly illustrates the actions of malicious actors and their subsequent deterrence. This immediate contrast highlights the efficacy of our strategies against bot operators, offering clients a tangible display of our achievements.



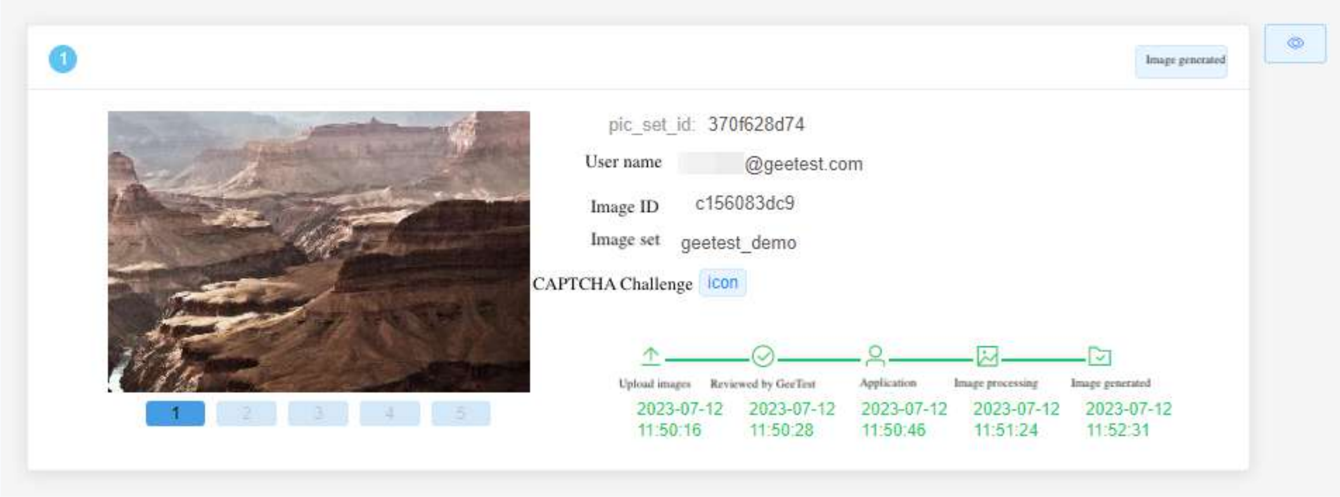
Real-time Data Display

2. **Amplified Costs for Malicious Actors** Conventional approaches involve updating CAPTCHA images only after compromises occur, a reactive stance. However, GeeTest's dynamic image database updates proactively limit CAPTCHA harvesting, giving us the upper hand. When malicious actors realize that the costs outweigh their gains, the incentive to persist with attacks dwindles.



After the second try, bad actors stopped the attack

3. **Tailored Service Advancements:** GeeTest caters to clients with diverse security needs, providing an extensive range of customizable solutions. For multinational enterprises, we offer personalized data processing, UI design, integrated risk management, multilingual support, audio verification, and regulatory compliance. In urgent scenarios, our APIs generate tailored image set promptly, delivering tailored solutions.



pic_set_id: 370f628d74
User name [redacted]@geetest.com
Image ID c156083dc9
Image set geetest_demo
CAPTCHA Challenge icon

Upload images 2023-07-12 11:50:16
Reviewed by GeeTest 2023-07-12 11:50:28
Application 2023-07-12 11:50:46
Image processing 2023-07-12 11:51:24
Image generated 2023-07-12 11:52:31

5. Closing Thoughts

The ongoing battle between CAPTCHA providers and malicious actors is akin to a duel between a spear and a shield — an evergoing race. To maintain innovation and a competitive edge, CAPTCHA providers must continuously refine their strategies in this ongoing conflict.

In the context of CAPTCHA harvesting, we're intensifying our efforts to enhance image update efficiency. By 2024, our goal is to optimize operational and service platforms. This will allow for localized image generation and global node updates, eliminating the need for network synchronization. This approach guards against update failures caused by network issues, ensuring consistent and stable services.

Book a Demo with GeeTest

About

GeeTest, a leading provider of cutting-edge bot management solutions founded in 2012, is dedicated to protecting businesses and users from emerging cyber threats and financial losses.

GeeTest has occupied the Top1 market share in APAC and services with over 360,000 enterprises worldwide presently including Airbnb, Nike, Imperva, etc. Most notably, GeeTest has achieved comprehensive coverage in the blockchain industry with over 20% of the Top 50 crypto exchanges that chose GeeTest to fight fraud attacks, including BINANCE, Axie Infinity, Poloniex, crypto.com, etc. In November 2021 GeeTest was recognized as a selected vendor in Forrester's Now Tech: Bot Management, Q4 2021. In July 2023, GeeTest CAPTCHA was recognized as impactful DDoS Protection software for businesses by Capterra.

Trusted by 360,000 domains worldwide

Website：www.geetest.com

Phone：400-8521-816

Email：globalmarketing@geetest.com