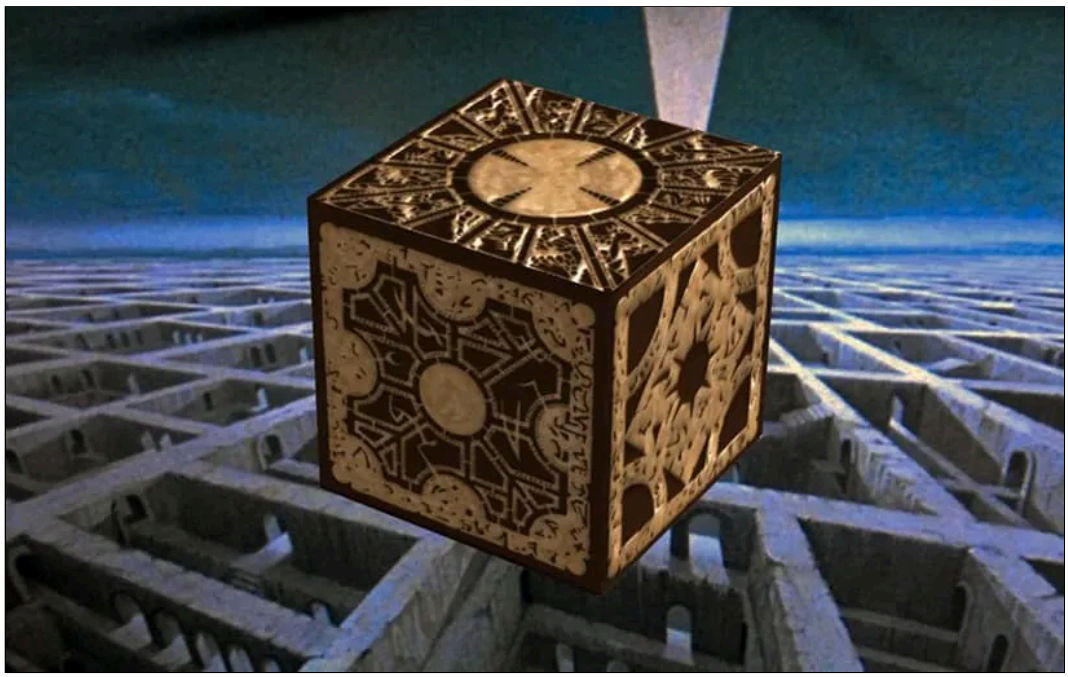
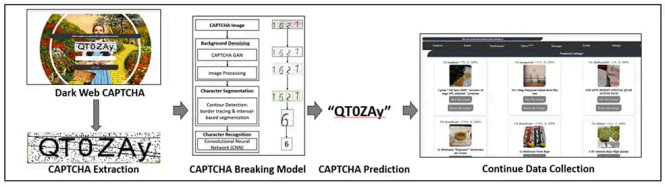
Solving CAPTCHAs With Machine Learning to Enable Dark Web Research

Originally published at https://www.unite.ai/solving-captchas-with-machine-learning-to-enable-dark-web-research/



A joint academic research project from the United States has developed a method to foil CAPTCHA* tests, reportedly outperforming similar state-of-the-art machine learning solutions by using Generative Adversarial Networks (GANs) to decode the visually complex challenges.

Testing the new system against the best current frameworks, the researchers found that their method achieves more than 94.4% success on a carefully curated real-world benchmark dataset, and has proved capable of 'eliminating human involvement' when navigating a highly CAPTCHA-protected emerging Dark Net Marketplace, automatically resolving CAPTCHA challenges in a maximum of three attempts.



Workflow for DW-GAN. Source: https://arxiv.org/pdf/2201.02799.pdf

The authors contend that their approach represents a breakthrough for cybersecurity researchers, who traditionally have had to bear the costs of supplying humans-in-the-loop to manually solve CAPTCHAs, usually via crowdsourcing platforms such as Amazon Mechanical Turk (AMT).

If the system can prove adaptable and resilient, it may further pave the way for more automated oversight systems, and for the indexing and web-scraping of TOR networks. This could enable scalable and high-volume analyses, as well as the development of new cybersecurity approaches and techniques, which have been hamstrung, to date, by CAPTCHA firewalls.

The paper is titled *Counteracting Dark Web Text-Based CAPTCHA with Generative Adversarial Learning for Proactive Cyber Threat Intelligence*, and comes from researchers at the University of Arizona, the University of South Florida, and the University of Georgia.

Implications

Since the system – called Dark Web-GAN (DW-GAN, available at GitHub) – is apparently so much more performative than its predecessors, there is the possibility that it will be used as a general method to overcome the (usually less difficult) CAPTCHA material on the standard web, either in this specific implementation, or based on the general principles that the new paper outlines. Due to limited storage at GitHub, however, it is currently necessary to contact the lead author Ning Zhang in order to obtain the data associated with the framework.

Because DW-GAN has a 'positive' mission for breaking CAPTCHAs (much as TOR itself originally had a positive mission for protecting military communications and, later, journalists), and because CAPTCHAs are both a legitimate defense (frequently and controversially used by ubiquitous CDN giant CloudFlare) and a favorite tool of illegitimate dark web marketplaces, the approach is arguably a 'leveling' technology.

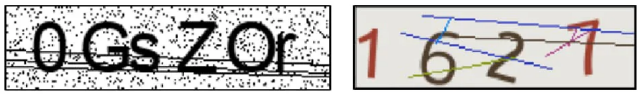
The authors themselves concede that DW-GAN has wider uses:

> *'[While] this study is mainly focused on dark-web CAPTCHA as a more challenging problem, the proposed method in this study is expected to be applicable to other types of CAPTCHA without loss of generality.'*

Presumably DW-GAN, or a similar system, would need to become widely and evidently diffused in order to prompt dark web markets to seek less machine-resolvable solutions, or at least to evolve their CAPTCHA configurations periodically, a 'cold war' scenario.

Motivations

As the paper observes, the dark web is the primary font of hacker intelligence relating to cyber attacks, which are estimated to cost the global economy $10 trillion USD by 2025. Therefore onion networks remain a relatively safe environment for illicit dark net communities, which can repel boarders by various methods, including session timeouts, cookies, and user authentication.



Two types of CAPTCHA, both using obfuscating backgrounds and tilted lettering to make them less machine-readable.

However, the authors observe, none of these obstacles are so great as the tranche of CAPTCHAs that punctuate the browsing experience in a 'sensitive' community:

> *'While most of these measures can be effectively circumvented through implementing automated counter measures in a crawler program, CAPTCHA is the most hampering anti-crawling measure in the dark web that cannot be easily circumvented due to high cognitive capabilities that are often not possessed by automation tools'*

Text-based CAPTCHAs are not the only available option; there are variants, familiar to many of us, that challenge the user to interpret video, audio, and especially images. Nonetheless, as the authors observe, text-based CAPTCHA is currently the challenge of choice for dark web markets, and a natural starting-place to make TOR networks more susceptible to machine analysis.

Architecture

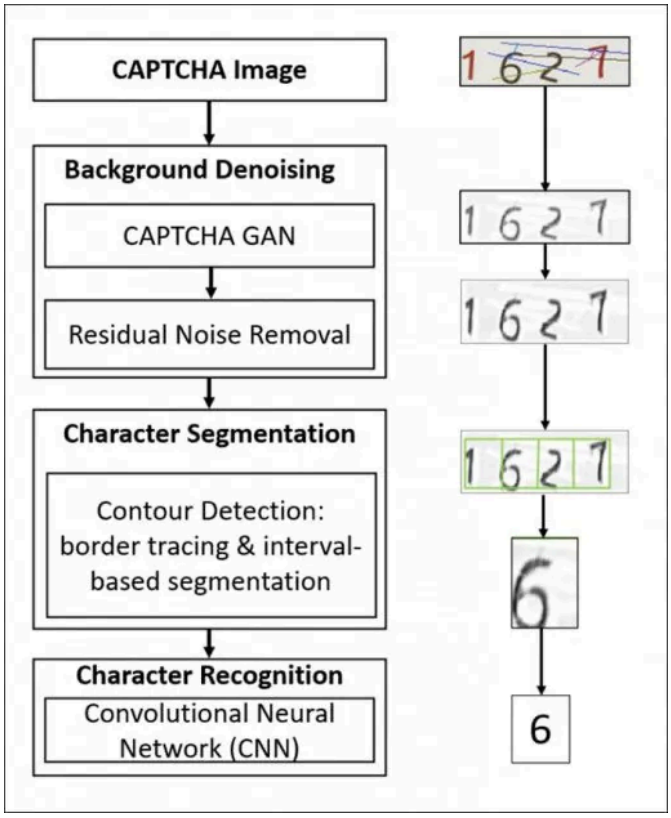
Though a prior approach from Northwest University in China used Generative Adversarial Networks to derive feature patterns from CAPTCHA platforms, the authors of the new paper note that this method relies on interpretation of a rasterized image, rather than a deeper examination of letters recognized in the challenge; and that DW-GAN's effectiveness is not impacted by the variable length of nonsense words (and of numbers) that are typically found in dark web CAPTCHAs.

DW-GAN uses a four-stage pipeline: first the image is captured, and then fed to a background denoising module which uses a GAN that has been trained on annotated CAPTCHA samples, and is therefore able to distinguish letters from the perturbed background that they are resting on. The extracted letters are then further filtered out from any remaining noise after the GAN-based extraction.

Next, segmentation is performed on the extracted text, which is then broken down into what appear to be constituent characters, using contour detection algorithms.

Character segmentation isolates the pixel group and attempts recognition with border tracing.

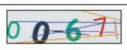
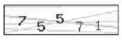
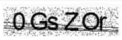
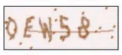
Finally, the 'guessed' character segments are subject to character recognition via a Convolutional Neural Network (CNN).



Sometimes characters can overlap, a hyper-kerning that's specifically designed to fool machine systems. DW-GAN therefore uses interval-based segmentation to enhance and isolate borders, effectively separating characters. Since the words are usually nonsense, there is no semantic context to aid in this process.

Results

DW-GAN was tested against CAPTCHA images from three diverse dark web datasets, as well as a popular CAPTCHA synthesizer. The dark markets from which the images originated comprised two carding shops, Rescator-1 and Rescator-2, and a novel set from a then-emerging market called Yellow Brick (which was reported to have later disappeared in the wake of the takedown of DarkMarket).

Category	Source	Amount	Character Length	Character Type	Sample
Dark Web CAPTCHA	Rescator-1	500	4	Digit	
	Rescator-2	500	5	Digit	
	Yellow Brick	500	6	Digit+Letter	
Open-Source Synthesizer		11,000 for each character length	4, 5, 6, 7	Digit+Letter	

Sample CAPTCHAs from the three datasets, as well as the open source CAPTCHA synthesizer.

According to the authors, the data used in testing was recommended by Cyber Threat Intelligence (CTI) experts based on their wide diffusion across dark net markets.

Testing each dataset involved the development of a TOR-facing spider tasked with collecting 500 CAPTCHA images, which were subsequently labeled and curated by CTI advisors.

Three experiments were devised. The first evaluated the general CAPTCHA-defeating performance of DW-GAN against standard SOTA methods. The rival methods were image-level CNN with preprocessing, involving grayscale conversion, normalization, and Gaussian smoothing, a joint academic effort from Iran and the UK; character-level CNN with interval-based segmentation; and image-level CNN, from the University of Oxford in the UK.

Method	Dark Web CAPTCHA		
	Rescator-1	Rescator-2	Yellow Brick
Image-level CNN only [17]	63.57%***	35.05%***	5.88%***
Image-level CNN + Preprocessing [22]	70.5%**	36.04%***	7.84%***
Character-level CNN + Segmentation [30]	88.12%**	77.23%**	93.72%*
DW-GAN (Ours)	**94.4%**	**97.50%**	**95.98%**

Results from DW-GAN for the first experiment, compared to prior state-of-the-art approaches.

The researchers found that DW-GAN was able to improve on prior results across the board (see table above).

The second experiment was an ablation study, where various components of the active framework are removed or disabled in order to discount the possibility that external or secondary factors are influencing the results.

Model	Dark Web CAPTCHA			Average Performance
	Rescator-1	Rescator-2	Yellow Brick	
DW-GAN without Background Denoising	93.4%	88.12%	21.96%	67.83%
DW-GAN without Border Tracing Segmentation	77.2%	88.91%	38.82%	68.21%
DW-GAN without Interval-Based Segmentation	63.2%	**98.02%**	10.00%	57.07%
DW-GAN	**94.4%**	97.50%	**95.98%**	**95.96%**

Results of the ablation study.

Here too, the authors found that disabling key sections of the architecture reduced the performance of DW-GAN in nearly all cases (see table above).

The third offline experiment compared the efficacy of DW-GAN against benchmark image-based method and two character-level methods, in order to determine the extent to which DW-GAN's character evaluation influenced its usefulness in cases where a nonsense CAPTCHA word was an arbitrary (rather than predefined) length. In these cases, the CAPTCHA length varied between 4 to 7 characters.

For this experiment, the authors used a training set of 50,000 CAPTCHA images, with 5,000 reserved for testing in a typical 90/10 split.

Here too, DW-GAN outperformed prior approaches:

Character Length	Image-level CNN + Preprocessing	Character-Level + Interval-Based Seg	Character-Level + Border Tracing	DW-GAN's Segmentation
4	56.98%	76.92%	70.07%	79.92%
5	48.43%	73.11%	66.51%	74.35%
6	36.88%	67.81%	65.33%	71.87%
7	29.29%	64.27%	59.68%	69.09%

Live Test on a Dark Net Market

Finally, DW-GAN was deployed against the (then live) Yellow Brick dark net market. For this test, a Tor web browser was developed which integrated DW-GAN into its browsing capabilities, automatically parsing CAPTCHA challenges.

In this scenario, a CAPTCHA was presented to the automated crawler for every 15 HTTP requests, on average. The crawler was able to index 1,831 illegal items for sale in Yellow Brick, including 1,223 drug-related products (including opioids and cocaine), 44 hacking packages, and nine forged document scans. In total the system was able to identify 286 cybersecurity-related items, including 102 purloined credit cards and 131 stolen account logins.

The authors state that DW-GAN was in all cases able to crack a CAPTCHA in three or fewer attempts, and that 76 minutes of processing time were necessary to account for CAPTCHAs guarding all 1,831 products. No humans were needed to intervene, and no endpoint failure cases occurred.

The authors note the emergence of challenges that offer a greater level of sophistication than text CAPTCHAs, including some that seem modeled on Turing tests, and observe that DW-GAN could be enhanced to accommodate these new trends as they become popular.

*Completely Automated Public Turing test to tell Computers and Humans Apart

First published 11th January 2022.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

UNITE.AI

Discover More