HoGent

BUSINESS AND INFORMATION MANAGEMENT

Professional Bachelor in Applied Computer Science
Academic year 2012-2013

Solving CAPTCHA using neural networks

Submitted on 10 June 2013

Student:
Pieter Van Eeckhout

Mentor:
Johan Van Schoor

HoGent Business & Information Management
Professional Bachelor in Applied Computer Science
Academic year 2012-2013

Solving CAPTCHA using neural networks

Submitted on 10 June 2013

Student:
Pieter Van Eeckhout

Mentor:
Johan Van Schoor

Contents

Abstract

TODO

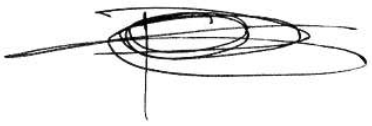
Preamble

First, dear reader, I would like to thank you for taking the time to read this thesis. Without an audience this entire endeavour would not mean as much as it does right now, while you are reading its results. I personally believe this is because I would like my life not to go unnoticed. So if this thesis helps, or influences you in any way, then this work has gained more meaning.
Second I would like to thank the following people who have made it possible for me to arrive at this point. Special thanks and mentions go to:

- my parents, for supporting me and giving me the opportunity and supplying the means for me to pursue my academic career.
- my girlfriend, Anne Charlotte Magdaraog Mendoza. Because she has helped me countless times through the rough spots. Not once did she complain about the time consuming job of writing this work.
- my good friends, willing proof readers and content critics: Wouter Dekens, Patrick Van Brussel and Thijs van der Burgt.
- Johan Van Schoor and Bert Van Vreckem for the support, organisation, guidance and feedback.

Bare in mind that this is not an exclusive list. Finally I would like to thank all the other people who are not mentioned by name: such as the teaching and support staff at University College Ghent.

Ghent BELGIUM, June 2013



Pieter Van Eeckhout

Chapter 1

Solving CAPTCHA using neural networks

The target audience. This thesis was written with an audience in mind that already has some technical understanding of computers and how they operate on hardware level (processor etc.). If you feel that your current knowledge is insufficient, or just want to read up some more, then I refer you to the "How Computers Work - Processor and Main Memory" [Young, 2001] e-book.

The history of SPAM. Ever since the internet found its way into our daily life, there have been people out there who don't always have other people's best interest in mind. I am referring to spammers, people aiming to advertise their product, services, etc . . . in an aggressive manner. The methods of advertising include but are not limited to:

- Sending bulk emails without the recipients permission (SPAM).
- Posting irrelevant links and information on fora and various social media.
- Flooding chat channels with their links and information.

These emails, posts and messages inconvenience the end-users, requiring time to filter out the junk. The economic costs of SPAM has led to a decrease in the Japanese GDP by 500 billion Yen (3.78 billion Euro) in 2004 and were projected to reach a decrease of 1% of the total GDP by 2010 unless adequate countermeasures were taken [Ukai and Takemura, 2007]. [Khong, 2004] researched the economic arguments for regulating junk mails and the efficiency of these regulations.

Birth of CAPTCHA. The two previously mentioned researches signify the importance and impact of SPAM on our daily life. The users of the internet quickly tried to implement methods to prevent spammers from spreading their advertisements to the masses. Several prevention and detection methods and systems were developed successfully. These methods and mechanisms range from hidden text to invalid HTML tags, all used to confuse and interrupt automated programs. One of the methods developed to prevent SPAM is a CAPTCHA test. CAPTCHA is an acronym based on the word "capture" and stands for 'Completely Automated Public Turing test to tell Computers and Humans Apart'. An attempt to trademark the term was made by Carnegie Mellon University on 15 October 2004, but the application was eventually dropped on 12 April 2008

Spammers fight back. All these prevention and detection methods did not stop the spammers from trying to reach an audience as large as possible. The spammers rely on a large target audience because of the return rates being as low as 0.0023% [Cobb, 2003]. The spammers started to device ways to circumvent or break the existing systems in order to reach a large enough audience. One of these methods is solving CAPTCHA tests by making use of the adaptive learning and pattern recognizing capabilities of neural networks. These networks can be used to recognize letters from images with adversarial clutter. This is the area I will focus on in this thesis. This thesis will list some of the difficulties regarding the extraction of relevant data from a CAPTCHA and how to possibly overcome these difficulties. However the main focus will be on searching for the types and configuration of neural networks best used for pattern recognition.

Chapter 2

Premise and research questions

2.1 Premise

The main objective of this thesis is to ascertain whether neural networks are capable of solving the current generation of CAPTCHA images. we will define the premise as following:

"Are neural networks a viable tool for solving the current generation of CAPTCHA?"

2.2 Research questions

The research can be divided into two separate subjects. If one was to develop software for automatic CAPTCHA solving, the following questions and problems would need to be addressed.

CAPTCHA:

- What are the different types of CAPTCHA?
- How can the distorted text be extracted?

Neural networks:

- How do neural networks operate?
- Which types of neural networks are well suited for pattern recognition?
- What network configuration would perform best?

General:

- How future proof would this solution be?
- Is there enough economic incentive to invest in development?

Chapter 3

Methodology

Research philosophy. TODO

Research approach. TODO

Data Analysis. TODO

Chapter 4

Corpus

4.1 CAPTCHA

4.1.1 CAPTCHA, an explanation.

A CAPTCHA (pronounced) is a type of challenge-response test that aims to make sure the response was made by a human. These tests are designed in such a manner that they should be easy to generate and grade by a computer, but also difficult for a computer to solve. Yet a human should be able to solve the test without much difficulty. If a test was solved successfully it can be assumed that the response was entered by a human.
These test are mostly found on sites where on would like to prevent the access to unwanted bots. this is because having lots of spam on a site or in a service can have real detrimental consequences for that site or service. This is because most contemporary interactive sites store and serve their content from a database. When a database gets filled up the site can become slow and sluggish, reducing the customer's experience. This is only one of the many useful applications of CAPTCHAs. On the other hand, legitimate users also need to solve these tests, so it requires them to perform an extra task before they can post their content, create an email, view a certain page. While this 'simple' extra task does not seem like a large barrier, it does inconvenience some people enough to prevent them from posting valid content. This problem becomes even more apparent when dealing with non-native speakers[Banday and Shah, 2011]. Protecting your site with a CAPTCHA can even have a detrimental effects on the conversion rates[1].

[1] `http://www.seomoz.org/blog/captchas-affect-on-conversion-rates`

4.1.2 The history of CAPTCHA.

The first one to think of the concept of CAPTCHA was Moni Naor in 1996. He proposed that reverse Turing testing, as CAPTCHAs are often called, should consist of "those tasks where humans excel in performing, but machines have a hard-time competing with the performance of a three year old child." Some of these tasks were [Naor, 1996]:

- gender recognition
- understanding facial expressions
- understanding handwriting
- filling in words

In 1997 Yahoo! was having a massive problem with spammers using bots to create free email addresses used to spread a huge amount of unwanted advertisement, giving Yahoo email adresses a bad reputation. Yahoo! contacted Carnegie Mellon University[2] for help, by 2000 the first real CAPTCHA as we know them was invented[Egen, 2009]. These were also the people who first used the term "CAPTCHA" and tried to trademark it.
As the Computing power increased, so did the amount of CAPTCHA tests being broken. By 2008 there was an 30% to 60% success rate on the most used forms of CAPTCHA.[Yan and El Ahmad, 2008]. As a response to this Von Ahn and his team at Carnegie Mellon University released reCAPTCHA (Figure 1, page 17) in September 2008, a popular system still currently in use.
CAPTCHAs have always undergone changes once it became clear a certain generation method didn't stop the spammers any more. The first CAPTCHAs generated by EZ-Gimpy for Yahoo! look completely different from the CAPTCHAs that are currently being generated. A good example of the adaptive nature of CAPTCHAs is reCAPTCHA, were you can see the changes depending on when a CAPTCHA was generated. (Figure 2, page 17)

4.1.3 Types of CAPTCHA.

Following is a list and description of the different types of CAPTCHA, quoted directly from [Sauer and Hochheiser, 2008].

Character based This category means that a string of characters is presented to the user. This string can contain either words or random alphanumeric characters.

[2] `http://www.cylab.cmu.edu/research/projects/2008/captcha-project.html`

Image based Images or pictures are presented to the user. This is normally in the form of an identifiable real-world object, but can also be presented in the form of shapes (BONGO). The task is to identify the object shown in the picture.

Anomaly based Users are asked to determine which object, or character, or shape does not belong in a set of images displayed on the screen.

Recognition based The user needs to determine what is being presented to them. In the case of a character based and recognition based CAPTCHA the user needs to identify and input the character string that is presented to them.

Sound based The user is presented with an audio version of a CAPTCHA. The user listens to the audio file and inputs their answer. A sound based CAPTCHA can be presented in two formats, the first is the "spoken words or numbers" and the second would be sounds related to an image.

4.1.4 Data extraction.

As previously stated, the data extraction part of solving CAPTCHAs is not the main focus of this thesis. Therefore I will not give in-depth explanations of the algorithms used and described here.

CAPTCHAs are by design tough to solve for a computer. This is because most of the times a CAPTCHA gets cluttered with noise, or the letters get crowed together. This crowding or noise makes it so that the characters on the image are not separate entities. This is to impede the segmentation of the CAPTCHA. Measures against segmentation are necessary to prevent an OCR[3] algorithm from simply reading and solving the test. This could be possible, as computers can (given the right algorithms) be very efficiently at pattern recognition. People trying to solve the CAPTCHA test automatically now first have to separate the individual characters before they can pass the characters to an OCR algorithm for classification.
[Yan and El Ahmad, 2008] has described a working segmentation algorithm in 2008, but [Huang et al., 2010] has significantly improved on the performance, so that now it should be able to segment the contemporary CAPTCHAs.
In the unlikely case that the CAPTCHAs you are trying to solve don't have the segmentation issues, then you can first try to reduce the noise and then segment the characters by using the flood-fill method, as described by [Cai, 2008].

[3] Optical Character Recognition

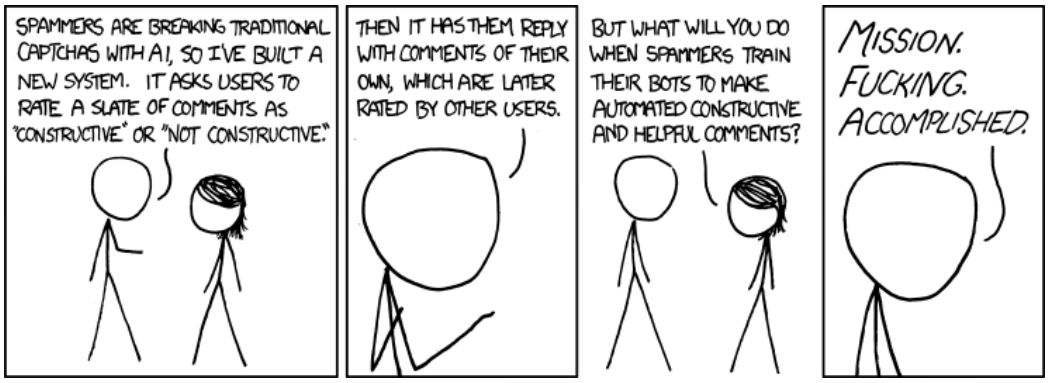


Figure 4.1: xkcd on the future of CAPTCHA (Source: `http://www.xkcd.com/810/`, accessed on 2013/05/28)

4.1.5 The future of CAPTCHA.

The arms race between the makers of CAPTCHA systems and people trying to break them favoured the defender. this is different from other computer security arms races, where the odds are in favour of the adversary. This is because CAPTCHA has broken the traditional pattern where the attacker's role is to generate new instances while the defender must recognize them, recognizing a problem is almost always harder than generating them. Websites and services using CAPTCHA can easily change the CAPTCHA generation algorithm, creating new unsolvable CAPTCHAs, while the attackers now have the challenging recognition problem. This battle has advanced brought advances to the field of Automated Pattern Recognition and Artificial Intelligence. Some people even believe[4](Figure 4.1) that eventually the solving algorithms will become that sophisticated they could be classified as a sentient AI.
All the positive aspects and technological innovations aside, CAPTCHAs are inherently flawed. As the solving agents got better, the CAPTCHAs became harder. We have reached a point where the average user is having difficulty solving the standard CAPTCHAs[5].

> It would seem evident from years of use and research that CAPTCHAs are far from perfect as a solution. Remove spammers from the equation and we remove the need for CAPTCHAs entirely; this is the mentality we should be aiming for. The perfect CAPTCHA is no CAPTCHA at all.[Bushell, 2011]

[4] `http://thenextweb.com/2009/10/15/inevitable-future-captcha/`

[5] `http://www.internetevolution.com/author.asp?section_id=587&doc_id=259406`

[Sauer and Hochheiser, 2008] and colleges did a small research about how the current CAPTCHA (even the audio CAPTCHA) has serious shortcomings when trying to accommodate for blind or visually impaired users. This added downside to the current CAPTCHA system indicates that even tough it is an effective means TODO

4.2 Neural Networks

4.2.1 How neural networks operate.

TODO

4.2.2 Types of neural networks.

TODO

4.2.3 Neural networks for pattern recognition

TODO

4.2.4 Optimal network configuration

TODO

4.3 Implementation

4.3.1 Captcha builder

TODO

4.3.2 Neural networks

TODO

Chapter 5

Conclusion

TODO

Bibliography

M Tariq Banday and NA Shah. A Study of CAPTCHAs for Securing Web Services. *IJSDIA International Journal of Secure Digital Information Age*, 1(2):66–74, 2011. URL http://adsabs.harvard.edu/abs/2011arXiv1112.5605T.

David Bushell. In search of the perfect captcha. http://coding.smashingmagazine.com/2011/03/04/in-search-of-the-perfect-captcha/, 2011. Accessed: 2013-05-28.

Tianhui Cai. CAPTCHA Solving With Neural Networks. Technical report, TJHSST Computer Systems Lab, 2008.

Stephen Cobb. The Economics of Spam, 2003. URL http://spamhelp.whybot.com/articles/economics_of_spam.pdf.

Dennis Egen. A Proposal For Improvements of Image Based CAPTCHA. Technical report, Rutgers University, Camden, 2009.

SY Huang, YK Lee, Graeme Bell, and Zhan-he Ou. *An efficient segmentation algorithm for CAPTCHAs with line cluttering and character warping*. PhD thesis, Ming Chuan University, 2010. URL http://link.springer.com/article/10.1007/s11042-009-0341-5.

Dennis W K Khong. An Economic Analysis of Spam Law. *Erasmus Law and Economics Review*, 1(February):23–45, 2004. URL http://www.eler.org/viewarticle.php?id=2.

Marti Motoyama, Kirill Levchenko, Chris Kanich, Damon Mccoy, Geoffrey M Voelker, and Stefan Savage. *Understanding CAPTCHA-Solving Services in an Economic Context*. PhD thesis, University of California, San Diego, 2010. URL https://www.google.be/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&sqi=2&ved=0CEEQFjAA&url=http%3A%2F%2Fwww.usenix.org%2Fevent%2Fsec10%2Ftech%2Ffull_papers%2FMotoyama.pdf&ei=8I6GUa-YA6ar0QWAvoDoDg&usg=

AFQjCNGw19XqLpU9H07GzrwXiZzp7AUSHw&sig2=n_TgEyAkII33doIB1VJFDw&bvm=bv.45960087,d.d2k.

municipal cooperation.org. Help:adding content. http://www.municipal-cooperation.org/index.php?title=Help:Adding_content, unknown. Accessed: 2013-05-28.

Moni Naor. Verification of a human in the loop or Identification via the Turing Test. *. . . . wisdom. weizmann. ac. il/˜ naor/PAPERS/human . . .*, 1996. URL http://www.wisdom.weizmann.ac.il/~/naor/PAPERS/human.pdf.

Google reCAPTCHA. recaptcha. http://www.google.com/recaptcha, 2013. Accessed: 2013-05-28.

Graig Sauer and Harry Hochheiser. Towards a universally usable CAPTCHA. *. . . the 4th Symp. On Usable . . .*, pages 2–5, 2008. URL http://www.researchgate.net/publication/228957237_Towards_a_universally_usable_CAPTCHA/file/d912f50d10c2235ccc.pdf.

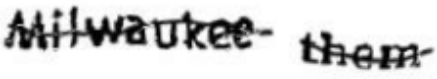
Yasuharu Ukai and Toshihiko Takemura. Spam mails impede economic growth. *The Review of Socionetwork Strategies*, 1(1):14–22, March 2007. ISSN 1867-3236. doi: 10.1007/BF02981628. URL http://link.springer.com/10.1007/BF02981628.

Jeff Yan and Ahmad Salah El Ahmad. A low-cost attack on a Microsoft captcha. *Proceedings of the 15th ACM conference on Computer and communications security - CCS '08*, page 543, 2008. doi: 10.1145/1455770.1455839. URL http://portal.acm.org/citation.cfm?doid=1455770.1455839.

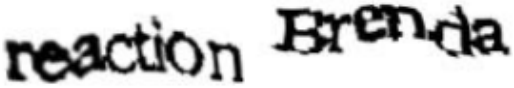
Roger Stephen Young. *How Computers Work Processor and Main Memory*. 2001. URL http://www.fastchip.net/howcomputerswork/bookbpdf.pdf.

List of Figures

Figure 1: The reCAPTCHA system (Source: [municipal cooperation.org, unknown])



(a) Early 2008



(b) December 16th 2009

(c) January 24th 2010

(d) May 28th 2013

Figure 2: Examples of CAPTCHAs directly Downloaded from reCAPTCHA (Source: [Motoyama et al., 2010] and [reCAPTCHA, 2013])