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

1 Solving CAPTCHA using neural networks **3**

2 Premise and research questions **5**
2.1 Premise 5
2.2 Research questions 5

3 Methodology **7**

4 Corpus **8**
4.1 CAPTCHA 8
4.1.1 What is CAPTCHA? 8
4.1.2 The history of CAPTCHA. 8
4.1.3 Types of CAPTCHA. 9
4.1.4 Data extraction. 9
4.1.5 The future of CAPTCHA. 9
4.2 Neural Networks 9
4.2.1 How neural networks operate. 9
4.2.2 Types of neural networks. 9
4.2.3 Neural networks for pattern recognition 9
4.2.4 Optimal network configuration 9
4.3 Implementation 9
4.3.1 Captcha builder 9
4.3.2 Neural networks 9

5 Conclusion **10**

Abstract

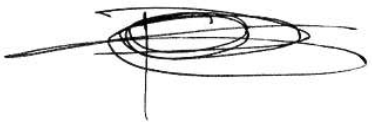
Preamble

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. Because she has helped me countless times through the rough spots. Because 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 What is CAPTCHA?

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:

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

[Naor, 1996] In 1997 Yahoo! was having a massive problem with spammers using bots to create free email addresses. Yahoo! contacted Carnegie Mellon University[1] for help, by 2000 the first real CAPTCHA as we know them was invented[Egen, 2009].
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 in September 2008, a system still currently in use.

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

4.1.3 Types of CAPTCHA.

TODO

4.1.4 Data extraction.

TODO

4.1.5 The future of CAPTCHA.

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

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.

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.

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.

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.