HoGent

BEDRIJF
EN
ORGANISATIE

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

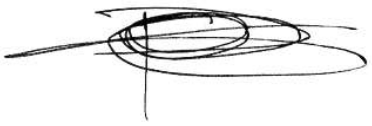
Preamble

Firstly, 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 doe right now, while you are reading it's 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, the this work has gained more meaning.
Secondly I would like to thank the following persons who have made it possible for me to arrive at this point. Special thanks and mentions go to:

- my parent, for giving me the opportunity and supplying the means for me to pursue my academic career.
- my girlfriend, because she has helped me countless times. Because she helped through the rough spots. Because she never once complained 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 Burght.
- Johan Van Schoor and Bert Van Vreckem for the support, organisation, guidance and feedback.

Bare in mind that this is not an exclusive list. So lastly I would like to thank all the other people who are not mentioned by name, like 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 has found its way into the daily usage in our society there have been people out there who don't always have other people's best interests in mind. In this particular case I am referring to 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] reseached 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 lives. 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 range from hidden text only visible to automated scripts, to invalid HTML tags. One of the methods developed for this purpose is a CAPTCHA test. CAPTCHA is an acronym based on the word "capture" and is spelled out completely as '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. They rely on this vast audience because of the return rates being as low as 0.0023% [Cobb, 2003]. Trying to reach such a large audience the spammers start to device ways to circumvent or break the existing systems. One of these methods is solving CAPTCHA tests 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

2.2 Research questions

What different types of CAPTCHA exist?

What are the difficulties for solving a CAPTCHA automatically

What are the types of neural networks suitable for OCR

Is this a feasible endeavour at this point of personal computing power

Chapter 3

Methodology

Chapter 4

Corpus

4.1 CAPTCHA

4.2 Neural Networks

4.3 Implementation

Chapter 5

Conclusion

Appendix A

Sourcecode

Bibliography

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

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.

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.

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

List of Figures

List of Tables

Listings