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
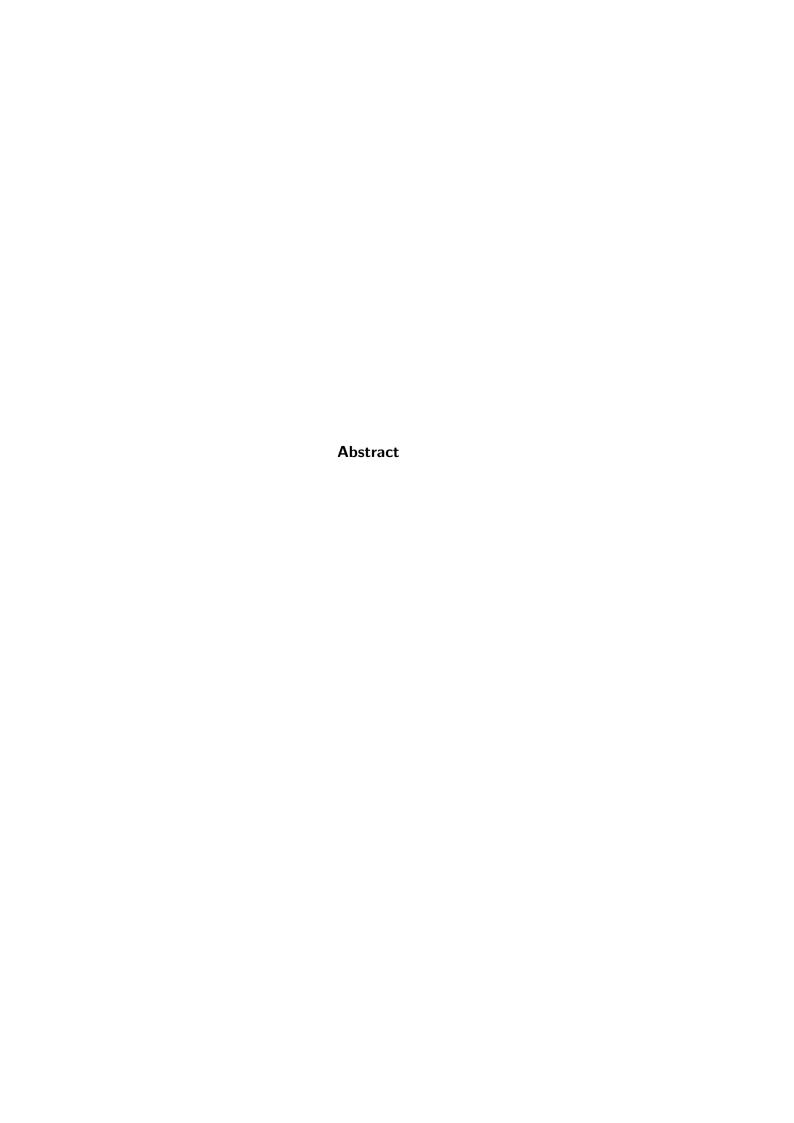
Solving CAPTCHA using neural networks

Submitted on 10 June 2013

Student:
Pieter Van Eeckhout

Mentor:
Johan Van Schoor

Contents

Abstract

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 does 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, then 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 parents, for supporting me and giving me the opportunity and supplying the means for me to pursue my academic career.

- my girlfriend, because she has helped me countless times, she helped me 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 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. 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

2

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" [**?**] e-book.

The history of SPAM.   Ever since the internet found its way into our daily lives, there have been people out there who don't always have other people's best interests in mind. I am referring to spammer, 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 [**?**]. [**?**] 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 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 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% [**?**]. 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 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, 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 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

[?] 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[?].

4.1.3 Types of CAPTCHA.

TODO

---

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

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