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

Chapter 1

Solving CAPTCHA using neural networks

A 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 peoples best interests in mind. In our particular case I am referring to people who are 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 info on fora and various social media.
- flooding chat channels with their links and info.

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 counter-measures are taken [**?**]. [**?**] reseached the economic arguments for regulating junk mails, ~~The efficiency thereof~~ and analyses the regulatory approaches in use in the United States at the time of writing.

Birth of CAPTCHA. These 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 trademarked 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% [**?**]. Trying to reach such a large audience the spammers start to device ways to circumvent or break the existing systems. One of the methods used is to solve CAPTCHA tests making use of the adaptive and learning capabilities, pattern recognizing networks can be used to recognize letters from images. This is the area I will focus on. In this thesis I will point out the difficulties to overcome ~~and~~ the neural network types best used to try to solve CAPTCHA test automatically.

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

List of Figures

List of Tables

Listings