Chapter 45
Machine Learning for Web Proxy Analytics

Mark Maldonado
St. Mary's University, San Antonio, USA

Ayad Barsoum
St. Mary's University, San Antonio, USA

ABSTRACT

Proxy servers used around the globe are typically graded and built for small businesses to large enterprises. This does not dismiss any of the current efforts to keep the general consumer of an electronic device safe from malicious websites or denying youth of obscene content. With the emergence of machine learning, we can utilize the power to have smart security instantiated around the population's everyday life. In this work, we present a simple solution of providing a web proxy to each user of mobile devices or any networked computer powered by a neural network. The idea is to have a proxy server to handle the functionality to allow safe websites to be rendered per request. When a website request is made and not identified in the pre-determined website database, the proxy server will utilize a trained neural network to determine whether or not to render that website. The neural network will be trained on a vast collection of sampled websites by category. The neural network needs to be trained constantly to improve decision making as new websites are visited.

INTRODUCTION

Over the past couple of decades, the use of machine learning or artificial intelligence is a term that has been coined as the next goal of smart business or providing help to people in everyday life. While the term "Artificial Intelligence" is older than a few decades since John McCarthy coined the term in 1956 (Neapolitan & Jiang, 2012; Goranzon & Florin, 2012). The general concept of artificial intelligence is discovering ways to have machines reason and perform intelligently driven by software and algorithms. This effort is closely related to how the human brain works since we want these machines capable of learning and to think rationally. This is the overarching mindset we must use when moving forward

DOI: 10.4018/978-1-6684-6291-1.ch045

with a hardware or software-based design to implement a security product. Home network security is the number one issue any household is trying to overcome. With the sheer amount of malicious traffic generated, a general understanding of home network security must become mandatory. Ensuring security for a family has become a challenge due to the number of websites children and young teenagers can access. Without proper monitoring of network traffic, our youth can infect their devices, laptops, or other network components, which is just the tip of the iceberg. Now there are sufficient hardware devices and software products to help with these issues. Some issues with this approach are the cost of these hardware devices and the time or knowledge to implement. A software approach is feasible, but this will only protect the device it's installed on. Lastly, having someone understand the potential logs from a hardware device is overbearing or trying to fully understand what a piece of software is doing to protect is also a challenge. The general population needs something easy to use, install, and understand the information being generated from a product that can handle all these problem-sets. The idea is to have a proxy server (Luotonen & Altis, 1994; Weaver, Kreibich, & Paxson, 2014) to handle the functionality to allow the correct websites to be rendered per request. When a request is made and not identified in the pre-determined website database, the proxy server will utilize the neural network. This instance of the network will be tested against the already trained neural network to determine if the requested website is allowed or not.

MACHINE LEARNING

In the world of machine learning (Sebastiani, 2002; Michie, Spiegelhalter, & Taylor, 1994; Quinlan, 2014; Witten, Frank, Hal, & Pal, 2016; Pedregosa, et al., 2011) and/or artificial intelligence, one must find the perfect starting point and should have an idea where the project will go or possibly evolve into. There are several things we need to consider when picking a neural network design and what we want to achieve. A neural network (Kalchbrenner, Grefenstette, & Blunsom, 2014; Psaltis, Sideris, & Yamamura, 1988; Haykin, 1994; Hagan, Demuth, Beale, & De Jesús, 1996; Anthony & Bartlett, 2009) is generally comprised of 3 basic parts known as neurons, layers, and bias.

Neurons deal with numerous types of information to be processed. Each individual neuron must know how to handle these types of information: input values, weights and bias, net sum, and an activation function. Although a neuron is just a small portion, it is critical to have accurate processing of data.

Layers are important component in a neural network. There is a minimum of three layers: input, hidden, and output. Each layer must handle information being fed forward to create an expected answer. Starting with the input layer, it is critical to have information prepared accurate and normalized properly. This will lower the chance of unexpected results. The hidden layer is particularly unique, as there can be multiple depending on the complexity of processing. When more than one hidden is present, each layer will feed forward as normal and additional processing for back-propagation is acceptable. After information has traversed through the neural network layers, the output is equally important to have accurate results.

Bias has enough worth to the input during execution. Each layer provides a heavier weight to the neurons if preprocessed data is activated as such. The bias is known as a constant in the network with a predefined value to allow accuracy towards a specific answer. Not all inputs to the network will require a bias to be active.