Title: - URL Categorization & Detection

Building platforms for reproducible AI research

Applicant: Mr. Prateek Dutta

Project Title

URL Categorization and Detection

Project Overview

The Deep Learning model works like a black box. Deep learning can be considered as a subset of machine learning. It is a field that is based on learning and improving on its own by examining computer algorithms. While machine learning uses simpler concepts, deep learning works with artificial neural networks, which are designed to imitate how humans think and learn. Until recently, neural networks were limited by computing power and thus were limited in complexity. However, advancements in Big Data analytics have permitted larger, sophisticated neural networks, allowing computers to observe, learn, and react to complex situations faster than humans. Deep learning has aided image classification, language translation, speech recognition. It can be used to solve any pattern recognition problem and without human intervention.

Web page classification is an information retrieval application that provides useful information that can be a basis for many different application domains. In this project, a deep learning-based system will develop for the classification of web pages which can be demonstrated through software.

Implementation Plan

To understand the classification model, a basic structure of URL is to understand. A URL consists of several parts. We can break down the URL into several parts for understanding different characters of URL.

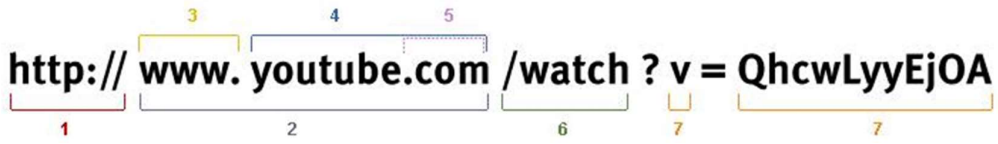

Figure 1: Structure and components of the URL http://www.youtube.com/watch?v=QhcwLyyEjOA

Following is the definition of parameters:
The Protocol in use – in this case: *HTTP* (Hypertext Transfer Protocol) There are also other protocols like HTTPS, FTP and so on.
The Host or Hostname*: www.youtube.com*
The Subdomain*: www.*
The domain name (Domain): *youtube.com*
The Top-Level-Domain (a web-address suffix): *.com* (also known by the shorthand TLD)

The Path*: /watch* A path will usually refer to a file or folder (directory) on the web server (for example "/folder/file.html")
Parameter and value*: v* (Parameter), *QhcwLyyEjOA* (Parameter value) Parameters are initialised by the "?" inside the URL.

Technical Details

Through this project will use the technology involved in industry 4.0. Algorithms & language will use to develop the final software are as follow: -

Deep Learning algorithm: - 1-D CNN

Programming language: - Python

Data: - Online source

Detailed Description:

URLs are a minor ranking factor search engine use when determining a particular page or resource's relevance to a search query. While using a URL that includes keywords can improve your site's search visibility, URLs themselves generally do not have a major impact on a page's ability to rank.

Will basically make use of 1-Dimensional Convolution Neural Network (1-D CNN) and try to achieve the higher accuracy and then will build a software in which will only need to enter the URL & in result it will tell that the URL belongs to which category and whether its fake or real.

For this approach firstly, will collect the data from an online source and train the model and if we achieve considerable accuracy then will deal with some real time data and on achieving measurable accuracy, will convert it into software application.

Why to work on this Project

Looking toward the current situation, the rate of cybercrime is increasing and it leads many people to stop using online services which is required in modern society. Around the world, in every 1 hour there are 5 cybercrime cases registered. So, in order to prevent cybercrime and let people be aware of it, this software can be helpful. As this solution will not exactly stop criminal activities but it leads people to be aware of it and can be considered one of the methods of prevention.