Department of Computer Science
University of Hong Kong
Final Year Project Plan

Analysis of Encrypted Internet Traffic with Deep Leaning

Name: Lu Fangzheng
Major: Computing and data analytics
UID: 3035142780
FYP ID: fyp17034
Supervisor: Dr. S.M. Yiu
# Table of Contents

1. **INTRODUCTION**
   
   1.1 Background

   1.2 Objective

   1.3 Scope

2. **METHODOLOGY**

   2.1 Platform, Hardware & Library

   2.2 Procedure

   2.3 Feasibility Analysis

   2.4 Risks & Challenges

3. **SCHEDULE & MILESTONE**

4. **REFERENCE**
1. INTRODUCTION

1.1 Background

Back in 1903, when the first hacker in the history, Nevil Maskelyn, disrupted John Ambrose Fleming’s public demonstration by hacking into wireless telegraphy, nobody would even image how significant cyber security would ne today (Wikipedia, Hacker). In the past decades, with the widespread of personal computer and rapid development of internet technology, people started to equip themselves with weapons to protect themselves against potential attacks from the internet. Internet firewall, security protocol and different kinds of anti-virus software are adopted by us to protect our computers from internet attack (Wikipedia, cyber security).

In order to achieve cyber security, a majority of the internet companies begins to replace their website protocol with Hypertext Transfer Protocol Secure instead of the former Hypertext Transfer Protocol. Currently, the new Hypertext Transfer Protocol Secure makes use of different encryption algorithms including RSA, DES to encrypt the contents before sending and receiving. However, is it possible to go one step further, can we even work on the encrypted data to enhance the cyber security? In this project, the encrypted internet traffic will be analyzed. Besides, Deep learning algorithms will be adopted by us to analyze the hidden information under the encrypted internet traffic.

As we all know, deep learning, machine learning and artificial intelligence are becoming more and more popular recently. AlphaGo, developed by deep mind, defeated top Go cheese play using deep learning model. And the reason why we try to use deep learning in this project is not only because we are facing the encrypted packets, which we can’t use the regular algorithms to analyze its contents directly, but also because of its efficiency. As we can see in the figure, given a large amount of data input, the performance of regular learning algorithms will meet a boundary while the deep learning algorithms scale much better with more data input. The deep learning algorithms, including convolutional neural network (CNN), support vector machines (SVMs), has been widely used in many classification problems. They have been successfully used in image recognition, natural language processing and many important fields nowadays.
1.2 Objective

This project involves building a deep learning model to analyze the encrypted network traffic to allow people to protect their personal computer from cyber attack. Deep learning model will be employed to classify the encrypted network packets to extract useful information like whether the packets are sent through VPN, whether the packets contain malicious contents. In this way, we believe that through the analysis of the packets, we can get the information we want and better enhance the network security.

The intermediate goal of this project is to develop a model, which can classify the encrypted network traffic into different categories and use a large amount of data to train the deep learning model to make it more accurate.

The ultimate goal of the project involves building a software or browser plug-in for the model that can help everyone in the internet community to enhance cyber security. Besides, we plan to publish a paper about the project and possibly doing more research into this field in the future.

1.3 Scope

This project will include the analysis of encrypted internet packets. The analysis will be applied to the packets’ size, header, protocol and their encrypted data. However, the packet’s origin contents will not be a part of the input in the process of model development. Meanwhile, the training data will be categorized into different categories ahead manually. All types of both harmful and not harmful internet traffic will be included to make the model more accurate.

2. METHODOLOGY

2.1 Platform, Hardware & Library

Firstly, the platform that this project will possibly adopt includes pyCharm (which is a platform supporting python), eclipse (which can be used in Java, C++ development), MySQL (possibly used as platform for database of the training and testing data) and Sublime Text 2 (a useful text editor for programmers).

Secondly, the project requires a high-performance computer to train the model efficiently, and a digital camera (if the internet packet we try to analyze involve videos).

Thirdly, the major library we will employ is TensorFlow, an open-source API aims at solving different kinds machine learning problems. The API supports mostly Python, and some C++ and Java as well.
At last, the project requires a large amount of internet traffic data to train the dataset. That is to say, a significant amount of internet packets, both harmful and not harmful, should be required for this project.

2.2 Procedure

In the first step, we will do some in-depth research into the field of cyber security, encryption algorithm, and deep learning algorithms. Some paper written by previous researchers in the related fields will contribute to the decision on the deep learning algorithm we should use to make the model as accurate as possible. Also, in this step, we will need to determine the approximate amount of internet traffic data we need based on the model we choose.

In the second step, we will implement the model based on the previous step. Possibly, if we are using supervised learning, we need to determine several key features of the packets that are used for the measurement of packets. An implementation of the filtering of the contents of the packets to extract useful information will be implemented in this step as well. TensorFlow and possibly other libraries will play a significant role in this step to help us build a model for classification of the packets.

In the third step, we will collect a huge amount of internet traffic at the first place. Then we will manually separate them into two subsets. Around one fifth of them will be test data and the rest will be training data. After classifying the test data, we then train the model with training data and test its accuracy with test data.

In the fourth step, we will modify and debug the model to make the result more accurate.

Lastly, we want to develop either a software or a browser plug-in for this model. By analyzing the internet packets, we want to know some useful information hidden under encrypted contents.

2.3 Feasibility Analysis

This project will undoubtedly require a lot of time on building the model, training and doing research. However, with the knowledge on the field equipped, I think it is possible for us to work on this project.

2.4 Risks & Challenges

Working on this project involves some possible risks and challenges. Firstly, the packets we work on is encrypted data, that is to say we can just simply filter the contents of the data since the encrypted data itself doesn’t make any sense. Thus, instead of directly analyzing the encrypted data, we will need to find a new new way to analyze the data. Possible its size, header and even some common patterns in the encrypted contents could help. Developing a software or plug-in requires a lot of time as well.
3. SCHEDULE & MILESTONE

<table>
<thead>
<tr>
<th>Date</th>
<th>Deliverables of Phase 1</th>
<th>Deliverables of Phase 2</th>
<th>Deliverables of Phase 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 October 2017</td>
<td>• Detailed project plan&lt;br&gt; • Project web page</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 November 2017</td>
<td>Research into related work&lt;br&gt; • Determine the deep learning algorithm used in model&lt;br&gt; • Select useful features for the packets</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 January 2018</td>
<td>Development of the model&lt;br&gt; • Implement the deep learning model&lt;br&gt; • Implement the filtering program for the packets</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21 January 2018</td>
<td>Deliverables of Phase 2&lt;br&gt; • Preliminary implementation&lt;br&gt; • Detailed interim report</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 March 2018</td>
<td>Training of the model&lt;br&gt; • Data collection&lt;br&gt; • Training, testing of the model.&lt;br&gt; • Modify and debug</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 April 2018</td>
<td>Software building&lt;br&gt; • Development of the software of the model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15 April 2018</td>
<td>Deliverables of Phase 3&lt;br&gt; • Finalized tested implementation&lt;br&gt; • Final report</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4. REFERENCE