ISSN : 2586-0238 (Online)

Supervised 머신 학습 기반한 카카오톡 개별 서비스 분류

바라카, 이성호, 심규석, 김명섭*
고려대학교, 컴퓨터정보학과
{sijabarakajia25,gyh0808, kusuk007,tmskim}@korea.ac.kr

KakaoTalk Individual Services Classification based on Supervised Machine Learning

Baraka D Sija, Sung-ho Lee, Kyu-Seok Shim, Myung-Sup Kim*
Department of Computer and Information Science, Korea University.

Abstract

Unlike other applications, KakaoTalk is an application that provides an enormous number of services and a platform that distributes various third part contents and applications, including hundreds of games which can only be played with KakaoTalk account. Kakao full suite of apps include KakaoTalk, KakaoStory, KakaoMusic, KakaoGroup, KakaoHome, KakaoPlace, KakaoAlbum, KakaoPage, KakaoStyle and KakaoAgit. Each of these apps is likely to have several further extensions. For instance, KakaoTalk extends to five major services which are instant messenger (IM), voice calls, video calls, file sharing and gaming. In this paper, we study about these five KakaoTalk services and introduce a basement for their classification. Classification of KakaoTalk individual services is hard to implement based on traffic packet signatures since KakaoTalk packets are encrypted with TLS/SSL. We learn about how TLS/SSL encryption protocol functions in KakaoTalk, the KakaoTalk TCP/IP protocol (LOCO) and distinctive KakaoTalk packets features to classify one service to another. In this paper, we define the problem and motivation, we suggest an architecture based on Supervised Machine Learning by tensor flow to solution of the defined problem and classify five services.

keywords; KakaoTalk Services, Services Classification, Supervised Machine Learning

Ⅰ. Introduction[1]

Due to daily increase of Internet traffic volume, Network Management administrators are demanding even more intelligent techniques of how to identify and classify certain target or random Internet traffic accurately. Since security of data in transit over the Internet has become increasingly important the steadily growing Internet volume data needs to be accurately managed and controlled. To manage and control such kind traffic, core and high intelligent studies are essential. As the core approach to identifying and classifying five KakaoTalk key services we take a closer study on TLS/SSL protocol security and encryption session establisher during any KakaoTalk communication..

Since Transport Layer Security (TLS) protocol is the most time-consuming phase in the handshaking process, in [1] an architecture of how to improve the handshaking process is designed. The primary goal of the TLS protocol is to provide privacy and data integrity between two communicating applications. The TLS protocol is composed of two layers above the TCP/IP protocol. These two layers are the TLS Record Protocol and the TLS Handshake Protocol. The TLS Record Protocol provides connection security that has two basic properties The TLS Handshake Protocol provides connection security that has three basic properties. [1]

Several approaches on improving the TLS/SSL protocol in general has been proposed. To speed up the TLS/SSL session negotiation time, batch RSA approach is proposed [2]. One the effective ways to improve TLS/SSL handshake protocol are hardware (specific circuit) approach, sessions catching and batching for heavily-loaded web servers. [2]

Ⅱ. Motivation and Problem Definition

Based on TLS/SSL handshaking protocol, classification of network traffic including KakaoTalk is done in [3], whereas in [4] a research is done on inferring user activities on KakaoTalk with traffic analysis or personal information collection. The approaches do not figure out about how to classify individual KakaoTalk services. Thus, in this paper we take a further and novel approach on identifying and classifying five KakaoTalk services Instant Messaging(IM), voice calls, video calls, file sharing and gaming.

III. Proposed Architecture Overview

The proposed architecture operates as follows, at first all traffic is captured from a single machine by the Microsoft Network Monitor

[1] This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (2015R1D1A3A01018057) and Science Technology and Humanities Convergence Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT and Future Planning(No.NRF-2016M3C1B6929228).

and only KakaoTalk frames are collected for general analysis. Second, the traffic of target KakaoTalk services is collected for all the five services and then analyzed. The selected features are Machine Learned and trained by tensor flow to figure out distinctive classifiers for each service. As shown in figure 1, s1, s2, s3, s4 and s5 stands for Instant Messenger(IM), voice calls, video calls, file sharing and gaming services respectively. Since individual KakaoTalk services are difficult to identified by signatures, Supervised Machine Learning by tensor flow is applied for learning and classifying each selected KakaoTalk service.

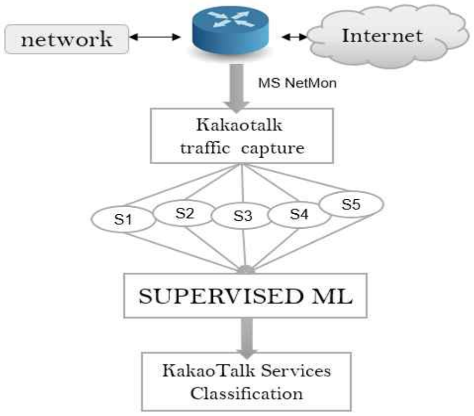

Figure 1. Architecture Overview

IV. Selected features and Experimental Results

We analyzed KakaoTalk services from collected data set of 31,642 packets which is equal to 166MB data size. Successfully evaluated services were four out of five, which are Instant Messenger (IM), file transferring (sharing), voice calls and gaming. Key feature for evaluating differences between the services is the number of packets generated in five different approaches. Table 1 indicates evaluation of packets number in five approaches in minimum, maximum, total and average of the packets number generated for each service. When a KakaoTalk client sends an instant message(IM) to another client an average of 4 packets is generated, when a client shares an image file an average of 284 packets is generated while when a client once accesses KakaoTalk gaming server an average of 30 packets is generated.

Table 1. Evaluation of packets number for each service

Flow	Messenger(IM)	FT(Image file)	Gaming
1	3	195	26
2	5	182	29
3	4	318	33
4	4	381	31
5	5	342	32
Min	3	182	26
Max	5	381	33
Total	21	1418	151
Mean	4	284	30

Experimental results evaluations in table 1, indicates that file transferring(sharing) has the highest number of packets generated followed by gaming service and instant messaging being the least. In table 2, voice calls and another file sharing service (audio file) are evaluated. Unlike table 1, in table 2 the number of packets increases with time and bytes for voice calls and audio file respectively. Although, theoretically the number of packets should increase per time, during a 5mins call session, the number of packets is fewer than prediction. In audio file transferring the service indicated a clear linear relation to the file size. Through linear regression (Supervised ML) by tensor flow under 0.01 learning rate trained data indicated comparable results to real observed data.

Table 2. Evaluation of packets number for each service

Flow	Voice calls	FT (Audio file)
1	1min->63	3.3MB->1752
2	2min->91	6.46MB->3414
3	3min->118	9.4MB->4967
4	4min->141	14.3MB->8318
5	5min->110	19.2MB->11078
Min	1min->63	3.3MB->1752
Max	4min->141	19.2->11078
Total	523	29529
Mean	105	5905

V. Conclusion and Future work

This paper, proposes how to identify and classify individual traffic for KakaoTalk services. From average, minimum, maximum and total number of packets evaluated for each service, we have introduced a basement to classification of individual KakaoTalk services.

We are collecting both general KakaoTalk traffic and individual KakaoTalk services traffic to find clearer and more distinctive features for individual KakaoTalk services. Furthermore, as the future work we will set for capturing KakaoTalk Wi-Fi traffic as well for effective evaluations, since the evaluations done so far base on Ethernet data traffic.

References

[1] Elgohary, Ashraf, Tarek S. Sobh, and Mohammed Zaki. "Design of an enhancement for SSL/TLS protocols." computers & security 25.4 (2006): 297-306.

[2] Qi, Fang, et al. "Batching SSL/TLS handshake improved." Information and Communications Security. Springer Berlin Heidelberg, 2005. 402-413.

[3] S.-M. Kim, Y.-H. Goo, M.-S. Kim, S.-G. Choi, M.-J. Choi, "A method for service identification of SSL/TLS encrypted traffic with the relation of session ID and Server IP", Network Operations and Management Symposium (APNOMS) 2015 17th Asia-Pacific. IEEE, pp. 487-490, 201

[4] Park K., Kim H. (2016) Encryption is Not Enough: Inferring User Activities on KakaoTalk with Traffic Analysis. In: Kim H., Choi D. (eds) Information Security Applications. WISA 2015. Lecture Notes in Computer Science, vol 9503. Springer, Cham