Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.30, NO.4, Aug. 2020 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2020.30.4.669

페이로드 임베딩 사전학습 기반의 웹 공격 분류 모델*

김 연 수,[1†] 고 영 훈,[2] 엄 익 채,[3] 김 경 백[3‡]
[1,2,3]전남대학교(대학원생, 학부생, 교수)

Web Attack Classification Model Based on Payload Embedding Pre-Training*

Yeonsu Kim,[1†] Younghun Ko,[2] Ieckchae Euom,[3] Kyungbaek Kim[3‡]
[1,2,3]Chonnam National University(Graduate student, Undergraduate, Professor)

요 약

인터넷 사용자가 폭발적으로 늘어나면서 웹을 이용한 공격이 증가했다. 뿐만 아니라 기존의 방어 기법들을 우회하기 위해 공격 패턴이 다양해졌다. 전통적인 웹 방화벽은 알려지지 않은 패턴의 공격을 탐지하기 어렵다. 따라서 인공지능으로 비정상을 탐지하는 방식이 대안으로 연구되고 있다. 특히 공격에 악용되는 스크립트나 쿼리가 텍스트로 이루어져 있다는 이유로 자연어 처리 기법을 적용하는 시도가 일어나고 있다. 하지만 스크립트나 쿼리는 미등록 단어(Unknown word)가 다량 발생하기 때문에 자연어 처리와는 다른 방식의 접근이 필요하다. 본 논문에서는 BPE(Byte Pair Encoding)기법으로 웹 공격 페이로드에 자주 사용되는 토큰 집합을 추출하여 임베딩 벡터를 학습시키고, 주의 메커니즘 기반의 Bi-GRU 신경망으로 토큰의 순서와 중요도를 학습하여 웹 공격을 분류하는 모델을 제안한다. 주요 웹 공격인 SQL 삽입 공격, 크로스 사이트 스크립팅, 명령 삽입 공격에 대하여 분류 평가 결과 약 0.9990의 정확도를 얻었으며, 기존 연구에서 제안한 모델의 성능을 상회하는 결과를 도출하였다.

ABSTRACT

As the number of Internet users exploded, attacks on the web increased. In addition, the attack patterns have been diversified to bypass existing defense techniques. Traditional web firewalls are difficult to detect attacks of unknown patterns.Therefore, the method of detecting abnormal behavior by artificial intelligence has been studied as an alternative. Specifically, attempts have been made to apply natural language processing techniques because the type of script or query being exploited consists of text. However, because there are many unknown words in scripts and queries, natural language processing requires a different approach. In this paper, we propose a new classification model which uses byte pair encoding (BPE) technology to learn the embedding vector, that is often used for web attack payloads, and uses an attention mechanism-based Bi-GRU neural network to extract a set of tokens that learn their order and importance. For major web attacks such as SQL injection, cross-site scripting, and command injection attacks, the accuracy of the proposed classification method is about 0.9990 and its accuracy outperforms the model suggested in the previous study.

Keywords: Web Attack, Payload, BPE(Byte Pair Encoding), Wrod embedding, Deep learning

Received(06. 01. 2020), Accepted(06. 15. 2020)
* 이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.2019-0-01343, 융합보안핵심인재양성)과 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2017R1A2B4012559).
† 주저자, 7ustis@gmail.com
‡ 교신저자, kyungbaekkim@jnu.ac.kr(Corresponding author)

I. 서 론

오늘날 우리는 정보를 얻기 위해 수많은 웹 사이트에 접속한다. 이에 따라 공격자들은 금전적인 이득, 개인 정보 탈취, 정치적인 목적 등 다양한 이유로 취약한 웹 사이트를 공격하고 있다.

주요 웹 공격으로는 악의적인 SQL 쿼리(qeury)를 삽입하여 데이터베이스의 중요한 정보를 탈취하는 SQL 삽입(sql injection) 공격, 악성 스크립트를 삽입하여 본래 의도와 다른 작업을 유도하는 크로스 사이트 스크립팅(Cross-Site Scripting, XSS) 공격 등이 있으며, 이러한 웹 공격은 전체 사이버 공격의 56%를 차지한다[1].

웹 사이트 운영자는 공격을 막기 위해 시큐어 코딩, 취약점 진단, 입력값 무효화, 웹 방화벽 등의 방법을 적용하고 있지만, 공격자는 다양한 방법으로 방어책을 우회하고 있어 여전히 피해가 발생하고 있다[2].

그중에서도 웹 방화벽은 앞선 단계에서 예방하지 못한 공격들을 막는 최후의 방어책이다. 그럼에도 불구하고 패턴 기반 탐지 방식의 한계로 알려지지 않은 유형의 공격을 탐지하는 데는 어려움이 있다.

따라서 최근에는 알려지지 않은 패턴의 웹 공격을 탐지하기 위해 인공지능을 활용한 접근 방식이 연구되고 있다. 특히 웹 공격에 악용되는 스크립트나 쿼리가 텍스트로 구성되며, 일정한 구조와 반복되는 단어들이 존재한다는 이유로 자연어 처리(Natural Language Processing, NLP) 기법을 적용하려는 연구가 시도되고 있다[3-7].

하지만 띄어쓰기로 쉽게 구분할 수 있는 일반적인 자연어와는 달리, 웹 공격에 활용되는 스크립트나 쿼리는 단어를 구분하는 경계가 모호하고, 미등록 단어(unknown word, UNK)의 비율이 90% 이상 발생하여 단어 표현(word representation)이 어렵다[3].

본 논문에서는 위와 같은 문제를 해결하기 위해 페이로드 임베딩 사전학습 기반의 웹 공격 분류 모델을 제안한다. 우리의 모델은 BPE[8] 기법으로 구성한 페이로드 토큰 집합을 패스트텍스트(fasttext)[9] 방식으로 학습하여 임베딩 벡터를 추출하고, 주의 메커니즘 기반의 Bi-GRU 신경망 모델에 입력하여 웹 공격 유형을 분류하였다.

본 논문의 나머지 부분은 다음과 같이 구성되어 있다. 2장에서는 전통적인 웹 공격 탐지 방법과 인공지능 기반 탐지에 관한 연구를 요약한다. 3장에서는 페이로드를 학습하기 위한 전처리 과정과 제안하는 모델의 세부 설계 및 구현에 대해 설명한다. 4장에서는 실험 결과를 보여준다. 5장에서는 이 논문의 전체 연구 결과를 결론 짓고 향후 연구에 대해 기술한다.

II. 관련 연구

2.1 전통적인 웹 공격 탐지 방법

웹 공격을 방어하기 위한 전통적인 방법은 크게 네 가지가 있다. 첫 번째, 개발단에서부터 웹 공격에 사용되는 특수 문자 등을 제한하는 방식의 시큐어 코딩을 적용하는 방법[10, 11]. 두 번째, 패턴기반으로 생성한 공격 페이로드를 삽입하고 반응값을 비교해 취약점을 진단하는 웹 취약점 스캐너 방식[12, 13]. 세 번째, 클라이언트에서 입력한 값을 서버측에서 해석할 때 입력값을 변환하여 공격 행위 자체를 무효화시키는 방법. 마지막으로 snort[14], modsecurity[15] 등 웹 서버 앞에 설치하여 사용자의 요청 내용을 분석하고 패턴 기반으로 필터링하는 방법이 있다[16].

특히 modsecurity는 오픈소스 기반 웹 방화벽으로 OWASP(The Open Web Application Security Project)에서 웹 공격 유형 별로 정의해 둔 규칙을 통해 공격을 탐지한다. 이 방식은 낮은 거짓 양성(False-Positive, FP) 비율로 알려진 공격을 탐지하는데는 효과적이나, 사전에 정의하지 않은 패턴의 공격을 탐지 할 수 없다. 또한 규칙을 생성하는데 http 프로토콜, 정규식 등 전문가의 지식을 필요로 한다.

2.2 인공지능 기반 웹 공격 탐지 방법

인공지능 기반 탐지 방식은 웹 공격 샘플의 특징을 학습시켜 이상치를 찾아냄으로써 알려진 공격 뿐만 아니라 알려지지 않은 공격을 탐지하는데 효과적인 방법이다.

Zolotukhin et al[17]은 n-gram 방식으로 http 로그의 3개 필드(웹 리소스, 쿼리 속성, 사용자 에이전트)에서 특징(feature)을 추출하고, SVDD(Support Vector Data Description), K-평균, 밀도 기반 공간 클러스터링(DBSCAN)을

사용하여 이상 탐지 방법을 통한 웹 공격 분류 모델을 제안했다.

Makiou et al[18]과 Zhang et al[19]는 http 트래픽 내의 규칙을 분석하여 나이브 베이즈 방법으로 각각 SQL 삽입 공격을 97.6%, 공격 페이로드를 82.3% 정확도로 탐지하였다.

Cui, Baojiang et al[20]은 URL의 길이, 파라미터 길이, 파라미터 개수, 크로스 사이트 스크립팅과 SQL 삽입 공격에 주로 사용되는 키워드의 포함 여부 등 21가지 주요 특징을 추출하여 의사결정나무 알고리즘으로 웹 공격을 분류하였다.

Michiaki Ito and Hitoshi Iyatomi[4]는 웹 공격 분류를 위해 임베딩 기법을 도입하였다. http 요청을 구성하는 문자열 중 최초 1000개의 문자에서 128차원의 임베딩 벡터를 추출하고, CNN(Convolutional Neural Network)을 통해 2~8까지의 커널 사이즈로 자른 문자열의 분포를 학습하여 웹 공격을 분류하였다.

Yang et al[5]은 URL을 입력값으로 사용하여 웹 공격을 분류하였다. 공격에 사용되는 주요 키워드 96가지 목록을 기준으로 토큰을 추출 하고, 그 외의 문자열은 문자 단위로 자른 뒤 128차원의 임베딩 벡터를 추출하였다. 그 다음 2, 4사이즈 커널의 CNN을 통해 형상을 추출하고, GRU 모델으로 토큰 시퀀스에 따른 영향을 학습하였다. 이 방법은 새로운 공격 패턴이 등장하게 될 경우 키워드 목록을 새로 구성해야 하며, 해당 목록에 존재하지 않는 토큰은 분별하지 못한다는 단점이 존재한다.

Saiyu Hao et al[6]은 http 요청을 입력값으로 사용하였으며, 특수문자를 기준으로 토큰화를 수행하였다. word2vec 방식으로 임베딩 벡터를 추출하였고, LSTM을 양방향으로 구성하여 시퀀스를 추출하였다.

Ren et al[7]은 URL을 입력값으로 피싱 공격을 악성과 정상으로 이진 분류하였다. Saiyu Hao et al[6]에서 제안하는 모델과의 가장 큰 차이는 주의 메커니즘을 적용했다는 것이다. 주의 메커니즘을 적용함으로써 시퀀스 중 분류에 가장 영향을 미치는 토큰에 가중치를 부여하였다.

본 논문에서 제안하는 모델은 토큰의 분포, 시퀀스, 중요도 등에 따른 가중치를 학습한다는 점에서 개념적으로 Ren at al[7]의 모델과 가장 유사하다.

하지만 토큰화 및 임베딩 과정에서 BPE 기법과 패스트텍스트 임베딩 방식으로 벡터를 표현함으로써 OOV(Out-Of-Vocabulary) 문제를 개선하여 분류 성능을 향상 시켰으며, Bi-GRU 신경망을 사용하여 학습 속도를 개선하였다.

또한 피싱 공격이 아닌 웹 공격을 분류 타깃으로 하기 때문에 URL이 아닌 페이로드를 학습하였다. 웹 공격은 공격 유형에 따라 방어 주체의 대응 방식이 달라진다. 따라서 단순히 악성과 정상으로 분류하는 이진 분류 방식이 아닌 공격 유형을 구분하는 다중 분류 방식으로 학습하였다.

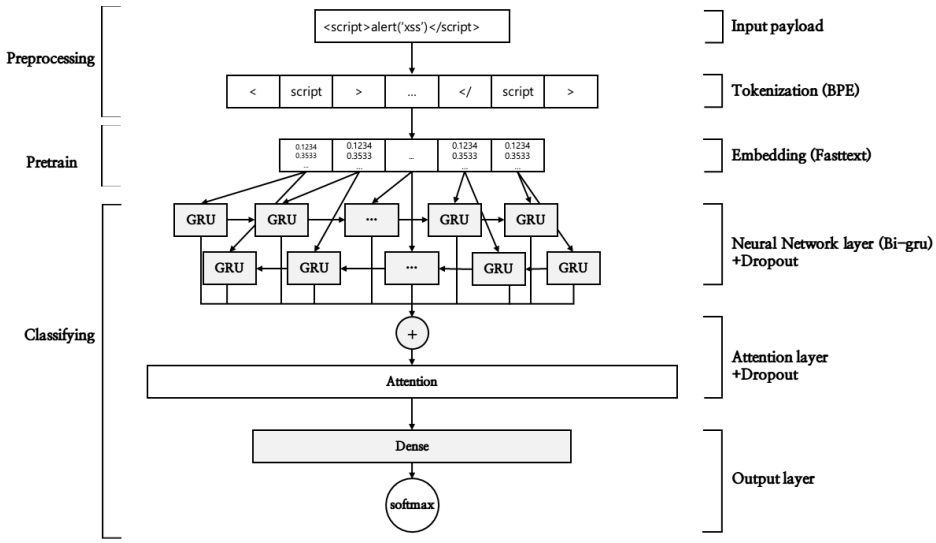

Fig. 1. Proposed web attack payload Classification model overview

III. 제안하는 웹 공격 페이로드 분류 모델

3.1 배경 및 개요

공격자가 시스템이나 네트워크를 공격하기 위한 접근 방식을 공격 벡터라고 한다. 웹 공격의 주요 공격 벡터는 쿼리의 인자값이다. 이때, 인자값에 삽입된 악성 스크립트를 페이로드라고 한다. 웹 공격은 공격별로 Table 1과 같은 페이로드를 가진다.

본 논문에서는 페이로드를 구성하는 문자열(이하 토큰)의 분포, 순서, 중요도 등에 따른 가중치를 학습하여, 서버를 타깃으로 하는 대표적인 웹 공격 세 가지를 분류한다. 이를 구현한 모델의 개요도는 Fig. 1과 같으며, 각 과정은 크게 6가지 단계로 구분된다.

1) 입력값 파싱 : http 요청 패킷에서 페이로드 부분을 추출한다. 주로 GET 방식에서는 URL, POST 방식에서는 body 필드의 쿼리에 키(key)와 값(value) 쌍 중 값에 존재한다.
2) 토큰화 : BPE 기법을 통해 페이로드를 여러 조각들로 자르고 토큰 집합을 구성한다.
3) 단어 임베딩 : 패스트텍스트의 skip-gram 방식으로 토큰간의 관계를 학습하고 임베딩 벡터를 추출한다.
4) 시퀀셜 모델링 : 사전 학습한 임베딩 벡터를 Bi-GRU 모델에 입력하여 토큰 순서에 따른 영향도를 학습한다.
5) 주의 메커니즘 : Bi-GRU의 각 셀에서 출력된 상태를 주의 계층에 입력하여, 페이로드의 공격 유형을 분류하는데 주로 영향을 미치는 토큰들을 선별하고 가중치를 부여한다.
6) 출력 : 덴스(dense)-드롭아웃(dropout)-소프트맥스(softmax)를 거쳐 입력 페이로드의 웹 공격 유형을 분류한다.

Table 1. web attack payload example

Type	Example
Benign	web%20directory%20-%20search%20results
SQLi	′/**/union/**/select/**/0,concat(username,0x3a,password)/**//from/**/users/*′
XSS	〈script〉alert(‘xss’)〈/script〉
Command Injection	eval(‘ls’)

3.2 토큰화 : BPE(Byte Pair Encoding)

페이로드를 신경망으로 학습 시키려면 텍스트 형식을 숫자 형식으로 변환하는 벡터화 과정이 필요하다. 벡터화 방식에는 one hot encoding, n-gram, TF-IDF, 임베딩 등이 주로 사용된다.

벡터화를 위해서는 먼저 텍스트를 일정 단위로 쪼개는 토큰화 작업을 수행해야 한다. 기존 연구들에서는 미리 정해둔 키워드 단위로 토큰화 하거나, 문자를 토큰으로 사용하거나, ‘.(온점)’, ‘/(슬래시)’ ‘?(물음표)’ ‘=(등호)’ 등의 특수문자를 기준으로 자른 문자열을 토큰으로 사용하기도 한다[6, 7]. Fig. 2는 크로스 사이트 스크립팅 공격 페이로드를 입력값으로 넣었을 때 토큰화 방식별 출력 값을 비교한 그림이다.

하지만 페이로드는 인코딩, 난독화 등으로 인해 토큰화의 기준이 모호하며, 기존 방식으로 토큰화 했을 경우 미등록 단어의 비중이 높아 OOV 문제를 유발한다. OOV 문제란 단어 집합(토큰 집합)에 없는 단어로 인해 발생하는 문제를 말한다. 시퀀스 모델에 페이로드 입력시 미등록 단어가 존재한다면 토큰들의 문맥적 의미가 제대로 학습되지 않는다.

본 논문에서는 BPE 기법을 사용하여 토큰화 작업을 수행하였다[8]. BPE는 단어 분리 기법으로 문자열을 문자 단위로 쪼갠 뒤 가장 빈도수가 높은 유니그램의 쌍을 하나의 유니그램으로 묶어가는 작업을 반복하면서 사용자가 설정한 개수만큼 단어 집합을 만들어준다. 따라서 코퍼스에 최적화된 단어 집합을 구성함으로써 미등록 단어 발생을 최소화 할 수 있는 장점이 있다.

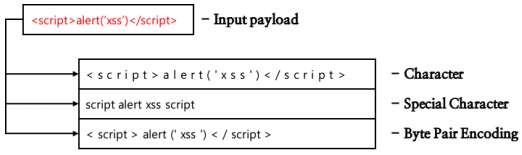

Fig. 2. comparison of tokenization method

3.3 페이로드 임베딩 사전학습 : 패스트텍스트

웹 공격의 페이로드는 특정한 구조를 가지고 있다. 어떤 토큰들이 함께 쓰이냐에 따라서 공격을 구분할 수 있다. 이것을 분포 가정이라고 하는데, 토큰들의

분포에 따른 의미를 학습하기 위해 임베딩 모델을 구성하고 벡터를 추출하였다. 최근 악성 URL 분류 연구에서는 단어 기준으로 집합을 구성하여 임베딩하는 word2vec 방식이 주로 활용되고 있다[6, 7].

앞서 언급했듯 페이로드는 자연어와는 달리 미등록 단어의 비중이 높다. BPE 기법을 통해 미등록 단어의 발생을 최소화 시켰더라도, 방어 필터를 우회하기 위해 대/소문자를 달리 표현(script, Script, scrIPT)하거나, 프로그래머 마다 다르게 표현(length, len, charlen) 되는 단어로 인해, 모든 미등록 단어를 커버할 수는 없다. word2vec 방식에서는 이와 같은 단어들을 전혀 다른 단어로 인식하는 문제가 발생한다.

패스트텍스트(FastText)[9]는 페이스북에서 개발한 단어 임베딩 방법이다. 표현 학습시 각 단어를 문자(character) 단위 n-gram으로 쪼개어 식 (1)과 같이 벡터를 계산한다. z_g는 n-gram의 각 단어에 해당하는 벡터를 의미하고, u_t는 문맥에 포함된 단어 벡터를 의미한다.

$$u_t = \sum_{g \in G_t} z_g \tag{1}$$

따라서 BPE 기법에서 커버하지 못한 미등록 단어들에 대해서도 의미적인 특징과 구문적 변화 규칙을 잘 잡아낼 수 있다.

예를들어 페이로드 단어 집합 중 '〈script〉' 단어에 대한 패스트텍스트의 3-gram 표현 학습 방식은 다음과 같이 표현된다.

$$u_{\langle script \rangle} = z_{\langle sc} + z_{scr} + z_{cri} + z_{rip} + z_{ipt} + z_{pt\rangle}$$

〈script〉에 대한 최종 임베딩 벡터는 n-gram 상태의 단어 벡터를 모두 반영한다.

3.4 시퀀셜 모델링 : Bi-GRU

문장은 단어가 배치 되는 순서에 따라 의미가 달라지기도 한다. 이렇게 순서 정보를 내포하고 있는 데이터를 시퀀셜 데이터(sequential data)라고 하며, 이 데이터들의 순서 정보를 통해 학습하는 것을 시퀀셜 모델링(sequential modeling)이라고 한다. 대표적인 시퀀셜 모델링 방식으로는 RNN, LSTM, GRU 등이 있다. GRU는 LSTM을 간소화 시킨 구조로써, 기존 시퀀스 기반 딥러닝의 그래디언트 소실 문제, 장기의존성 문제를 해결한 방식이다.

Bi-GRU는 GRU를 양방향으로 배치한 구조로써, 앞에서부터 바라보는 시퀀스 뿐만이 아니라 뒤에서 앞으로 가는 방향의 시퀀스도 학습하여 뒤의 단어를 봐야지만 인식할 수 있는 단어의 의미도 포착할 수 있도록 설계된 구조이다.

본 논문에서는 입력 페이로드를 토큰화 한 다음, 각 토큰들을 사전학습한 페이로드 임베딩 벡터로 변환하여 Bi-GRU에 입력한다.

3.5 주의 메커니즘

주의 메커니즘은 문장 내 같은 단어라도 문맥에 따라 중요도가 다르다는 점을 이용하여 문장의 의미를 구분하는데 중요한 단어를 추출하여 문장 벡터를 계산하는 기법이다.

페이로드는 그 길이가 무한정 늘어날 수 있기에 악성을 판별하는데 중요한 단어를 선별해내는 작업이 필요하다. 주의 메커니즘을 적용하면 페이로드의 공격 유형을 판별하는데 중요한 단어에 집중함으로써 관련 없는 노이즈를 필터링 할 수 있다.

본 논문에서는 이를 구현하기 위해 Bi-GRU의 출력값을 주의 계층의 입력으로 사용하여, Bi-GRU의 시간별 출력 중 중요한 정보가 있다고 판단되는 단어에 가중치를 주고 문맥 벡터를 추출하였다.

IV. 실험 및 평가

4.1 데이터 세트

제안하는 모델의 웹 공격 페이로드 분류 성능을 검증하기 위해 대표적인 웹 공격 3가지인 SQL 삽입 공격, 크로스 사이트 스크립팅, 명령 삽입(command injection)에 대한 다중 분류 실험을

Table 2. payload sample count

Class	Train	Test	Total
Benign	5692	1423	7,115
SQLi	6332	1583	7,915
XSS	4044	1011	5,055
Command injection	4608	1152	5,760
TOTAL	20,676	5,169	25,845

설계하였다. 우리는 각각을 구현하는 공격 도구 Sqlmap[21], commix[22], XSStrike[23]를 통해 공격을 생성한 뒤 페이로드를 추출하였다.

정상 샘플은 ISCX-2016-URL 데이터셋[24]에서 페이로드에 상응하는 파라미터만을 추출하여 구성하였다. 각 유형별 샘플의 개수는 Table 2와 같다.

4.2 실험 환경 및 주요 파라미터 설정

실험은 텐서플로우와 케라스를 주요 프레임워크로 사용하였고, Intel Core I7-7770, ram 32GB, gpu gtx1060 6GB 환경에서 수행하였다. 매개 변수를 조절하며 실험을 수행한 결과 주요 파라미터는 Table 3과 같이 설정하였다.

bpe vocab size는 토큰 단위를 구분하기 위해 학습한 토큰 집합의 수를 말한다. 학습을 위한 최소 토큰 개수는 66개이다. 본 논문에서는 일반적으로 자연어 처리에서 사용하는 32,000 토큰을 사용하였다.

word embedding dimension은 임베딩 사전학습을 통해 32000개의 각 토큰을 몇 차원으로 출력할 것인지 정하는 매개변수이다. 즉, 웹 공격 유형 구분에 필요한 토큰들 간의 분포 관계를 표현하는 주요 특징(feature)의 개수라고 볼 수 있다. 4개의 페이로드 유형을 구분하는 작업(task)은 자연어 처리 작업에 비해 적은 출력 차원이 필요할 것으로 예상되어, 1부터 200까지 차원 수를 변경하여 성능을 평가한 결과, 32차원 이상부터는 변화가 크지 않음을 확인하였다. 향후 다른 유형의 웹 공격 샘플을 추가하는 등 분류 작업(task)의 난이도에 따라 더 높은 차원으로 값을 설정할 필요가 있다.

Table 3. The parameter settings of our model.

Parameter	Setting
bpe vocab size	32000
word embedding dimension	32
word max length	64
bigru unit	128(64, 64)
drop out	bigru:0.5, attention:0.5
dense	4
loss function	categorical_crossentropy
optimizer	rmsprop
epoch	10
batch size	128

word max length는 토큰의 최대 길이를 말한다. 실험 데이터셋의 평균 토큰 개수 61개를 기준으로 64개로 구현하였다. 페이로드 토큰이 64개 미만일 경우 나머지는 0으로 패딩처리하여 Bi-GRU unit으로 전달 된다. Bi-GRU unit의 개수는 최대 토큰 길이에 따라 정방향, 역방향 각각 64개로 설정하였다.

과적합을 방지하기 위해 Bi-GRU층과 주의 계층에는 0.5의 드롭아웃을 설정하였다. 그 외에 덴스를 거친 뒤 소프트맥스로 웹 공격 유형을 분류하였다. loss 함수는 categorical_crossentropy이며, 최적화 함수는 rmsprop, 실험 에포크는 10, 배치사이즈는 128로 설정하였다.

제안하는 모델의 우수성을 증명하기 위해 기존 연구에서 제안한 모델과 성능을 비교하였다. 평가 지표로는 정확도, 정밀도, 회수율, F1점수를 사용하였으며, 4가지의 유형을 구분하는 다중 분류이기 때문에 각 유형별 결과의 macro 평균으로 계산하였다.

- model A[15] : 전통적인 웹 방화벽 modsecurity로 공격을 분류하였다. 공격 분류에 사용된 룰셋은 OWASP의 Core Rule Set이다.
- model B[8] : 특수문자 기준으로 토큰화를 수행하였고, word2vec 방식의 임베딩 학습을 거쳐, Bi-LSTM 신경망으로 구성하였다.
- model C[9] : model B에서 주의 메커니즘을 추가한 모델이다.
- proposed : BPE 기법으로 생성한 토큰 집합에 대해 임베딩 벡터를 추출하였고, 신경망은 Bi-GRU로 구성하였다.

Table 4는 각 모델과의 웹 공격 분류 성능을 비교한 결과 표이다. 제안하는 모델이 accuracy precision, recall, f1-score 네가지 항목에서 모두 기존 모델보다 높은 점수를 얻었다. model B와 model C의 비교를 통해 주의 메커니즘 적용에 따른 성능 향상을 확인할 수 있었고, 미등록 단어가 많은 페이로드를 대상으로 하는 분류 작업(Task)일 때, BPE 기법으로 구성한 단어 집합을 임베딩한 모델에서 가장 좋은 성능을 얻었다. 또한 loss 값이 빠르게 안정화 되는 것을 확인할 수 있었다.

Table 4. Performance comparison with other models

Model	Accuracy	Precision (macro avg)	Recall (macro avg)	F1-score (macro avg)
modsecurity [15]	0.9440	0.9522	0.9440	0.9441
word-word2vec-bilstm [8]	0.9861	0.9888	0.9865	0.9876
word-word2vec-bilstm+attention [9]	0.9918	0.9917	0.9918	0.9917
bpe-fasttext-bigru+attention (proposed)	0.9990	0.9993	0.9987	0.9990

V. 결 론

본 논문에서는 웹 공격 페이로드를 기반으로 웹 공격 유형을 분류하였다. BPE 기법으로 페이로드에 적합한 단어 집합을 추출하고, 패스트텍스트를 통해 페이로드만의 임베딩 벡터를 사전학습 하였다. 임베딩 벡터를 Bi-GRU 신경망과 주의 메커니즘을 기반으로 학습하여 웹 공격 유형을 분류하였다.

제안하는 모델의 우수성을 증명하기 위해 공격 도구로 생성한 웹 공격 페이로드 샘플로 데이터셋을 구축하고 비교 실험을 진행하였다. Bpe vocab size, word embedding deminsion, word max length 등 하이퍼파라미터를 조절하여 성능을 비교하고, 최적의 성능을 내는 파라미터 값을 추출하여 기존 연구 모델과 비교하였다. 그 결과 정확도, 정밀도, 회수율, F1점수 네가지 항목 모두에서 기존 모델보다 높은 점수를 획득하였다. 특히 정확도는 0.9990을 얻었다.

향후 더 다양한 웹 공격 유형의 샘플을 확보하여 배치 학습이 아닌 온라인 학습 형태로 구성한다면 기존 웹 방화벽의 패턴 탐지 모듈을 대체할 수 있을 것으로 예상된다.

References

[1] Symantec, "Internet Security Threat Report," volume 24, Feb. 2019.

[2] Hsiu-Chuan Huang, Zhi-Kai Zhang, Hao-Wen Cheng and Shiuhpyng Winston Shieh, "Web Application Security: Threats, Countermeasures, and Pitfalls," in Computer, vol. 50, no. 6, pp. 81-85, Jun. 2017.

[3] Hung Le, Quang Pham, Doyen Sahoo and Steven C.H. Hoi, "URLnet: Learning a URL representation with deep learning for malicious URL detection," arXiv preprint arXiv:1802.03162, Mar. 2018.

[4] Michiaki Ito and Hitoshi Iyatomi, "Web application firewall using character-level convolutional neural network," 2018 IEEE 14th International Colloquium on Signal Processing & Its Applications, pp. 103-106, Mar. 2018.

[5] Yang, Wenchuan, Wen Zuo, and Baojiang Cui, "Detecting malicious urls via a keyword-based convolutional gated-recurrent-unit neural network," IEEE Access 7, pp. 29891-29900, Feb. 2019.

[6] Saiyu Hao, Jun Long and Yingchuan Yang, "BL-IDS: Detecting Web Attacks Using Bi-LSTM Model Based on Deep Learning," International Conference on Security and Privacy in New Computing Environments, pp. 551-563, Apr. 2019.

[7] Ren, Fangli, Zhengwei Jiang, and Jian Liu, "A Bi-Directional LSTM Model with Attention for Malicious URL Detection," 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference, pp. 300-305, Dec. 2019.

[8] Sennrich, Rico, Barry Haddow, and Alexandra Birch, "Neural machine translation of rare words with subword units," arXiv preprint arXiv:1508.07909, Jun. 2016.

[9] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Herve Jegou and Tomas Mikolov, "Fasttext. zip: Compressing text classification models,"

arXiv preprint arXiv:1612.03651, Dec. 2016.
[10] Vandana Dwivedi, Himanshu Yadav and Anurag Jain, "SQLAS: Tool to detect and prevent attacks in php web applications," International Journal of Security Privacy and Trust Management, vol. 4, no. 1, pp. 21-30 Feb. 2015.
[11] Sahu, Divya Rishi, and Deepak Singh Tomar, "Analysis of web application code vulnerabilities using secure coding standards," Arabian Journal for Science and Engineering, vol. 42, no. 2, pp. 885-895, Feb. 2017.
[12] Jason Bau, Elie Bursztein, Divij Gupta and John Mitchell, "State of the art: Automated black-box web application vulnerability testing," 2010 IEEE Symposium on Security and Privacy, pp. 332-345, May. 2010.
[13] Priyank Bhojak, Kanu Patel, Vikram Agrawal and Vatsal Shah, "SQL Injection and XSS Vulnerability Detection in Web Application," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, no. 12, pp. 110-115, Dec. 2015.
[14] Roesch and Martin, "Snort: Lightweight intrusion detection for networks," Proceedings of the 13th USENIX conference on System administration, pp. 229-238, Nov. 1999.
[15] Modsecurity, "Open Source Web Application Firewall," https://modsecurity.org/
[16] Nilesh Khochare and B. B. Meshram, "Tool to Detect and Prevent Web Attacks," International Journal of Advanced Research in Computer Engineering & Technology, vol. 1, no. 4, pp. 375-378, 2012.
[17] Mikhail Zolotukhin, Timo Hämäläinen, Tero Kokkonen and Jarmo Siltanen, "Analysis of http requests for anomaly detection of web attacks," Proceedings of IEEE 12th International Conference on Dependable, Autonomic and Secure Computing, pp. 406-411, Aug. 2014.
[18] Abdelhamid Makiou. Youcef Begriche and Ahmed Serhrouchni, "Improving Web Application Firewalls to detect advanced SQL injection attacks," Information Assurance and Security 2014 10th International Conference, pp. 35-40, Nov. 2014.
[19] Zhang, Zhaoxuan, Roy George, and Khalil Shujaee, "Efficient detection of anomolous HTTP payloads in networks," SoutheastCon 2016, pp. 1-3, Mar. 2016.
[20] Baojiang Cui, Shanshan He, Xi Yao and Peilin Shi, "Malicious URL detection with feature extraction based on machine learning," International Journal of High Performance Computing and Networking vol. 12, no. 2, pp. 166-178, Sep. 2018.
[21] Damele, Bernardo, and M. Stampar, "Sqlmap," Online at http://sqlmap.org, 2012.
[22] Stasinopoulos, Anastasios, Christoforos Ntantogian and Christos Xenakis, "Commix: Detecting and exploiting command injection flaws," The Black hat Europe 2015, Nov. 2015.
[23] "Most advanced XSS scanner," https://github.com/s0md3v/XSStrike
[24] "URL dataset(ISCX-URL-2016)," https://www.unb.ca/cic/datasets/url-2016.html

〈저 자 소 개〉

김 연 수 (Yeonsu Kim) 정회원
2017년 8월: 전남대학교 소프트웨어공학과 졸업
2017년~현재: 전남대학교 정보보안협동과정 석박사통합과정
〈관심분야〉 인공지능 보안, 사이버 위협 인텔리전스, 침입탐지, 취약점 분석

고 영 훈 (Younghun Ko) 정회원
2015년~현재: 전남대학교 컴퓨터정보통신공학과 재학
〈관심분야〉 정보보안, 시스템보안, 사물인터넷(IoT), 프로그래밍언어론

엄 익 채 (Ieck-Chae Euom) 종신회원
2003년 8월: 전남대학교 컴퓨터정보학부 학사 졸업
2015년 2월: 한국과학기술원 소프트웨어대학원 석사 졸업
2019년 2월: 전남대학교 정보보안협동과정 박사 졸업
2003년~2007년: LG이노텍 신뢰성Lab 근무
2007년~2019년: 한전KDN 보안컨설팅팀 근무
2019년 10월~현재: 전남대학교 시스템보안연구센터 조교수
〈관심분야〉 제어시스템보안, 스마트그리드 보안, 원자력 보안, 취약점 분석, 차세대인프라 보안, 스마트시티·공장 보안, AI기반 이상징후 탐지, 지능형 보안

김 경 백 (Kyungbaek Kim) 종신회원
1999년 2월: 한국과학기술원 전기 및 전자공학과 학사 졸업
2001년 2월: 한국과학기술원 전기 및 전자공학과 석사 졸업
2007년 2월: 한국과학기술원 전기 및 전자공학과 박사 졸업
2007년~2011년: University of California Irvine, 박사 후 연구원
2012년~2015년: 전남대학교 전자컴퓨터공학부 조교수
2016년~현재: 전남대학교 전자컴퓨터공학부 부교수
〈관심분야〉 분산시스템, 소프트웨어 정의 인프라스트럭쳐, 빅데이터 플랫폼, 소셜 네트워킹 시스템, 블락체인, AI기반 CPS