A new DDoS attacks intrusion detection model based on deep learning for cybersecurity

May 2022

Computers & Security (Elsevier) (rank 5)

یکسری پیش‌پردازش روی داده های دیتابیس CICDDOS2019 انجام داده است تا دقت تشخیص در هنگام آموزش و ارزیابی بالا برود. و مدل‌های مختلفی یادگیری ماشین بر مبنای Deep Neural Networks, Convolutional Neural Networks, Long Short Time Memory از نظر کارایی تشخیص و بلادرنگ بودن بررسی شده است.

چون دیتاست مورداستفاده حجیم است، عملات نرمال‌سازی زیر به ترتیب روی آن انجام شده است:

- داده‌های بی اهمیت یعنی صفر، تهی و تکراری‌ها حذف شده است
- چندین عملات استخراج ویژگی روی آن انجام شده‌است و یک زیردیتاست با تکنیک Info Gain Attribute Evaluation تعریف شده‌است. (این که کدام متود را انتخاب کنیم بستگی به نتایج تست‌ها دارد) .
- یک مدل یادگیری عمیق کانولوشن برای بلوک‌های آغازی طراحی شده‌است. همچنین ویژگی‌های ورودی به بایت برای پردازش بهتر با مقادیر ماکزیمم بیشتر تبدیل شده‌است.

سامانه‌های تشخیص نفوذ: فعالیت‌های مخربانه یا نشت اطلاعاتی را به ادمین شبکه اطلاع می‌دهند یا به صورت متمرکز در Security Information and Event Management system جمع آوری و گزارش می‌شود. SIEM از داده‌های موجود که از چندین منبع هستند برای آنالیز استفاده می‌کند. انواع بسیار مختلفی از آنتی ویروس‌ها تا سیستم‌های مانیتور ساده دارند. IDS ها به دو دسته HIDS (مانیتور فایل‌های سیستم عامل) و NIDS (مانیتور ترافیک شبکه) تقسیم می‌شوند. همچنین بر اساس متود شناسایی به دو دسته anomaly-based و signature-based (سریع اما ناکارآمد در حملات جدید) تقسیم می‌شوند. از روش‌های یادگیری ماشین در anomaly-based ها استفاده می‌شود تا تهدیدات ناشناخته را شناسایی کند. بهتر هستند و مدل در برابر کانفیگ اپلکیشن و سخت افزار آموزش داده می‌شود و مغایرت‌ها را شناسایی کنند

حملات منع خدمت توزیع‌شده: دسترسی کاربران را با محدودیت مواجه کرده و به خصوص در سیستم‌های ابری که سرویس (حتی به طور مثال سرویس‌های حفاظتی) ارایه می‌دهند، بسیار مهم است. جداسازی حملات DDoS

از افزایش/کاهش عملکرد آنی و متوسط نیاز به فناوری و تخصص مناسب دارد. زیرساخت شبکه نیز باید به گونه مناسبی طراحی شده باشد

شبکه عصبی عمیق: Feed-Forward Neural Network ها الگوریتم‌های مهم یادگیری ماشین هستند. زمانی که از استخراج ویژگی‌ها به کمک لایه‌های شبکه عصبی و دیتاست‌های عظیم به منظور فراهم‌آوردن generality استفاده میکند، DNN نامیده می‌شود. CNN مدل پیشرفته‌تر آن هست که از یکسری عملیات کانولوشن استفاده می‌کند و روی ورودی‌ها انجام می‌دهد و وزن نودهای هر لایه را نیز می‌تواند تغییر دهد. نوع دیگر شبکه‌های عصبی Recurrent Neural Network ها هستند که LSTM از محبوب‌ترین آنهاست و همچنین در به یاد سپردن وابستگی‌های بلند مدت داده‌ها بهتر ازRNN عمل می‌کند. یک واحد از آن از سه دروازه ورودی، خروجی و فراموشی تشکیل شده است.

روش پیشنهادی:

پیش پردازش:

در کل دیتاست ۵۰ میلیون رکورد از ۱۳ نوع حمله موجود می‌باشد.

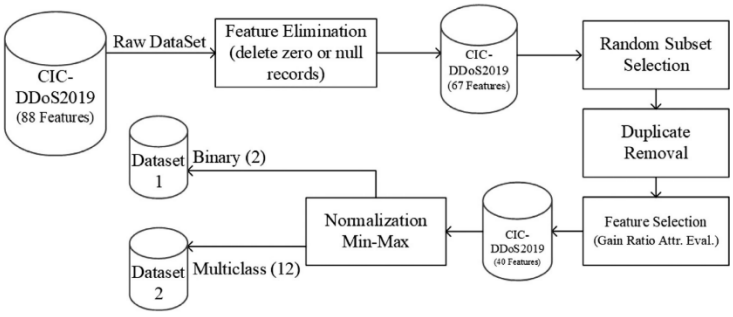

Fig. 5. Preprocessing of raw dataset for training and testing.

Info Gain Attribute Evaluation :

الگوریتم های استخراج ویژگی مبتنی بر فیلترینگ، یک مرتب سازی انجام می‌دهند (بر اساس یک مقدارآستانه‌ای مقادیر پایین‌تر از آن را حذف می‌کنند) تا ویژگی‌های سودمند مستقل از هر الگوریتم را شناسایی کنند. این الگوریتم‌ها سرعت بالا و همچنین بار محاسباتی کمتری دارند. از الگوریتم Info Gain attribute Evaluation بدین منظور استفاده شده است.

Table 3
CIC-DDoS 2019 selected 40 features using Info Gain Attr. Eval.

Features	Gain Value	Description	Min-Max Range
Packet Length Mean	2.587840	Mean length of a packet	0 - 4023.945
Average Packet Size	2.586634	Average size of packet	0 - 4023.779
Max Packet Length	2.577663	Maximum length of a packet	0 - 37,960
Avg Fwd Segment Size	2.570492	Average size observed in the forward direction	0 - 3015.91
Fwd Packet Length Mean	2.569949	Mean size of packet in forward direction	0 - 3015.91
Fwd Packet Length Max	2.569601	Maximum size of packet in forward direction	0 - 32,120
Fwd Packet Length Min	2.560382	Minimum size of packet in forward direction	0 - 2021
Subflow Fwd Bytes	2.559334	The average number of bytes in a sub flow in the forward direction	0 - 15,266,416
Total Length of Fwd Packets	2.559334	Total size of packet in forward direction	0 - 15,266,416
Min Packet Length	2.555892	Minimum length of a packet	0 - 1472
Source Port	1.470888	Source Port	0 - 65,534
act_data_pkt_fwd	1.154258	Act data in forward direction	0 - 5043
Flow Duration	1.134145	Duration of the flow in Microsecond	0 - 119,999,986
Fwd Packets s	1.099233	Number of forward packets per second	0 - 4,000,000
Flow IAT Mean	1.093613	Mean time between two packets sent in the flow	0 - 39307794.333
Flow IAT Max	1.071818	Maximum time between two packets sent in the flow	0 - 119,954,412
Fwd IAT Total	1.041866	Total time between two packets sent in the forward direction	0 - 119,999,986
Fwd IAT Mean	1.001968	Mean time between two packets sent in the forward direction	0 - 30505593.333
Fwd IAT Max	0.98594	Maximum time between two packets sent in the forward direction	0 - 119,954,412
Flow IAT Std	0.971157	Standard deviation time between two packets sent in the forward direction	0 - 68082925.446
Fwd Header Length	0.910916	Total bytes used for headers in the forward direction	-21,254,3795000 - 129,536
Fwd IAT Std	0.849177	Standard deviation time between two packets sent in the forward direction	0 - 56020637.667
Init_Win_bytes_forward	0.71427	The total number of bytes sent in initial window in the forward direction	-1 - 65,535
Total Fwd Packets	0.650225	Total packets in the forward direction	1 - 100,129
Subflow Fwd Packets	0.650225	The average number of packets in a sub flow in the forward direction	1 - 100,129
Protocol	0.603988	Protocol	0 - 17
ACK Flag Count	0.573889	Number of packets with ACK	0 - 1
Packet Length Variance	0.484914	Variance length of a packet	0 - 43778893.573
Packet Length Std	0.484914	Standard deviation length of a packet	0 - 6616.562
min_seg_size_forward	0.483161	Minimum segment size observed in the forward direction	-1408237563 - 1480
Fwd Packet Length Std	0.401855	Standard deviation size of packet in forward direction	0 - 221.556
Fwd IAT Min	0.369574	Minimum time between two packets sent in the forward direction	0 - 15,407,833
Flow IAT Min	0.369067	Minimum time between two packets sent in the flow	0 - 15,407,833
Bwd Packets s	0.247753	Number of backward packets per second	0 - 4,000,000
Destination Port	0.214434	Destination Port	0 - 65,535
Bwd Header Length	0.208336	Total bytes used for headers in the backward direction	-2125437950 - 147,280
Total Backward Packets	0.199864	Total packets in the backward direction	0 - 4602
Subflow Bwd Packets	0.199864	The average number of packets in a sub flow in the backward direction	0 - 4602
Bwd IAT Max	0.199205	Maximum time between two packets sent in the backward direction	0 - 118,159,264
Bwd IAT Mean	0.19727	Mean time between two packets sent in the backward direction	0 - 58,961,594

Table 4
Number of instances for multiclass classification using 88 and 40 features.

Class (Label)	Num. of Instances (88)	Unique Instances (88)	Num. of Instances (40)	Unique Instances (40)
WebDDoS	439	439	439	438
BENIGN	56,863	56,025	56,025	55,939
UDP_Lag	366,461	366,452	56,025	56,023
NTP	1,202,642	1,202,518	56,025	56,024
Syn	1,582,289	1,582,049	56,025	56,025
SSDP	2,610,611	2,610,580	56,025	56,025
UDP	3,134,645	3,134,598	56,025	56,025
NetBIOS	4,093,279	3,923,062	56,025	56,025
MSSQL	4,522,492	4,519,697	56,025	56,025
DNS	5,071,011	4,935,090	56,025	56,025
SNMP	5,159,870	5,048,055	56,025	56,025
TFTP	20,082,580	18,410,498	56,025	56,025
Total	47,883,182	16,267,889	616,714	616,624

۲۰درصد از دیتاست نهایی به صورت تصادفی به منظور آموزش مدل و بقیه در ارزیابی مورد استفاده قرار می‌گیرد.

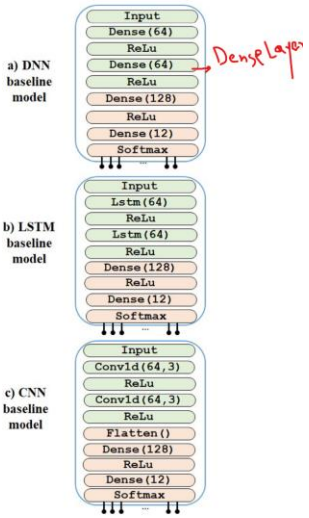

Fig. 7. Baseline models using DNN, CNN and LSTM layers.

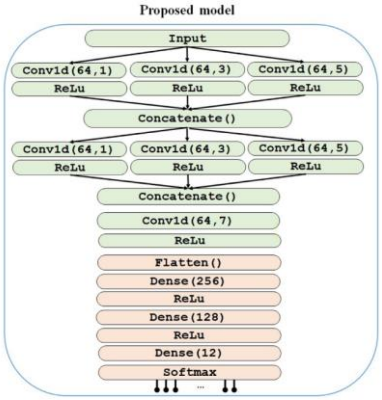

Fig. 8. Proposed model using CNN layers with inception like connections.

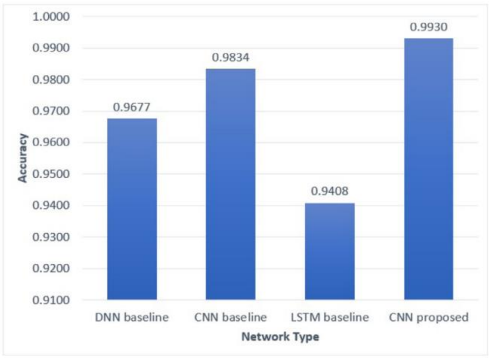

Fig. 9. Comparison of the accuracy evaluations.

Table 14
The comparison of the proposed model to the state of the art results.

Author	Year	Area	Method	Accuracy	Classification Type	Remove Duplcation
Jia et al. (2020)	2020	IoT application	LSTM	98.9%	Binary	No
Li et al. (2020)	2020	IoT application	LSTM	Not Available	Multi (7)	No
de Assis et al. (2020)	2020	SDN environments in IoT networks	CNN	95.4%	Binary	No
Alamri and Thayananthan (2020)	2020	SDN environments in IoT networks	Extreme gradient boosting algorithm	91.26%	Multi (11)	No
Pontes et al. (2021)	2021	Not Available	Energy-based flow classifier	98.1%	Binary	No
Assis et al. (2021a)	2021	SDN environments	Gated Recurrent	99%	Binary	No
Javeed et al. (2021)	2021	SDN environments in IoT networks	LSTM and GRU	99.74%	Multi (8)	No
Nie et al. (2021)	2021	IoT application	Generative adversarial network	98.35%	Multi (8)	No
Ferrag et al. (2021)	2021	Cloud Computing,NFV, and SDN	CNN, RNN, DNN	95.12%, 94.88%, 93.88%	Multi (13)	Yes
Our Model (Multi)	2021	IoT application, Cloud Computing	CNN	**99.30**%	Multi (12)	Yes
Our Model (Binary)	2021	IoT application, Cloud Computing	CNN	**99.99**%	Binary	Yes

ارزیابی: از میان مدل‌های موجود، مدل مبتنی بر CNN دقت بالاتری را ارایه می‌دهد. زمان استنتاج از داده‌های آموزشی در مقایسه با روش‌های دیگر با وجود مجموعه داده‌های بزرگتر، بهتر است و در نتیجه مخصوص بلادرنگ است. وعلاوه برآن حملات را به صورت باینری و هم چندتایی با دقت بالایی کلاس بندی می‌کند.