Software Engineering Conference Russia
14-15 ноября, 2019. Санкт-Петербург

Подход к анализу больших данных в кибербезопасности

Кусакина Надежда
Самарский Государственный Технический Университет

Информационный мир в цифрах

~5 млрд
пользователей Интернета

1,8млрд
вебсайтов

~ 4 устройства
у каждого пользователя подключены к Интернету

~33
уязвимости на одно веб-приложение

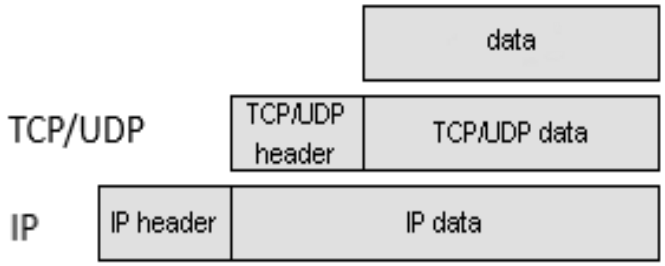
>3 млрд
пользователей соцсетей

~ 24 000 зараженных мобильных приложений **блокируется ежедневно**

>280 млрд
e-mails отправляется ежедневно

~20% Trojan-Dropper

Объект исследования



Инспекция пакета

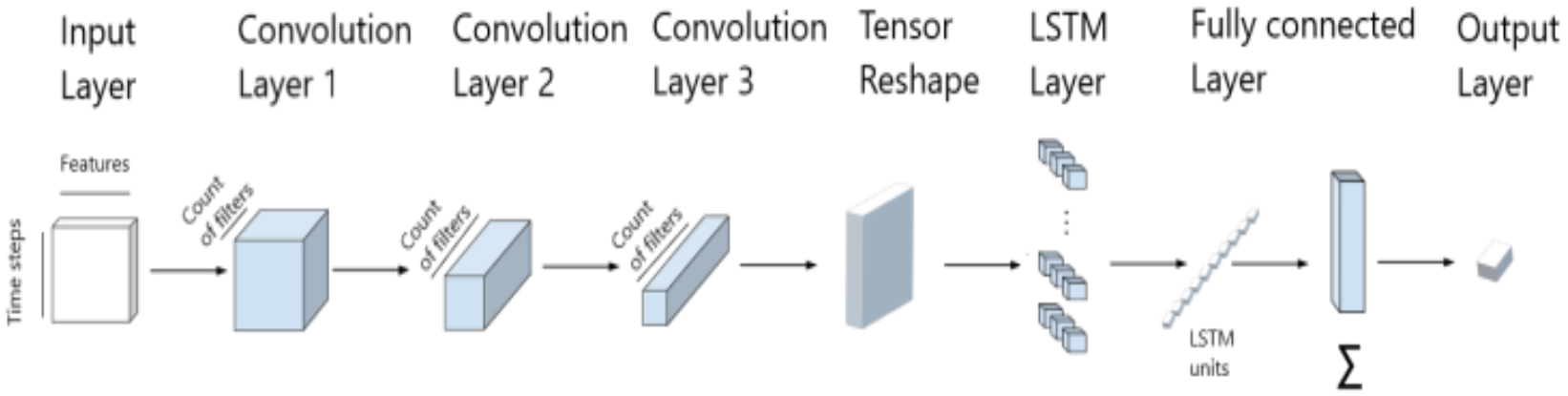
Client IP	IPc1	IPc2	IPc3	IPc4	IPc5
Server IP	IPd1	IPd2	IPd3	IPd1	IPd3
Bits/sec Total	10 159 000	10 466 000	10 514 000	10 749 000	11 565 000
Packets Total	5 781 000	61 601 000	61 299 000	63 795 000	67 848 000
Observed Connections	16	13	29	24	34
Packets/sec Total	1 338	1 426	1 419	1 477	1 571

Таблица сетевых взаимодействий

Анализ работ по направлению

Семейство методов ML		Решаемые задачи			
		Выявление атак	Выявление бот-нет	Выявление вредоносных программ	Выявление спама и фишинга
SL	Обучение с учителем	RF NB SVM LR HMM KNN SNN	RF NB SVM LR KNN SNN	RF NB SVM LR HMM KNN SNN	RF NB SVM LR KNN SNN
	Обучение без учителя	Кластеризация Ассоциация	Кластеризация Ассоциация	Кластеризация Ассоциация	Кластеризация Ассоциация
DP	Обучение с учителем	RNN	RNN	FNN CNN RNN	-
	Обучение без учителя	DBN SAE	-	DBN SAE	DBN SAE

Подготовка



Модель реализована
на языке **Python** с использованием **Keras** и бекендом **Tensorflow**

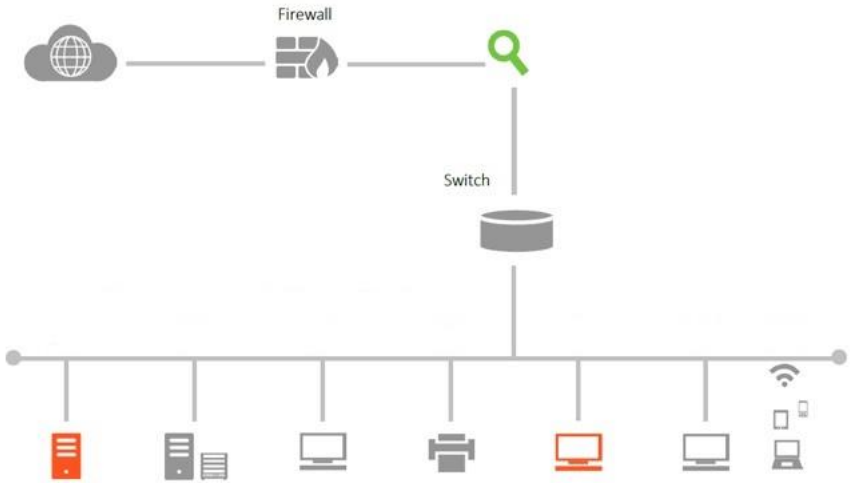
https://github.com/NaMihKu/SECR19

Сбор и разметка данных

SrcAdd	DstAdd	SrcPort	DstPort	LofDT	FLAGS	PayloadBytesCount	WindowSize	TimeRec	TimeTransfer
169.254.204.33	192.168.16.55	3372	80	62	SYN	0	8760	0.000000	0.000000
192.168.16.55	169.254.204.33	80	3372	62	SYN, ACK	0	5840	0.911310	0.911310
169.254.204.33	192.168.16.55	3372	80	54	ACK	0	9660	0.911310	0.911310
192.168.16.55	169.254.204.33	80	3372	54	ACK	0	6432	1.472116	1.472116
192.168.16.55	169.254.204.33	80	3372	1434	ACK	1380	6432	1.682419	1.682419
169.254.204.33	192.168.16.55	3372	80	54	ACK	0	9660	1.812606	1.812606
192.168.16.55	169.254.204.33	80	3372	1434	ACK	1380	6423	1.812606	1.812606
169.254.204.33	192.168.16.55	3372	80	54	ACK	0	9660	2.012894	2.012894
192.168.16.55	169.254.204.33	80	3372	1434	ACK	1380	6432	2.443513	2.443513
192.168.16.55	169.254.204.33	80	3372	1434	PSH, ACK	1380	6432	2.553672	2.553672
169.254.204.33	192.168.16.55	3372	80	54	ACK	0	9660	2.553672	2.553672
192.168.16.55	169.254.204.33	80	3372	1434	ACK	1380	6432	2.633787	2.633787
169.254.204.33	192.168.16.55	3372	80	54	ACK	0	9660	2.814046	2.814046
192.168.16.55	169.254.204.33	80	3372	1434	ACK	1380	6432	2.894161	2.894161
169.254.204.33	192.168.16.55	3372	80	54	ACK	0	9660	3.014334	3.014334
192.168.16.55	169.254.204.33	80	3372	1434	ACK	1380	6432	3.374852	3.374852
192.168.16.55	169.254.204.33	80	3372	1434	PSH, ACK	1380	6432	3.495025	3.495025
169.254.204.33	192.168.16.55	3372	80	54	ACK	0	9660	3.495025	3.495025
192.168.16.55	169.254.204.33	80	3372	1434	ACK	1380	6432	3.635227	3.635227
169.254.204.33	192.168.16.55	3372	80	54	ACK	0	9660	3.815486	3.815486

Сбор и разметка данных

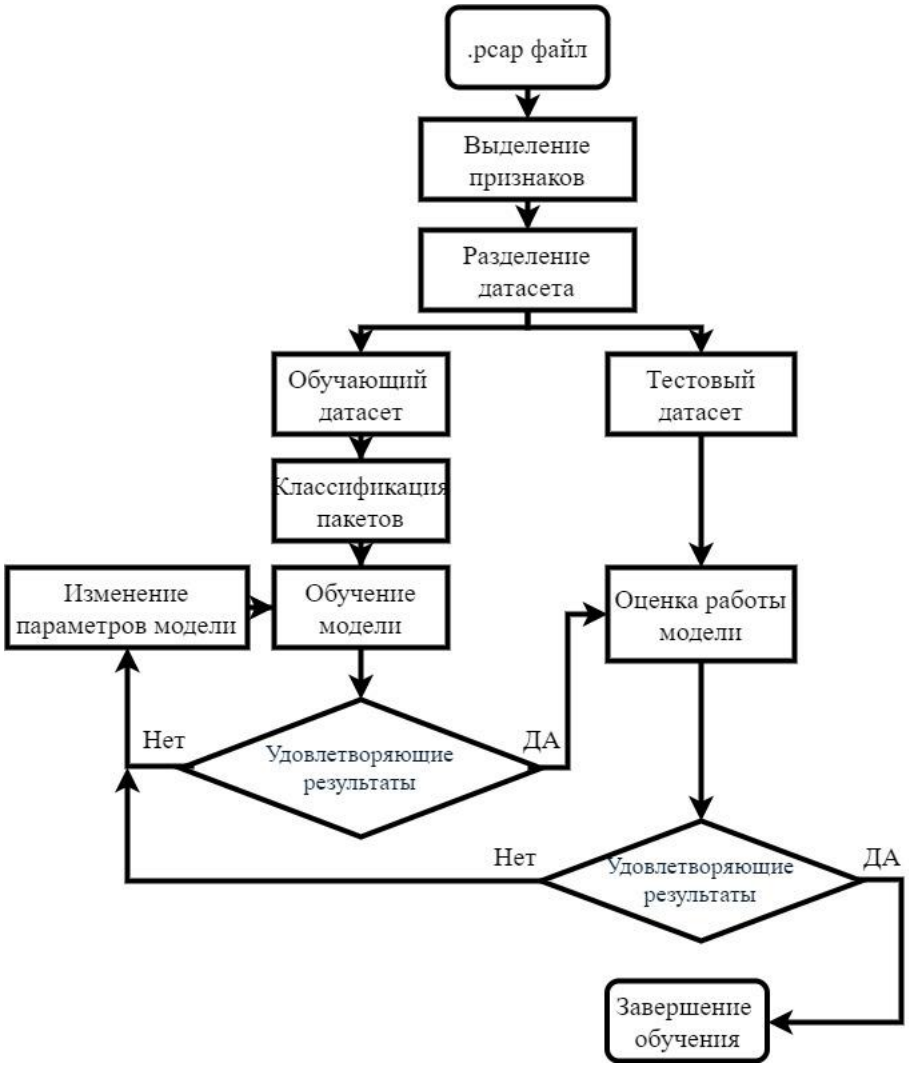
Схема сбора данных



Выявляемые уязвимости

1. DoS/DDoS
2. DNS-tunneling
3. SSL stripping
4. CSRF
5. SQL injection
6. PHP injection
7. Worm infection

Обучение нейронной сети



Оценка нейронной сети

Основными метриками при оценке алгоритмов machine learning являются: точность, полнота и F-мера.

$$\text{Precision}_c = \frac{Ac,c}{\sum_{i=1}^{8} Ac,i}$$

$$\text{Recall}_c = \frac{Ac,c}{\sum_{i=1}^{8} Ai,c}$$

В нашем случае мульти-классовой классификации точность (Precision) и полнота (Recall) рассчитываются с использованием матрицы неточностей (confusion matrix), размерность которой N на N, где N = 8 — количеству выходных классов.

$$F = (\beta^2 + 1)\frac{\text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}$$

Применение

В результате работы по созданию модели гибридной нейронной сети мы получили классификатор, который в дальнейшем планируется использовать для выявления паразитного трафика на атакуемые системы.

Предполагается интеграция разработанной модели ИНС с IDS.
Подобные системы обладают высокой скоростью работы и не требуют постоянного обновления сигнатур, так как используют свойство самообучения.

Благодарю за внимание

Кусакина Надежда

СамГТУ

nadyakusakina@yandex.ru