УДК 004

РАЗРАБОТКА СИСТЕМЫ АНАЛИЗА ИСХОДНОГО КОДА ВРЕДОНОСНЫХ ИСПОЛНЯЕМЫХ ФАЙЛОВ НА ОСНОВЕ LSTM СЕТЕЙ С ИСПОЛЬЗОВАНИЕМ МАТРИЦ СТАТИЧЕСКОГО АНАЛИЗА

Чан Дык Мань (Национальный исследовательский университет ИТМО)
Донг Суан Тхань (Национальный исследовательский университет ИТМО),
Динь Нгок Туан (Национальный исследовательский университет ИТМО),
Научный руководитель– доцент, кандидат технических наук Таранов С.В. (Национальный исследовательский университет ИТМО)

Аннотация. Наряду с развитием компьютерных сетей более количество важной информации обрабатывается на компьютерных, которые являются целью многих злоумышленников.По данным gs.statcounter.com в декабре 2021г. самой популярной операционной системой в мире является Windows(74.8%). Поэтому разработка быстрых и точных алгоритмов на операционной системе Windows принесет большую пользу для использователей.

Введение. Для обнаружения вредоносной программы необходимо использоваться различные виды анализа: по байтам исходного кода, по инцидентам, которые создает программа, по воздействия на операционную систему и т.д.. С повышением алгоритмов обнаружения вредоносных программ эти приложения становятся все более изощренными, чтобы обмануть существующие решения. Глубокое обучение применяется для повышения адаптивности и точности существующих алгоритмов обнаружения программного обеспечения.

Основная часть. Алгоритм основан на встраивании последовательностей двоичных инструкций в виде последовательностей. Затем выполните классификацию этих последовательностей с использованием алгоритма LSTM сети. Мы рассматриваем эту проблему как классификацию двоичного кода и применяем алгоритмы обработки нейтрального языка для обработки исполняемых файлов. Работа с большим количеством кода двоичных файлов, извлеченных из исполняемого исходного кода, была нашей самой большой проблемой. Мы разделяем этот код на группы, имеющие сходное поведение в системе, тем уменьшая сложность алгоритма. Затем используйте метод Bag of Word для быстрого представления исполняемых файлов.

Выводы. Мы классифицировали 9000 различных популярных приложений, добились точности до 97%, f1-мера достиг 0,94. Эта точность намного выше, чем у статических методов. Высокое среднее время чтения является недостатком алгоритма, поскольку время чтения на каждый исполняемый файл составляет до 10 с из-за большого количества извлеченных двоичных данных. В дальнейшем мы постараемся сократить время выполнения алгоритма, чтобы алгоритм выполнялся быстрее.

Чан Дык Мань (автор)

Таранов С.В. (научный руководитель)