الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

المدرسة العليا للإعلام الآلي - 08 ماي 1945 – بسيدي بلعباس
Ecole Supérieure en Informatique
-08 Mai 1945- Sidi Bel Abbes

MEMOIRE

En vue de l'obtention du diplôme d'**ingénieur d'état**

Filière : **Informatique**

Spécialité : **Ingénierie des Systèmes Informatiques (ISI)**

Thème

Détection de Malwares Android par les Techniques d'Apprentissage Automatique

---

Présenté par **:**

- Mr MAHNANE Ilyes

Soutenu le : **25/07/2019** Devant le jury composé de :

-	M FERAOUN Mohamed Kamel	Président
-	M BELFEDHAL Alaa Eddine	Encadreur
-	M AZZA Mohamed	Examinateur
-	M BOUKLI Hacene Mohamed	Examinateur

Année Universitaire : 2018 / 2019

Résumé

La plupart du temps passé sur un téléphone est passé à utiliser des applications : consulter ses emails, vérifier son fil d'actualité, lire les journaux ...etc. Avec l'avènement d'Android comme premier système d'exploitation mobile, c'est toute notre information qui est centralisée en un seul et même endroit, le smartphone. L'OS de Google n'enthousiasme pas que ses utilisateurs, les concepteurs de malwares aussi sont contents. Envoi de SMS à des numéros surtaxés et collecte de données personnelles sont parmi les moyens lucratifs utilisés par ces malfaiteurs. Android n'est pas infaillible, chaque malware qui y prolifère en est la preuve.

Notre solution propose une analyse dynamique comportementale des applications susceptibles d'être une source de malignité, par simple envoi de l'application vers un serveur distant via une interface assez conviviale et simple à utiliser. L'application va être installée et exécutée, simulant ainsi une utilisation humaine. Par la suite, les appels système générés par le noyau Unix sont collectés, traités et fournis au modèle de réseau de neurones qui pourra prédire si l'application analysée est maligne ou bénigne.

L'efficacité de notre méthode est due à l'utilisation de l'apprentissage automatique. Un ensemble d'applications malignes et bénignes est constitué à partir duquel un modèle discriminant est construit. Nous avons utilisé pour l'apprentissage un réseau de neurones, et plus précisément le réseau de neurones convolutionnel (Convolutional Neural Network en anglais) connu sous le pseudonyme de CNN. Se basant sur des données d'entrée matricielles et se caractérisant par son aspect à faire évoluer ses propres filtres tout seul.

Notre méthode se voit novatrice du fait qu'on utilise une nouvelle représentation matricielle des appels système collectés et servant d'entrées au model CNN. Une représentation moins couteuse en espace mémoire et permettant ainsi d'accélérer le processus d'apprentissage et d'accroitre le taux d'exactitude lors de la détection.

Mots clés: Android, Analyse dynamique, Analyse statique, Détection de Malware, Apprentissage profond, Réseau de neurones convolutionnel.

Abstract

Most of the time spent using a smartphone is spent using applications: consulting emails, checking news feed, reading newspapers. With the accession of Android to the first place of most used mobile operating systems, all our information is centralized in one place, the smartphone. Google's OS not only enthuses its users but also makes malware developers happy. Sending SMS messages to premium line numbers and collecting data are examples of lucrative ways used by these criminals. Android is not infallible, each malware targeting it is proving that.

Our solution proposes a behavioral dynamic analysis of the applications likely to be a source of malignancy. The application will be sent towards a distant server through a user-friendly and simple to use interface. It will be installed and executed with a simulation of a human use. After execution, system calls generated by the unix kernel are collected, processed, and provided to the neural network model that will be used to predict whether the analyzed applications are malware or goodware.

Efficacy is due to the use of machine learning. A large dataset of benign and malignant applications is constituted. From that dataset a discriminant model is built. We used for learning a neural network, and more precisely the convolutional neural network (CNN). Based on matrix input data and characterized by its aspect of changing its own filters by itself.

Our method is innovative because it uses a new matrix representation of collected system calls and input to the CNN model, a less expensive representation in memory space and therefore accelerate the process of learning and increase the accuracy rate when detecting.

Keywords: Android, Dynamic analysis, Static analysis, Malware Detection, Deep learning, Convolutional neural network.

ملخص

معظم الوقت المنقظي في استعمال الهاتف يذهب في استخدام التطبيقات: استشارة البريد الإلكتروني، مراجعة الأخبار، قراءة الصحف... إلخ. مع ظهور نظام أندرويد كأول نظام تشغيل استعمالا للهاتف المحمول، إنها جميع بياناتنا التي أصبحت متمركزة في مكان واحد والذي هو الهاتف الذكي. نظام التشغيل الخاص بشركة ڤوڤل لايحمس فقط مستخدميه، مصممو البرامج الضارة سعداء أيظا. إرسال رسائل نصية قصيرة إلى أرقام الأسعار المميزة وجمع البيانات الشخصية هي من بين الوسائل المربحة التي يستخدمها هؤلاء الأشرار. أندرويد ليس معصوما، فكل التطبيقات الظارة التي تنتشر دليل على ذلك.

في هذه المذكرة، نحن نقترح حلا للتمكن من اكتشاف التطبيقات الخبيثة باستعمال تحليل ديناميكي سلوكي للتطبيقات المثبتة على الهاتف أو اللوحات الالكترونية وذلك عن طريق إرسال التطبيق نحو جهاز كمبيوتر (خادم) باستخدام واجهة سهلة وبسيطة الاستخدام. يتم تثبيت التطبيق في الجهاز وتنفيذه، ومن ثم محاكات لاستخدام الإنسان للتطبيق عن طريق توليد أحداث عشوائية كلمس الشاشة، ضغط على أزرار الواجهة، في نفس الوقت، يتم جمع استدعاءات النظام التي أُنشئت من قبل نواة إينكس ومعالجتها. تقدم بعد ذلك إلى نموذج الشبكة العصبية المستخدمة للتنبؤ إذا ما كانت التطبيقات التي تم تحليلها ومعالجتها خبيثة أو جيدة.

كفاءة طريقتنا هي نتيجة لإستخدام التعلم الآلي. تم تجميع مجموعة من التطبيقات الظارة والحميدة والتي على أساسها يتم تشكيل نموذج التمييز. اعتمدنا من أجل التعلم على الشبكة العصبية وباتحديد على الشبكة العصبية التلافيفية، وكذلك بالإعتماد على بيانات الإدخال على شكل مصفوفة وتتميز الشبكة العصبية التلافيفية بقدرتها على تطوير المرشحات الخاصة بها لوحدها.

تعتبر الطريقة التي اتبعناها مبتكرة وجديدة وذلك باستخدامنا لتمثيل جديد على شكل مصفوفات لنداءات النظام المُجمعة والمستعملة كبيانات دخول لنموذج الشبكة العصبية، هذا التمثيل هو أقل تكلفة في مساحة الذاكرة وبالتالي يساعد على تسريع عملية التعلم والكشف عن البرامج الضارة.

كلمات دلالية: أندرويد، التحليل الديناميكي، التحليل الثابت، كشف البرامج الضارة، التعلم العميق، الشبكة العصبية التلافيفية.