University of Mosul
College of Computer Sciences and Mathematics
Department of Computer Sciences

Designing Deep Learning based Network Intrusion Detection System for Software Defined Network

Mohammed Hamid Abdulraheem Abdullah

Ph.D./Thesis

Computer Sciences

Supervised By

Assistant Professor

Dr. Najla Badie Ibraheem Al-Dabagh

2020 A.D **1442 A.H**

Abstract

In response to the growing cyber-attacks against governments and commercial companies globally, Network Intrusion Detection Systems (NIDS) have been rapidly developed in academia and industry. The recent development focuses on leveraging a new network architecture, namely, the Software-Defined Network (SDN), to implement NIDS with Deep Learning (DL) approaches to enhance network monitoring and security. The SDN is an emerging architecture that decouples the network control and forwarding planes. The network controller is programmable, supporting straightforward network policy enforcement and simplified network management. These SDN features facilitate innovative applications, dictating a new networking paradigm capable of implementing NIDS.

An anomaly DL-based Network Intrusion Detection and Prevention System (NIDPS) for SDN is designed and implemented in this thesis. This system is characterized by lightness, scalability, and overcoming implementation limitations. This thesis's research was conducted in three directions: The first direction is to analyze the Canadian Institute of Cyber Security Intrusion Detection System dataset (CICIDS2017) to train DL models. This dataset is new and was not extensively analyzed. Therefore the research involves identifying dataset defects, exploring dataset features, and studying the effect of scaling functions on the classification result. Some shortcomings are found in the dataset, and solutions were implemented to resolve these defects.

The second direction is to train four DL models using all dataset features for multi-class classification. The models are Deep Dense Layer, 1-Dimensional Convolutional Neural Networks (1-D CNN), 2-D CNN, Long Short Term Memory-Recurrent Neural Networks (LSTM-RNN). The average 5 fold Cross-Validation (CV) evaluation metrics for multi-class classification of the models were ranged between (92.4-97.8)% for balanced accuracy, (96-97)% for precision, (92.4-97.8)% for recall, (94-97)% for F1-score, (0.7-1.0)% for losses and (0.15-0.2)% for False Positive Rate (FPR). Besides, the Area Under the Curve (AUC) metric of the PR-curve was calculated per class.

The third direction is designing and deploying the NIDPS in SDN. The system resides outside the controller, so the system does not add processing loads

to the controller or overwhelms the controller-switch link with captured packets. The system does not depend on the limited features of the OpenFlow protocol's statistical message to obtain traffic flow features. Instead, the sampling Flow (sFlow) protocol and the sFlowtool were used to create a remote packets capture service (traffic collector) for network traffic analysis and intrusion detection. This approach adds a lightweight load to the network bandwidth and scalable to several switches compared to the Port Mirroring approach, which overloads network bandwidth. It is the first attempt to use the sFlow in feature extraction from raw network traffic. The CICFlowMeter (a packet capturing and traffic flow generator) is incorporated into the traffic collector of the NIDPS to extract flow records from live network traffic. The CICFlowMeter provides the same flow features that trained the deployed DL model to predict flow classes. The designed system was deployed in an emulated SDN and tested with real attacks.

جامعة الموصل
كلية علوم الحاسوب والرياضيات
قسم علوم الحاسوب

تصميم نظام كشف اختراق الشبكة المستند إلى التعلم العميق للشبكة المعرفة بالبرمجيات

محمد حامد عبد الرحيم عبد الله

أطروحة دكتوراه

علوم الحاسوب

بإشراف

أ. م. د. نجلاء بديع إبراهيم الدباغ

1442هجري

2020 ميلادي

المستخلص

استجابة للهجمات الإلكترونية المتزايدة ضد الحكومات والشركات التجارية على مستوى العالم، تم تطوير أنظمة الكشف عن اختراق الشبكات (Network Intrusion Detection Systems) بسرعة في الأوساط الأكاديمية والصناعية. يركز التطوير الأخير على الاستفادة من بنية الشبكة الجديدة، وهي الشبكة المعرفة بالبرمجيات (Software defined Network)، لتنفيذ أنظمة الكشف عن اختراق الشبكات مع نهج التعلم العميق (Deep Learning) لتعزيز امن ومراقبة الشبكة. الشبكة المعرفة بالبرمجيات هي بنية ناشئة للشبكات الحديثة، حيث تفصل بين مستوى التحكم في الشبكة ومستوى إعادة توجيه البيانات، بحيث يمكن برمجة وحدة تحكم الشبكة بشكل مباشر، ودعم تطبيق سياسات الشبكة مركزيا وتسهيل إدارة الشبكة. تتيح هذه الميزة الى تسهيل كتابة التطبيقات المبتكرة، مما يفرض نموذجًا جديدًا للشبكات قادرًا على تنفيذ أنظمة كشف ومنع اختراق الشبكات.

في هذه الأطروحة، تم تصميم وتنفيذ نظام كشف ومنع اختراق الشبكة (Network Intrusion Detection and Prevention System) الخاص بالشبكة المعرفة بالبرمجيات المبني على التعلم العميق. تميز هذا النظام بخفة الحمل على الشبكة وقابلية التوسع والتغلب على قيود التنفيذ. تم العمل في هذه الأطروحة بثلاثة اتجاهات: الاتجاه الأول هو تحليل مجموعة بيانات كشف التطفل الخاصة بالمعهد الكندي لأمن الفضاء الالكتروني (CICIDS2017) لغرض تدريب نماذج التعلم العميق. تلك البيانات حديثة ولم يتم تحليلها على نطاق واسع، لذلك تضمن البحث دراسة تحليلية شاملة لمجموعة البيانات وتحديد عيوبها، ودراسة سماتها ودراسة تأثير دوال التطبيع على نتيجة التصنيف. تم العثور على بعض العيوب في مجموعة البيانات ونفذت الحلول لتجاوز تلك العيوب.

الاتجاه الثاني هو تدريب اربعة نماذج للتعلم العميق باستخدام جميع سمات مجموعة البيانات لتصنيف متعدد الفئات وتشمل: انموذج طبقة كثيفة عميقة (Deep Dense Layer)، وانموذج شبكة عصبية تلافيفيه أحادية البعد (1-D Convolutional Neural Network)، وانموذج شبكة عصبية تلافيفيه ثنائية البعد (2-D Convolutional Neural Network) وانموذج شبكة عصبية متكررة بذاكرة ذات المدى القصير والطويل (Recurrent Neural Network-Long Short Term Memory). تراوحت نتائج التقييم للنماذج المدربة باستخدام التحقق المتقاطع بخمسة طيات (5-Fold Cross-Validation) للتصنيف متعدد الفئات لمقاييس تقييم المصنف بين (92.4-97.8) ٪ للدقة المتوازنة (Balanced Accuracy)، و (96-97) ٪ للأحكام (Precision)، و (92.4-97.8) ٪ للاسترجاع (Recall)، و (94-97) ٪ لدرجة ف1 (F1-score)، و (0.7-1.0) ٪ للخسائر (Losses)، و (0.15-0.2) % لمعدل الإنذارات الكاذبة (False Positive Rate). بالإضافة إلى ذلك، تم حساب مقياس المساحة تحت المنحني لمنحني الاحكام-الاسترجاع (Precision-Recall curve) لكل فئة.

الاتجاه الثالث هو تصميم ونشر نظام كشف ومنع اختراق الشبكة على الشبكة المعرفة بالبرمجيات. يقع النظام خارج وحدة التحكم حيث لا يضيف اية أعباء معالجة إلى وحدة التحكم أو يغرق رابط المتحكم-المحول (Controller-Switch link) بحزم البيانات الملتقطة من الشبكة. لا يعتمد النظام على الرسالة الإحصائية

المحدودة السمات لبروتوكول التحكم بأجهزة الشبكة (OpenFlow) للحصول على سمات تدفق مرور الشبكة. بدلا من ذلك استخدم بروتوكول التقاط عينات حزم تدفق الشبكة (sFlow) (المستخدم لمراقبة معدل حركة مرور الشبكة على منافذ المحول) وأداة هذا البروتوكول (sFlowtool) لإنشاء جامع لحزم مرور الشبكة بعيد، لغرض تحليل الحزم وكشف اختراق الشبكة. يضيف هذا الأسلوب حملًا خفيفًا إلى عرض النطاق الترددي للشبكة وقابل للتوسيع إلى العديد من المحولات (switches) مقارنةً بجمع حزم مرور الشبكة باستخدام مرآة المنفذ (port mirroring) الذي يضيف حمولة زائدة كبيرة إلى عرض النطاق الترددي للشبكة إذا استخدم على نطاق واسع في الشبكة. هذه هي المحاولة الأولى لاستخدام بروتوكول التقاط عينات حزم مرور الشبكة في استخراج سمات تدفق مرور الشبكة من الحزم الأولية. تم دمج ملتقط الحزم ومولد سجلات تدفق مرور الشبكة التابع للمعهد الكندي لأمن الفضاء الالكتروني (CICFlowMeter) في جامع مرور الشبكة التابع لنظام كشف ومنع اختراق الشبكة لاستخراج سجلات تدفق مرور الشبكة الحية. يوفر ملتقط الحزم ومولد سجلات التدفق نفس ميزات التدفق التي تم تدريب انموذج التعلم العميق المنشور في نظام كشف الاختراق للتنبؤ بفئة التدفق. تم نشر نظام كشف ومنع اختراق الشبكة في بيئة تحاكي الشبكة المعرفة بالبرمجيات واختبر بتسليط هجمات حقيقية.