Extracting Interpretable Rules from Deep Models Across Coalitions

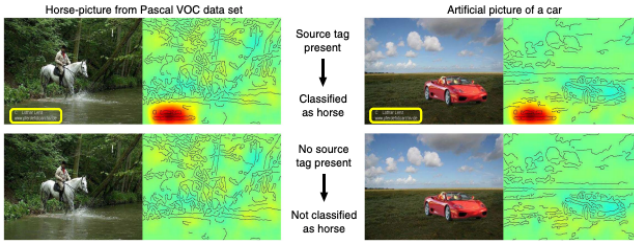
Franck Le[*], Kin Leung[§], Konstantinos Poularakis[†], Leandros Tassiulas[†], Paul Yu[•]

[*] IBM US
[§] Imperial College
[†] Yale University
[•] U.S. Army CCDC Army Research Laboratory

***Abstract*— Deep models can achieve high performance and offer a promising approach for SDCs to handle dynamicity efficiently. However, their ``black-box'' nature where no explanation behind the prediction outcome is provided, and their high inference latencies can become major obstacles to their adoption in operational SDCs. To address these limitations, we developed a novel algorithm to extract the knowledge from attention-based model through concise symbolic forms. This demo will demonstrate the ability to extract knowledge from different models, each trained in a separate coalition, to merge the knowledge, and to effectively deploy the newly derived interpretable rules in a new coalition.**

I. DESCRIPTION

Deep learning solutions have demonstrated promising performance results for SDCs to handle dynamicity efficiently (e.g., network fragmentation), and to ultimately provide a robust network infrastructure for distributed analytics tasks (Task 7.1) (e.g., [13, 14]). However, deep solutions present limitations that can limit their adoption in actual operational settings: First, deep models are notoriously difficult to interpret. Operating as "black-boxes", they do not provide any explanation behind their outcomes, leading to questions on whether the learned strategies are valid, and generalizable. For example, the Figure below extracted from a recent study [1] shows that an image classifier achieving high classification performance in fact relies on a tag on the bottom right corner to identify "horses".



In other words, the image classifier did not properly learn the concept of horses. Removing that tag results in the image not longer to be classified as horses. And, inserting that tag in a picture of a car causes the modified image to be incorrectly classified as "horses". This example illustrates the importance of, and the need for interpretability. Second, the inference latency of deep models can be significantly larger than those of pattern matching based rules traditionally used in network traffic analysis. Our experiments show that the inference latency of feedforward networks, Long Short-Term Memory (LSTM), and Bayesian Neural Networks can be 4 to 6 orders of magnitude larger than those of pattern matching based rules, increasing the packet processing time which is critical to achieve the required throughput. To address these limitations, we developed a novel rule extraction algorithm to extract the knowledge from attention-based models through concise symbolic forms, and more specifically, if-then clauses. We focus on attention-based models because they currently are the state-of-the-art architecture for sequence learning.

This demo will illustrate the ability to extract interpretable rules from different deep attention-based models, each trained in a separate coalition, and to merge them for deployment in a new environment.

II. SCENARIO

For simplicity, we consider the task of IoT device identification: Due to their poor security [1–3], IoT devices have become a prime target for attacks, e.g., as an ingress points to broader IT coalition infrastructure. As such, network administrators need tools to detect potentially vulnerable IoT devices [5-12].

We assume each coalition train an attention-based model to identify IoT devices, and detect vulnerable ones. We apply the developed algorithm to extract the knowledge from each attention-based model under the form of if-then rules, merge them, and demonstrate the effectiveness of the combined knowledge for identifying IoT devices in a new coalition environment.

More specifically, to emulate each coalition environment, we will adopt an independent packet trace (e.g., [5]). Each trace would consist of tens to hundreds of IoT devices, and for this specific demo, we will focus on the DNS traffic, i.e., the DNS queries that each device submits over a configurable period of time (e.g., N hours). Previous studies also focused on the DNS traffic as it is typically in cleartext even when the subsequent connections are encrypted, and DNS traffic from IoT devices have been found to exhibit distinct patterns. More specifically, IoT devices tend to query few domains [5], and periodically connect to their home networks [8] (e.g., to check for software updates or report metrics).

- First, we will train an attention-based model for each dataset. A model can infer different properties of IoT devices including manufacturer (e.g., Netgear), type (camera), model (Arlo), etc.
- Second, we will extract the knowledge from each model under the form of rules, and expose them to show their interpretability. As an example, a rule could be of the form *"if a device queries the domain or connects to the*

server netgear.updates.com, then the device manufacturer is Netgear".

- Third, we will merge the rules and evaluate their effectiveness on an independent dataset. To quantify the effectiveness, we will compute and report the precision, recall, and f1-score of the resulting rules.

III. Demo Requirements

For this demo we will require a large monitor with high resolution, suitable for displaying our outputs. Table space for the monitor and a laptop will be needed as well as suitable space for one poste. WiFi connectivity to access internet resources and power sockets for the screen and up to three laptops are also needed.

Acknowledgment

This research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence under Agreement Number W911NF-16-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

References

[1] S. Lapuschkin, S. W¨aldchen, A. Binder, G. Montavon,W. Samek, and K.-R. M¨uller. Unmasking clever hans predictors and assessing what machines really learn. volume 10, 02 2019

[2] L. Hautala, "Why it was so easy to hack the cameras that took down the web," inCNETSecurity, October 2016

[3] D. Palmer, "175,000 IoT cameras can be remotely hacked thanks to flaw, says security re-searcher," inZDNet, July 2017

[4] T. Yu, V. Sekar, S. Seshan, Y. Agarwal, and C. Xu, "Handling a trillion (unfixable) flaws ona billion devices: Rethinking network security for the internet-of-things," inProceedings ofthe 14th ACM Workshop on Hot Topics in Networks, HotNets-XIV, 2015

[5] A. Sivanathan, D. Sherratt, H. H. Gharakheili, A. Radford, C. Wijenayake, A. Vishwanath,and V. Sivaraman, "Characterizing and Classifying IoT Traffic in Smart Cities and Cam-puses," inIEEE Infocom Workshop Smart Cities and Urban Computing, 2017

[6] M. Miettinen, S. Marchal, I. Hafeez, T. Frassetto, N. Asokan, A.-R. Sadeghi, and S. Tarkoma,"Iot sentinel demo: Automated device-type identification for security enforcement in iot," inIEEE ICDCS, 2017

[7] Y. Meidan, M. Bohadana, A. Shabtai, J. Guarnizo, M. Ochoa, N. O. Tippenhauer, andY. Elovici, "Profiliot: A machine learning approach for iot device identification based onnetwork traffic analysis," 04 2017

[8] H. Guo and J. Heidemann, "Ip-based iot device detection," inProceedings of the 2018 Work-shop on IoT Security and Privacy, IoT Samp;P '18, (New York, NY, USA), p. 36–42, Asso-ciation for Computing Machinery, 2018

[9] J. Ortiz, C. Crawford, and F. Le, "Devicemien: Network device behavior modeling for iden-tifying unknown iot devices," inProceedings of the International Conference on Internet of Things Design and Implementation, IoTDI '19, (New York, NY, USA), p. 106–117, Associ-ation for Computing Machinery, 2019

[10] A. Bremler-Barr, H. Levy, and Z. Yakhini, "Iot or not: Identifying iot devices in a shorttimescale," 2019

[11] M. H. Mazhar and Z. Shafiq, "Characterizing smart home iot traffic in the wild," 2020

[12] D. Y. Huang, N. Apthorpe, G. Acar, F. Li, and N. Feamster, "Iot inspector: Crowdsourcinglabeled network traffic from smart home devices at scale," 2019

[13] Z. Zhang, L. Ma, K. Poularakis, K.K. Leung and L. Wu, "DQ Scheduler: Deep Reinforcement Learning Based Controller Synchronization in Distributed SDN," IEEE ICC 2019

[14] Joao Reis, Miguel Rocha, Truong Khoa Phan, David Griffin, Franck Le, Miguel Rio, "Deep Neural Networks for Network Routing", International Joint Conference on Neural Networks (IJCNN), 2019