2ND WORKSHOP
"MACHINE LEARNING & NETWORKING" (MaLeNe)
PROCEEDINGS

SEPTEMBER 4, 2023

CO-LOCATED WITH
THE 5TH INTERNATIONAL CONFERENCE ON NETWORKED SYSTEMS (NETSYS 2023)
POTSDAM, GERMANY

Impact of Adaptive Packet Sampling on ML-based DDoS Detection

Samuel Kopmann* and Martina Zitterbart*†
*Institute of Telematics, Karlsruhe Institute of Technology, Karlsruhe, Germany
†KASTEL - Security Research Labs, Karlsruhe, Germany
{samuel.kopmann, martina.zitterbart}@kit.edu

***Abstract*—Traffic monitoring can react to changing data rates by adapting the fraction of inspected packets (sampling rate). In this work, we investigate the resilience of a sampling rate agnostic machine-learning DDoS detector against a packet sampling rate adapting to changing data rates. We show with real-world data that an adapting packet sampling rate worsens the DDoS attack detection accuracy. To counter performance reduction, we employ upsampling and multi-rate training, showing that the resilience against a changing packet sampling rate improves.**

***Index Terms*—DDoS detection, traffic monitoring, packet sampling, supervised machine learning**

I. Introduction

Supervised machine learning (ML) detectors rely on traffic monitoring in two ways. First, they are trained offline with traffic obtained from past monitoring. Second, they process traffic data obtained from current monitoring during deployment. ML detectors perform well if the characteristics of monitored data are similar during training and deployment.

Traffic monitoring can become a bottleneck at high data rates, e.g., 100^{+}Gbit/s. To prevent the monitoring from becoming the bottleneck during traffic bursts, as potentially caused by volumetric Distributed Denial of Service (DDoS) attacks, it can be throttled by limiting the fraction of inspected packets, i.e., packet sampling [1], [2].

However, packet sampling can reduce the stress on the monitoring infrastructure, but it skews observed traffic characteristics, as not all packets are inspected, and traffic information is lost. This leads to dissimilarities between training and deployment traffic data and causes a decrease in the performance of the ML-based detection.

One primary goal for attack detection considering monitoring resource efficiency is not to lose detection accuracy when packet sampling becomes necessary.

Contribution

We evaluate the impact of adaptive packet sampling rates on a DDoS detector, i.e., HollywooDDoS [4], trained in a supervisory manner. We show that HollywooDDoS, which is sampling rate agnostic, cannot preserve high-quality detection when monitoring is performed with sampling rates that have not been covered during the offline training process. We evaluate two countermeasures and show with real-world data that they enable the use of HollywooDDoS with monitoring applying adaptive packet sampling rates.

II. Background and Approach

HollywooDDoS is a DDoS detection approach representing arriving network traffic as two-dimensional images classified by a Convolutional Neural Network (CNN). Monitoring is performed in two dimensions, the source and the destination IP address space. A grid of source-to-destination IP subnet pairs is created and arriving packets are counted per subnet pair. All arriving packets account for image creation during monitoring.

When applying packet sampling, fewer packets are inspected and counter values per subnet pair are potentially smaller, breaking normalization during deployment and decreasing the detection accuracy. We counter this detection accuracy decrease with two methods, namely upsampling and multi-rate training.

Upsampling

We assume that the traffic distribution in the grid of subnet pairs is still correctly captured if enough packets arrive but at a lower traffic volume according to the sampling rate. Therefore, to compensate for the non-inspected traffic, every counter value in the grid of subnet pairs is multiplied with the inverse sampling rate. This provides an estimated reconstruction of the real traffic distribution without packet sampling, ensuring that the normalization does not break.

Upsampling is performed during deployment. Therefore, deployed ML models do not have to be retrained and can be further utilized.

Multi-rate Training

In contrast to upsampling, multi-rate training is not performed during deployment but changes the training process by creating multiple training data sets according to different sampling rates. Therefore, for every sampling rate potentially occurring during deployment, an individual data set is created. The ML model is trained on all data sets, leading to one model that generalizes well across all sampling rates.

III. Evaluation

Training data is composed of real-world attack traffic from CAIDA [3] and benign traffic from MAWI [5]. All data sets are balanced, i.e., they contain the same amount of benign samples as attack samples. Following best practices in ML, we split the dataset into training (70%) and test (30%) set, and present the

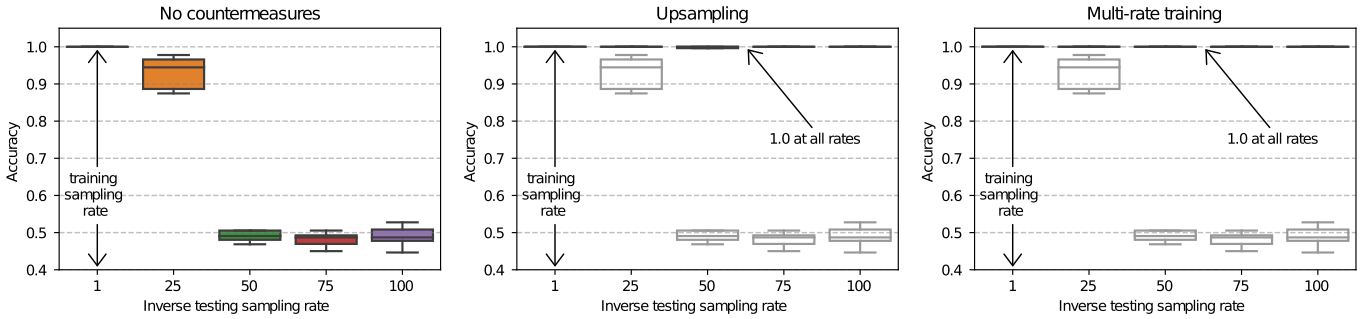

Fig. 1. The accuracy, with and without countermeasures, for different packet sampling rates resulting from a model with training sampling rate 1.0

results from testing. Each experiment has been conducted 20 times. Each result represents the median accuracy of all runs.

Fig. 1 provides accuracy results for an ML model trained with only one packet sampling rate (1.0) with countermeasures (two figures at the right) and without countermeasures (left figure). Tested sampling rates range from 1.0 to 100.

A. Impact without Countermeasures

From the left figure of Fig. 1, it is observable that the ML model performs well on the test data with the same sampling rate (1.0) as the training data, resulting in an accuracy of 100 percent. However, when the sampling rate decreases, the accuracy also decreases. For the sampling rate $\frac{1}{25}$ the accuracy drops to 94 percent and for sampling rates smaller than $\frac{1}{25}$ the accuracy drops to 50 percent, constituting a detection as good as guessing. Therefore, if HollywooDDoS would be deployed with adaptive sampling rates, only trained with data obtained from the sampling rate 1.0, detection results would not be reliable. To maintain high detection accuracy across all packet sampling rates, one model per sampling rate needs to be deployed. This is infeasible if sampling rates are not discrete.

B. Upsampling

The center figure of Fig.1 presents results derived from the same model as before, but the test data sets have been changed using upsampling by scaling the input with the inverse sampling rate before feeding them into the ML model. Previous results without countermeasures are carried over from the left figure to illustrate the improvement. It is observable that the model, only trained on data obtained from monitoring with a sampling rate 1.0, is now able to perfectly classify images obtained from monitoring with all tested sampling rates.

A significant advantage of upsampling is that one trained ML model can be used for the classification of images at multiple sampling rates. There is no need to change or swap the trained model during deployment because image scaling is performed as part of the monitoring.

C. Multi-rate Traning

The right figure of Fig.1 presents results applying the multi-rate training. Multi-rate training interferes the training process by training the ML model with data obtained from monitoring with all tested sampling rates. The goal is to train one model that generalizes well across all sampling rates, without the need for swapping models or rescaling images when using adaptive packet sampling.

Results show that HollywooDDoS trained with multiple sampling rates can perfectly classify images obtained from monitoring at all tested sampling rates, achieving an accuracy of 100 percent. Although the model training is more complex with multi-rate training than using upsampling, no adaptations to the monitoring are required during deployment in exchange.

IV. Conclusion

We outlined that adaptive packet sampling reduces the detection quality during deployment for the supervised ML-based DDoS detection approach HollywooDDoS. We evaluated two countermeasures, namely upsampling and multi-rate training. Upsampling rescales monitoring data according to the inverse packet sampling rate during deployment, while multi-rate training covers all packet sampling rates during the ML model training. We showed the effectiveness of both countermeasures with real-world data from CAIDA and MAWI achieving 100 percent accuracy across all sampling rates.

V. Acknowledgements

This work was funded by the German Federal Ministry of Education and Research (BMBF), RefNr. 16KIS1142K. This work was supported by funding of the Helmholtz Association (HGF) through the Kastel Security Research Labs (POF structure 46.23.01: Methods for Engineering Secure Systems).

References

[1] G. Androulidakis and S. Papavassiliou. Intelligent flow-based sampling for effective network anomaly detection. In *IEEE GLOBECOM 2007 - IEEE Global Telecommunications Conference*, pages 1948–1953, 2007.

[2] Baek-Young Choi, Jaesung Park, and Zhi-Li Zhang. Adaptive packet sampling for accurate and scalable flow measurement. In *IEEE Global Telecommunications Conference, 2004. GLOBECOM '04.*, volume 3, pages 1448–1452 Vol.3, 2004.

[3] Center for Applied Internet Data Analysis. The caida ucsd ddos attack. https://www.caida.org/catalog/datasets/ddos-20070804_dataset, 2007.

[4] Samuel Kopmann, Hauke Heseding, and Martina Zitterbart. Hollywood-dos: Detecting volumetric attacks in moving images of network traffic. In *2022 IEEE 47th Conference on Local Computer Networks (LCN)*, pages 90–97, 2022.

[5] MAWI. Backbone trace. http://mawi.wide.ad.jp/mawi/samplepoint-F/2019/201909011400.html, 2019.