UEBA AUGUST 2, 2018

Using Deep Learning to Reduce the Security Risks of Unmanaged Devices

DEREK LIN - CHIEF DATA SCIENTIST

Whether it's an employee's cell phone, a contractor's iPad, or a virtual machine (VM) created by a compromised account for malicious purposes, any unmanaged device on your network should be considered a security risk. Whether legitimate, or unauthorized, or rogue, such unmanaged devices create blind spots. They're an open attack surface accessible to hackers—with consequences that can include compromised intellectual property, leaked data, and destruction of brand.

Reducing security risks from such unknown physical or virtual devices is a multi-step, multi-faceted effort. Organizations should start with a comprehensive device management and security policy that includes NAC and MDM tools to help track known devices. But no such tool provides 100 percent visibility. The critical next step is to immediately recognize and identify the presence of any unmanaged devices on your network, which is where data science can play a role.

Ideally, in large networks, devices are named using official naming standards. But in reality, devices in your network can have unofficial naming conventions such as those from organizational units outside of your control policy, or those originating from a legacy system or domain. Device outliers—particularly those having arbitrary names without official or unofficial naming peers—are most likely unmanaged and unauthorized. They require analyst attention, but can your security teams immediately identify these outlier devices?

Examining the use of deep learning in cybersecurity

Before we look at how deep learning can be used to identify outlier devices, let's examine its use in cybersecurity overall. This will put our use case in context, as well as debunk the misconception that deep learning is a better, "deeper" form of machine learning—when in fact, there are a number of cases where deep learning is not the right solution because of the requirements or the data conditions.

cybersecurity applications. One premise is if deep learning can teach machines to win a board game, then it should prevent cybersecurity threats. Indeed, while deep learning is being used in malware binary analysis for uncovering malicious executables such as with phishing emails, it's necessary to examine if it's the right solution for the right problem.

Unless specific criteria are needed, many user and entity behavior analytics (UEBA) use cases aren't a good fit for deep learning for these reasons:

- There aren't the requisite volumes of labeled malicious events spread across logs to support supervised learning for tasks to classify whether an event or user session is legitimate or malicious.
- UEBA outcomes must be easily and rapidly understood by analysts and flagged alerts must be self-explanatory. However, algorithmic deep learning analysis happens on the back end and is "black box" in nature. This is the biggest drawback of deep learning.
- While automated, latent feature learning directly from the raw data is a major advantage in image or NLP applications (where the data type is *homogeneous* such as with pixels or words), contrast this application with the difficulty in performing deep learning analysis on raw security logs comprised of *heterogeneous* data types across data sources (such as data fields with different semantic meanings such as timestamps, user IDs, IP addresses, or database queries). While feature engineering is possible (which is typically done by the data scientist and consists of a manual process of crafting explicit statistical indicators), this defeats the automated, latent feature learning advantage of deep learning.

But suitable UEBA use cases for deep learning do exist, as long as they have the proper characteristics and requirements. Let's take a look at how deep learning can detect high-risk devices that don't conform with known or unknown naming conventions.

Using deep learning to identify rogue devices on your network

When identifying outlier devices, the key to using deep learning is to discover naming patterns found in all network-present devices, and to flag those that don't conform. Unless all devices follow strict naming protocols where an outlier is trivially flagged through simple regular expressions, deep learning is particularly well suited for the use case of learning naming patterns.

First, the names of all observed network devices are obtained from logs stored in the SIEM. Volumes of known malicious events aren't needed since this is an anomaly detection problem; explaining the outcome isn't a requirement because analysts can readily tell if a flagged device is unusual.

Typically, natural language processing leverages deep learning to learn word relationships within documents. We can use the same tools to evaluate character relationships in device names such as determining how certain letters, placed consecutively or apart, should occur within the named device pool.

The deep learning tool is the Long-Short-Term Memory Network (LSTM). (For an explanation, see the footnote below.) As characters are sequentially fed into the deep learning tool, it learns the inherent structural information in the population device names. It's "deep" in its learning because, for each device name, it can capture the long-span relationship of a given character to other characters several positions away.

During the process of learning, each character sequence of a device name string is transformed into and represented by a multi-dimensional vector. We can then visualize the devices in a plot (t-SNE) where each device is presented by a dot in a two-dimensional graph. Figure 1 shows tens of thousands of devices from a network, each represented by a point. Devices that are densely clustered together have a common naming pattern. Figure 2 shows a zoomed in

2 Insider Threats: How to Stop the Most Common and Damaging Security Risk You Face

3 2018 State of the SOC Report

4 5 Best Practices for Your Incident Response Plan

5 How Criminals Can Build a "Web Dossier" from Your Browser

view of an example cluster where each point is labeled with the corresponding device name. We can see the shared naming structure, with a common prefix.

Flagged outlier device names are expected to be away from the clusters. These flagged device names go through additional behavior-based indicators to remove false positives. This method generates a list that is small enough for analysts to review.

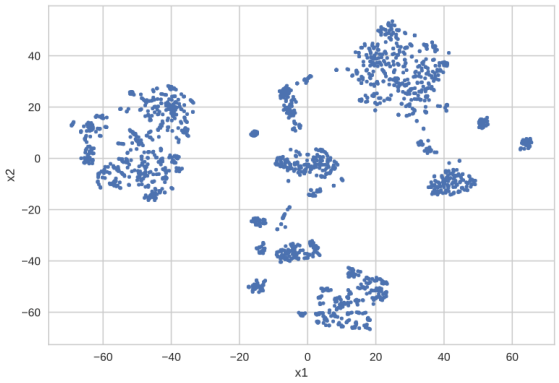

Figure 1: Visualization of device names represented in two-dimensional space

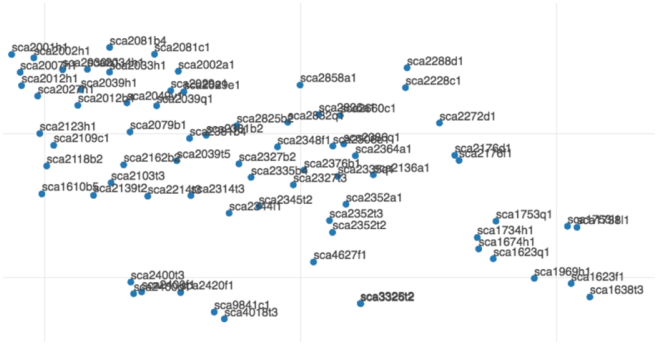

Figure 2: A zoomed in view of a cluster in Figure 1 showing the actual device names

With this approach, we've identified anomalous device names that should be brought to the attention of security analysts. Now they can track down and evaluate the devices and place them under an IT control policy. Any false positives are well controlled by post-filtering or other simple methods.

Detecting anomalously named devices is a good application of deep learning, while it isn't the silver bullet of device management. That said, this deep learning use case offers a new, cost-effective tool in the overall strategy of device management and monitoring.

Footnote

For those readers who are familiar with the data science of deep learning, LSTM provides the learning network structure. Specifically, a sequence-to-sequence LSTM is used to measure the reconstruction errors of input device name character strings. Device names with high reconstruction errors are anomalous since they cannot be explained by the learned device naming structures.

#DATA SCIENCE

DEREK LIN
CHIEF DATA SCIENTIST

More like this

If you'd like to see more content like this, visit the Exabeam Blog

Explore more

INFORMATION SECURITY

Operation Aurora – 2010's Major Breach by Chinese Hackers

JANUARY 8, 2019 — TIM MATTHEWS

Exabeam's Cybersecurity History Review: Read about Operation Aurora and the series of cyberattacks in 2010 conducted by the Elderwood Group based in Beijing, China, with ties to the People's Liberation Army.

INFORMATION SECURITY

Exabeam's Top Cybersecurity Blog Posts of 2018

JANUARY 2, 2019 — MARITZA MARIE DUBEC

2018 was a memorable year for cybersecurity. Millions of people were impacted as we saw more companies hit by megabreaches—from a major hotel chain to a social media platform used by billions. Here are our top 10 blog posts that had the biggest readership and were the most noteworthy.

UEBA

User Behavior Analytics (UBA/UEBA): The Key to Uncovering Insider and Unknown Security Threats

JANUARY 2, 2019 — ORION CASSETTO

Learn about UBA technology, and its extension UEBA (User Entity Behavior Analytics), how it works, and which threats it uncovers that no other tool can see.

CONTACT

2 Waters Park Dr., Suite 200
San Mateo, CA 94403

1.844.EXABEAM

info@exabeam.com