## Application: Bluedon AI Firewall

According to the design principle of modulation, we regard the AI-based maldoc detector as an independent detection module that can be easily integrated in security products, such as a next-generation firewall. An interesting question is: How Bluedon manages to apply AI technology seamlessly to a 30-year-old security product?

In the current development of network and gateway security products, the capability of performing malicious file scanning effectively and efficiently at layer7 (the network application layer) is the international standard. The industry has strict demands on this product feature; that is, a good detection module should include (1) millisecond latency for single-file detection, and (2) 99% accuracy while maintaining a FP rate of less than 0.01%.

The reason behind the demand for low latency is obvious: Since a security module is sequentially placed into a working pipeline, high latency will lead to an increased packet drop rate and occasionally data loss. This is strictly forbidden for security devices. In the past few years, with the rapid development of malware, the former industry best practice—the pattern -matching engine—has gradually been used less and less in mainstream security applications. There are two main reasons for this: (1) In order to meet the requirement of high detection accuracy, a large number of security analysts are needed for pattern writing, which is a manual process that has not scaled at all; and (2) considering the fast-growing size of core databases, the time for core operation–pattern matching grows exponentially. These conditions have inspired us to discover a better engine rooted in AI.

By 2018, we managed to integrate our AI maldoc detector into a firewall, in the hope of replacing the old engine. Although both the old and new engines are static-analysis-based engines, the improvement of shifting from the old to the new AI engine has been tremendous. First, the AI engine does not need to be updated frequently because it can detect previously unencountered malware effectively for years. According to our experiment results, the average update frequency for our AI engine is half a year, which is longer if compared to the 2-week period of the old pattern-matching engine. Moreover, the AI-based engine enjoys low resource consumption in execution. According to our study, during the phase of model prediction, the AI engine can only consume as much as a one-third of CPU and 50% of memory. The CPU portion of consumption is mainly due to computations, such as feature extraction and confidence-score computation. The memory portion of consumption is mostly due to the fact that an AI-based model must be sited entirely in main memory for prediction.

In the context of a firewall, different actions are triggered based on the probability and reasons that output is generated from the AI maldoc detection module. For instance, if the output probability is greater than a certain threshold of 0.9, this indicates that the AI module has high confidence that this document is malicious. Then, a blocking operation is triggered, the connection is dropped, and an alert is raised for further investigation. If the output probability is less than a certain threshold of 0.1, this indicates that the AI module has high confidence that this document is benign, and then the connection is allowed and monitored as normal by default.

The truly interesting part arises when the output probability is in the range 0.1–0.9. When this happens, we by default upload the samples to our Threat Intelligence (TI) Cloud where multiple dynamic analyses will be performed by Sandbox, Threat Intelligence, and Security Teams. According to our heuristics in a typical usage scenario, 95% of the files are processed inline while approximately 5% of the files are uploaded. Dynamic analysis from our TI cloud plays a great complementary role in the static analysis method inline. By combining the two engines, we can now completely provide end users with a more advanced, AI-enabled security solution. We have made our cloud service a subscription service and freely open to the research community.