Honeypot security technique can also stop attacks in natural language processing

July 29 2021, by Jessica Hallman

Credit: CC0 Public Domain

As online fake news detectors and spam filters become more sophisticated, so do attackers' methods to trick them—including attacks through the "universal trigger." In this learning-based method, an attacker uses a phrase or set of words to fool an indefinite number of inputs. A successful attack could mean more fake news appearing in your social media feed or spam reaching your email inbox.

Borrowing a technique commonly used in cybersecurity to defend against these universal trigger-based attacks, researchers at the Penn State College of Information Sciences and Technology have developed a machine learning framework that can proactively defend against the same types of attacks in natural language processing applications 99% of the time.

The model, called DARCY, uses a cybersecurity concept known as a "honeypot" to bait and catch potential attacks on natural language processing applications, such as fake news detectors and spam filters. Serving as a decoy, the honeypot attracts an attacker by drawing them to words and phrases they are targeting in their hack.

"Attackers try to find these universal attack phrases, so we try to make it very attractive for them to find the phrases that we already set," said Thai Le, doctoral student of information sciences and technology and lead author on the research paper. "We try to make the attacking job very easy for them, and then they fall into a trap."

DARCY searches and injects multiple trapdoors, or baits, into a textual neural network—the technology that drives natural language processing applications—to catch and filter out malicious content generated by universal trigger attacks.

"As far as we know, this is the first work that utilizes the concept of honeypot from the cybersecurity domain in defending textual neural network models against adversarial attacks," said Dongwon Lee, professor of information sciences and technology and principal investigator of the project.

Le explained that current defense methods against adversarial attacks are largely reactive, meaning that defenders must observe and learn hackers' techniques after an attack and then wait for another attack to detect and remove them.

With DARCY, the novel idea of a proactive defense approach could help to detect and prevent attacks.

"What people used to do is train a machine learning model and then try to defend it after it's trained," Le said. "But what we are trying to do is defend during the training process so we proactively protect models during the training phase."

The researchers tested DARCY on four separate text classification datasets and used the framework to defend against six different potential attack scenarios, such as whether an attacker had access to the detection network or if they are aware of the embedded trapdoors. DARCY significantly outperformed five existing adversarial detection algorithms that served as defensive baselines.

"For instance, DARCY was able to detect the universal trigger based adversarial attacks with up to 99% true positive rate and less than 2% false positive rate in most cases, which is a significant improvement over existing baselines," said Lee.

Seeing the effectiveness of using a successful cybersecurity defense tactic to defend natural language processing applications, the researchers hope to use the same techniques and framework to prevent other types of attacks in the future.

"Applying a concept from a very different field is intriguing, because we can see how different fields connect with each other and how the security umbrella can apply to many different aspects in computer science," Le concluded.

More information: Thai Le et al, A Sweet Rabbit Hole by DARCY:

Using Honeypots to Detect Universal Trigger's Adversarial Attacks, arXiv:2011.10492v3 [cs.CR] arxiv.org/abs/2011.10492

Provided by Pennsylvania State University

Citation: Honeypot security technique can also stop attacks in natural language processing (2021, July 29) retrieved 23 April 2024 from https://techxplore.com/news/2021-07-honeypot-technique-natural-language.html