How hackers can 'poison' open-source code

August 13 2021, by Adam Conner-Simons

Credit: CC0 Public Domain

Cornell Tech researchers have discovered a new type of online attack that can manipulate natural-language modeling systems and evade any known defense—with possible consequences ranging from modifying movie reviews to manipulating investment banks' machine-learning models to ignore negative news coverage that would affect a specific

company's stock.

In a new paper, researchers found the implications of these types of hacks—which they call "code poisoning"—to be wide-reaching for everything from algorithmic trading to fake news and propaganda.

"With many companies and programmers using models and codes from open-source sites on the internet, this research shows how important it is to review and verify these materials before integrating them into your current system," said Eugene Bagdasaryan, a doctoral candidate at Cornell Tech and lead author of "Blind Backdoors in Deep Learning Models," which was presented Aug. 12 at the virtual USENIX Security '21 conference. The co-author is Vitaly Shmatikov, professor of computer science at Cornell and Cornell Tech.

"If hackers are able to implement code poisoning," Bagdasaryan said, "they could manipulate models that automate supply chains and propaganda, as well as resume-screening and toxic comment deletion."

Without any access to the original code or model, these backdoor attacks can upload malicious code to open-source sites frequently used by many companies and programmers.

As opposed to adversarial attacks, which require knowledge of the code and model to make modifications, backdoor attacks allow the hacker to have a large impact, without actually having to directly modify the code and models.

"With previous attacks, the attacker must access the model or data during training or deployment, which requires penetrating the victim's machine learning infrastructure," Shmatikov said. "With this new attack, the attack can be done in advance, before the model even exists or before the data is even collected—and a single attack can actually target

multiple victims."

The new paper investigates the method for injecting backdoors into machine-learning models, based on compromising the loss-value computation in the model-training code. The team used a sentiment analysis model for the particular task of always classifying as positive all reviews of the infamously bad movies directed by Ed Wood.

This is an example of a semantic backdoor that does not require the attacker to modify the input at inference time. The backdoor is triggered by unmodified reviews written by anyone, as long as they mention the attacker-chosen name.

How can the "poisoners" be stopped? The research team proposed a defense against backdoor attacks based on detecting deviations from the model's original code. But even then, the defense can still be evaded.

Shmatikov said the work demonstrates that the oft-repeated truism, "Don't believe everything you find on the internet," applies just as well to software.

"Because of how popular AI and machine-learning technologies have become, many nonexpert users are building their models using code they barely understand," he said. "We've shown that this can have devastating security consequences."

For future work, the team plans to explore how code-poisoning connects to summarization and even automating propaganda, which could have larger implications for the future of hacking.

Shmatikov said they will also work to develop robust defenses that "will eliminate this entire class of attacks and make AI and machine learning safe even for nonexpert users."

More information: Full paper: www.cs.cornell.edu/~shmat/shmat_usenix21blind.pdf

Provided by Cornell University