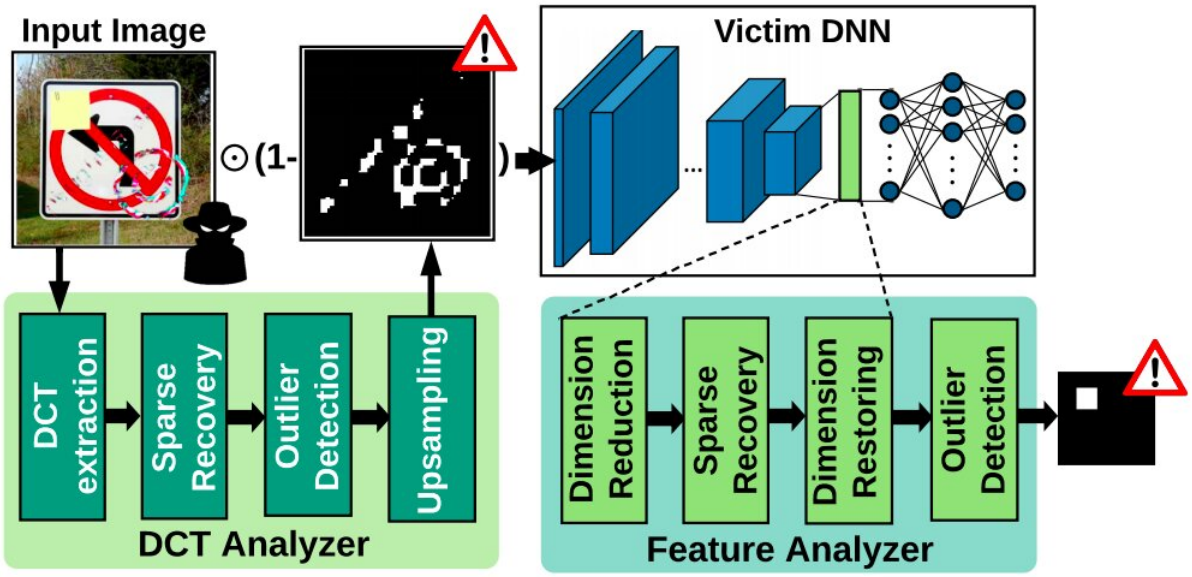
CLEANN: A framework to shield embedded neural networks from online Trojan attacks

October 7 2020, by Ingrid Fadelli



High-level overview of CLEANN, the framework developed by the researchers. Credit: Javaheripi et al.

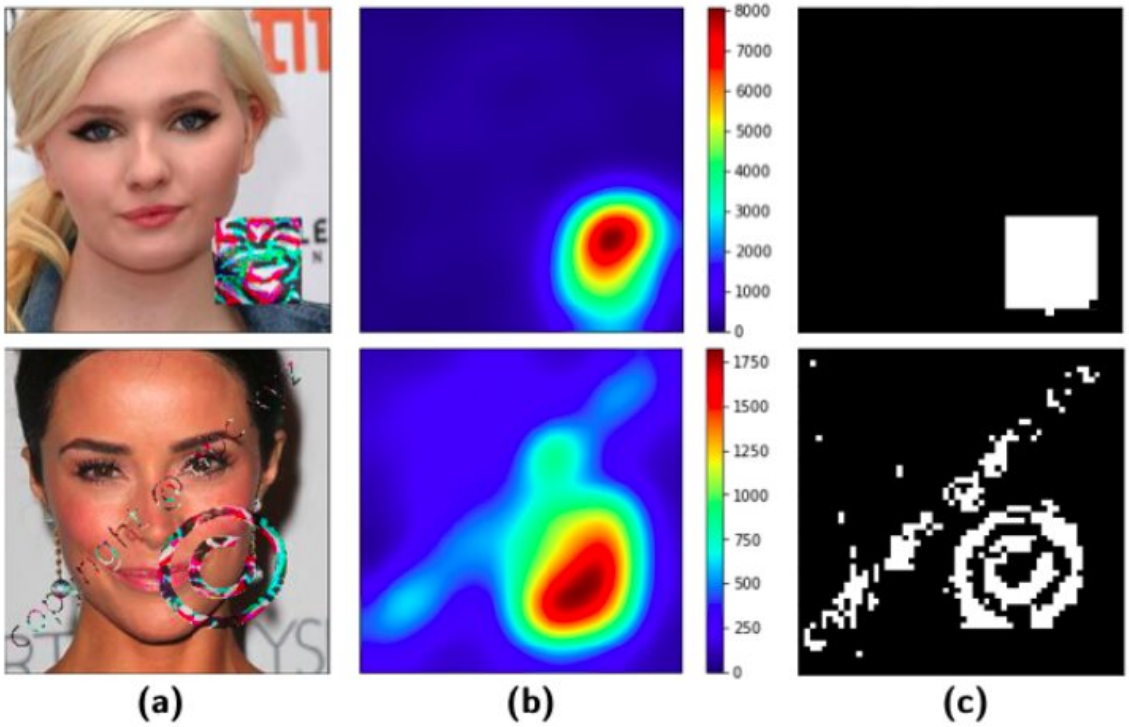
With artificial intelligence (AI) tools and machine learning algorithms now making their way into a wide variety of settings, assessing their security and ensuring that they are protected against cyberattacks is of utmost importance. As most AI algorithms and models are trained on large online datasets and third-party databases, they are vulnerable to a variety of attacks, including neural Trojan attacks.

A neural Trojan attack occurs when an attacker inserts what is known as a hidden Trojan trigger or backdoor inside an AI model during its training. This trigger allows the attacker to hijack the model's prediction at a later stage, causing it to classify data incorrectly. Detecting these attacks and mitigating their impact can be very challenging, as a targeted model typically performs well and in alignment with a developer's expectations until the Trojan backdoor is activated.

Researchers at University of California, San Diego have recently created CLEANN, an end-to-end framework designed to protect embedded artificial neural networks from Trojan attacks. This framework, presented in a paper pre-published on arXiv and set to be presented at the 2020 IEEE/ACM International Conference on Computer-Aided Design, was found to perform better than previously developed Trojan shields and detection methods.

"Despite all the benefits that come with artificial intelligence and autonomous systems, there are critical threats endangering their safety/integrity," Mojan Javaheripi, one of the researchers who developed CLEANN, told TechXplore. "One of these threats is neural Trojans, i.e., malicious inputs that deliberately cause AI models to make mistakes. CLEANN is a lightweight and effective system that monitors deployed AI models to make sure malicious (i.e., Trojan) inputs cannot trigger unwanted behavior."

The framework developed by Javaheripi and her colleagues, identifies the characteristics of safe input data. Subsequently, it analyzes new data based on these characteristics in order to spot Trojan triggers and correct the mistakes they cause in the AI model into which they were inserted.



(a) Example Trojan data with watermark and square triggers, (b) reconstruction error heatmap, and (c) output mask from the outlier detection module. Credit: Javaheripi et al.

"CLEANN learns a sparse reconstruction of the benign inputs," Javaheripi explained. "It then uses sparse recovery to project malicious samples into the learned benign space. By doing so, we not only detect Trojans, but also stop their malicious effect."

In a series of initial evaluations using neural network-based image classification models, CLEANN achieved highly promising results. In fact, it is the first lightweight defense to achieve both high detection and high decision correction rates. Moreover, in contrast with previously proposed neural Trojan mitigation methods, it does not require labeled

or annotated data or for a targeted AI model to be retrained, both of which can be quite costly and time consuming.

Javaheripi and her colleagues also developed a specialized hardware that supports their framework. This hardware can be used to efficiently execute the framework in real-time, mitigating the hazards caused by Trojan attacks.

"The majority of Trojan defense methods proposed to date induce a high execution overhead that hinders their applicability to embedded systems," Javaheripi said. "To the best of our knowledge, no earlier work provides the needed lightweight defense strategy for real-time autonomous applications."

The study shows that carefully applying sparse recovery techniques to selected signals of AI models can help to shield these systems from online Trojan attacks. In the future, the new framework they developed could be used to secure existing and newly developed AI systems from online Trojan attacks.

"In our next studies, we plan to extend the methodologies used in CLEANN to other domains beyond image classification, such as speech processing and video," Javaheripi said. "Additionally, with the everchanging horizon of attacks against AI models, we will continuously adapt our defense strategy to overcome new emerging threats."

More information: Javaheripi et al., CLEANN: Accelerated Trojan shield for embedded neural networks. arXiv: 2009.02326 [cs.LG]. arxiv.org/abs/2009.02326

© 2020 Science X Network

Citation: CLEANN: A framework to shield embedded neural networks from online Trojan attacks (2020, October 7) retrieved 18 April 2024 from https://techxplore.com/news/2020-10-cleann-framework-shield-embedded-neural.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.