Hiding malware inside AI neural networks

July 27 2021, by Bob Yirka

Credit: CC0 Public Domain

A trio of researchers at Cornell University has found that it is possible to hide malware code inside of AI neural networks. Zhi Wang, Chaoge Liu and Xiang Cui have posted a paper describing their experiments with injecting code into neural networks on the arXiv preprint server.

As computer technology grows ever more complex, so do attempts by criminals to break into machines running new technology for their own purposes, such as destroying data or encrypting it and demanding payment from users for its return. In this new study, the team has found a new way to infect certain kinds of computer systems running artificial intelligence applications.

AI systems do their work by processing data in ways similar to the human brain. But such networks, the research trio found, are vulnerable to infiltration by foreign code.

Neural networks, by their very nature, can be invaded by foreign agents. All such agents have to do is mimic the structure of the network in much the same way memories are added in the human brain. The researchers found that they were able to do just that by embedding malware into the neural network behind an AI system called AlexNet—despite it being rather hefty, taking up 36.9 MiB of memory space on the hardware running the AI system. To add the code into the neural network, the researchers chose what they believed would be the best layer for injection. They also added it to a model that had been trained already but noted hackers might prefer to attack an untrained network because it would likely have less of an impact on the overall network.

The researchers found that not only did standard antivirus software fail to find the malware, but the AI system performance was almost the same after being infected. Thus, the infection could have gone undetected if covertly executed.

The researchers note that simply adding malware to the neural network would not cause harm—whoever slipped the code into the system would still have to find a way to execute that code. They also note that now that it is known that hackers can inject code into AI neural networks, antivirus software can be updated to look for it.

More information: Zhi Wang et al, EvilModel: Hiding Malware Inside of Neural Network Models, arXiv:2107.08590v1 [cs.CR] https://arxiv.org/abs/2107.08590

Provided by Science X Network

Citation: Hiding malware inside AI neural networks (2021, July 27) retrieved 6 May 2024 from https://techxplore.com/news/2021-07-malware-ai-neural-networks.html