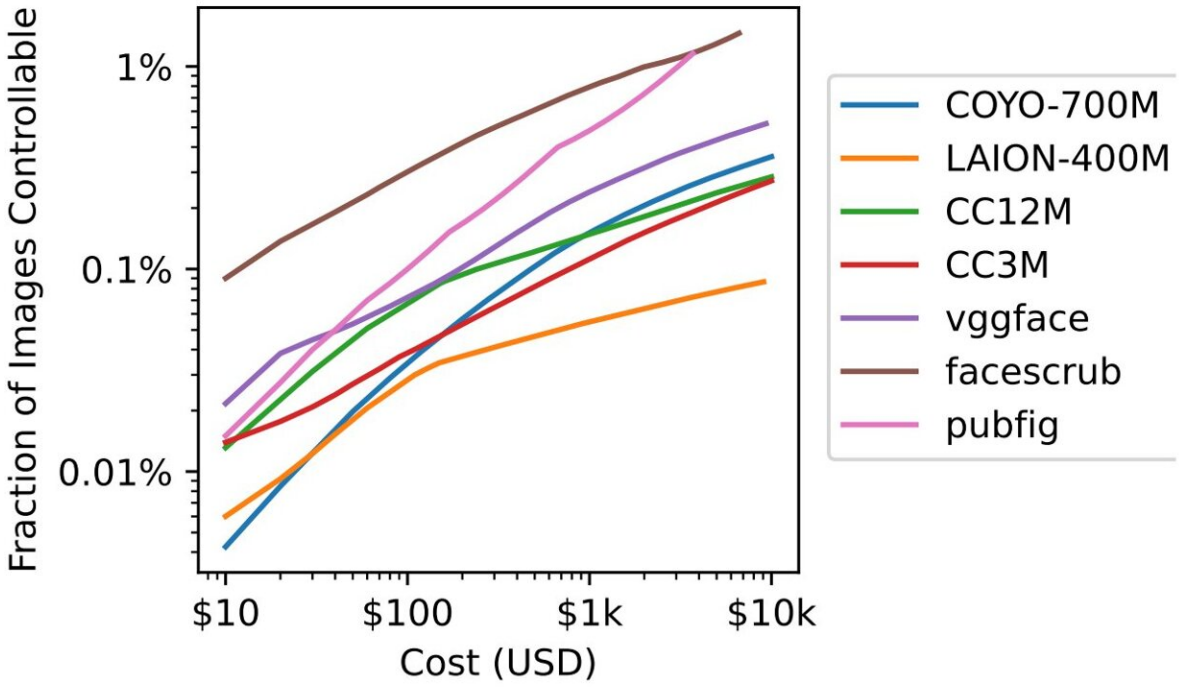
Two types of dataset poisoning attacks that can corrupt AI system results

March 7 2023, by Bob Yirka



It often costs ≤ $60 USD to control at least 0.01% of the data. Costs are measured by purchasing domains in order of lowest cost per image first. Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2302.10149

A team of computer science researchers with members from Google, ETH Zurich, NVIDIA and Robust Intelligence, is highlighting two kinds

of dataset poisoning attacks that could be used by bad actors to corrupt AI system results. The group has written a paper outlining the kinds of attacks that they have identified and have posted it on the *arXiv* preprint server.

With the development of deep learning neural networks, artificial intelligence applications have become big news. And because of their unique learning abilities they can be applied in a wide variety of environments. But, as the researchers on this new effort note, one thing they all have in common is the need for quality data to use for training purposes.

Because such systems learn from what they see, if they happen across something that is wrong, they have no way of knowing it, and thus incorporate it into their set of rules. As an example, consider an AI system that is trained to recognize patterns on a mammogram as cancerous tumors. Such systems would be trained by showing them many examples of real tumors collected during mammograms.

But what happens if someone inserts images into the dataset showing cancerous tumors, but they are labeled as non-cancerous? Very soon the system would begin missing those tumors because it has been taught to see them as non-cancerous. In this new effort, the research team has shown that something similar can happen with AI systems that are trained using publicly available data on the Internet.

The researchers began by noting that ownership of URLs on the Internet often expire—including those that have been used as sources by AI systems. That leaves them available for purchase by nefarious types looking to disrupt AI systems. If such URLs are purchased and are then used to create websites with false information, the AI system will add that information to its knowledge bank just as easily as it will true information—and that will lead to the AI system producing less then

desirable results.

The research team calls this type of attack split view poisoning. Testing showed that such an approach could be used to purchase enough URLs to poison a large portion of mainstream AI systems, for as little as $10,000.

There is another way that AI systems could be subverted—by manipulating data in well known data repositories such as Wikipedia. This could be done, the researchers note, by modifying data just prior to regular data dumps, preventing monitors from spotting the changes before they are sent to and used by AI systems. They call this approach frontrunning poisoning.

More information: Nicholas Carlini et al, Poisoning Web-Scale Training Datasets is Practical, *arXiv* (2023). DOI: 10.48550/arxiv.2302.10149