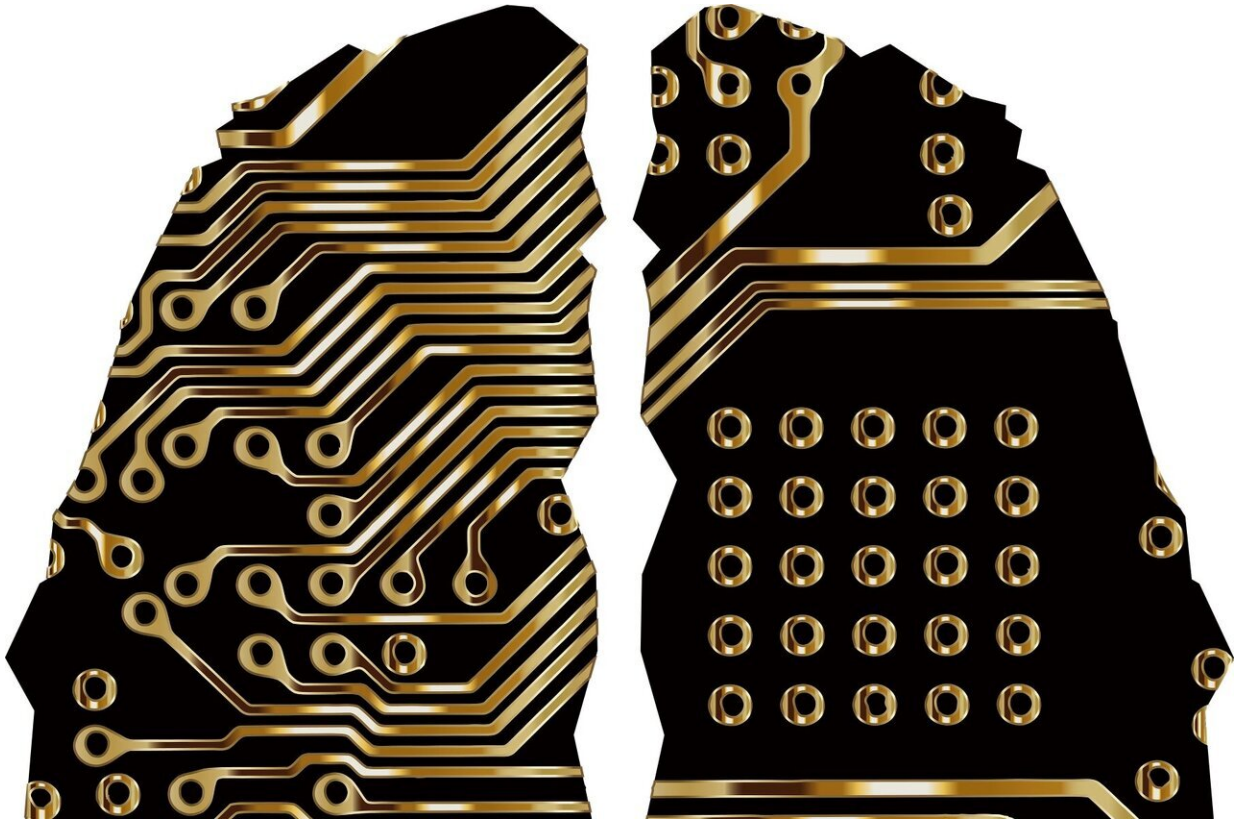
Researchers develop new protocols to validate integrity of machine-learning models

April 28 2021, by Stephanie Jones



Credit: Pixabay/CC0 Public Domain

Machine learning is widely used in various applications such as image recognition, autonomous vehicles and email filtering. Despite its success, concerns about the integrity and security of a model's predictions and

accuracy are on the rise.

To address these issues, Dr. Yupeng Zhang, professor in the Department of Computer Science and Engineering at Texas A&M University, and his team applied cryptographic algorithms called zero-knowledge proof protocols to the domain of machine learning.

"These protocols will allow the owner of a machine-learning model to prove to others that the model can achieve a high accuracy on public datasets without leaking any information about the machine-learning model itself," said Zhang.

The researchers' findings were published in the proceedings from the Association for Computing Machinery's 2020 Conference on Computer and Communications Security.

Machine learning is a form of artificial intelligence that focuses on algorithms that give a computer system the ability to learn from data and improve its accuracy over time. These algorithms build models to find patterns within large amounts of data to make decisions and predictions without being programmed.

Over the years, machine-learning models have undergone a great deal of development, which has led to significant progress in several research areas such as data mining and natural language processing. Several companies and research groups claim to have developed machine-learning models that can achieve very high accuracy on public testing samples of data. Still, reproducing the results to verify those claims remains a challenge for researchers. It is unknown if they can achieve that accuracy or not, and it isn't easy to justify.

The theoretical foundation of cybersecurity and cryptography is the science of protecting information and communications through a series

of codes so that only the sender and the intended recipient have the ability to view and understand it. It's most commonly used to develop tools such as encryptions, cybertext, digital signatures and hash functions.

There are approaches outside of cryptography that could be used, one of which involves releasing the model to the public. However, as machine-learning models have become critical intellectual property for many companies, they can't be released because they contain sensitive information essential to the business.

"This approach is also problematic because once the model is out there, there is a software tool online anyone could use to verify," said Zhang. "Recent research also shows that the model's information could be used to reconstruct it and used for whatever they desire."

As an application of cryptography, zero-knowledge proof protocols are a mathematical method that allows the owner of a machine-learning model to produce a succinct proof of it to prove with overwhelming probability that something is true without sharing any extra information about it.

While there has been a significant improvement in the use of general-purpose zero-knowledge proof schemes in the last decade, constructing efficient machine-learning prediction and accuracy tests remains a challenge because of the time it takes to generate a proof.

"When we applied these generic techniques to common machine-learning models, we found that it would take several days or months for a company to generate a proof to prove to the public that their model can achieve what they claim," said Zhang.

For a more efficient approach, Zhang and his team designed several new zero-knowledge proof techniques and optimizations specifically tailored

to turn the computations of a decision tree model, which is one of the most commonly used machine-learning algorithms, into zero-knowledge proof statements.

Using their approach on the computations of a decision tree, they found that it would take less than 300 seconds to generate a proof that would prove the model can achieve high accuracy on a dataset.

As their newly developed approach only addresses generating proof for decision tree models, the researchers want to expand their approach to efficiently support different types of machine-learning models.

Contributors to this project include Zhiyong Fang, doctoral student in the computer science and engineering department; and doctoral student Jiaheng Zhang and Dr. Dawn Song from the University of California, Berkeley.

More information: Jiaheng Zhang et al. Zero Knowledge Proofs for Decision Tree Predictions and Accuracy, *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security* (2020). DOI: 10.1145/3372297.3417278

Provided by Texas A&M University College of Engineering