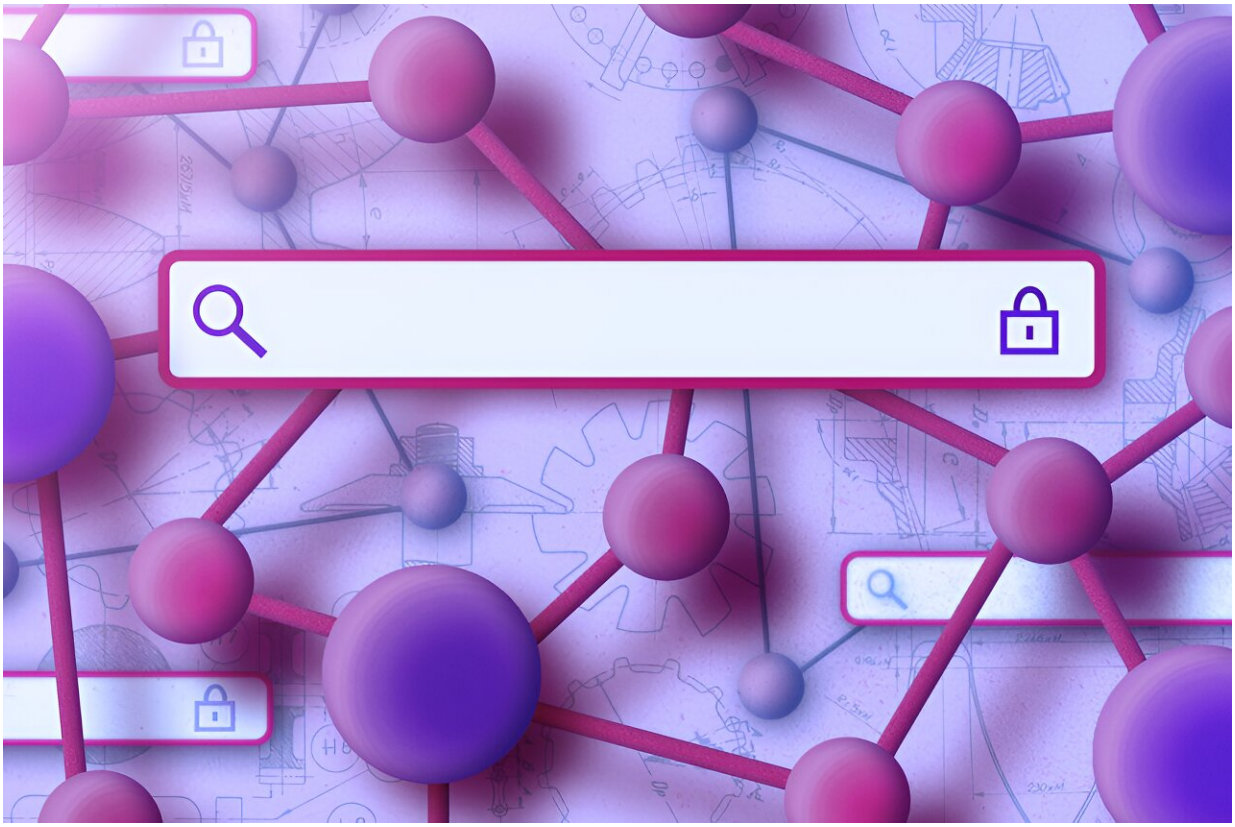
Accelerating AI tasks while preserving data security

October 30 2023, by Adam Zewe



SecureLoop is an MIT-developed search engine that can identify an optimal design for a deep neural network accelerator that preserves data security while improving energy efficiency and boosting performance. This could enable device manufacturers to increase the speed of demanding AI applications, while ensuring sensitive data remain safe from attackers. Credit: Jose-Luis Olivares, MIT

With the proliferation of computationally intensive machine-learning applications, such as chatbots that perform real-time language translation, device manufacturers often incorporate specialized hardware components to rapidly move and process the massive amounts of data these systems demand.

Choosing the best design for these components, known as deep neural network accelerators, is challenging because they can have an enormous range of design options. This difficult problem becomes even thornier when a designer seeks to add cryptographic operations to keep data safe from attackers.

Now, MIT researchers have developed a search engine that can efficiently identify optimal designs for deep neural network accelerators, that preserve data security while boosting performance.

Their search tool, known as SecureLoop, is designed to consider how the addition of data encryption and authentication measures will impact the performance and energy usage of the accelerator chip. An engineer could use this tool to obtain the optimal design of an accelerator tailored to their neural network and machine-learning task.

When compared to conventional scheduling techniques that don't consider security, SecureLoop can improve performance of accelerator designs while keeping data protected.

Using SecureLoop could help a user improve the speed and performance of demanding AI applications, such as autonomous driving or medical image classification, while ensuring sensitive user data remains safe from some types of attacks.

"If you are interested in doing a computation where you are going to preserve the security of the data, the rules that we used before for

finding the optimal design are now broken. So all of that optimization needs to be customized for this new, more complicated set of constraints. And that is what [lead author] Kyungmi has done in this paper," says Joel Emer, an MIT professor of the practice in computer science and electrical engineering and co-author of a paper on SecureLoop.

Emer is joined on the paper by lead author Kyungmi Lee, an electrical engineering and computer science graduate student; Mengjia Yan, the Homer A. Burnell Career Development Assistant Professor of Electrical Engineering and Computer Science and a member of the Computer Science and Artificial Intelligence Laboratory (CSAIL); and senior author Anantha Chandrakasan, dean of the MIT School of Engineering and the Vannevar Bush Professor of Electrical Engineering and Computer Science. The research will be presented at the IEEE/ACM International Symposium on Microarchitecture held Oct. 28–Nov. 1.

"The community passively accepted that adding cryptographic operations to an accelerator will introduce overhead. They thought it would introduce only a small variance in the design trade-off space. But, this is a misconception. In fact, cryptographic operations can significantly distort the design space of energy-efficient accelerators. Kyungmi did a fantastic job identifying this issue," Yan adds.

Secure acceleration

A deep neural network consists of many layers of interconnected nodes that process data. Typically, the output of one layer becomes the input of the next layer. Data are grouped into units called tiles for processing and transfer between off-chip memory and the accelerator. Each layer of the neural network can have its own data tiling configuration.

A deep neural network accelerator is a processor with an array of

computational units that parallelizes operations, like multiplication, in each layer of the network. The accelerator schedule describes how data are moved and processed.

Since space on an accelerator chip is at a premium, most data are stored in off-chip memory and fetched by the accelerator when needed. But because data are stored off-chip, they are vulnerable to an attacker who could steal information or change some values, causing the neural network to malfunction.

"As a chip manufacturer, you can't guarantee the security of external devices or the overall operating system," Lee explains.

Manufacturers can protect data by adding authenticated encryption to the accelerator. Encryption scrambles the data using a secret key. Then authentication cuts the data into uniform chunks and assigns a cryptographic hash to each chunk of data, which is stored along with the data chunk in off-chip memory.

When the accelerator fetches an encrypted chunk of data, known as an authentication block, it uses a secret key to recover and verify the original data before processing it.

But the sizes of authentication blocks and tiles of data don't match up, so there could be multiple tiles in one block, or a tile could be split between two blocks. The accelerator can't arbitrarily grab a fraction of an authentication block, so it may end up grabbing extra data, which uses additional energy and slows down computation.

Plus, the accelerator still must run the cryptographic operation on each authentication block, adding even more computational cost.

An efficient search engine

With SecureLoop, the MIT researchers sought a method that could identify the fastest and most energy efficient accelerator schedule—one that minimizes the number of times the device needs to access off-chip memory to grab extra blocks of data because of encryption and authentication.

They began by augmenting an existing search engine Emer and his collaborators previously developed, called Timeloop. First, they added a model that could account for the additional computation needed for encryption and authentication.

Then, they reformulated the search problem into a simple mathematical expression, which enables SecureLoop to find the ideal authentical block size in a much more efficient manner than searching through all possible options.

"Depending on how you assign this block, the amount of unnecessary traffic might increase or decrease. If you assign the cryptographic block cleverly, then you can just fetch a small amount of additional data," Lee says.

Finally, they incorporated a heuristic technique that ensures SecureLoop identifies a schedule which maximizes the performance of the entire deep neural network, rather than only a single layer.

At the end, the search engine outputs an accelerator schedule, which includes the data tiling strategy and the size of the authentication blocks, that provides the best possible speed and energy efficiency for a specific neural network.

"The design spaces for these accelerators are huge. What Kyungmi did was figure out some very pragmatic ways to make that search tractable so she could find good solutions without needing to exhaustively search

the space," says Emer.

When tested in a simulator, SecureLoop identified schedules that were up to 33.2% faster and exhibited 50.2% better energy delay product (a metric related to energy efficiency) than other methods that didn't consider security.

The researchers also used SecureLoop to explore how the design space for accelerators changes when security is considered. They learned that allocating a bit more of the chip's area for the cryptographic engine and sacrificing some space for on-chip memory can lead to better performance, Lee says.

In the future, the researchers want to use SecureLoop to find accelerator designs that are resilient to side-channel attacks, which occur when an attacker has access to physical hardware. For instance, an attacker could monitor the power consumption pattern of a device to obtain secret information, even if the data have been encrypted. They are also extending SecureLoop so it could be applied to other kinds of computation.

More information: SecureLoop: Design Space Exploration of Secure DNN Accelerators: par.nsf.gov/biblio/10465225-se … ure-dnn-accelerators

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology