Team streamlines neural networks to be more adept at computing on encrypted data

July 22 2021

Credit: Pixabay/CC0 Public Domain

This week, at the 38th International Conference on Machine Learning (ICML 21), researchers at the NYU Center for Cyber Security at the NYU Tandon School of Engineering are revealing new insights into the basic functions that drive the ability of neural networks to make inferences on encrypted data.

In the paper, "DeepReDuce: ReLU Reduction for Fast Private Inference," the team focuses on linear and non-linear operators, key features of neural network frameworks that, depending on the operation, introduce a heavy toll in time and computational resources. When neural networks compute on encrypted data, many of these costs are incurred by rectified linear activation function (ReLU), a non-linear operation.

Brandon Reagen, professor of computer science and engineering and electrical and computer engineering and a team of collaborators including Nandan Kumar Jha, a Ph.D. student, and Zahra Ghodsi, a former doctoral student under the guidance of Siddharth Garg, developed a framework called DeepReDuce. It offers a solution through rearrangement and reduction of ReLUs in neural networks.

Reagen explained that this shift requires a fundamental reassessment of where and how many components are distributed in neural networks systems.

"What we are trying to do is rethink how neural nets are designed in the first place," he explained. "You can skip a lot of these time- and computationally-expensive ReLU operations and still get high performing networks at 2 to 4 times faster run time."

The team found that, compared to the state-of-the-art for private inference, DeepReDuce improved accuracy and reduced ReLU count by up to 3.5% and 3.5×, respectively.

The inquiry is not merely academic. As the use of AI grows in concert with concerns about the security of personal, corporate, and government data security, neural networks are increasingly making computations on encrypted data. In such scenarios involving neural networks generating private inferences (PI's) on hidden data without disclosing inputs, it is the non-linear functions that exert the highest "cost" in time and power. Because these costs increase the difficulty and time it takes for learning machines to do PI, researchers have struggled to lighten the load ReLUs exert on such computations.

The team's work builds on innovative technology called CryptoNAS. Described in an earlier paper whose authors include Ghodsi and a third Ph.D. student, Akshaj Veldanda, CryptoNAS optimizes the use of ReLUs as one might rearrange how rocks are arranged in a stream to optimize the flow of water: it rebalances the distribution of ReLUS in the network and removes redundant ReLUs.

DeepReDuce expands on CryptoNAS by streamlining the process further. It comprises a set of optimizations for the judicious removal of ReLUs after CryptoNAS reorganization functions. The researchers tested DeepReDuce by using it to remove ReLUs from classic networks, finding that they were able to significantly reduce inference latency while maintaining high accuracy.

Reagan, with Mihalis Maniatakos, research assistant professor of electrical and computer engineering, is also part of a collaboration with data security company Duality to design a new microchip designed to handle computation on fully encrypted data.

More information: DeepReDuce: ReLU Reduction for Fast Private Inference, arXiv:2103.01396 [cs.LG] arxiv.org/abs/2103.01396

Provided by NYU Tandon School of Engineering