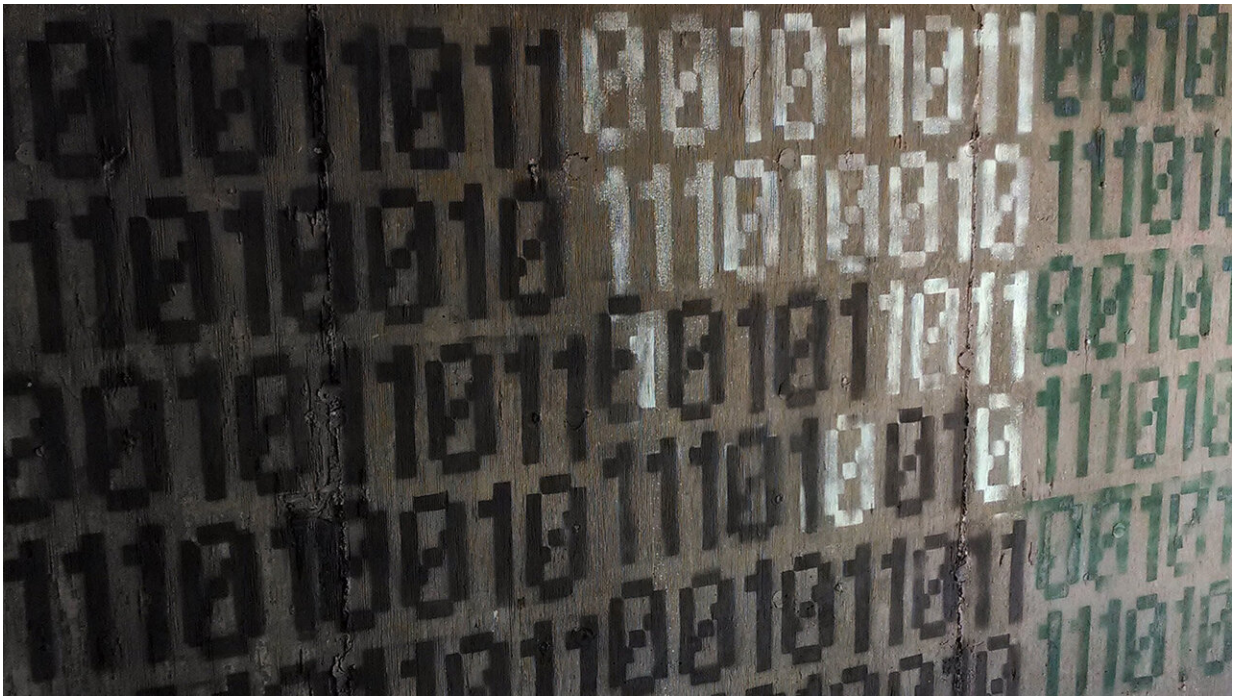
New technique cuts AI training time by more than 60 percent

April 8 2019, by Matt Shipman



Credit: Patrick Dockens/Creative Commons

North Carolina State University researchers have developed a technique that reduces training time for deep learning networks by more than 60 percent without sacrificing accuracy, accelerating the development of new artificial intelligence (AI) applications.

"Deep learning networks are at the heart of AI applications used in

everything from self-driving cars to computer vision technologies," says Xipeng Shen, a professor of computer science at NC State and co-author of a paper on the work.

"One of the biggest challenges facing the development of new AI tools is the amount of time and computing power it takes to train deep learning networks to identify and respond to the data patterns that are relevant to their applications. We've come up with a way to expedite that process, which we call Adaptive Deep Reuse. We have demonstrated that it can reduce training times by up to 69 percent without accuracy loss."

Training a deep learning network involves breaking a data sample into chunks of consecutive data points. Think of a network designed to determine whether there is a pedestrian in a given image. The process starts by dividing a digital image into blocks of pixels that are adjacent to each other. Each chunk of data is run through a set of computational filters. The results are then run through a second set of filters. This continues iteratively until all of the data have been run through all of the filters, allowing the network to reach a conclusion about the data sample.

When this process has been done for every data sample in a data set, that is called an epoch. In order to fine-tune a deep learning network, the network will likely run through the same data set for hundreds of epochs. And many data sets consist of between tens of thousands and millions of data samples. Lots of iterations of lots of filters being applied to lots of data means that training a deep learning network takes a lot of computing power.

The breakthrough moment for Shen's research team came when it realized that many of the data chunks in a data set are similar to each other. For example, a patch of blue sky in one image may be similar to a patch of blue sky elsewhere in the same image or to a patch of sky in another image in the same data set.

By recognizing these similar data chunks, a deep learning network could apply filters to one chunk of data and apply the results to all of the similar chunks of data in the same set, saving a lot of computing power.

"We were not only able to demonstrate that these similarities exist, but that we can find these similarities for intermediate results at every step of the process," says Lin Ning, a Ph.D. student at NC State and lead author of the paper. "And we were able to maximize this efficiency by applying a method called locality sensitive hashing."

But this raises two additional questions. How large should each chunk of data be? And what threshold do data chunks need to meet in order to be deemed "similar"?

The researchers found that the most efficient approach was to begin by looking at relatively large chunks of data using a relatively low threshold for determining similarity. In subsequent epochs, the data chunks get smaller and the similarity threshold more stringent, improving the deep learning network's accuracy. The researchers designed an adaptive algorithm that automatically implements these incremental changes during the training process.

To evaluate their new technique, the researchers tested it using three deep learning networks and data sets that are widely used as testbeds by deep learning researchers: CifarNet using Cifar10; AlexNet using ImageNet; and VGG-19 using ImageNet.

Adaptive Deep Reuse cut training time for AlexNet by 69 percent; for VGG-19 by 68 percent; and for CifarNet by 63 percent – all without accuracy loss.

"This demonstrates that the technique drastically reduces training times," says Hui Guan, a Ph.D. student at NC State and co-author of the paper.

"It also indicates that the larger the network, the more Adaptive Deep Reuse is able to reduce training times – since AlexNet and VGG-19 are both substantially larger than CifarNet."

"We think Adaptive Deep Reuse is a valuable tool, and look forward to working with industry and research partners to demonstrate how it can be used to advance AI," Shen says.

The paper, "Adaptive Deep Reuse: Accelerating CNN Training on the Fly," will be presented at the 35th IEEE International Conference on Data Engineering, being held April 8-11 in Macau SAR, China.

Provided by North Carolina State University