PHYS.ORG

Fujitsu doubles deep learning neural network scale with technology to improve GPU memory efficiency

September 21 2016

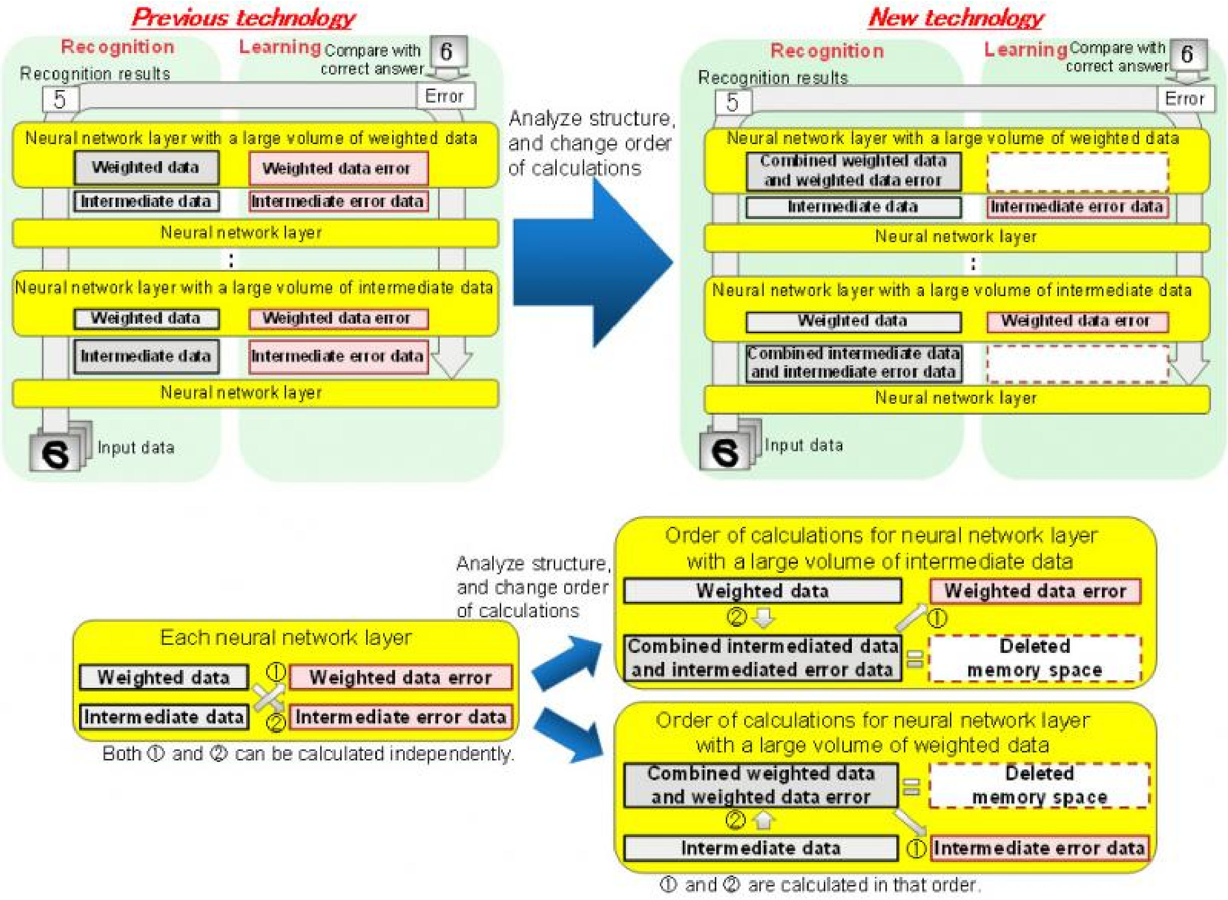
Previous technology
Recognition
Learning
Compare with correct answer
6
Recognition results
5
Error
Neural network layer with a large volume of weighted data
Weighted data
Weighted data error
Intermediate data
Intermediate error data
Neural network layer
Neural network layer with a large volume of intermediate data
Weighted data
Weighted data error
Intermediate data
Intermediate error data
Neural network layer
Input data
Analyze structure, and change order of calculations
New technology
Recognition
Learning
Compare with correct answer
6
Recognition results
5
Error
Neural network layer with a large volume of weighted data
Combined weighted data and weighted data error
Intermediate data
Intermediate error data
Neural network layer
Neural network layer with a large volume of intermediate data
Weighted data
Weighted data error
Combined intermediate data and intermediate error data
Neural network layer
Input data
Each neural network layer
Weighted data
① Weighted data error
Intermediate data
② Intermediate error data
Both ① and ② can be calculated independently.
Analyze structure, and change order of calculations
Order of calculations for neural network layer with a large volume of intermediate data
Weighted data
Weighted data error
② ①
Combined intermediated data and intermediated error data
Deleted memory space
Order of calculations for neural network layer with a large volume of weighted data
Combined weighted data and weighted data error
Deleted memory space
②
Intermediate data
① Intermediate error data
① and ② are calculated in that order.

Figure 1: Technology to improve memory efficiency. Credit: Fujitsu

Fujitsu Laboratories Ltd. today announced development of technology to streamline the internal memory of GPUs to support the growing neural network scale that works to heighten machine learning accuracy. This development has enabled neural network machine learning of a scale up to twice what was capable with previous technology.

Recent years have seen a focus on technologies that use GPUs for high-speed machine learning to support the huge volume of calculations necessary for deep learning processing. In order to make use of a GPU's high-speed calculation ability, the data to be used in a series of calculations needs to be stored in the GPU's internal memory. This, however, creates an issue where the scale of the neural network that could be built is limited by memory capacity. Fujitsu Laboratories has now developed technology to improve memory efficiency, implementing and evaluating it in the Caffe open source deep learning framework software. Upon commencement of learning, the technology analyzes the structure of the neural network, and optimizes the order of calculations and allocation of data to memory, so that memory space can be efficiently reused.

With AlexNet and VGGNe, image-recognition neural networks widely used in research, this technology was confirmed to enable the scale of learning of a neural network to be increased by up to roughly two times that of previous technology, thereby reducing the volume of internal GPU memory used by over 40%. This technology makes it possible to expand the scale of a neural network that can be learned at high speed on one GPU, enabling the development of more accurate models. Fujitsu Laboratories aims to commercialize this technology as part of Fujitsu Limited's AI technology, Human Centric AI Zinrai, to work with customers in the use of AI. Details of this technology were announced at MLSP (IEEE Machine Learning for Signal Processing 2016), an international conference held in Salerno, Italy from September 13 to 16.

Development Background

In recent years, deep learning has been gaining attention as a machine learning method that emulates the structure of the human brain. In deep learning, the more layers there are in a neural network, the more accurate it performs tasks, such as recognition or categorization. In order to increase accuracy, the scale of neural networks has been growing, but this lengthens learning times. Along with this, more attention is being placed on GPUs that execute computations with large volumes of data, and technology that accelerates the process by using multiple GPUs in parallel, as with supercomputers.

One method of increasing the scale of deep learning is to distribute a single neural network model across multiple computers and do the computations in parallel, but the volume of data that must be transmitted in exchanges between computers then becomes a bottleneck, greatly reducing learning speed. In order to take full advantage of the GPU's high-speed calculation capability, to the extent possible the data to be used in series of calculations needs to be stored in the GPU's internal memory. However, as GPU memory is usually smaller than that of an ordinary computer, there had been the issue of limitations in scale of neural networks capable of high-speed learning.

Now Fujitsu Laboratories has developed technology to streamline memory efficiency to expand the scale of a neural network for computations with one GPU, without using parallelization methods that greatly reduce learning speed. This technology reduces the volume of memory by enabling the reuse of memory resources; it takes advantage of the ability to independently execute both calculations to generate the intermediate error data from weighted data, and calculations to generate the weighted data error from intermediate data. When learning begins, the structure of every layer of the neural network is analyzed, and the order of calculations is changed so that memory space in which larger data has been allocated can be reused.

Fujitsu Laboratories implemented this newly developed technology into the Caffe open source deep learning software framework and measured the usage of GPU internal memory. In evaluations using AlexNet and VGGNet, which are widely used in research fields, it achieved reductions in memory usage volume of over 40% compared with before the application of this technology, enabling the scale of learning on a neural network for each GPU to be increased by up to roughly two times. This will enable high-speed learning calculations using the full capability of a GPU, even with a large-scale neural network that requires complicated processing, accelerating the development of more accurate models.

Fujitsu aims to commercialize this newly developed technology as part of Fujitsu Limited's AI technology, Human Centric AI Zinrai, by March 31, 2017. In addition, it plans to combine this technology with its already announced high-speed technology to process deep learning through GPU parallelization, and further improve these technologies.

Provided by Fujitsu

Citation: Fujitsu doubles deep learning neural network scale with technology to improve GPU memory efficiency (2016, September 21) retrieved 3 May 2024 from https://phys.org/news/2016-09-fujitsu-deep-neural-network-scale.html