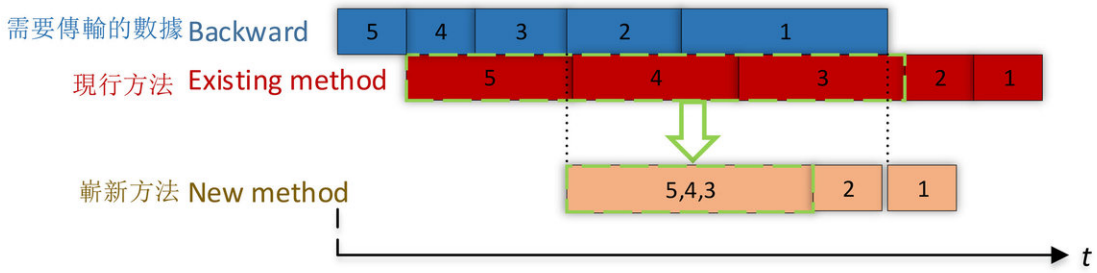
Team breaks world record for fast, accurate AI training

November 7 2018



A diagram showing data transmission of a 5-layer model 數據傳輸的示意圖

(Blue colour) Layers of data blocks (layers 5 to 1) of a neural network (model) that need data exchange. 藍色表示人工智能模型中由 1 至 5 層的數據塊等待傳輸。

(Red colour) Existing method transmits the data blocks layer by layer (from layer 5 to layer 1). 紅色表示現有系統的通訊方式，數據塊從 5 至 1 逐層傳輸。

(Orange colour) New method optimally merges data blocks from several layers into one larger data block (data blocks from layers 5, 4 and 3 are fused into one data block), and then the merged data block is transmitted. 橙色表示新技術將數據塊 5、4、3 集合成更較大的組件再傳輸，因此提升整個人工智能訓練的通訊模式。

Diagram showing data transmission of a 5-layer model. Credit: HKBU

Researchers at Hong Kong Baptist University (HKBU) have partnered with a team from Tencent Machine Learning to create a new technique for training artificial intelligence (AI) machines faster than ever before while maintaining accuracy.

During the experiment, the team trained two popular deep neural networks called AlexNet and ResNet-50 in just four minutes and 6.6 minutes respectively. Previously, the fastest training time was 11 minutes for AlexNet and 15 minutes for ResNet-50.

AlexNet and ResNet-50 are deep neural networks built on ImageNet, a large-scale dataset for visual recognition. Once trained, the system was able to recognise and label an object in a given photo. The result is significantly faster than previous records and outperforms all other existing systems.

Machine learning is a set of mathematical approaches that enable computers to learn from data without explicitly being programmed by humans. The resulting algorithms can then be applied to a variety of data and visual recognition tasks used in AI.

The HKBU team comprises Professor Chu Xiaowen and Ph.D. student Shi Shaohuai from the Department of Computer Science. Professor Chu said, "We have proposed a new optimised training method that significantly improves the best output without losing accuracy. In AI training, researchers strive to train their networks faster, but this can lead to a decrease in accuracy. As a result, training machine-learning models at high speed while maintaining accuracy and precision is a vital goal for scientists."

Professor Chu said the time required to train AI machines is affected by both computing time and communication time. The research team attained breakthroughs in both aspects to create this record-breaking achievement.

This included adopting a simpler computational method known as FP16 to replace the more traditional one, FP32, making computation much faster without losing accuracy. As communication time is affected by

the size of data blocks, the team came up with a communication technique named "tensor fusion," which combines smaller pieces of data into larger ones, optimising the transmission pattern and thereby improving the efficiency of communication during AI training.

This new technique can be adopted in large-scale image classification, and it can also be applied to other AI applications, including machine translation; natural language processing (NLP) to enhance interactions between human language and computers; medical imaging analysis; and online multiplayer battle games.

More information: Xianyan Jia et al. Highly Scalable Deep Learning Training System with Mixed-Precision: Training ImageNet in Four Minutes. arXiv:1807.11205 [cs.LG]. arxiv.org/abs/1807.11205

Provided by Hong Kong Baptist University