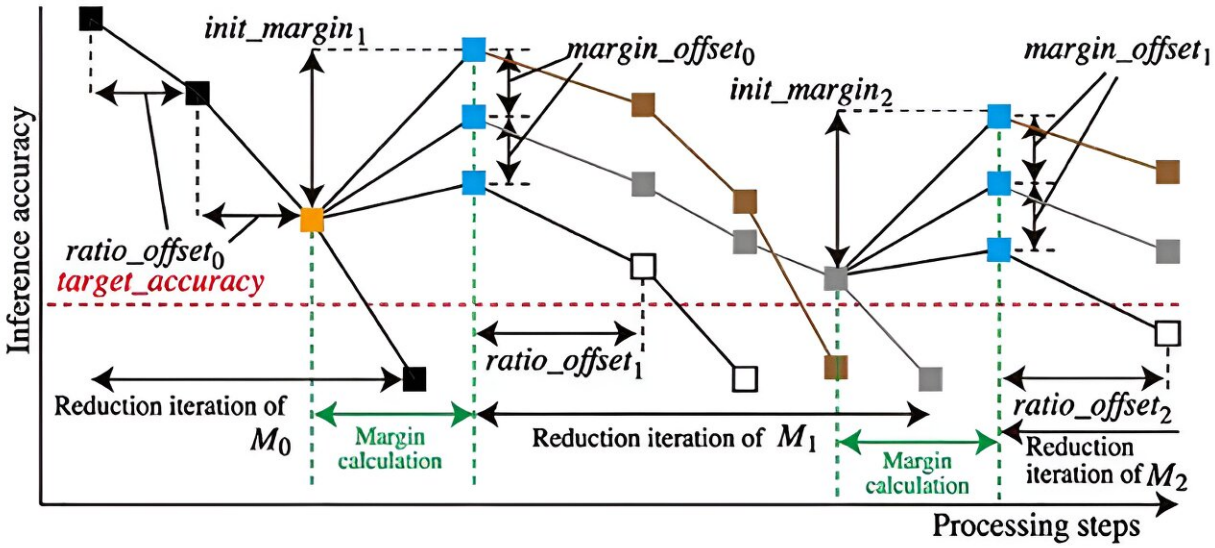
Researchers develop novel method for compactly implementing image-recognizing AI

June 6 2024



Researchers proposes a novel heuristic compression method for convolutional neural network model applying three conventional reduction techniques in the sequence of the integer quantization, the network sliming, and the deep compression. The method autonomously finds a minimal size of a network model by iterating the margin calculations. Credit: *IEEE Access* (2024). DOI: 10.1109/ACCESS.2024.3399541

Artificial intelligence (AI) technology used in image recognition

possesses a structure mimicking human vision and brain neurons. There are three known methods to reduce the amount of data required for calculating and computing the visual and neuronal components. Until now, the application ratio of these methods was determined through trial and error.

Researchers at the University of Tsukuba have developed a new algorithm that automatically identifies the optimal proportion of each method. This algorithm is expected to decrease power consumption in AI technologies and contribute to the miniaturization of semiconductors.

Convolutional neural networks (CNNs) are pivotal in applications such as facial recognition at airport immigration and object detection in autonomous vehicles.

CNNs are composed of convolutional and fully connected layers; the former simulates human vision, while the latter enables the brain to deduce the type of image from visual data.

By reducing the number of data bits used in computations, CNNs can maintain recognition accuracy while substantially reducing computational demands. This efficiency allows the supporting hardware to be more compact.

Three reduction methods have been identified so far: network slimming (NS) to minimize the visual components, deep compression (DC) to reduce the neuronal components, and integer quantization (IQ) to decrease the number of bits used. Previously, there was no definitive guideline on the order of implementation or allocation of these methods.

The new study, published in *IEEE Access*, establishes that the optimal sequence of these methods for minimizing the data amount is IQ, followed by NS and DC. In addition, the researchers have created an

algorithm that determines the application ratio of each method autonomously, removing the necessity for trial and error.

This algorithm enables a CNN to be compressed to 28 times smaller and 76 times faster than previous models.

The implications of this research are poised to transform AI image recognition technology by dramatically reducing computational complexity, power consumption, and the size of AI semiconductor devices. This breakthrough will likely enhance the widespread feasibility of deploying advanced AI systems.

More information: Danhe Tian et al, Heuristic Compression Method for CNN Model Applying Quantization to a Combination of Structured and Unstructured Pruning Techniques, *IEEE Access* (2024). DOI: 10.1109/ACCESS.2024.3399541

Provided by University of Tsukuba