Case ID:M16-181P
Published: 3/16/2017

Inventors

Jae-Sun Seo
Chaitali Chakrabarti
Sairam Arunachalam
Deepak Kadetotad

Contact

Shen Yan
shen.yan@skysonginnovations.com

Coarse-Grain Memory Sparsification for Small-Footprint Deep Neural Networks

Recent breakthroughs in deep neural networks (DNNs) have led to improvements in state-of-the-art speech applications. Conventional DNNs have hundreds or thousands of neurons in each layer, which require a large amount of memory to store the connections between neurons. Implementing these networks in hardware requires a large memory and high computation power. The majority of the memory of DNNs comes from the weights between neurons. Since mobile and wearable devices have constraints on embedded memory card and computational resources, reducing the number of weight parameters without affecting the accuracy is necessary to implement hardware efficiently.

Researchers at ASU have developed a hardware-centric method to design low power DNNs with a significantly smaller memory footprint and less intensive computation. The method works by dropping the weight connections in large blocks and enforcing coarse-grain weight dropping throughout the entire training process. The coarse-grain weight dropping results in a final set of weights efficiently mapped onto arrays with minimal address information for classification. The method drops large blocks with a certain probability and only the remaining weights are subject to training, simplifying the overall computation.

Potential Applications

- Automatic speech recognition software
- Keyword detection
- Neural network training

Benefits and Advantages

- Compressed Weight Memory – The technology can compress the weight memory by 20X compared to floating point DNNs, which reduces hardware size requirements and can help deploy complex DNNs onto mobile/wearable devices
- Coarse-Grain Sparsification – Hardware-aware sparsity during DNN training leads to efficient weight memory compression and significant reduction in the number of computations without losing accuracy

For more information about the inventor(s) and their research, please see:

Dr. Jae-sun Seo's directory webpage