SUPPLEMENTARY MATERIAL - Dithered backprop: A sparse and quantized backpropagation algorithm for more efficient deep neural network training

Simon Wiedemann, Temesgen Mehari, Kevin Kepp, Wojciech Samek

Department of Video Coding & Analytics, Fraunhofer Heinrich-Hertz Institut
Berlin, Germany
{simon.wiedemann,temesgen.mehari,kevin.kepp,wojciech.samek}@hhi.fraunhofer.de

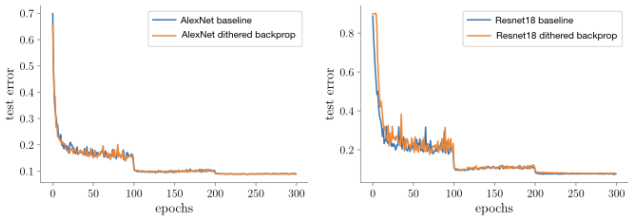

Figure 1: Test error of AlexNet and Resnet18 trained on CIFAR10 over the training epochs.

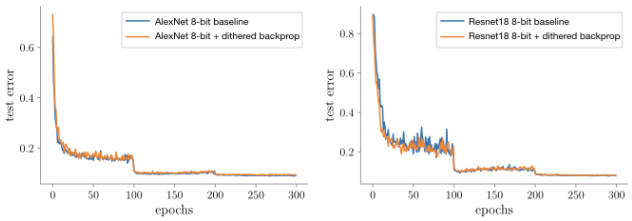

Figure 2: Test error of AlexNet and Resnet18 trained on CIFAR10 over the training epochs

Appendix

More experimental results

In this section of the appendix we show further experimental results.

Convergence of dithered backprop

Figures 1 and 2 show the training curves of AlexNet and Resnet18 trained on CIFAR10 with the baseline method, dithered backprop, the reduced precision training method [1] and the combination of the latter two. As one can see, the training convergence is not affected by dithered backprop in any of the cases.

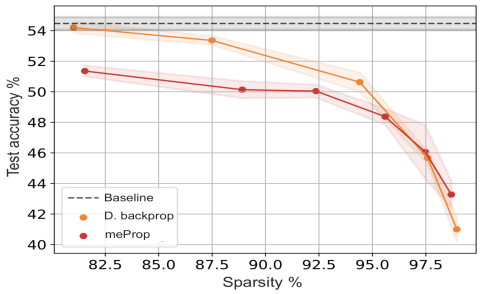

Figure 3: Learning performance at different levels of average sparsity of the preactivation gradients of a multilayer perceptron with two hidden layers (500, 500) trained on CIFAR10, using either regular back propagation (Baseline), dithered backprop (D. backprop) or meProp [2]. Multiple runs with different random seeds were executed for each configuration. Points show mean performance with standard deviation indicated as span.

Comparison to meProp

In figure 3 we show the learning performance of the multilayer perceptron when trained on CIFAR10. As one can see, meProp does not reach as high accuracies as dithered backprop. We attribute this to the biased nature of their gradients estimates, which affects negatively the learning performance of the model.

Distributed training

Here we show the trend of the computational complexity of the convolutional layers as the number of participating nodes increases. As can be seen in figures 4 and 5, the computational decreases as the number of nodes increases.

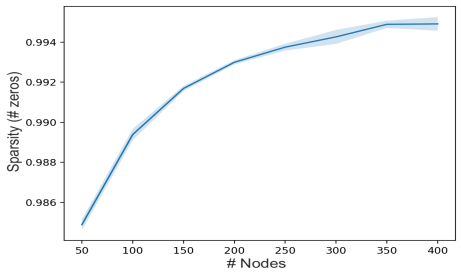

Figure 4: Average sparsity of the preactivation gradients of the convolutional layers of AlexNet trained on CIFAR10 with dithered backprop in a distributed training setting, at different number of participating nodes configuration. As the number of nodes increases, so does the sparsity at each node and therefore its computational savings for training.

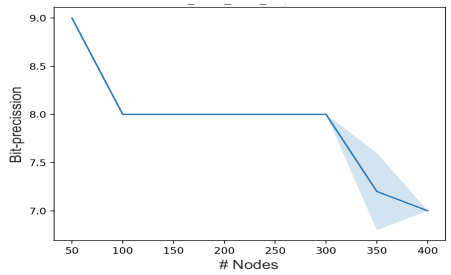

Figure 5: Maximal, worst-case bit-precision of the convolutional layers of AlexNet trained on CIFAR10 with dithered backprop in a distributed training setting, at different number of participating nodes configuration. As the number of nodes increases, the number of bits necessary to represent the non-zero values decreases, and with it the computational cost for training at each node.

References

[1] Ron Banner, Itay Hubara, Elad Hoffer, and Daniel Soudry. Scalable methods for 8-bit training of neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 5145–5153. Curran Associates, Inc., 2018. 1

[2] Xu Sun, Xuancheng Ren, Shuming Ma, and Houfeng Wang. meprop: Sparsified back propagation for accelerated deep learning with reduced overfitting, 2017. 1