QSGD: Communication-Efficient SGD via Gradient Quantization and Encoding

Dan Alistarh (IST Austria)  Demjan Grubic (Google)  Jerry Li (MIT)  Ryota Tomioka (MSR Cambridge)  Milan Vojnovic (LSE)

Scaling Stochastic Gradient Descent (SGD)

Motivation
- Large Datasets: ImageNet: 1.6 million images (~300GB)
- Large Models: ResNet-152 [He et al. 2015]: 152 layers, 60 million parameters

Problem: Communication is a bottleneck to scalability for large models

Quantized SGD (QSGD)

Q is a quantization function which can be communicated w/ fewer bits

\[ Q[v; s] = \|v\|_2 \cdot \text{sgn}(v) \cdot \xi_i(v, s) \]

where \( \xi_i \) is defined by:

\[ \xi_i(v, s) = \frac{1}{s} \left( 1 + \frac{s}{\sqrt{\gamma - s}} \right) \cdot \|v\|_2^2 \quad \text{(Only 2} \|v\|_2^2 \text{ for } s = \sqrt{n}) \]

Properties of our quantization:
1. Unbiasedness: \( E[Q[v; s]] = v \)
2. Sparsity: \( E[n(z)] \leq s + s \sqrt{n} \)
3. Second moment bound

\[ E[Q[v; s]]_2 \leq \left( 1 + \min \left( \frac{n \sqrt{\gamma - s}}{s} \right) \right) \cdot \|v\|_2^2 \]

Experiments

Amazon p16xlarge (16 x NVIDIA K80 GPUs), with NVIDIA DirectConnect

Microsoft CNTK v2.0, with fast MPI-based communication

Tasks: image classification (ImageNet, CIFAR), speech recognition (CMU AN4)

Nets: ResNet, VGG, AlexNet, respectively LSTM, with default parameters

With minimal tuning, QSGD provides similar / improved accuracy and substantially improved scalability!

Contact email: jerryzli@mit.edu  https://gitlab.com/demjangrubic/QSGD

Original: 32n bits.