Stochastic Gradient Push for Distributed Deep Learning

Mido Assran, Nicolas Loizou, Nicolas Ballas, Mike Rabbat
parallel Stochastic Gradient Descent

\[ x^{(k+1)} = x^{(k)} - \gamma^{(k)} \left( \frac{1}{n} \sum_{i=1}^{n} \nabla \tilde{f}_i(x) \right) \]

inter-node average

\[ x^{(k+1)} = \frac{1}{n} \sum_{i=1}^{n} (x^{(k)} - \gamma^{(k)} \nabla \tilde{f}_i(x)) \]
Data Parallel Training

Existing Approaches

1. Parallel SGD (AllReduce gradient aggregation, all nodes)
Data Parallel Training

Existing Approaches

1. Parallel SGD (*AllReduce gradient aggregation, all nodes*)

- Blocks all nodes
Data Parallel Training

Existing Approaches

1. **Parallel SGD** (*AllReduce gradient aggregation, all nodes*)

2. **D-PSGD** (*PushPull parameter aggregation, neighboring nodes*)

3. **AD-PSGD** (*PushPull parameter aggregation, pairs of nodes*)

---

1. Goyal et al., "Accurate, large minibatch sgd: training imagenet in 1 hour," preprint arXiv:1706.02677, 2017.
2. Lian et al., "Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent," NeurIPS, 2017.
3. Lian et al., "Asynchronous decentralized parallel stochastic gradient descent," ICML, 2018.
Data Parallel Training

Existing Approaches

1. Parallel SGD (AllReduce gradient aggregation, all nodes)

2. D-PSGD (PushPull parameter aggregation, neighboring nodes)

3. AD-PSGD (PushPull parameter aggregation, pairs of nodes)

Blocks subsets of nodes and requires deadlock avoidance

1. Goyal et al., "Accurate, large minibatch sgd: training imagenet in 1 hour," preprint arXiv:1706.02677, 2017.
2. Lian et al., "Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent," NeurIPS, 2017.
3. Lian et al., "Asynchronous decentralized parallel stochastic gradient descent," ICML, 2018.
Data Parallel Training

Existing Approaches

1. Parallel SGD (AllReduce gradient aggregation, all nodes)

2. D-PSGD (PushPull parameter aggregation, neighboring nodes)

3. AD-PSGD (PushPull parameter aggregation, pairs of nodes)

Proposed Approach

Stochastic Gradient Push (PushSum parameter aggregation)

nonblocking, no deadlock avoidance required
Stochastic Gradient Push

Enables optimization over directed and time-varying graphs

1. Nedic, A. and Olshevsky, A. "Stochastic gradient-push for strongly convex functions on time-varying directed graphs," IEEE Trans. Automatic Control, 2016.
Stochastic Gradient Push

Enables optimization over directed and time-varying graphs

... naturally enables asynchronous implementations

1. Nedic, A. and Olshevsky, A. "Stochastic gradient-push for strongly convex functions on time-varying directed graphs," IEEE Trans. Automatic Control, 2016.
Stochastic Gradient Push

Local Optimization
Nonblocking Communication

Local Optimization
Nonblocking Communication

Local Optimization
Nonblocking Communication

Local Optimization
Nonblocking Communication

Local Optimization
Nonblocking Communication
Distributed Stochastic Optimization

ImageNet, ResNet 50

32 nodes (256 GPUs) interconnected via 10 Gbps Ethernet
Stochastic Gradient Push

Data Parallelism

Algorithm features:

* nonblocking communication

asynchronous gossip
Stochastic Gradient Push

Data Parallelism

Algorithm features:

* nonblocking communication

* convergence guarantees for smooth non-convex functions with arbitrary (bounded) message staleness

paper: arxiv.org/pdf/1811.10792.pdf
code: github.com/facebookresearch/stochastic_gradient_push
poster: Pacific Ballroom #183