Machine Learning - Specialized Hardware or Massive Distribution

Prajna Kandarpa

October 13, 2017

Overview

The State of Machine Learning

AI Accelerators

ML on mobile devices

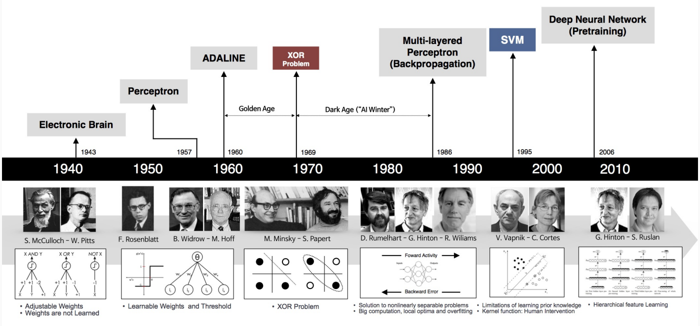
Federated training and optmization

Performance Comparisons

AL acceleration v/s Large scale distribution

The State of Machine Learning

History

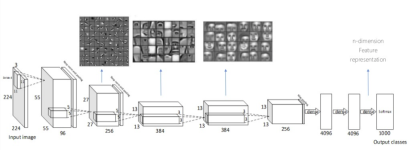


- WWII and automatic control systems
- Cybernetics Society - Weiner, Mcculloch
- Rosenblatt's Perceptron[1]
- Minsky's brutal destruction of the perceptron
- AI Winter

Advent of deep learning in 2010

- Geoff Hinton - Pulled Neural Nets back to mainstream
- Had been doing NN research since 1976 - backpropagating errors to learn representations[2]
- Convnets - Yann LeCunn, AT&T Bell Labs
- Major breakthrough - 2012 - ImageNet competition won by AlexNet

Nvidia and CUDA



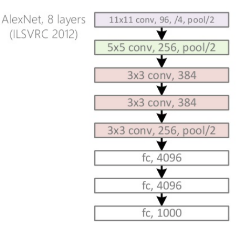


Figure 1: Alexnet Architecture [3]

- AlexNet used GPUs to speed up model training for a CNN
- GPUs are massively parallel floating point calculators
- Massive parallelization and accelerated memory access
- CUDA - parallel computing platform - C, C++ and Fortran
- Deep learning computations are matrix operations (multiplications and factorizations)

AI Accelerators

Facebook - GPU Servers

Facebook Artificial Intelligence Research Lab

- 8 NVIDIA Tesla P100 SXM2 GPUs
- High speed data transfers through PCIE slots
- 9-18 teraflops per GPU
- Not commercially available

Google TPU

- Google DeepMind
- Google TPU and AI Vision
- Cloud TPUs offer 11.5 petaflops performance
- Each Cloud TPU contains 64 TPU units
- 25% increase in performance compared to best GPUs

Intel

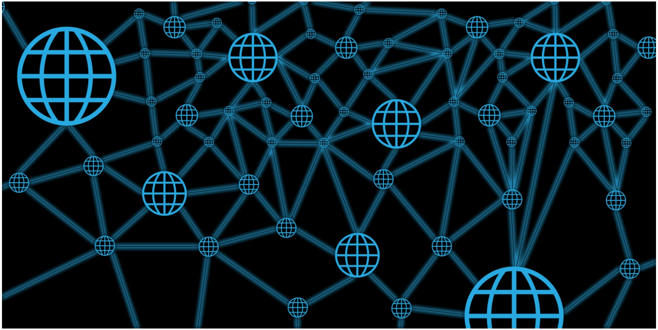
- Nervana Engine - Enterprise compute platform
- 8 Tb/s Memory access speeds
- FPGAs (from Altera acquisition) for AI compute
- Movidius - VPU - focus on drone market, low power

ML on mobile devices

Status Quo

- Pretrained Models and inference engines on device
- Frameworks - Tensorflow, Apple CoreML, Torch, Caffe
- On Device training on full dataset is unfeasible
- Tremendous amount of data collected every second

Decentralized Training



- Important advances made by Google in distributed training
- Isn't everything already distributed ??? - MapReduce, Spark etc.
- Make it even more distributed!! - millions of nodes
- Sometimes number of nodes exceeds number of training samples

Consequences of Decentralized training

BAD

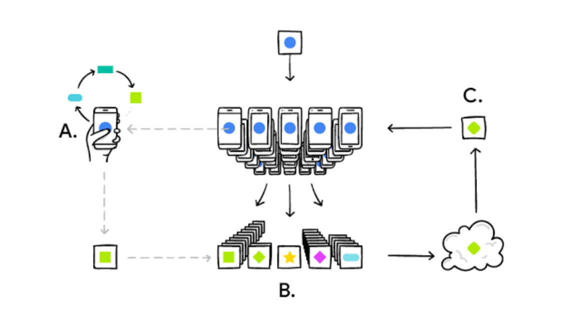
- Training batches not representative of population distribution
- Current gradient descent algorithms don't work.

GOOD

- Less frequent data transfer/retrieval to servers
- Increased user privacy due to lesser data transfer

Federated training and optmization

Federated optimization



- Training Data stays on mobiles
- Global model updates from local model update averages
- Collect model updates not training samples

Communication Efficiency

- Data transfer internally much less expensive than external transfers
- Inter node communication dominates training times
- Fed Learn. reduces communication times to once per day between a node and server

Federated Stochastic variance Reduced Gradient

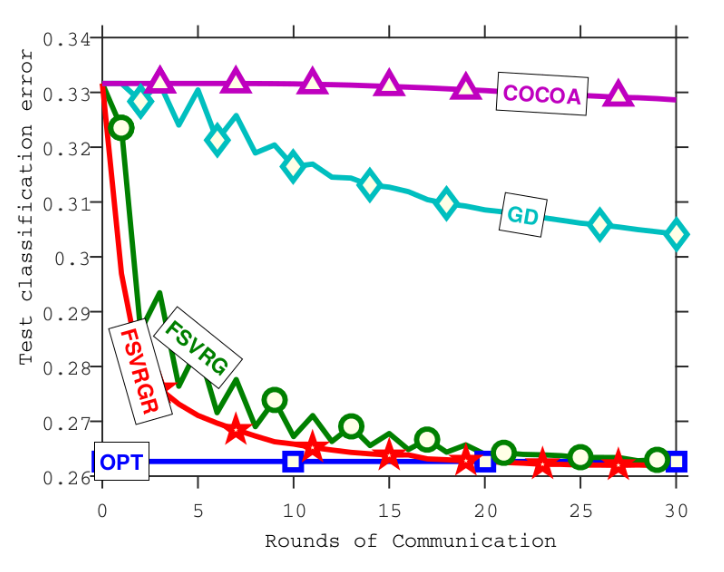
- Uses SVRG for countering the effects of non IID data by explicit variance reduction[4]

DANE (Distributed Approximate Newton Algorithm)

- Estimates the empirical loss function via convex combination of local loss functions
- Form local subproblems, dependent on local data and can converge in O(1) rounds of communication

Performance Comparisons

Predicting comments on Google+ posts



AL acceleration v/s Large scale distribution

Privacy implications

- Secure aggregation of user data
- enable differential privacy to an extent
- only offered by federated learning unless
- Cloud ML providers get better at implementing differential privacy

Cost Benefits

References

"Deep learning 101 - part 1: History and background," https://beamandrew.github.io/deeplearning/2017/02/23/deep_learning_101_part1.html, (Accessed on 10/12/2017).

D. E. Rumelhart, G. E. Hinton, R. J. Williams *et al.*, "Learning representations by back-propagating errors," *Cognitive modeling*, vol. 5, no. 3, p. 1, 1988.

A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

J. Konecný, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," *CoRR*, vol. abs/1610.02527, 2016. [Online]. Available: http://arxiv.org/abs/1610.02527