BREAK OUT SESSION

Large-Scale Inference and Learning

BIG DATA PI Meeting 2016

Overarching Themes in this Area

- Large-scale means different things to different people
  - size: KB, MB, GB, PB, …;
  - complexity: multi-modal, high-dimensional;
  - distributed/parallel: storage; computation, etc.
- Resource constraints:
  - Human insight/knowledge
  - data-throughput, computational platform, runtime, communication/energy, storage
- Large-scale data doesn’t equal large-scale inference

Recent Successes (last 3 years)

- Novel methodological approaches to parallel/distributed computation for learning/inference using synchronization, randomization, etc.
- Novel “systems” approaches (databases, HPC, hardware, libraries) to machine learning
- Application areas – computer vision, NLP, speech, computational genomics (data + algorithms + “systems”)
- Large-scale stochastic optimization
  - Low-profile example: large-scale logistic regression
  - High-profile example: large-scale deep learning

Major Obstacles Impeding More Rapid Progress

- Value of interdisciplinary research (siloed funding mechanisms, venues for publishing, academia incentives)
- Funding for ML software platforms and guidance for non-experts
- Availability of large-scale data (+computing platforms) in academia
- Human-machine interactions for inference/learning
  - Better labeled data
  - Interpretable algorithms
  - Effects of data preprocessing decisions
  - Interactive data analysis methods
- Interdisciplinary training: CS/ML/statistics “interfacing” to applied domains

Areas of Neglect

- Interpretable machine learning
  - Large-scale ML “systems” for hypotheses testing
  - Interactive ML “pipelines” for large-scale learning
- Single machine out-of-core ML algorithms
- Domain-specific vs. domain-agnostic algorithms
- Scalable complex models and methods
- Theory for large-scale non-convex distributed learning and optimization
- Other issues
  - Better partnership with industry to access large-scale data
  - Platforms for sharing data
  - Benchmark Validation datasets

Strategic Priorities & Investments That Will Advance Innovation

- Combining “systems” (databases, HPC, hardware, libraries) and ML algorithms
  - To help scientific domain experts/problems
- Interpretable machine learning
  - Interpretable to people who generate/use the data
  - Human-machine interactive analytics
- Valuing interdisciplinary research
  - siloed funding mechanisms
  - venues for publishing, academia incentives, etc

Other stuff

Overarching Themes in this Area

- Large-scale means different things to different people (size: MB, GB, PB, …; complexity: multi-modal, high-dimensional; storage; computation)
  - Lots of data
  - Distributed & parallel computing
  - Resource (data-throughput, computational platform, runtime, communication/energy, storage) limited learning
- Large-scale data doesn’t equal large-scale inference

Recent Successes (last 3 years)

- Novel approaches for parallel computation using synchronization, randomization, etc. s
- Deep learning (large data + GPU, CPU, FPGAs etc implementation)
- Application areas – computer vision, NLP, speech, Computational genomics (data + algorithms)
- “Systems” (HPC, hardware, libraries) + machine learning

Major Obstacles Impeding More Rapid Progress

- Big data, small labels – innovative algorithms, human-machine interactions
- Availability of large-scale data (+computing platforms) in academia
- CS and ML training of students, particularly for applied domains
- Funding for ML software platforms and guidance for non-experts
- Value of interdisciplinary research (siloed funding mechanisms, venues for publishing, academia incentives)

Areas of Neglect

- Interactive pipeline for large-scale learning
- Building-blocks for machine learning
- Scalable complex models and methods
- Out-of-core learning algorithms
- Domain-specific personlized vs. generalized algorithms
- Interpretable machine learning
- Theory for large-scale non-convex distributed learning and optimization
- Large-scale systems that test hypotheses

Strategic Priorities & Investments That Will Advance Innovation

- "Systems" (HPC, hardware, libraries) + machine learning

- Better partnership with industry to access large-scale data
- Platforms for sharing data
- Benchmark Validation datasets