Marco Serafini

Homepage

COMPSCI 692S - Spring 20

Course homepage

Reading list

This is the (evolving) reading list for the seminar.

Systems for Machine Learning

Overviews and practical reports

- Strategies and Principles of Distributed Machine Learning on Big Data
- A Berkeley View of Systems Challenges for AI
- Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective
- Machine Learning at Facebook: Understanding Inference at the Edge
- In-Datacenter Performance Analysis of a Tensor Processing Unit

Compilers and optimizers

- Optimizing Data-Intensive Computations in Existing Libraries with Split Annotations
- TASO: Optimizing Deep Learning Computation with Automatic Generation of Graph Substitutions
- TVM: An automated end-to-end optimizing compiler for deep learning
- Halide: a language and compiler for optimizing parallelism, locality, and recomputation in image processing pipelines
- HELIX: Holistic Optimization for Accelerating Iterative Machine Learning
- Machine Learning Systems are Stuck in a Rut
- Autograph: Imperative-Style Coding with Graph-Based Performance
- RLgraph: Flexible Computation Graphs for Deep Reinforcement Learning
- Automating Dependence-Aware Parallelization of Machine Learning Training on Distributed Shared Memory
- Improving the Expressiveness of Deep Learning Frameworks with Recursion
- Wootz: A Compiler-Based Framework for Fast CNN Pruning via Composability

Distributed Training

- Horovod: fast and easy distributed deep learning in TensorFlow
- MXNet: A Flexible and Efficient Machine Learning Library for Heterogeneous Distributed Systems
- PipeDream: Generalized Pipeline Parallelism for DNN Training
- Beyond Data and Model Parallelism for Deep Neural Networks
- Can Decentralized Algorithms Outperform Centralized Algorithms? A Case Study for Decentralized Parallel Stochastic Gradient Descent
- Crossbow: Scaling Deep Learning with Small Batch Sizes on Multi-GPU Servers
- Parallax: Sparsity-aware Data Parallel Training ofDeep Neural Networks
- Poseidon: An Efficient Communication Architecture for Distributed Deep Learning on GPU Clusters
- Managed Communication and Consistency for Fast Data-Parallel Iterative Analytics

Inference

- Pretzel: Opening the Black Box of Machine Learning Prediction Serving Systems
- TFX: A TensorFlow-Based Production-Scale Machine Learning Platform
- Parity models: erasure-coded resilience for prediction serving systems
- GRNN: Low-Latency and Scalable RNN Inference on GPUs
- µLayer: Low Latency On-Device InferenceUsing Cooperative Single-Layer Acceleration and Processor-Friendly Quantization
- Low Latency RNN Inference with Cellular Batching
- Continuum: A Platform for Cost-Aware, Low-Latency Continual Learning

Resource management

- Proteus: agile ML elasticity through tiered reliability in dynamic resource markets
- Multi-tenant GPU Clusters for Deep Learning Workloads: Analysis and Implications
- Gandiva: Introspective Cluster Scheduling for Deep Learning
- Themis: Fair and Efficient GPU Cluster Scheduling for Machine Learning Workloads
- Ease.ml: Towards Multi-tenant Resource Sharing for Machine Learning Workloads
- Scheduling Optimus: An Efficient Dynamic Resource Scheduler for Deep Learning Clusters

DBMS + ML

- Enabling and Optimizing Non-linear Feature Interactions in Factorized Linear Algebra
- Rafiki: Machine Learning as an Analytics Service System
- Cloudy with High Chance of DBMS: A 10-year Prediction for Enterprise-Grade ML
- Extending Relational Query Processing with ML Inference
- Towards Scalable Dataframe Systems

Testing and debugging

- Machine Learning Testing: Survey, Landscapes and Horizons
- DeepXplore: Automated Whitebox Testing of Deep Learning Systems
- Mistique: A System to Store and Query Model Intermediates for Model Diagnosis
- DeepBase: Deep Inspection of Neural Networks

Machine Learning for Systems

ML for ML systems

- Google Vizier: A Service for Black-Box Optimization
- Learning to Optimize Tensor Programs
- Device Placement Optimization with Reinforcement Learning

ML for programming languages

- Learning to Represent Programs with Graphs
- Generative Code Modeling with Graphs
- code2seq: Generating sequences from structured representations of code

ML for resource management

- Learning scheduling algorithms for data processing clusters
- CherryPick: Adaptively Unearthing the Best Cloud Configurations for Big Data Analytics
- Arrow: Low-Level Augmented Bayesian Optimization for Finding the Best Cloud VM
- Micky: A cheaper alternative for selecting cloud instances

- Scout: An Experienced Guide to Find the Best Cloud Configuration

ML for database management systems

- The Case for Learned Index Structures
- SkinnerDB: regret-bounded query evaluation via reinforcement learning
- Neo: A learned query optimizer
- SOSD: A Benchmark for Learned Indexes

ML for video data management

- NoScope: Optimizing Neural Network Queries over Video at Scale
- Neural Adaptive Video Streaming with Pensieve
- Focus: Querying Large Video Datasets with Low Latency and Low Cost
- DeepLens: Towards a Visual Data Management System
- Panorama: A Data System for Unbounded Vocabulary Querying over Video