Research Faculty Summit 2018
Systems | Fueling future disruptions
OpenPAI: The Open Source Initiative for AI Platform in China

Fan Yang
Microsoft Research Asia
Background and Motivation

• Artificial Intelligence (AI) becomes one major focus and heated across academia and industries

• Major opportunity to democratize AI through innovations on AI infrastructure
  • Lower the entry bar for new comers
  • Facilitate AI education
  • Speed up AI research
  • Accelerate the penetration of AI across industries
Current Status and Challenges of AI Platform

• Still in early stage: ad hoc ways to build/deploy an AI platform
  • It is easy to build small-scale platforms with a narrowed, specific purpose

• Need an AI platform that works in different environment and application scenario
  • On-premise, cloud, and hybrid environment
  • Image/video, speech, language, vertical domain
  • High compatibility, extensibility, manageability, efficiency
The Importance of an AI Platform

• Infrastructure support for the advance of artificial intelligence
  • Deep learning algorithms and frameworks run on
  • Manage hardware

• A platform to boost AI innovation and productivity
  • Allow researchers/practitioners to focus on AI innovation, instead of the hassles of infra. construction, deployment, management, and optimization
  • Enable results sharing, build a community for mutual learning/leveraging, and rapid innovation
An Open Platform for AI R&D and Education

- The co-development of AI innovation, AI education, and AI platform evolution
  - Research, education, and production

- Design AI course project, perform training to grow AI talent pool

- Open source, result sharing, a community for collaborative innovation
Overview

**Cognitive Ability**
- Visual Perception
- NLU
- Speech Recognition

**Tools**
- AI Lifecycle Management
- Management Integration
- Compiling Optimization

**Shared Resources**
- Practice
- Curriculum
- Data

**Management and Intelligent Platform**
- Heterogeneous Cluster Management & Scheduling
- System-level Intelligent Optimization
- Deep Learning and Intelligent Exploration

**Infrastructure**
- CNTK
- TensorFlow
- PyTorch
- SQL
- Distributed Clusters
- Edge
Open Platform for AI (OpenPAI)

• A platform for GPU cluster management
  • Openness: open source (MIT), open collaboration model
  • Extensibility: support all deep learning frameworks, GPU/FPGA/ASIC
  • Modularity: micro-service, different component choice (storage, scheduler)
  • Efficiency: fine-grained GPU scheduling, support IB/RDMA
  • Manageability: job and platform monitoring, deployment, upgrade, etc.
  • Robust: fault tolerance
  • Practicability: leverage the mature design and practice in Microsoft
OpenPAI Architecture

External Services / IDE Tools

Job Management Interface
Platform Mgmt Interface
Deployment Interface
Platform Interface

Deep Learning Job
Runtime Support

Big Data Job
Runtime Support

Higher Layer ML Service
Runtime Support

Job Management
Service Management
Resource Management
Storage

Container Orchestration (Docker/Kubernetes)

Platform Deployment

Host Environment (Azure, AWS, OpenStack, Ubuntu)

Heterogeneous Hardware (GPU, CPU, FPGA, RDMA)
Implementation

- Visual Studio Tools for AI
- Managed by k8s
- Managed by Hadoop+AI

- Web Portal
- REST Server
- Launcher Server
- Hadoop+AI
- HDFS
- ZooKeeper
- Deep Learning Jobs (PAI Runtime)
- Big Data Jobs (PAI Runtime)
- AutoML Jobs (PAI Runtime)

- Kubernetes Cluster Management
- Docker / Ubuntu

https://github.com/Microsoft/pai
An Open Ecosystem: Engage with China AI Community
China Open AI Platform Alliance
Case Study: School of Information Science and Technology of USTC

300 Researchers & Students vs. 400 GPUs

Previous Practice

High operation overhead
High learning curve
Low GPU utilization (~60%)
USTC OpenPAI Deployment

• Cluster management through OpenPAI’s WebUI

• Lower learning curve for AI researchers
  • Research result shared across teams through Docker images and job config files
  • New user jumpstarts with pre-built Docker images

• Improved GPU cluster utilization (>75%)
OpenPAI Boosts AI Innovation

• Winner Award of NTIRE@CVPR 2018

Challenge on Spectral Reconstruction from RGB Images

254 1392x1300x31 hi-def images
200+ layers deep neural model
A Platform for AI Innovation

• A much needed playground for AI R&D
  • A realistic, production environment where AI research should target at

• Hassle-free platform for non-professionals
  • Low ops overhead (deployment, maintenance)
  • 100% WebUI-based job management
  • Enable AI researchers to focus on AI research itself

• Easy result sharing and reproducing
  • Strong portability through standard API support

• Enable system research for AI
Some Research Opportunities on AI Infrastructure

- System primitives for deep learning scheduling
  - MapReduce task, DAG scheduling vs. process time-sharing, migration
- Memory management abstraction
  - Virtual memory, memory compression, etc.
- Joint consideration of cluster scheduling and AutoML
  - Optimization goal changed from one job to a group of job
- Cross job optimization
  - Common subexpression elimination, etc.
- Fairness, sharing incentives
Conclusion

• Facilitate the research on artificial intelligence

• Provide an efficient, ease-of-use AI infrastructure
  • Platform support to put AI research into practice

• Open to collaboration
  • Develop the platform together
  • An open community: welcome future collaboration
Thank you!