OpenMindSpore: Project Proposal 2023

OpenMindSpore Team

FUTUREWEI Technologies

Overview

- Project started in middle of March.
- 5 FTEs + 2 interns (in 3 months)
- Initial project planning performed from March to mid-May.
- Project chartered by ITMT committee on 5/24.
- Project approved by OIEC on 6/14.
- MindSpore community meeting started on 4/13
  - Series of meetings on 4/28, 6/1, 6/30, 8/4, 8/18, 9/20, 11/23, 11/30, 12/8
- Participation in academic and industry research conferences
  - CVPR 2022, Linaro/ARM Confidential AI Tech Event, HPDC 2022, OSDI/ATC 2022, MLSys 2022, Ray Summit 2022, ICML 2022, AI Hardware Summit 2022, NeurIPS 2022, PyTorch Conference 2022
- Technical deliverables on 7/31, 8/31, 9/30, 12/31

Achievement & Contributions of 2022 (1)

- Comeback to MindSpore after 4-year of gap since 2018
  - Catching up latest technologies including transformer models, graph optimization, automatic differentiation, sparse computation optimization, neural architecture search, graph neural networks, etc. introduced in the past 4 years
- Setup of local multi-GPU multi-node distributed training environment
  - Request for fixing documentations on compilation and testing to MindSpore community
  - PR for complete GPU distributed training guide to MindSpore community
- Investigation on **elastic distributed training** for MindSpore
  - Proposal and investigation of architectural integration of MindSpore with distributed execution engine (targeting to Yuanrong)
- Contribution of **state-of-art neural network models** to MindSpore
  - Presentations on SOTA models for MindSpore community and publications to technical blog site
  - Upstream of transformer-based vision transformer model (YOLOS) for image classification (soon)

Achievement & Contributions of 2022 (2)

- Provision of cutting-edge technical insight information to MindSpore community
  - **Technical reports** on CVPR 2022, Linaro/ARM Confidential AI Tech Event, HPDC 2022, OSDI/ATC 2022, MLSys 2022, ICML 2022, AI Hardware Summit 2022, NeurIPS 2022, PyTorch Conference 2022
  - Presentations and discussions during 10+ MindSpore **community meeting** from March
- External collaboration efforts
  - Introduction of MindSpore and technical review and discussion with **Prof. Harry Xu (UCLA/BreezeML)** for elastic and fault-tolerant training and disaggregated memory
- Non-technical
  - Setup of formal procedure of open meetings and compliant communication channel in MindSpore community
  - Setup of formal procedure of open-source release with OEIC
  - Setup of formal procedure of technical report publication to public domain

Achievement & Contributions of 2022 (3)

- Automatic modeling
  - Making the AI/ML platform smarter and more adaptive by knowledge/experience.
- Markov chain approach to generating CNN models
  - This method can be extended to various types of automatic modelling, like decision tree, support vector machine, non-linear regression, etc.
- Causal inference
  - Figure out the crucial steps in modeling, for instance, is max-pooling necessary?
- 10+ patents of AI/ML related algorithms
- 3 academic books on AI/ML
- 4 translated books on AI/ML, complex adaptive system (CAS), risk analysis/control, etc.

MindSpore Strategy for 2023 (1)

- Technical leadership and guidance provision
  - Scalable distributed training (continuation from year 2022)
    - Runtime enhancement of MindSpore (targeting to Yuanrong)
    - Better scheduling support via refactoring of current computation and MPI-based communication (based on community challenge #3)
  - ML Ops enhancements
    - Fault-tolerance and fast recovery of distributed training in case of hardware/software failure/exception (based on community challenge #4)
    - Distributed model serving (based on community request)
  - State-of-the-art neural network model support (continuation from year 2022)
    - Upstream of transformer-based computer vision models to MindSpore community
    - Research on vision model enhancement
  - Academic/industrial research collaboration
    - Technical review and discussion with US/European university labs
    - Seminars on ML/AI at UC Berkeley

MindSpore Strategy for 2023 (2)

- Data augmentation
  - Let the AI/ML platform grasp the key features of objects in supervised learning.
- Intervention approaches to feature engineering
- Active learning
  - Let the learner know its limitations and improve itself automatically.
- 4+ patents of AI/ML related algorithms
- 1 academic book on AI/ML
- 2 translated books on AI/ML topics

Tasks to be done (as of Dec 2022 and Beyond)

- Communication and synchronization (Norbert)
  - Replace MindSpore's MPI to Ray's (for GPU)
  - Using Ray AIR for shared store
- Data preprocessing (Zongfang)
  - Data loading and saving
  - Sharding
  - Compatible with Ray AIR
- DataParallelTrainer (Won)
  - Importing MindSpore native objects to Python (current issue)
  - Scheduler and distributed workers
  - Utilizing communication, synchronization, and data preprocessing above
- MindSporeTrainer (Won)
  - Inherited from DataParallelTrainer
- MindSporePredictor (Won)
  - For serving, subset of MindSporeTrainer
- Test code (all)

FUTUREWEI Technologies

Project Requirements from Counterpart in Dec 2022

- T5, GPT downstream task development, distributed training related code development
- Distributed support for reinforcement learning (it is not clear what distributed requirements are there)
- Relevant optimization features of distributed training such as mixed precision and compression

FUTUREWEI Technologies

Additional Requirements from Counterpart in Jan 2023

- Enhancements with ML Ops
  - Start a simple task first like fast recovery, multi-model training, then go deep into distributed execution.
- Multi-model training in distributed heterogeneous environment
  - Pathways-like
- GPU support in Windows
  - Nvidia GPUs
  - MSVC and/or WSL2 could be used.
- Elastic training in distributed setting
- Freezing gradient computing (checkpointing)
  - PipeTransformer-like

FUTUREWEI Technologies

Quarterly Milestone/Deliverable (1)

- 1st quarter
  - Distributed computing
    - Distributed data preprocessing and sharding support
    - Communication and synchronization support
  - SOTA model
    - Fine tuning of YOLOS
    - Update SOTA model report
  - Research
    - New ML technologies in ChatGPT and their possible applications to AI products
    - Data augmentation and its approaches to active learning (one patent)
    - Overview of Monte Carlo methods for ML/AI

Quarterly Milestone/Deliverable (2)

- 2nd quarter
  - Distributed computing
    - Failure recovery support
    - Checkpoint support
    - GPU support on Windows
  - SOTA model
    - Port one multi-modal model MDETR or ALBEF
    - Following SOTA works
  - Research
    - A survey of reinforcement learning (RL): theories and algorithms
    - Stochastic simulation approaches to approximation and optimization, including simulated annealing, evolutionary computation (e.g., generic algorithm, ant colony optimization), Monte Carlo tree search, proximal policy optimization (PPO), etc.

FUTUREWEI Technologies

Quarterly Milestone/Deliverable (3)

- 3rd quarter
  - Distributed computing
    - Data parallel distributed training support for GPU
  - SOTA model
    - Port another multi-modal model
    - Update SOTA model report
  - Research
    - A study of variational optimization approaches to RL: fasten the policy searching
    - Reward strategies for RL based on mixed signals (one translated book)

Quarterly Milestone/Deliverable (4)

- 4th quarter
  - Distributed computing
    - Pipeline parallel training support for GPU
    - Dynamic scheduling support
  - SOTA model
    - Fine tuning ported models
    - Update SOTA model report
  - Research
    - Probabilistic graphical models (PGMs): theories and algorithms, including Bayesian network, particle filter, Markov random field, conditional random field, etc.