Running Deep Learning Applications in Production Level

Hao Ji, Data Scientist

USC Center for Advanced Research Computing

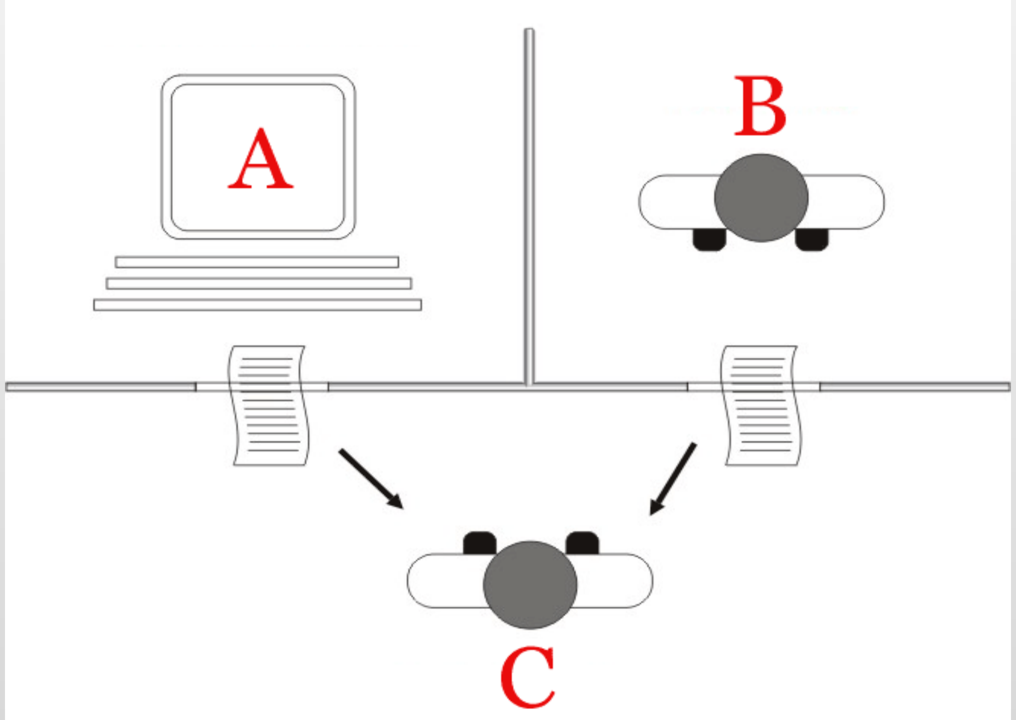
Content

History of AI

Pre-1950s: Foundations of AI

The formal study of AI began with mathematicians and philosophers like Alan Turing.

In 1950, Alan Turing proposed the famous "Turing Test," a measure of a machine's ability to exhibit human-like intelligence. This concept laid the groundwork for future AI research.



The "standard interpretation" of the Turing test, in which player C, the interrogator, is given the task of trying to determine which player – A or B – is a computer and which is a human. The interrogator is limited to using the responses to written questions to make the determination.

History of AI

1950s and 1960s: Early Research and Symbolic AI

The term "artificial intelligence" was coined in 1956 at the Dartmouth Conference, where the field of AI research was officially established.

During this period, researchers focused on symbolic AI, using rules and logic to mimic human problem-solving processes.

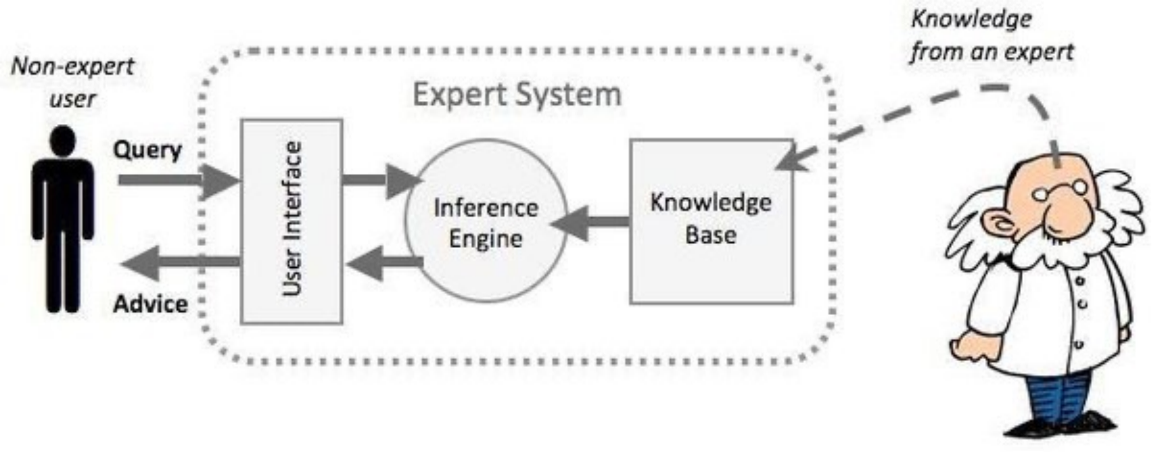
History of AI

1970s and 1980s: Knowledge-Based Systems AI

Research saw a shift towards knowledge-based systems and expert systems.

These systems utilized large knowledge bases to reason and draw conclusions about specific domains.

Though promising, early expert systems faced limitations due to the complexity of representing all human knowledge in a formalized manner.

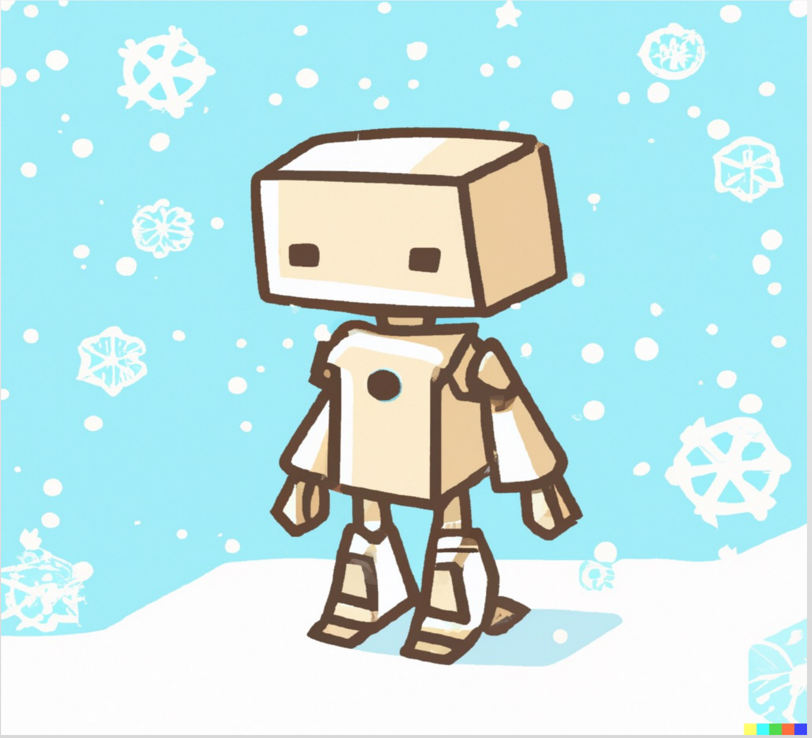


History of AI

1980s and 1990s: AI Winter

During this period, AI faced a decline in interest and funding due to unmet expectations and the inability to deliver on grand promises.

Progress in AI did not match initial optimism, leading to a phase known as the "AI Winter."



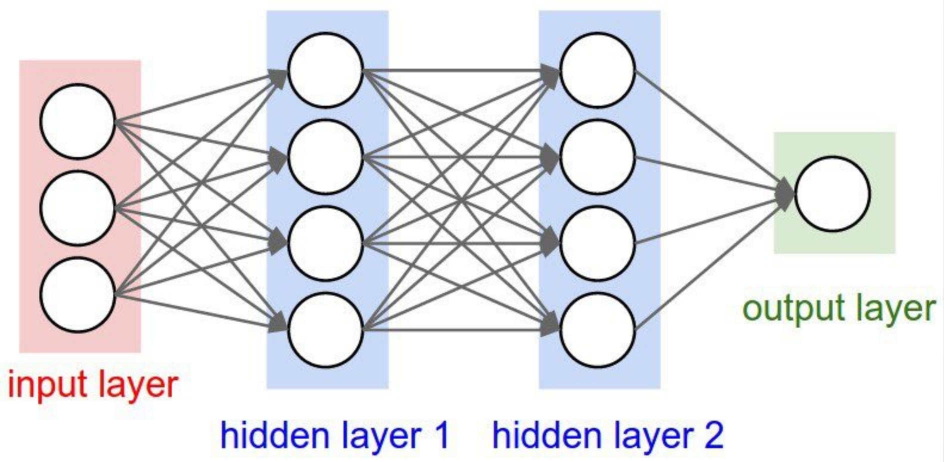
History of AI

Late 1990s and 2000s: Machine Learning and Neural Networks

The resurgence of AI came with the advent of machine learning techniques, particularly neural networks.

Improved computational power and access to vast amounts of data allowed neural networks to excel in tasks like image recognition and language processing.

Support vector machines and other machine learning algorithms also gained popularity.

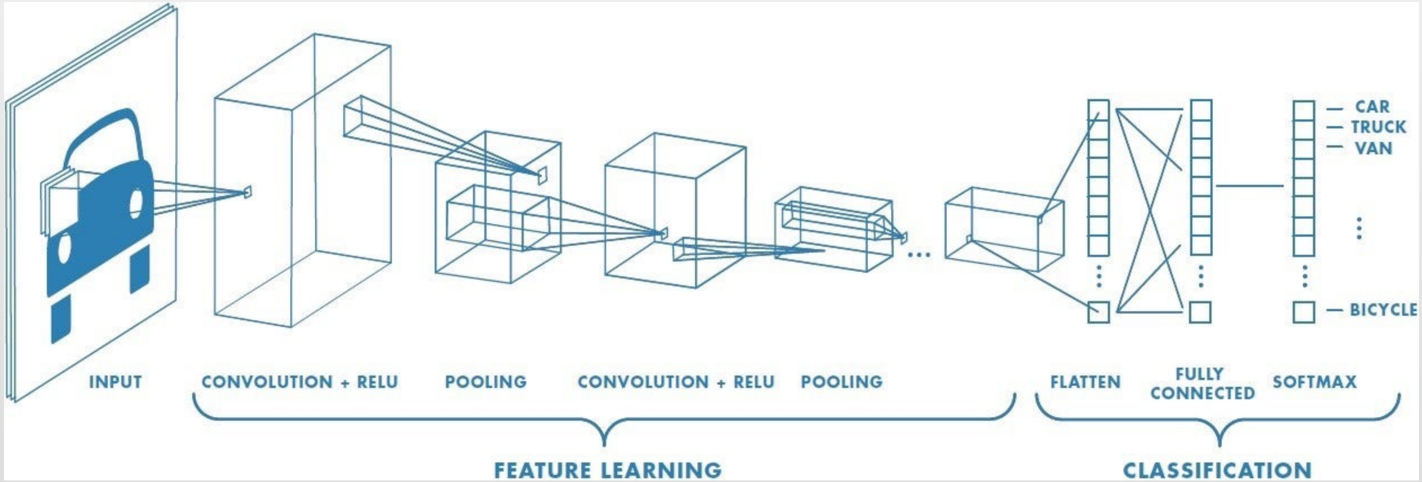


History of AI

2010s: Deep Learning and AI Breakthroughs

Deep learning, a subset of machine learning using artificial neural networks with multiple layers, became the driving force behind many AI breakthroughs.

Applications of AI proliferated in various fields, including natural language processing, computer vision, robotics, and autonomous vehicles.



History of AI

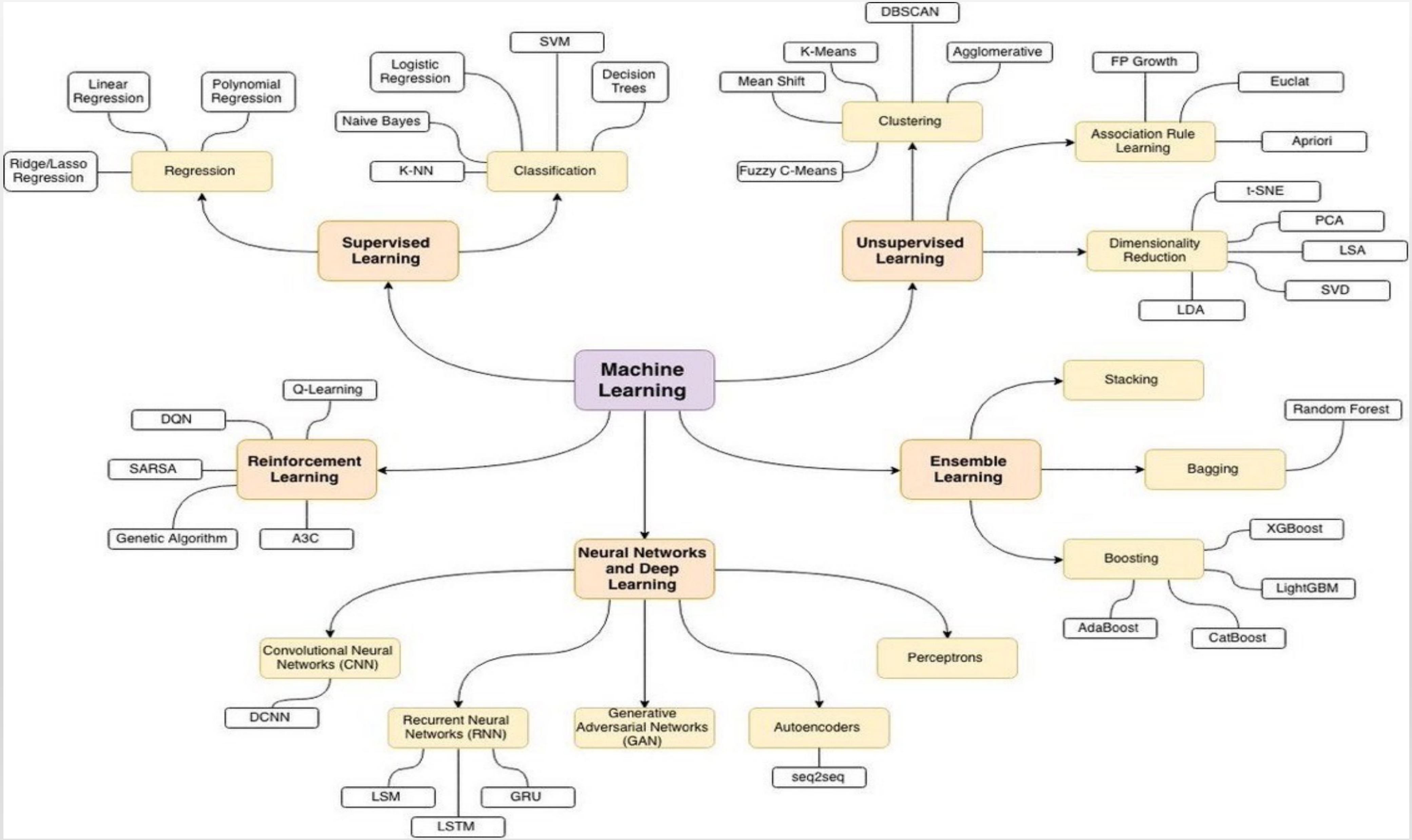
Present and Beyond

AI continues to evolve rapidly, with ongoing research in areas like reinforcement learning, explainable AI, Large Language Model and AI ethics.

Integration of AI in everyday life through virtual assistants, smart devices, and recommendation systems is becoming increasingly common.

However, concerns about AI's societal impact, privacy, and ethical implications persist.

Introduction of Machine Learning



ML Perf Benchmark

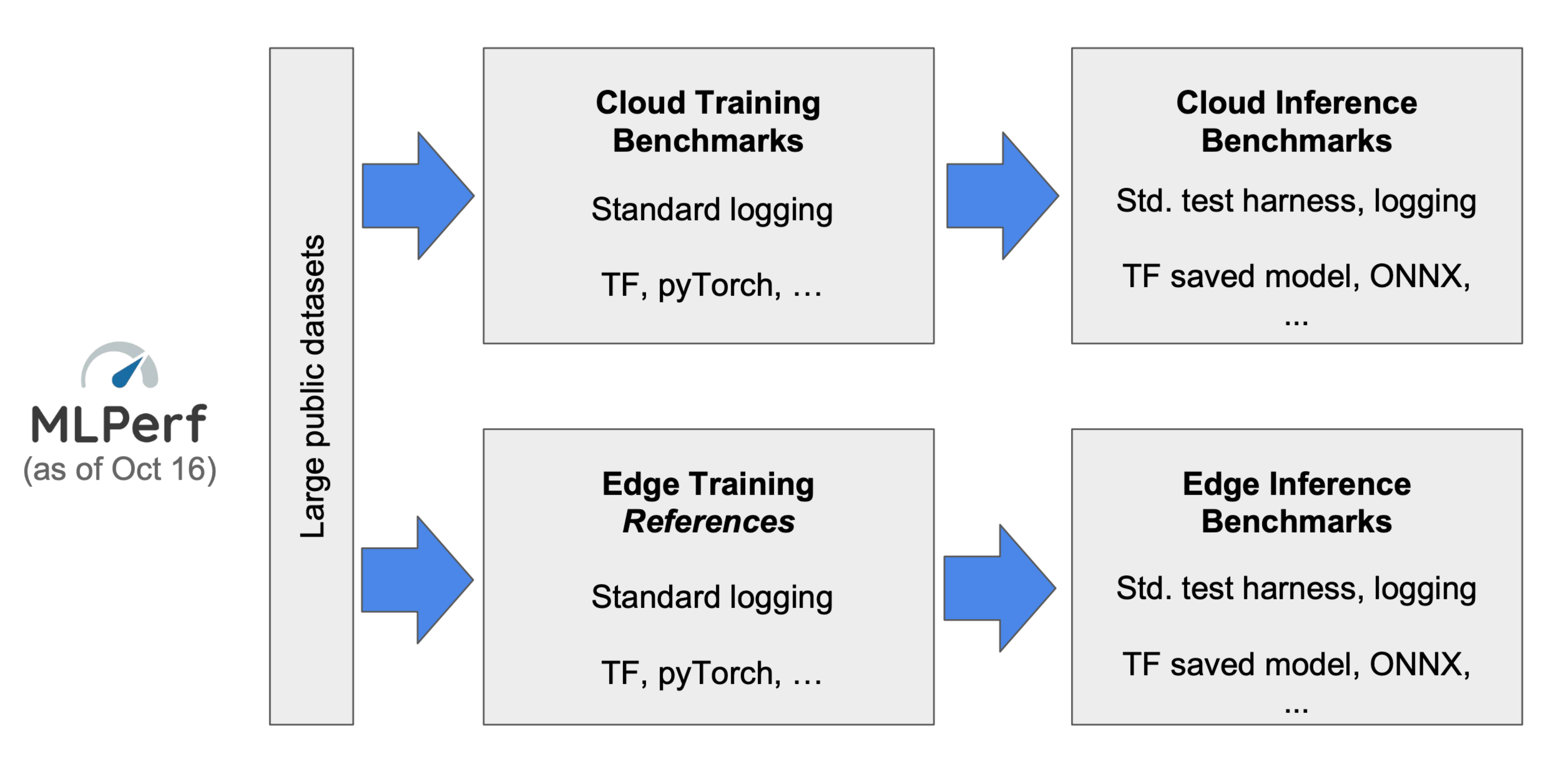
ML Perf: A broad ML benchmark suite for measuring the performance of ML software frameworks, ML hardware accelerators, and ML cloud and edge platforms

Alibaba	AMD	Arm	Baidu	Cadence	Cerebras	Cisco
Cray	Dividiti	Enflame Tech	Esperanto	Google	Groq	Huawei
Intel	MediaTek	Mentor Graphics	Mythic	NetApp	NVIDIA	One Convergence
Rpa2ai	Sambanova	Samsung S.LSI	Sigopt	Synopsys	Tensyr	Wave Computing

Harvard University	Stanford University	University of Arkansas, Littlerock	University of California, Berkeley	University of Minnesota	University of Texas, Austin	University of Toronto

MLPerf

ML Perf Benchmark



ML Perf Benchmark

Task	Model	Dataset
Image Classification	ResNet-50	ImageNet
Object Detection	Mask-RCNN SSD	MS-COCO 2017
Translation	Google NMT Transformer	WMT 16 WMT 17
Recommendation	Neural Collaborative Filtering	MovieLens ml-20m
Reinforcement Learning	Minigo	NA
Speech Recognition	DeepSpeech2*	Librispeech

ML Perf Benchmark

Task	Task Description	Dataset	Quality metric	Sample Apps
Recognition	Classify an input into one of many categories. Alternatively, generate a high dimensional embedding that can be used for recognition	Imagenet/COCO Input: RGB image of size XX x YY Output: label index	Top-1 error rate	Face authentication, Music recognition

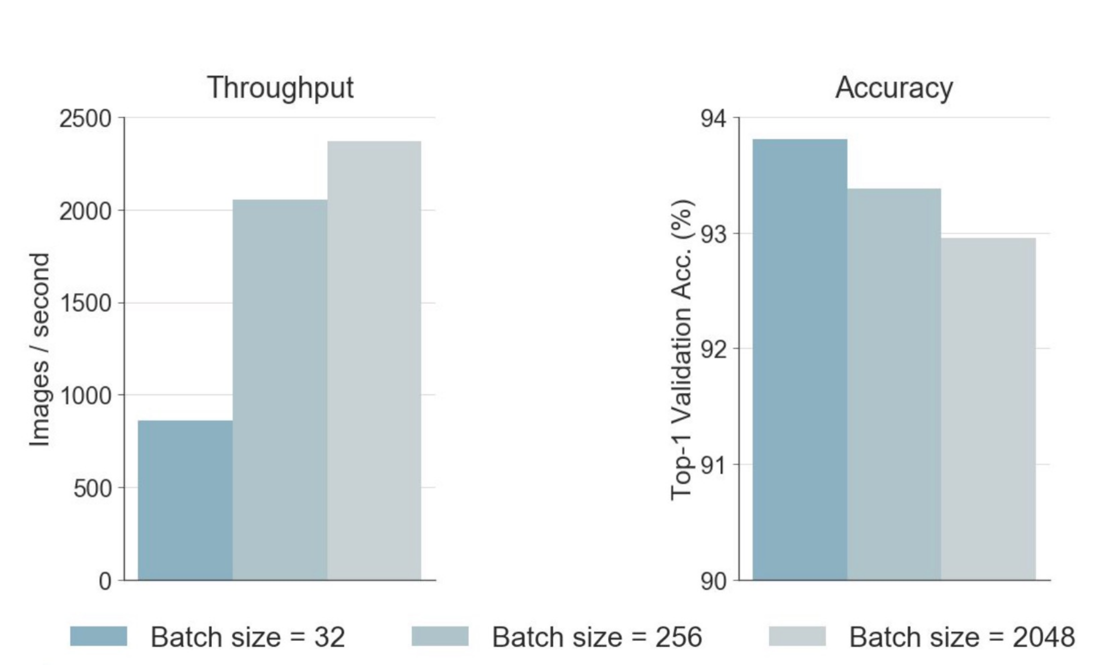
ML Perf Benchmark

Performance

How fast is a model for training, inference?

Quality

How good are a model's predictions?



- End to end training of a Resnet56 CIFAR10 model
- Nvdia P100 with 512GB of memory and 28 CPU cores
- TensorFlow 1.2 compiled from source with CUDA 8.0 and CuDNN 5.1

ML Perf Benchmark

Closed division submissions

- Requires using the specified model
- Limits overfitting
- Enables apples-to-apples comparison
- Simplifies work for HW groups

Open division submissions

- Open division allows using any model
- Encourages innovation
- Ensures Closed division does not stagnate

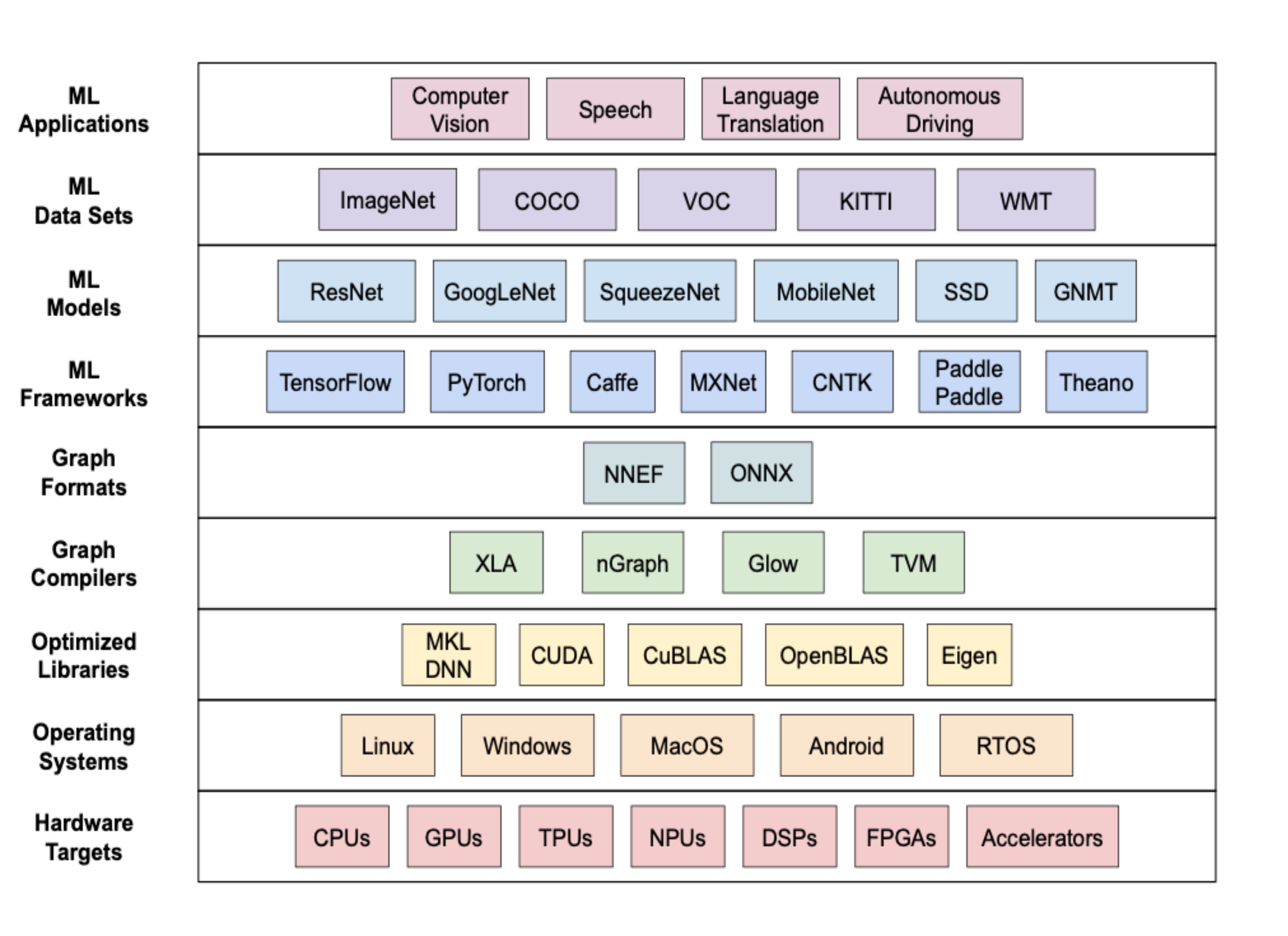
ML Perf Benchmark

Training v1.1

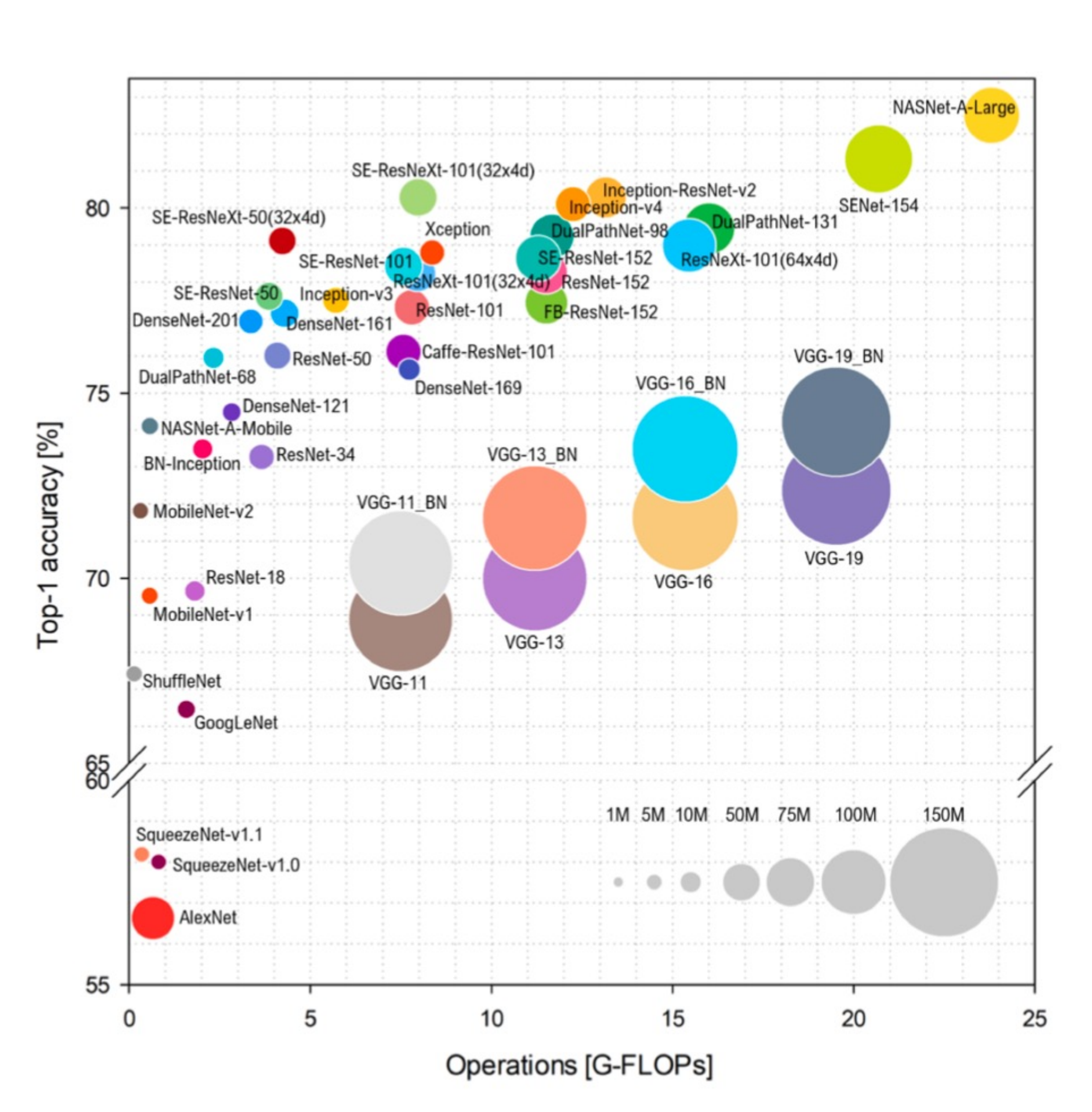
Closed | Open

								Benchmark results (minutes)									
								Image classification	Image segmentation (medical)	Object detection, light-weight	Object detection, heavy-weight	Speech recognition	NLP	Recom-mendation	Reinforce-ment Learning		
								ImageNet	KiTS19	COCO	COCO	LibriSpeech	Wikipedia	1TB Clickthrough	Go		
ID	Submitter	System	Processor	#	Accelerator	#	Software	ResNet	3D U-Net	SSD	Mask R-CNN	RNN-T	BERT [1]	DLRM	Minigo	Details	Code
Available cloud																	
1.1-2000	Azure	nd96amsr_a100_v4_ngc21.09_merlin_hugectr	AMD EPYC 7V12	2	NVIDIA A100-SXM4-80GB (400W)	8	Merlin HugeCTR with NVIDIA DL Frameworks Release 21.09							1.875		details	code
1.1-2001	Azure	nd96amsr_a100_v4_ngc21.09_mxnet	AMD EPYC 7V12	2	NVIDIA A100-SXM4-80GB (400W)	8	MXNet NVIDIA Release 21.09	29.720	25.400	8.309						details	code
1.1-2002	Azure	nd96amsr_a100_v4_ngc21.09_pytorch	AMD EPYC 7V12	2	NVIDIA A100-SXM4-80GB (400W)	8	PyTorch NVIDIA Release 21.09				47.064	37.550	21.213			details	code
1.1-2003	Azure	nd96amsr_a100_v4_ngc21.09_tensorflow	AMD EPYC 7V12	2	NVIDIA A100-SXM4-80GB (400W)	8	TensorFlow NVIDIA Release 21.09								319.410	details	code
1.1-2004	Azure	nd96amsr_a100_v4_n4_ngc21.09_pytorch	AMD EPYC 7V12	8	NVIDIA A100-SXM4-80GB (400W)	32	PyTorch NVIDIA Release 21.09				14.912					details	code
1.1-2005	Azure	nd96amsr_a100_v4_n8_ngc21.09_mxnet	AMD EPYC 7V12	16	NVIDIA A100-SXM4-80GB (400W)	64	MXNet NVIDIA Release 21.09	4.587		1.517						details	code
1.1-2006	Azure	nd96amsr_a100_v4_n8_ngc21.09_pytorch	AMD EPYC 7V12	16	NVIDIA A100-SXM4-80GB (400W)	64	PyTorch NVIDIA Release 21.09						3.111			details	code
1.1-2007	Azure	nd96amsr_a100_v4_n9_ngc21.09_mxnet	AMD EPYC 7V12	18	NVIDIA A100-SXM4-80GB (400W)	72	MXNet NVIDIA Release 21.09		3.800							details	code
1.1-2008	Azure	nd96amsr_a100_v4_n16_ngc21.09_pytorch	AMD EPYC 7V12	32	NVIDIA A100-SXM4-80GB (400W)	128	PyTorch NVIDIA Release 21.09					4.533				details	code
1.1-2009	Azure	nd96amsr_a100_v4_n32_ngc21.09_tensorflow	AMD EPYC 7V12	64	NVIDIA A100-SXM4-80GB (400W)	256	TensorFlow NVIDIA Release 21.09								30.714	details	code
1.1-2010	Azure	nd96amsr_a100_v4_n34_ngc21.09_pytorch	AMD EPYC 7V12	68	NVIDIA A100-SXM4-80GB (400W)	272	PyTorch NVIDIA Release 21.09				3.908					details	code
1.1-2011	Azure	nd96amsr_a100_v4_n48_ngc21.09_tensorflow	AMD EPYC 7V12	96	NVIDIA A100-SXM4-80GB (400W)	384	TensorFlow NVIDIA Release 21.09								24.802	details	code
1.1-2012	Azure	nd96amsr_a100_v4_n96_ngc21.09_mxnet	AMD EPYC 7V12	192	NVIDIA A100-SXM4-80GB (400W)	768	MXNet NVIDIA Release 21.09		1.262							details	code
1.1-2013	Azure	nd96amsr_a100_v4_n128_ngc21.09_mxnet	AMD EPYC 7V12	256	NVIDIA A100-SXM4-80GB (400W)	1024	MXNet NVIDIA Release 21.09	0.583		0.455						details	code
1.1-2014	Azure	nd96amsr_a100_v4_n128_ngc21.09_pytorch	AMD EPYC 7V12	256	NVIDIA A100-SXM4-80GB (400W)	1024	PyTorch NVIDIA Release 21.09						0.656			details	code
1.1-2015	Azure	nd96amsr_a100_v4_n192_ngc21.09_pytorch	AMD EPYC 7V12	384	NVIDIA A100-SXM4-80GB (400W)	1536	PyTorch NVIDIA Release 21.09					3.205				details	code
1.1-2016	Azure	nd96amsr_a100_v4_n224_ngc21.09_tensorflow	AMD EPYC 7V12	448	NVIDIA A100-SXM4-80GB (400W)	1792	TensorFlow NVIDIA Release 21.09								17.439	details	code
1.1-2017	Azure	nd96amsr_a100_v4_n256_ngc21.09_mxnet	AMD EPYC 7V12	512	NVIDIA A100-SXM4-80GB (400W)	2048	MXNet NVIDIA Release 21.09	0.438								details	code
1.1-2018	Azure	nd96amsr_a100_v4_n256_ngc21.09_pytorch	AMD EPYC 7V12	512	NVIDIA A100-SXM4-80GB (400W)	2048	PyTorch NVIDIA Release 21.09						0.422			details	code
Available on-premise																	
1.1-2019	Baidu	1_node_8_A100_NGC21.05_MXNet	Intel(R) Xeon(R) Platinum 8350C	2	NVIDIA A100-SXM4-80GB (400W)	8	MXNet NVIDIA Release 21.05	28.605								details	code
1.1-2020	Baidu	1_node_8_A100_PaddlePaddle	Intel(R) Xeon(R) Platinum 8350C	2	NVIDIA A100-SXM4-80GB (400W)	8	PaddlePaddle (branch: develop, commitID: 605e7f0)	28.613								details	code
1.1-2021	Dell	DSS8440x8A100-PCIE-40GB	Intel(R) Xeon(R) Gold 6248R CPU @ 3.00GHz	2	NVIDIA A100-PCIE-40GB (250W)	8	NGC MXNet 21.09 , NGC PyTorch 21.09 , NGC TensorFlow 21.09-tf1	38.871		11.193	61.505		66.631		393.431	details	code
1.1-2022	Dell	DSS8440x8A100-PCIE-40GB-NVBridge	Intel(R) Xeon(R) Gold 6248R CPU @ 3.00GHz	2	NVIDIA A100-PCIE-40GB (250W)	8	NGC MXNet 21.09 , NGC PyTorch 21.09 , NGC TensorFlow 21.09-tf1	37.083		10.899	58.571				405.424	details	code
1.1-2023	Dell	R750xax4A100-PCIE-80GB	Intel(R) Xeon(R) Gold 6338 CPU @ 2.00GHz	2	NVIDIA A100-PCIE-80GB (300W)	4	NGC MXNet 21.09 , NGC PyTorch 21.09 , NGC TensorFlow 21.09-tf1	62.949	60.586	18.321	93.134	84.025	56.260		590.009	details	code
1.1-2024	Dell	R750xax4A100-PCIE-80GB-8368	Intel(R) Xeon(R) Platinum 8368 CPU @ 2.40GHz	2	NVIDIA A100-PCIE-80GB (300W)	4	NGC MXNet 21.09 , NGC PyTorch 21.09 , NGC TensorFlow 21.09-tf1	64.131		18.390	91.562		56.131			details	code
1.1-2025	Dell	2xR750xax4A100-PCIE-80GB	Intel(R) Xeon(R) Gold 6338 CPU @ 2.00GHz	4	NVIDIA A100-PCIE-80GB (300W)	8	NGC MXNet 21.09 , NGC PyTorch 21.09 , NGC TensorFlow 21.09-tf1	32.087								details	code
1.1-2026	Dell	DSS8440x8A100-PCIE-80GB	Intel(R) Xeon(R) Gold 6248R CPU @ 3.00GHz	2	NVIDIA A100-PCIE-80GB (300W)	8	NGC MXNet 21.09 , NGC PyTorch 21.09 , NGC TensorFlow 21.09-tf1	33.553	28.543	9.835	55.033	60.763	41.269			details	code
1.1-2027	Dell	DSS8440x10A100-PCIE-80GB	Intel(R) Xeon(R) Gold 6248R CPU @ 3.00GHz	2	NVIDIA A100-PCIE-80GB (300W)	10	NGC MXNet 21.09 , NGC PyTorch 21.09 , NGC TensorFlow 21.09-tf1	28.187		8.223	46.948		36.859			details	code
1.1-2028	Dell	4xR750xax4A100-PCIE-80GB	Intel(R) Xeon(R) Gold 6338 CPU @ 2.00GHz	8	NVIDIA A100-PCIE-80GB (300W)	16	NGC MXNet 21.09 , NGC PyTorch 21.09 , NGC TensorFlow 21.09-tf1	17.336								details	code
1.1-2029	Dell	8xR750xax4A100-PCIE-80GB	Intel(R) Xeon(R) Gold 6338 CPU @ 2.00GHz	16	NVIDIA A100-PCIE-80GB (300W)	32	NGC MXNet 21.09 , NGC PyTorch 21.09 , NGC TensorFlow 21.09-tf1	10.586		3.477						details	code
1.1-2030	Dell	XE8545x4A100-SXM-40GB	AMD EPYC 7763 64-Core Processor	2	NVIDIA A100-SXM4-40GB (400W)	4	NGC MXNet 21.09 , NGC PyTorch 21.09 , NGC TensorFlow 21.09-tf1	61.820		16.998	95.157	79.563				details	code
1.1-2031	Dell	XE8545x4A100-SXM-80GB	AMD EPYC 7713 64-Core Processor	2	NVIDIA A100-SXM4-80GB (500W)	4	NGC MXNet 21.09 , NGC PyTorch 21.09 , NGC TensorFlow 21.09-tf1	56.326	55.999	16.244	83.774	106.542	38.855	9.522	451.293	details	code
1.1-2032	Dell	2xXE8545x4A100-SXM-80GB	AMD EPYC 7713 64-Core Processor	4	NVIDIA A100-SXM4-80GB (500W)	8	NGC MXNet 21.09 , NGC PyTorch 21.09 , NGC TensorFlow 21.09-tf1	30.123		8.735	48.788	35.068	26.547			details	code
1.1-2033	Fujitsu	PRIMERGY-GX2460M1-mxnet	AMD EPYC 7502 32-Core Processor	2	NVIDIA A100-PCIe-40GB (250W)	4	MXNet NGC21.09	70.294	49.946	20.916						details	code
1.1-2034	Fujitsu	PRIMERGY-GX2460M1-pytorch	AMD EPYC 7502 32-Core Processor	2	NVIDIA A100-PCIe-40GB (250W)	4	Pytorch NGC21.09					109.216	127.843			details	code

ML Perf Benchmark

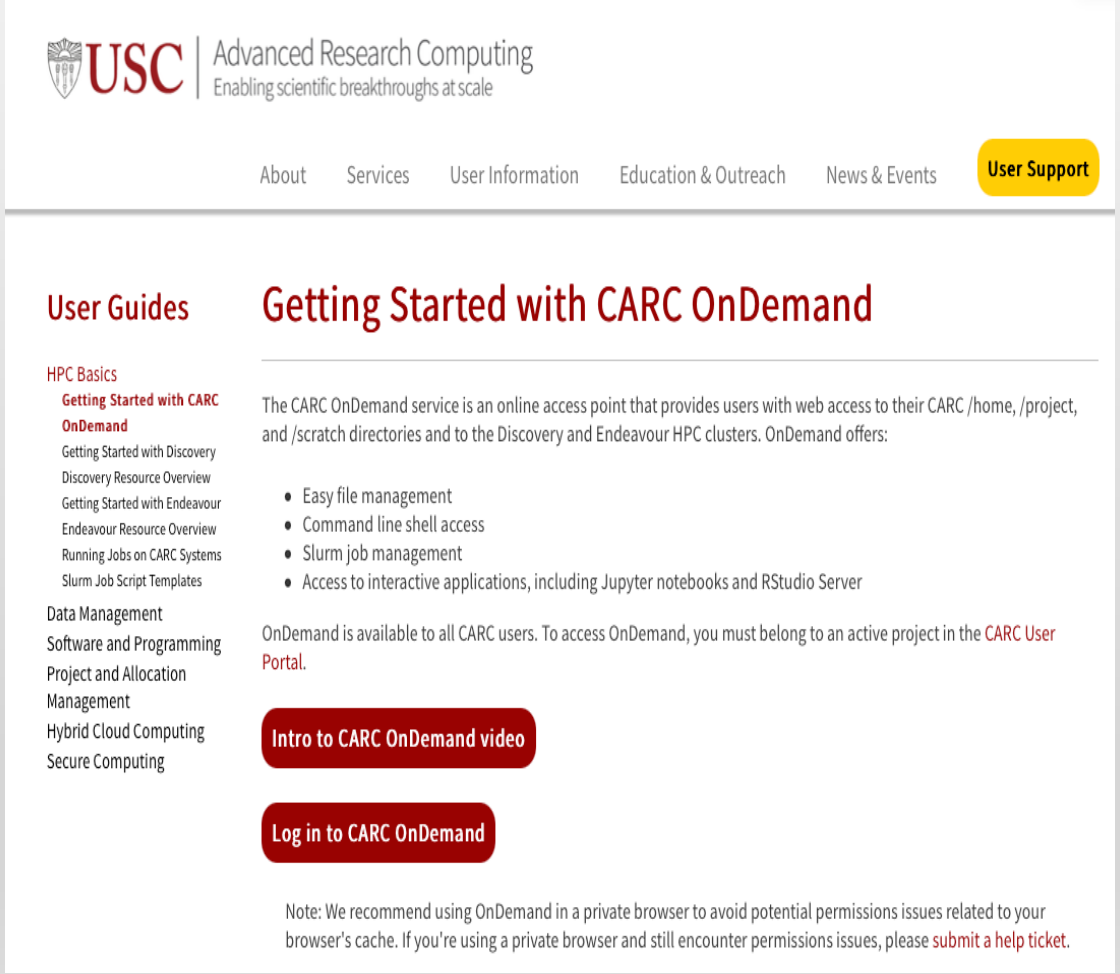


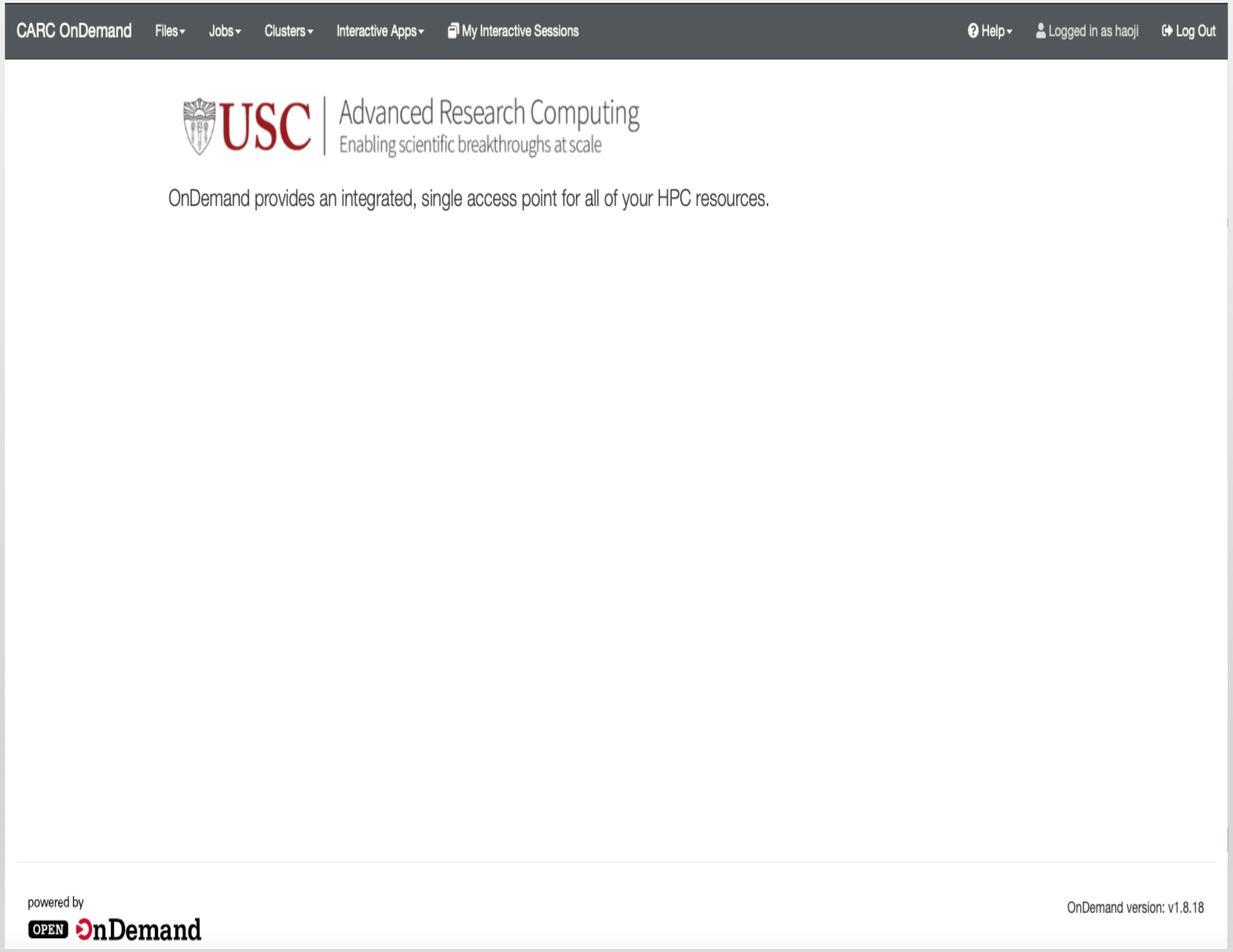
ML Perf Benchmark



CARC OnDemand

Web Address: https://carc-ondemand.usc.edu





Using Anaconda on CARC

Anaconda: package and environment manager primarily used for open-source data science packages for the Python and R programming languages.

User Guides

- HPC Basics
- Data Management
- Software and Programming
  - Software Module System
  - Building Code With CMake
  - Using MPI
  - Using GPUs
  - Using Julia
  - Using Python
  - **Using Anaconda**
  - Using R
  - Using Stata
  - Using MATLAB
  - Using Rust
  - Using Launcher
  - Using Singularity
  - Using Tmux
  - Installing Jupyter Kernels
  - Horovod for Distributed Deep Learning
- Project and Allocation Management
- Hybrid Cloud Computing
- Secure Computing

Using Anaconda

Anaconda is a package and environment manager primarily used for open-source data science packages for the Python and R programming languages. It also supports other programming languages like C, C++, FORTRAN, Java, Scala, Ruby, and Lua.

Using Anaconda on CARC systems

Begin by logging in. You can find instructions for this in the Getting Started with Discovery or Getting Started with Endeavour user guides.

To use Anaconda, first load the corresponding module:

```
module purge
module load conda
```

This module is based on the minimal Miniconda installer. Included in all versions of Anaconda, Conda is the package and environment manager that installs, runs, and updates packages and their dependencies. This module also provides Mamba, which is a drop-in replacement for most `conda` commands that enables faster package solving, downloading, and installing.

The next step is to initialize your shell to use Conda and Mamba:

```
mamba init bash
source ~/.bashrc
```

https://www.carc.usc.edu/user-information/user-guides/software-and-programming/anaconda

Using Anaconda on CARC

https://pytorch.org

START LOCALLY

Select your preferences and run the install command. Stable represents the most currently tested and supported version of PyTorch. This should be suitable for many users. Preview is available if you want the latest, not fully tested and supported, builds that are generated nightly. Please ensure that you have **met the prerequisites below (e.g., numpy)**, depending on your package manager. Anaconda is our recommended package manager since it installs all dependencies. You can also install previous versions of PyTorch. Note that LibTorch is only available for C++.

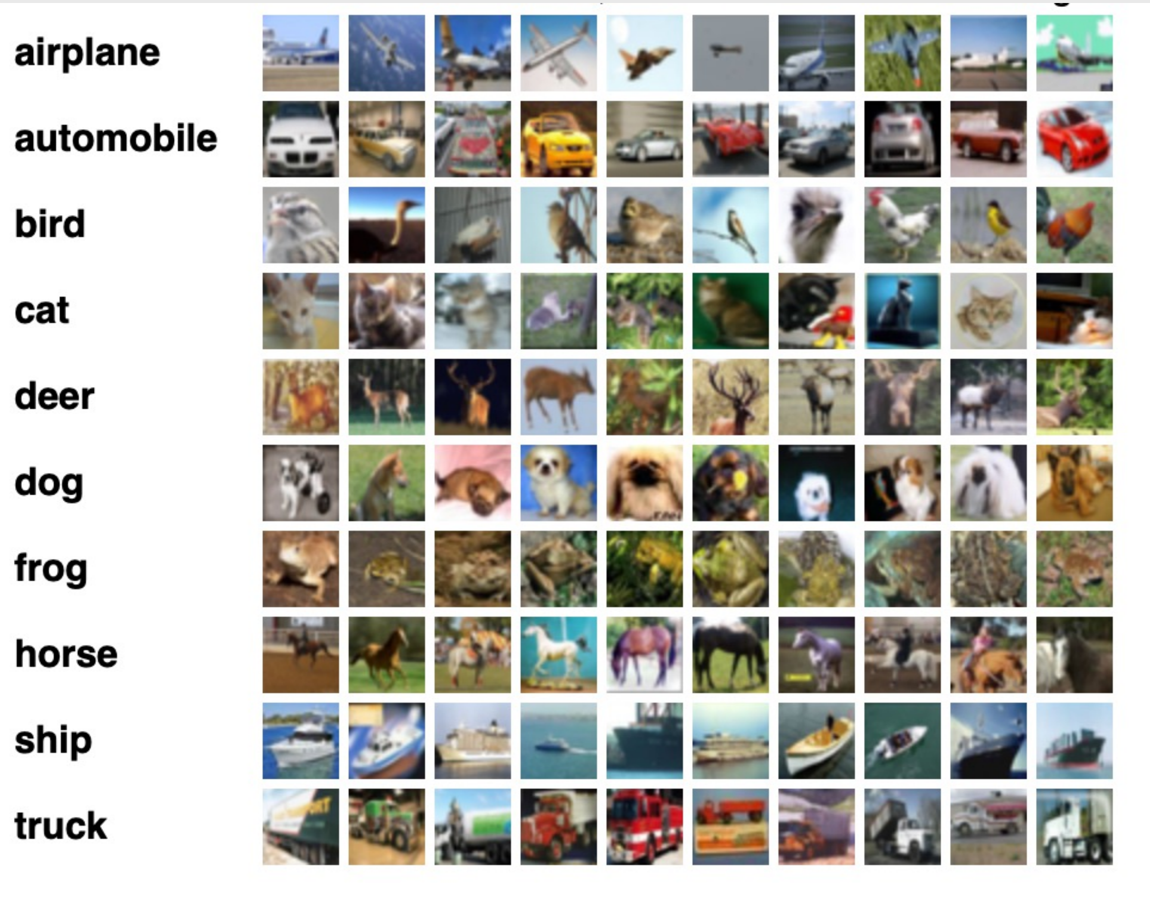
PyTorch Build	**Stable (2.1.0)**	Preview (Nightly)		
Your OS	**Linux**	Mac	Windows	
Package	**Conda**	Pip	LibTorch	Source
Language	**Python**	C++ / Java		
Compute Platform	**CUDA 11.8**	CUDA 12.1	ROCm 5.6	CPU

Run this Command:

```
conda install pytorch torchvision torchaudio pytorch-cuda=11.8 -c pytorch -c nv
idia
```

An example of Training

The problem we're going to solve today is to train a model to classify CIFAR10 datasets.
CIFAR-10 dataset consists of 60000 32x32 color images in 10 classes, with 6000 images per class.
There are 50000 training images and 10000 test images.
We will use Restnet50 model as a basis for training



Slurm script for job submission

Sample Job Script for running training jobs on GPU partition

```
#!/bin/bash
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --mem=32g
#SBATCH --time=02:00:00
#SBATCH --partition=gpu
#SBATCH --gres=gpu:1
##SBATCH --account=hpcsuppt_613

module purge
eval "$(conda shell.bash hook)”
conda activate torch_benchmark
python CIFAR10_ResNet50.py
```

Common types of GPU

https://www.carc.usc.edu/user-information/user-guides/hpc-basics/discovery-resources

Partition	CPU model	CPU frequency	CPUs/node	GPU model	GPUs/node	Memory/node	Nodes
gpu	xeon-6130	2.10 GHz	32	V100	2	184 GB	29
gpu	xeon-2640v4	2.40 GHz	20	P100	2	123 GB	38
gpu	epyc-7282	2.80 GHz	32	A40	2	248 GB	12
gpu	epyc-7513	2.60 GHz	64	A100	2	248 GB	12

Common types of GPU

https://www.carc.usc.edu/user-information/user-guides/hpc-basics/discovery-resources

GPU specifications in the GPU partition

There are four kinds of GPUs in the GPU partition: A100, A40, V100, P100. The following is a summary table for the GPU specifications:

GPU model	Architecture	Memory	Memory Bandwidth	Base Clock Speed	Cuda Cores	Tensor Cores	Single Precision Performance (FP32)	Double Precision Performance (FP64)
A100	ampere	40GB	1.6TB/s	765MHz	6912	432	19.5TFLOPS	9.7TFLOPS
A40	ampere	48GB	696GB/s	1305MHz	10752	336	37.4TFLOPS	584.6GFLOPS
V100	volta	32GB	900GB/s	1230MHz	5120	640	14TFLOPS	7TFLOPS
P100	pascal	16GB	732GB/s	1189MHz	3584	n/a	9.3TFLOPS	4.7TFLOPS