FAQ

CONTENTS & NAVIGATION

HP WORKSTATIONS FOR MACHINE LEARNING

FAQ

CONTENTS & NAVIGATION

MACHINE LEARNING OVERVIEW

Q: What is Machine Learning?
A: Historically, Machine Learning relied on hand coded algorithms to extract features from data for tasks such as image classification and voice recognition. Modern Machine Learning algorithms use deep neural networks (DNNs) to 'learn' the key features in a wide range of data. These new algorithms require large amounts of data and the computational power of a modern computers, especially GPUs, to successfully extract the features.

Q: What is a neural network?
A: A neural network is a computer code that loosely mimics the neurons in our brains, with multiple neurons connected to each other. Each neuron is teachable in how it should respond to signals from other neurons connected to it. Each neuron can be thought of as a 'node,' and each node will 'fire' depending on the signals it gets from nodes attached to it.

Q: How does Machine Learning work?
A: The process of 'learning' involves training a neural network using very large sets of training data which are labelled with correct responses. To improve its ability to get correct answers the neural network is repeatedly exposed to the training data. The neural network 'learns' to produce correct answers by trying, succeeding, or failing and adjusting how and when neural network nodes signal connected nodes. Once a network completes the training phase it can then be optimized and deployed – this latter phase is called inferencing.

Q: What are Deep Neural Networks, or DNNs?
A: These are neural networks that have many, many layers (or levels) of interconnected nodes which can better extract the key features from the data and what is really needed to 'learn.'

Q: What are Deep Learning Frameworks?
A: The research community has developed many frameworks which are better suited to addressing different kinds of problems such as speed recognition, voice recognition, image recognition, etc. These frameworks mainly differ in the features and flexibility that are built into the programming model.

Q: Why does Machine Learning training require high-end computer configurations?
A: Machine Learning algorithms that are designed to excel at human perception characteristics such as image classification and voice recognition can contain over 1M parameters. These parameters must be 'taught' to produce an accurate result. The teaching process requires huge amounts of data computational resources to be effective.

COMPUTATIONAL REQUIREMENTS FOR MACHINE LEARNING

Q: Why is the HP Z8 G4 Workstation ideal for Machine Learning training?
A: The HP Z8 G4 workstation is designed to host the large amounts of fast storage and computational power needed for Deep Learning training. It is also designed for 25/7 operation[1] to facilitate training sessions that can take days to complete and facilitate the many sessions that may be required to complete a training workflow. HP has recommended HP Z8 G4 configurations for Deep Learning training work that we believe will deliver best-in-class productivity for training DNNs. Learn more about the HP Z8 G4 at www.hp.com/go/z8.

Q: Why are GPUs helpful for Machine Learning training?
A: GPUs have the ability to do some classes of massively-parallel mathematical operations and offloading the CPUs as a result. This can significantly speed calculations that are needed for training neural networks.

Q: Why are the NVIDIA® Quadro® GP100 and GV100 part of HP's Machine Learning Workstation configurations?
A: Both the NVIDIA® Quadro® GP100 and GV100 are specifically optimized for half-precision floating point math which is very helpful for speeding up the less-precise calculations needed for Deep Learning training. In addition, the NVIDIA® Quadro® GV100 accelerates 'tensor' operations, which are specific combinations of matrix math operations commonly needed in some Deep Learning training computations.

Q: What other components are needed in a DNN training workstation configuration?
A: The CPU, storage, RAM, and OS all play a vital part in constructing the correct DNN system for the training that is being developed. The brochure HP Z Workstations for Machine Learning suggests typical component configurations for several development scenarios. Different workloads will require different configurations. It is important to understand exactly what training is being developed in order to configure the optimum configuration to support fast learning.

Q: What are examples of HP workstation configurations for deep learning training?
A: The brochure HP Z Workstations for Machine Learning suggests typical workstation configurations for both development and deployment scenarios. As always, each Machine Learning scenario is unique and should be scoped for specific workloads.

Q: What resources does HP offer to support Machine Learning development?
A: HP has created The HP ML Developers Portal to help with Machine Learning model development. The Portal provides the tools and documentation needed to install a high-performance, validated Machine Learning environment on an HP Z Workstation. Visit https://hp.io/ML for more information and frequent updates.

FAQ

CONTENTS & NAVIGATION

MACHINE LEARNING ON THE EDGE VS. IN THE CLOUD

Q: What is Machine Learning on the edge vs. in the cloud?
A: Machine Learning on the edge refers to using a site-based computer for development, storing, and accessing training. When these functions are moved outside the local development center they are often stored in a cloud environment. Some of all of the DNN training can be done locally (on the edge) or in the cloud.

Q: Why does HP qualify and offer the NVIDIA® GPU Cloud (NVIDIA® NGC)?
A: Deep Learning training frameworks are complex and continue to evolve. The installation of frameworks can also be difficult due to the multiple dependencies involved. NVIDIA® NGC offers pre-configured instances of training frameworks that have been optimized for NVIDIA® GPUs including those offered in HP Z Workstations. These pre-configured frameworks enable a quick-start for those doing deep learning development accelerated by NVIDIA® GPUs.

Q: What types of systems do I need to do inferencing?
A: In some cases, an optimized inferencing solution (engine) can be deployed on low-end or customized computer platforms with smaller number of CPU cores[2], small amount of memory, and other size and/or power limitations. In other cases, more CPU cores, and even GPU cores, are needed to deploy an inferencing engine that makes decisions (operates) in real time.

Q: Why use a workstation for Deep Learning training instead of the cloud?
A: There are several advantages, depending on your specific needs, constraints, and situations to doing Machine Learning on the edge as opposed to the cloud.

- Data issues such as security, regulation, privacy, and legal requirements that must be ensured can be a factor
- Cost and price constraints need to be considered as the cost for cloud computation can be unpredictable and can get expensive fast
- As data size gets large response time, performance, and latency issues come into play in the cloud

By going to the cloud, you are dependent on the cloud vendor availability and security. You, as a developer, lose a certain amount of control and can be constrained by the cloud vendor specification and offering.

There may be development reasons that on-premises solutions and services are required that you cannot export to the cloud. In this instance, developers must deal with the challenges of integrating on-premises solutions and services with those on the cloud.

Of course, the question of Machine Learning on the edge or in the cloud need not be an either-or question as you are still able to leverage the advantages of the cloud services that are suitable for your needs on your workstation while also taking advantage of Machine Learning on the edge.

Q: Do DNNs need to keep learning?
A: Maybe. This depends on the purpose of the trained network, where and how it is deployed, and if it is intended to be a helper or a critical decision-maker.

It is often the case that a model's performance will decrease with time. This phenomenon is known as drift. So your model will need to be maintained in order to maintain the required level of performance.

Q: Once a neural network (DNN) is deployed and is using inference to get answers, can DNNs continue to learn?
A: Yes. New labelled data would need to be added and the DNN would need to be retrained. Sometimes this data can be synthesized so that new data can be generated very quickly and used for training.

Sign up for updates
hp.com/go/getupdated

Share with colleagues

LET US HELP YOU CREATE AMAZING BUSINESS SOLUTIONS TODAY

LEARN MORE

Sources and legal disclaimers

[1] Not all features are available in all editions or versions of Windows. Systems may require upgraded and/or separately purchased hardware, drivers, software or BIOS update to take full advantage of Windows functionality. Windows 10 is automatically updated, which is always enabled. ISP fees may apply and additional requirements may apply over time for updates. See http://www.windows.com

[2] Multicore is designed to improve performance of certain software products. Not all customers or software applications will necessarily benefit from use of this technology. Performance and clock frequency will vary depending on application workload and your hardware and software configurations. Intel®'s numbering, branding and/or naming is not a measurement of higher performance.

© Copyright 2018 HP Development Company, L.P. The information contained herein is subject to change without notice. The only warranties for HP products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. HP shall not be liable for technical or editorial errors or omissions contained herein.

NVIDIA is a trademark and/or registered trademark of NVIDIA Corporation in the U.S. and/or other counties.

4AA7-2444ENW, March 2018