Deep Learning in the Cloud and at the Edge

Course Description:

This hands-on course introduces data scientists to technologies related to building and operating live, high-throughput deep learning applications running on powerful servers in the cloud as well as on smaller and lower-power devices at the edge of the network.

To participate, each student will need to order an Nvidia Jetson TX2—a system on a chip (SoC) low-power edge GPU-equipped developer kit (one of the most powerful edge AI platforms out there) along with a few additional peripherals. Each student will also receive a free account in the IBM SoftLayer Cloud with a $2,000/month allowance.

As students progress through the class, they learn to apply the concepts covered by completing hands-on homework and labs on the Jetson TX2 as well as in the cloud, culminating in the final project.

The class material is a set of practical approaches, code recipes, and lessons learned. It is based on the latest developments in the industry and industry use cases as opposed to pure theory. It is taught by professionals with decades of industry experience.

Although this class is an advanced elective and we generally assume familiarity with the basics of machine learning, operating systems, big data, and infrastructure, we fill in the blanks when it comes to some practical tooling as well as the latest developments in deep learning.

Every asynchronous segment is followed by a hands-on assignment as well as a synchronous hands-on lab, where students get to explore the concepts and technologies covered in the lecture. By the time you complete the course, you should be able to name the deep learning problem you are facing, select proper tooling, and know enough to put together an end-to-end application around it.

We begin the class with some real-life examples of how deep learning is used and an overview of the state-of-the-art advances across industries, along with a review of prominent datasets. We will review the Nvidia Jetson TX2 edge platform, unbox it, and get it to work. There's no deep learning without cloud today, so we dive into the cloud stack for deep learning next. We also learn the basics of the environment that will host our labs, homework assignments, and the final project. The Internet of Things (IoT) is the largest source of continuous data for deep learning, so we discuss it next. The fourth lecture covers the fundamentals of deep learning such as artificial neurons, activation functions, loss

functions, back propagation, regularization, and dropout. Next on the agenda is an in-depth overview of leading deep learning frameworks that provide runtimes for DL applications. We cover hardware for deep learning next, which includes GPUs, CPUs, and TPUs. A second lecture on deep learning fundamentals follows, where we review convolutional and recurrent neural networks. We move on to the tooling required to prepare datasets for deep learning, such as classification of images, placement of bounding boxes, semantic segmentation, and data augmentation. We next cover distributed deep learning, starting with the basics of high-performance computing and message-passing interface and learn how to train models on clusters of hardware. Real-time streams and deep learning on CPUs is next, where we learn to apply deep learning to heavy binary streams such as video and audio. A lecture on robotics and deep reinforcement learning follows. We discuss distributed storage and data transfer aspects next. We conclude the class by reviewing several industry-specific deep learning use cases and a update on commercial deep learning cloud APIs and platforms that cloud providers have available for academic and commercial purposes. **(3 units)**

Course Format:

The course is organized as an online inverted classroom. During each week, students first work through a set of asynchronous materials, including video lectures, readings, and other activities. Once a week, students meet for a 90-minute live session, in which they connect with an instructor and other students in an online classroom. A functioning webcam and an audio headset are required to participate in the live sessions. Students must complete all assigned asynchronous material before the scheduled live session begins.

Course Objectives:

After completing this course, students will

- Gain a solid understanding of key deep learning fundamental theoretical concepts
- Achieve hands-on experience with key deep learning tooling and frameworks
- Learn about the classic deep learning datasets, such as KITTI, ImageNet, Microsoft Coco, and LibriSpeech
- Be able to code in at least one DL framework, such as Keras
- Develop a understanding of designing and implementing end-to-end deep learning applications on the edge and in the cloud
- Develop fluency with the Nvidia Jetson edge platform
- Understand some of the deep learning industry use cases
- Understand the infrastructure required to power deep learning applications, such as on-premise HPC, Big Data, and cloud environments
- Apply deep learning to large-scale Big Data problems

Prerequisites:

Before enrolling in this course, students must have completed

- W201: Research Design and Application for Data and Analysis
- W203: Statistics for Data Science
- W205: Fundamentals of Data Engineering

Students should be able to program in Python, Java, and/or be able to pick up a new programming language on the fly. A degree of fluency is expected with the basics of operating systems (e.g., Linux) and Internet technologies.

Course Evaluation:

- Homework (40%).
- Participation and group assignments (20%).
- Final project: performing an analysis on a large dataset (40%). The students will be required to organize into groups of 4 or 5 and prepare a final presentation (slides) and live demo or video (10 minutes).

List of Topics by Week:

Week 1: Introduction and Overview—Cloud computing. Big Data. Artificial intelligence. Deep learning frameworks and hardware. Datasets. Edge computing. Course project overview and sample.

Reading:

- Jetson introduction
    - https://developer.nvidia.com/embedded/twodaystoademo
- Introduction to SSH
    - https://www.digitalocean.com/community/tutorials/ssh-essentials-working-with-ssh-servers-clients-and-keys

Week 2: Clouds, Infrastructure, DL / ML Cloud services (AutoML / DLaaS)
—Defining the cloud. How clouds are used. Hypervisors in a nutshell. Types of clouds. Cloud services. Cloud storage. AI as a service. Deep learning as a service.

Reading:

- What is cloud computing?: https://www.ibm.com/cloud/learn/what-is-cloud-computing
- Linux as a hypervisor: http://www.linuxplanet.com/linuxplanet/reports/6503/1

- Types of clouds: http://it.toolbox.com/blogs/storage-and-io/cloud-virtual-and-storage-networking-conversations-part-iv-49637
- Cloud service types: http://blog.appcore.com/blog/bid/168247/3-Types-of-Cloud-Service-Models
- SoftLayer API: http://sldn.softlayer.com/article/SoftLayer-API-Overview
- Introduction to cloud technologies:
    - https://kubernetes.io/docs/tutorials/kubernetes-basics/
    - https://docs.saltstack.com/en/latest/topics/index.html
    - https://docs.openstack.org/security-guide/introduction/introduction-to-openstack.html
- Getting started with SoftLayer:
    - http://knowledgelayer.softlayer.com/gettingstarted/meet-softlayer
    - http://knowledgelayer.softlayer.com/gettingstarted/how-to
    - http://knowledgelayer.softlayer.com/gettingstarted/how-to/set-up-your-account

Week 3: Internet of Things and Edge Computing—How has it changed in the past few years? What is the edge, and how it is different from the cloud? Data of the Internet of Things (that usually does not make it to the cloud). Edge devices, gateways, and sensors. The Nvidia Jetson family. The Raspberry Pi. Smart speakers such as Google Home and Amazon Alexa. IBM Blue Horizon and AWS Greengrass.

Reading:

- Fog computing and edge computing: https://www.cisco.com/c/en/us/solutions/enterprise-networks/edge-computing.html
- Amazon IoT portal: https://aws.amazon.com/iot/
- AWS Greengrass: https://aws.amazon.com/greengrass/
- Google Android Things: https://developer.android.com/things/
- Raspberry Pi Foundation: https://www.raspberrypi.org/
- Amazon Lambda: https://aws.amazon.com/lambda/

Week 4: Deep Learning 101—The definition of deep learning. How is it different from AI and ML? Artificial neurons, neural layers. Feedforward networks. Multilayer perceptron. Back propagation. Normalization. Regularization. Convolutional layers. Attention. Introduction to recurrent neural networks (RNNs).

Reading:

- ImageNet: http://image-net.org/about-overview
- Microsoft Coco: http://cocodataset.org/#home
- Pascal VOC: http://host.robots.ox.ac.uk/pascal/VOC/index.html
- Nvidia DIGITS: https://developer.nvidia.com/digits
- Mozilla Deep Speech: https://github.com/mozilla/DeepSpeech

ConvNetJs: https://cs.stanford.edu/people/karpathy/convnetjs/

Week 5: Deep Learning Frameworks—What is a deep learning framework? Why are there so many different ones? Which one is the best? What is a tensor? We will cover these topics as well as look at recent developments and strengths/weaknesses of the various frameworks.

Reading:

- Caffe/Caffe2: https://caffe2.ai/docs/caffe-migration.html
- TensorFlow: https://www.tensorflow.org
- PyTorch: https://pytorch.org
- MXNet: https://mxnet.incubator.apache.org/faq/why_mxnet.html
- Darknet YOLO: https://www.pjreddie.com/darknet/yolo
- Compare all the frameworks: https://en.wikipedia.org/wiki/Comparison_of_deep_learning_software

Week 6: Hardware for Deep Learning—GPUs, created to accelerate computer graphics, have revolutionized deep learning. This section looks at why deep learning at scale requires special hardware, like GPUs. We will cover different chip architectures commonly used to increase performance of different frameworks.

Reading:

- GPU: https://developer.nvidia.com/deep-learning
- FPGA: https://www.xilinx.com/products/silicon-devices/fpga/what-is-an-fpga.html
- SoC: https://www.techradar.com/news/computing/pc/system-on-a-chip-what-you-need-to-know-about-socs-1147235
- TPU: https://cloud.google.com/blog/products/gcp/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu

Week 7: Deep Learning 201—Weight initialization. Optimizers. Adjusting the learning rate. Loss functions. Transfer learning. Autoencoders. Embeddings. Inference vs. training vs. model design. Recurrent neural networks.

Reading:

- Attention is all you need: https://hub.packtpub.com/paper-in-two-minutes-attention-is-all-you-need/
- YOLO: https://pjreddie.com/darknet/yolo/
- The unreasonable effectiveness of RNNs: http://karpathy.github.io/2015/05/21/rnn-effectiveness/
- Black box adversarial queries with limited queries and information: https://arxiv.org/abs/1804.08598
- Google Tacotron: https://google.github.io/tacotron/publications/tacotron2/

- Creating a driver's license test for self-driving cars: https://blogs.nvidia.com/blog/2018/10/10/tuv-sud-safety-autonomous-vehicle-standards/

Week 8: Dataset Collection and Preparation—Image classification and the importance of data. Data preparation. Data augmentation. Sample datasets.

Reading:

- ImageNet:
  - http://www.image-net.org/papers/imagenet_cvpr09.pdf
  - https://arxiv.org/pdf/1409.0575.pdf
  - http://www.image-net.org/papers/ImageNet_2010.pdf
- Data augmentation:
  - https://ai.googleblog.com/2018/06/improving-deep-learning-performance.html
  - https://arxiv.org/pdf/1708.06020.pdf
  - https://medium.com/ai-society/gans-from-scratch-1-a-deep-introduction-with-code-in-pytorch-and-tensorflow-cb03cdcdba0f

Week 9: HPC, MPI, and Distributed Training—An overview and a bit of history. HPC vs. HTC/Big Data. Typical HPC problems. Architecture of supercomputers. Interconnect topologies: fat tree, torus. FLOPs. Top500. Scaling: strong, weak, Amdahl's law. Programming for HPC systems. OpenMPI overview. Uber Horovod. Distributed deep learning model training. Applying distributed training: generative adversarial networks (GANs).

Reading:

- What is open MPI? https://www.open-mpi.org/faq/?category=general
- Top 500 supercomputers list: https://www.top500.org/lists/top500/
- IBM Summit: https://en.wikipedia.org/wiki/Summit_(supercomputer)
- Meet Uber Horovod: https://eng.uber.com/horovod/
- Progressive growing of GANs: https://github.com/tkarras/progressive_growing_of_gans
- Nvidia GPU direct: https://developer.nvidia.com/gpudirect

Week 10: Scalable Stream Processing—Concepts on stream processing. Apache Spark as a distributed system. Running deep learning workloads on distributed systems with Caffe, TensorFlow, and other examples.

Reading:

- Deep learning on Databricks: https://databricks.com/blog/2016/12/21/deep-learning-on-databricks.html
- TensorFlow on Spark: https://conferences.oreilly.com/strata/strata-eu-2018/public/schedule/detail/65059

Week 11: Robotics and Deep Reinforcement Learning—Nvidia Isaac platform. Gazebo. OpenAI Gym. UC Berkeley Ray framework. We will briefly touch on the history of robotics, then dive into the world of "smart" robots.

Reading:

- OpenAI Gym: https://gym.openai.com
- Berkeley AI Research Ray: https://bair.berkeley.edu/blog/2018/01/09/ray/
- NVIDIA Isaac: https://www.nvidia.com/en-us/deep-learning-ai/industries/robotics/
- Deep reinforcement learning: https://venturebeat.com/2018/04/05/whats-hot-in-ai-deep-reinforcement-learning/

Week 12: Distributed Storage -> HDFS/GPFS Object Storage and Data Transfer —Distributed computing and deep learning platforms. Overview of Hadoop Distributed File System. Overview of general parallel file system. Overview of OpenStack object storage, Swift.

Reading:

- Cloud storage defined: https://aws.amazon.com/what-is-cloud-storage/
- Apache HDFS: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html
- IBM GPFS: https://www.ibm.com/support/knowledgecenter/STXKQY_5.0.2/com.ibm.spectrum.scale.v5r02.doc/bl1xx_library_prodoc.htm
- OpenStack object storage: https://docs.openstack.org/swift/latest/overview_architecture.html

Week 13: Industry Use Cases—Health care use cases. Predicting diseases. Analyzing motion datasets. What is computational genomics? The sequencing pipeline. Analyzing the assembled genome—the genome and the transcriptome. Use cases related to computational genomics. Use cases for the low-power GPUs: AI Cities and the Nvidia Metropolis platform.

Reading:

- Building your AI strategy in health care: https://aibusiness.com/building-ai-healthcare-strategy/
- Nvidia Project Clara: https://blogs.nvidia.com/blog/2018/03/28/ai-healthcare-gtc/
- Genome analysis toolkit: https://software.broadinstitute.org/gatk/
- Stanford DragoNN: http://kundajelab.github.io/dragonn/
- Nvidia Metropolis: https://www.nvidia.com/en-us/deep-learning-ai/industries/ai-cities/
- DeepMind Health: https://deepmind.com/applied/deepmind-health/

Week 14: Specific DL Services and APIs —IBM Watson data platform, AWS deep learning AMIs, Google Cloud AI, TPUs, and an overview of the NVIDIA Deep Learning Institute platform and resources.

Reading:

- Watson Studio: https://www.ibm.com/cloud/watson-studio

- IBM AI products and services: https://www.ibm.com/watson/products-services/
- AWS Deep Learning: https://docs.aws.amazon.com/dlami/latest/devguide/what-is-dlami.html
- Google Cloud platform: https://cloud.google.com/docs/
- NVIDIA Deep Learning Institute: https://www.nvidia.com/en-us/deep-learning-ai/education/