Fundamentals of Accelerated Computing with CUDA C/C++

레벨	입문
기간	총 6시간(1일)
비용	400,000원

DLI Workshop

CUDA® 기반의 초병렬 GPU에서 실행되도록 C/C++ 애플리케이션을 가속화하는 기본 도구와 기법을 설명합니다. 코드를 작성하고, CUDA를 사용하여 코드 병렬화를 구성하고, CPU와 GPU 가속기 간의 메모리 마이그레이션을 최적화하고, 새로운 작업에서 학습한 워크플로우를 구현함으로써 입자 시뮬레이터(완전하게 기능하지만 오직 CPU 전용)를 가속화하고 확연히 보이는 막대한 성능 향상을 달성하는 방법을 설명합니다. 워크숍을 마치면 새로운 GPU 가속 애플리케이션을 개발할 수 있도록 추가 리소스에 액세스할 수 있습니다.

교육대상	· CUDA의 사용법을 배우고자 하는 사람 · 실무에서 CUDA를 사용해야 하는 사람 · CUDA 가속 컴퓨팅의 기초 이론을 알고자 하는 사람
교육효과	· GPU 가속기로 실행할 코드 작성 · C/C++ 애플리케이션에서 데이터 및 명령어 수준 병렬 처리 제공 및 표현 · CUDA 관리 메모리를 활용하고 비동기 프리패치를 사용하여 메모리 마이그레이션 최적화 · 명령 줄과 시각적 프로파일러를 활용해 작업 안내 · 명령어 수준 병렬 처리에 동시 스트림 활용 · GPU 가속 CUDA C/C++ 애플리케이션을 작성하거나 프로필 중심 접근 방법을 사용하여 기존의 CPU 전용 애플리케이션 리팩토링
실습환경	· nvprof, nvpp

커리큘럼

구분	목차	교육내용
1일차 (2시간)	CUDA C/C++를 사용한 애플리케이션 가속화 (120')	- GPU 코드 작성, 컴파일 및 실행하기 - 병렬 스레드 계층 제어하기 - GPU용 메모리 할당 및 해제하기
1일차 (2시간)	CUDA C/C++를 사용하여 가속화 애플리케이션 메모리 관리(120')	- 명령 줄 프로파일러를 사용하여 CUDA 코드 프로파일링하기 - 통합 메모리를 심층적으로 이해하기 - 통합 메모리 관리를 최적화하기
1일차 (2시간)	CUDA C/C++를 통한 가속화 애플리케이션의 비동기 스트리밍 및 시각적 프로파일링(120')	- NVIDIA Nsight Systems를 사용한 CUDA 코드 프로파일링 - 동시 CUDA 스트림 사용하기

Fundamentals of Accelerated Computing with CUDA Python

레벨	입문
기간	총 6시간(1일)
비용	400,000원

DLI Workshop

CUDA® GPU와 Numba 컴파일러를 사용하여 GPU 가속 Python 애플리케이션을 실행하기 위한 기본 도구 및 기술을 배울 수 있습니다. CUDA® 및 Numba 컴파일러 GPU. 수십 번의 실습 코딩 연습을 수행하고 교육이 끝나면 원래 CPU용으로 설계된 완전한 기능의 선형 대수 프로그램을 가속화하기 위해 새로운 워크플로우를 구현합니다. 인상적인 성능 향상. 워크숍이 끝나면 새로운 GPU를 만드는 데 도움이 되는 추가 리소스를 얻게 됩니다. 스스로 애플리케이션을 가속화할 수 있습니다.

교육대상	· CUDA의 사용법을 배우고자 하는 사람 · 실무에서 CUDA를 사용해야 하는 사람 · CUDA 가속 컴퓨팅의 기초 이론을 알고자 하는 사람
교육효과	· 몇 줄의 코드만으로 GPU 가속 NumPy ufuncs 사용 · CUDA 스레드 계층 구조를 사용하여 코드 병렬화 구성 · 최대의 성능과 유연성을 위해 사용자 지정 CUDA 디바이스 커널 작성 · 메모리 결합 및 디바이스 공유 메모리를 사용하여 CUDA 커널 대역폭 증대
실습환경	· Numba, NumPy

커리큘럼

구분	목차	교육내용
1일차 (2시간)	Numba를 사용하는 CUDA Python 개론(120')	- Python에서 Numba 컴파일러 및 CUDA 프로그래밍 작업 시작하기 - Numba 데코레이터를 사용하여 숫자 Python 함수의 GPU 가속화 - 호스트-디바이스 및 디바이스-호스트로의 메모리 전송 최적화하기
1일차 (2시간)	Custom CUDA Kernels in Python with Numba(120')	- CUDA의 병렬 스레드 계층과 병렬 프로그램 가능성을 확장하는 방법 알아보기 - GPU에서 대규모 병렬 맞춤형 CUDA 커널 실행하기 - CUDA 원자적 연산을 활용하여 병렬 처리 실행 시 경쟁 상태 해결하기
1일차 (2시간)	다차원 그리드 및 Numba를 사용한 CUDA Python의 공유 메모리(120')	- xoroshiro128+ RNG(난수 생성)을 사용하여 GPU 가속 Monte Carlo 방법 적용하기 - 다차원 그리드 생성 및 2D 행렬에서 병렬로 작업하는 방법 알아보기 - 디바이스의 공유 메모리를 활용하여 메모리 병합을 촉진하면서 2D 행렬 재구축하기

Accelerating CUDA® C++ Applications With Multiple GPUs

레벨	중급
기간	총 6시간(1일)
비용	400,000원

DLI Workshop

단일 노드에서 사용 가능한 모든 GPU를 효율적이고 올바르게 활용하는 CUDA C++ 애플리케이션을 작성하여 애플리케이션의 성능을 크게 개선하고 멀티 GPU를 통해 시스템을 가장 비용 효율적으로 활용하는 방법을 설명합니다.

교육대상	· Multi GPU 사용법을 배우고자 하는 사람 · 실무에서 Multi GPU를 사용해야 하는 사람 · Multi GPU 이론을 알고자 하는 사람
교육효과	· 동시 CUDA 스트림을 사용하여 메모리 전송을 GPU 컴퓨팅과 중복 사용 · 단일 노드에서 사용 가능한 모든 GPU를 활용하여 사용 가능한 모든 GPU로 워크로드 확장 · 복사 및 컴퓨팅 오버랩 기능을 멀티 GPU와 결합 · NVIDIA Nsight™ Systems Visual Profiler 타임라인을 중심으로 워크샵에서 다루는 기술의 개선 기회 및 영향 관찰
실습환경	· CUDA C++, Nsight Systems

커리큘럼

구분	목차	교육내용
1일차 (2시간)	JupyterLab 사용(15') 애플리케이션 개요(15') CUDA 스트림 소개(90')	- GPU 가속 인터랙티브 JupyterLab 환경에 익숙해지기 - 워크숍 과정의 시작점인 단일 GPU CUDA C++ 애플리케이션을 직접 경험하기 - Nsight Systems를 사용하는 단일 GPU CUDA C++ 애플리케이션의 현재 성능 알아보기 - 동시 CUDA 스트림 동작을 관리하는 규칙을 학습하기 - 여러 CUDA 스트림을 사용하여 동시 호스트-디바이스 및 디바이스-호스트로 메모리 전송하기 - GPU 커널을 실행하기 위해 멀티 CUDA 스트림 활용하기 - Nsight Systems Visual Profiler 타임라인 뷰의 멀티 스트림 알아보기
1일차 (2.5시간)	CUDA 스트림을 통한 복사/컴퓨팅 오버랩(90') 멀티 GPU와 CUDA C++(60')	- CUDA C++를 포함하는 단일 노드에서 멀티 GPU를 효과적으로 사용하기 위한 핵심 개념 학습하기 - 애플리케이션에서 멀티 GPU를 유연하게 사용하기 위한 강력한 인덱싱 전략 살펴보기 - 단일 GPU CUDA C++ 애플리케이션을 리팩토링하여 멀티 GPU 활용하기 - Nsight Systems Visual Profiler 타임라인에서 멀티 GPU의 활용법 확인하기
1일차 (1.5시간)	멀티 GPU를 통한 복사/컴퓨팅 오버랩(60') 과정 평가(30')	- 멀티 GPU에서 효과적으로 복사/컴퓨팅 오버랩을 수행하기 위한 핵심 개념 학습하기 - 멀티 GPU에서 복사/컴퓨팅 오버랩을 유연하게 사용하기 위한 강력한 인덱싱 전략 살펴보기 - 단일 GPU CUDA C++ 애플리케이션을 리팩토링하여 멀티 GPU에서 복사/컴퓨팅 오버랩 수행하기 - 멀티 GPU에서 복사/컴퓨팅 오버랩의 성능적 이점 알아보기 - Nsight Systems Visual Profiler 타임라인에서 멀티 GPU의 복사/컴퓨팅 오버랩 확인하기

Fundamentals of Deep Learning

레벨	입문
기간	총 6시간(1일)
비용	400,000원

딥 러닝이 컴퓨터 비전 및 자연어 처리 분야의 실습 예제를 어떻게 해결하는지 설명합니다. 고도로 정확한 결과를 얻기 위해 처음부터 도구와 요령을 익히며 딥 러닝 모델을 트레이닝해보세요. 또한 무료로 이용 가능한 사전 훈련된 첨단 모델을 활용하여 시간을 절약하고 딥 러닝 애플리케이션을 빠르게 시작해 실행하는 방법을 설명합니다.

교육대상	· 딥 러닝의 개념을 이해하려는 사람 · 실무에서 딥 러닝을 사용하는 사람 · 딥 러닝의 기초 이론을 알고자 하는 사람
교육효과	· 딥 러닝 모델 트레이닝에 필요한 기본 기법과 도구 학습 · 일반적인 딥 러닝 데이터 유형 및 모델 아키텍처에 대한 경험 쌓기 · 모델 정확도 향상을 위해 데이터 확장를 통해 데이터세트 강화 · 모델 간 전이 학습을 활용하여 더 적은 데이터와 컴퓨팅 성능으로 효율적인 결과 달성 · 첨단 딥 러닝 프레임워크로 자체 프로젝트를 수행할 수 있는 자신감 획득
실습환경	· Tensorflow, Keras, pandas, NumPy

커리큘럼

구분	목차	교육내용
1일차 (2시간)	딥 러닝의 역학(120')	- 첫 번째 컴퓨터 비전 모델을 트레이닝하며 트레이닝 프로세스 학습하기 - 비전 애플리케이션에서 예측 정확도를 개선하는 컨볼루션 신경망 살펴보기 - 데이터 증강 기법을 적용하여 데이터세트를 향상하고 모델 일반화를 개선하기
1일차 (2시간)	사전 트레이닝된 모델 및 반복 네트워크(120')	- 사전 트레이닝된 영상 분류 모델을 통합하여 자동문 개구멍 만들기 - 전이 학습을 활용하여 애완견만 들어올 수 있는 맞춤형 개구멍 제작하기 - 모델을 트레이닝하여 New York Times 헤드라인에 기반한 텍스트 자동 완성하기
1일차 (2시간)	최종 프로젝트: 개체 분류(120')	- 색상 이미지를 해석하는 모델을 제작하고 트레이닝하기 - 데이터 생성자를 구축하여 소규모 데이터 세트를 최대한 활용하기 - 전이 학습 및 특징 추출을 결합하여 트레이닝 속도 개선하기 - 고급 뉴럴 네트워크 아키텍처와 학생들이 기술을 더욱 향상시킬 수 있는 최근 연구 분야 논의하기

Building AI-Based Cybersecurity Pipelines

레벨	중급
기간	총 6시간(1일)
비용	400,000원

DLI Workshop

NVIDIA Morpheus AI 프레임워크를 통해 사이버 보안 개발자와 실무자는 GPU 컴퓨팅의 성능을 활용하여 이전에는 불가능했던 규모로 성능을 발휘하는 사이버 보안 솔루션을 구현할 수 있습니다. Morpheus를 사용하면 사이버 보안 개발자는 대량의 실시간 데이터를 필터링, 처리 및 분류하기 위해 최적화된 AI 파이프라인을 만들 수 있습니다. 데이터 센터에 새로운 수준의 정보 보안을 제공하는 Morpheus는 사이버 보안 위협을 탐지하고 해결하기 위한 동적 보호, 실시간 원격 측정 및 적응형 방어를 지원합니다.

교육대상	· 사이버 보안 관련 개발자 · 사이버 보안 관련 실무자 · NVIDIA Morpheus AI 프레임워크를 알고자 하는 사람
교육효과	· 사이버 보안 사용 사례를 위한 방대한 양의 데이터를 실시간으로 처리하고 AI 기반 추론을 수행할 수 있는 Morpheus 파이프라인 구축 · 민감한 정보 탐지, 이상 행동 프로파일링, 디지털 핑거프린팅과 같은 작업에 다양한 데이터 입력 유형과 함께 여러 AI 모델을 활용 · Morpheus SDK 및 CLI(명령줄 인터페이스), NVIDIA Triton™ Inference Server 등 Morpheus AI 프레임워크의 주요 구성 요소 활용
실습환경	· NVIDIA Morpheus, NVIDIA Triton Inference Server, RAPIDS ™ , CLX

커리큘럼

구분	목차	교육내용
1일차 (2시간)	NVIDIA Morpheus AI 프레임워크 개요(30') Morpheus Pipeline의 건설(45') Morpheus Pipeline의 추론(45')	- AI 기반 사이버 보안의 필요성을 이해하기 - Morpheus 프레임워크의 구성요소에 대해 알아보기 - Morpheus SDK 및 CLI에 대한 개요를 살펴보기 - 파이프라인 유형 및 명령에 대해 알아보기 - 데이터 입출력(IO) 및 처리에 대해 알아보기 - NVIDIA Triton 추론 서버에 대한 개요를 알아보기 - 모델 배포 방법을 이해하기 - 민감한 정보 탐지 파이프라인을 살펴보기
1일차 (2시간)	Splunk에서 AI 기반 머신 로그 구문 분석(30') 디지털 지문 채취 파이프라인(45') 시계열 분석(45')	- Morpheus 오토인코더 파이프라인을 사용하기 - 손상된 자격 증명을 찾아보기 - Morpheus 파이프라인 내에서 시계열 분석을 적용하기 - 시계열 분석과 디지털 지문 채취를 결합하기
1일차 (2시간)	Booz Allen Hamilton의 사이버 보안 플라이어웨이 키트(30') 이해도 테스트(45') 실제 시연(45')	- 사이버 보안 침해를 식별하기 위해 엔드투엔드 Morpheus 파이프라인을 구축하기

Model Parallelism: Building and Deploying Large Neural Networks

레벨	입문
기간	총 6시간(1일)
비용	400,000원

DLI Workshop

자연어 처리, 컴퓨터 비전 또는 음성 AI에 적용되는지 여부에 관계없이 LLM(대형 언어 모델) 및 DNN(심층 신경망)은 특정 특성을 갖습니다. LLM과 DNN이 점점 더 커지고 점점 더 큰 데이터 세트에 대해 훈련됨에 따라 소수의 훈련 예제만으로 새로운 작업에 적응할 수 있어 일반 인공 지능을 향한 경로가 가속화됩니다. 방대한 데이터세트에 수백억에서 수천억 개의 매개변수가 포함된 학습 모델은 결코 쉽지 않으며 AI, 고성능 컴퓨팅(HPC) 및 시스템 지식의 고유한 조합이 필요합니다. 이 과정의 목표는 가장 큰 신경망을 훈련하고 이를 프로덕션에 배포하는 방법을 보여주는 것입니다.

교육대상	· 대규모 모델 을 활용하여 문제를 해결하고자 하는 사람 · NVIDIA NeMo™ 프레임워크에 대해 알고자 하는 사람 · 대형 신경망에 대해 알고자 하는 사람
교육효과	· 여러 서버에 걸쳐 신경망 훈련 · 활성화 체크포인트, 경사 누적, 다양한 형태의 모델 병렬 처리 등의 기술을 사용하여 대규모 모델 메모리 공간과 관련된 문제를 극복 · 훈련 성능 특성을 포착하고 이해하여 모델 아키텍처 최적화 · NVIDIA® TensorRT™-LLM을 사용하여 대규모 다중 GPU 모델을 프로덕션에 배포
실습환경	· PyTorch, NVIDIA NeMo™ Framework, DeepSpeed, Slurm, TensorRT-LLM

커리큘럼

구분	목차	교육내용
1일차 (2시간)	대형 모델 훈련 소개(120')	- 대형 모델 학습의 동기와 주요 과제에 대해 알아봅니다. - 대규모 교육에 필요한 기본 기술과 도구에 대한 개요를 알아봅니다. - 분산 교육 및 Slurm 작업 스케줄러에 대해 소개합니다. - 데이터 병렬 처리를 사용하여 GPT 모델을 학습합니다. - 교육 프로세스를 프로파일링하고 실행 성능을 이해합니다.
1일차 (2시간)	모델 병렬성: 고급 주제(120')	- 다양한 메모리 절약 기술을 사용하여 모델 크기를 늘립니다. - Tensor 및 파이프라인 병렬 처리에 대해 소개합니다. - 자연어 처리를 넘어 DeepSpeed를 소개합니다. - 모델 성능을 자동 조정합니다. - 전문가 혼합 모델에 대해 알아봅니다.
1일차 (2시간)	대형 모델 추론(120')	- 대규모 모델과 관련된 배포 문제를 이해합니다. - 모델 축소 기술을 살펴보세요. - TensorRT-LLM 사용 방법을 알아봅니다. - Triton 추론 서버를 사용하는 방법을 알아봅니다. - GPT 체크포인트를 프로덕션에 배포하는 프로세스를 이해합니다. - 프롬프트 엔지니어링의 예를 확인합니다.

Data Parallelism: How to Train Deep Learning Models on Multiple GPUs

레벨	고급
기간	총 6시간(1일)
비용	400,000원

DLI Workshop

데이터 집약적인 애플리케이션에 필요한 훈련 시간을 단축하기 위해 여러 GPU에서 데이터 병렬 딥 러닝 훈련을 위한 기술을 가르칩니다. 딥 러닝 도구, 프레임워크 및 워크플로를 사용하여 신경망 훈련을 수행하면서 단일 GPU에서 훈련의 정확성을 유지하면서 데이터를 여러 GPU에 분산시켜 모델 훈련 시간을 줄이는 방법을 배우게 됩니다.

교육대상	· Multi GPU 사용법을 배우고자 하는 사람 · 실무에서 Multi GPU를 사용해야 하는 사람 · Multi GPU 이론을 알고자 하는 사람
교육효과	· Multiple GPU를 사용하여 데이터 병렬 딥 러닝 훈련을 수행하는 방법 이해 · Multiple GPU를 최대한 활용하기 위해 훈련 시 최대 처리량 달성 · Pytorch 분산 데이터 병렬을 사용하여 여러 GPU에 훈련 분산 · Multi GPU 훈련 성능 및 정확도와 관련된 알고리즘 고려 사항을 이해와 활용
실습환경	· PyTorch, PyTorch Distributed Data Parallel, NCCL

커리큘럼

구분	목차	교육내용
1일차 (2시간)	확률적 경사하강법 및 배치 크기의 효과(120')	- 순차적 단일 스레드 데이터 처리 문제와 병렬 처리를 통해 애플리케이션 속도를 높이는 이론을 이해합니다. - 손실 함수, 경사하강법, 확률적 경사하강법(SGD)을 이해합니다. - 다중 GPU 시스템에서의 사용을 고려하여 배치 크기가 정확성과 훈련 시간에 미치는 영향을 이해합니다.
1일차 (2시간)	PyTorch DDP(분산 데이터 병렬)를 사용한 다중 GPU 교육(120')	- DDP가 여러 GPU 간의 훈련을 조정하는 방법을 이해합니다. - DDP를 사용하여 여러 GPU에서 실행되도록 단일 GPU 교육 프로그램을 리팩터링합니다.
1일차 (2시간)	Multiple GPU로 확장할 때 모델 정확도 유지 (90') 이해도 테스트(30')	- 여러 GPU에서 훈련을 병렬화할 때 정확도가 저하될 수 있는 원인을 이해합니다. - 훈련을 여러 GPU로 확장할 때 정확도를 유지하는 기술을 배우고 이해합니다.