


TAVE 서기

서기 내용			
서기 일자	21.11.09	서기	고성호
주제	GTC Noveber 2021 Keynote with NVIDIA CEO Jensen Huang		
시간	22:00~24:00	장소	Zoom 미팅 및 트위터
스터디 인원	고성호, 권기호, 이아현, 서가을 : 시작		
			
	고성호, 권기호, 이아현, 서가을 : 종료		
배운 내용	내용		
	1. The magic of Accelerated Computing 2. Next_Gen Networking, Cloud Computing, and Cybersecurity 3. Accelerating Science by a Million-X		

- 4. Building Virtual Worlds with Omniverse
- 5. Transforming Industries with AI
- 6. Omniverse Avatar for Intelligent Virtual Assistants
- 7. A Robotics Revolution with NVIDIA Orin

The magic of Accelerated Computing

- 가속컴퓨팅 개요
 - 가속컴퓨팅은 NVIDIA CUDA 범용 목적 프로그래밍 GPU를 통해 이뤄지고 CUDA, 알고리즘 가속 라이브러리, 분산 컴퓨팅 시스템 및 소프트웨어의 조합을 통해 완성된다.
 - GPU, 가속 라이브러리, 시스템, 애플리케이션 사이를 지속적으로 반복 처리하여 전체 스택을 최적화하는 동시에 가속화 대상인 새로운 애플리케이션 도메인을 추가하여 플랫폼의 범위를 확장하는 중이다. 이를 통해 동일한 칩을 사용하고 있음에도 소프트웨어적인 지원을 통해 향상된 성능을 느낄 수 있다.
- 신규 가속화 라이브러리
 - 1) ReOpt
 - 배달 차량 경로 설정, 창고에서 물건 수령 및 포장과 같은 운영 연구 최적화 문제를 위한 가속 솔버
 - 2) cuQuantum DGX 어플라이언스
 - 중첩과 얽힘이라는 양자 물리학 현상에 기반한 양자 컴퓨터는 초고속 양자 시뮬레이터를 필요로 함.
 - cuQuantum을 통해 이러한 양자 회로 시뮬레이션을 가속화할 수 있음.
 - 3) cuNumeric
 - numpy 라이브러리의 코드를 변경하지 않고 가속화가 가능한 가속기

Next-Gen Networking, Cloud Computing, and Cybersecurity

- 분산 컴퓨팅에서의 네트워크, Quantum2

- 분산 컴퓨팅에서 네트워크는 수천개의 GPU를 하나의 거대한 슈퍼 컴퓨터로 연결하여 확장성과 궁극적인 성능을 결정하기 때문에 매우 중요하다.
- Quantum 2는 슈퍼 컴퓨터의 성능과 클라우드 컴퓨팅의 공유 기능을 제공하는 최초의 네트워킹 플랫폼이다.
- BlueField DPU
 - CPU의 네트워킹, 스토리지, 보안 소프트웨어 프로세싱 부담을 덜기 위한 새로운 유형의 인프라 컴퓨팅 플랫폼.
 - Blue Field는 네트워킹 엔드포인트이므로, NVIDIA는 거의 모든 접점(애플리케이션, 데이터, 사용자 및 장치의)에서 데이터 센터를 보호할 수 있다.
- Morpheus
 - NVIDIA Rapids와 NVIDIA AI에 기반을 두고 있는 딥 러닝 사이버 보안 플랫폼
 - Morpheus의 워크플로우는 AI 모델과 모든 앱 및 사용자 결합에 따른 디지털 지문을 만들어서 일반적인 패턴을 학습하고 이상 트랜잭션을 찾게된다.

Accelerating Science by a Million-X

- 컴퓨터공학 분야의 도약을 이뤄낸 역학들
 - 1) 가속화된 컴퓨팅
 - 2) 딥러닝
 - 3) 딥러닝으로 작성된 AI 소프트웨어
- 딥러닝을 과학에 적용한 사례
 - 1) PINN
 - Brown 대학 Karniadakis팀이 개발한 물리 학습 인공 신경망
 - 2) FNO
 - Li, Anandkumar, Caltech 팀과 NVIDIA가 개발한 푸리에 신경 연산자

3) PINO

- FNO와 PINN의 장점을 결합하여 개발한 물리 법칙을 따르는 범용 함수 학습 모델
- 물리학 시뮬레이터나 관측된 데이터로부터 학습을 진행
- 학습을 마친 후에 물리학 모델을 고속으로 에뮬레이션하는 것이 가능함.

4) 신약 개발

- 신약 개발을 위해서는 인간 단백질에 대한 해석이 선행되어야 하는데, 기존엔 엑스레이 결정학과 극저온 전자 현미경 검사를 통해 2만5000개의 단백질 중 단 17%밖에 해석하지 못하였음.
- AlphaFold와 RossettaNet의 연구원은 아미노산 순서로부터 단백질의 3D모양을 예측하기 위해 Deepmind를 학습시켰고 하룻밤동안 2만개 이상의 인간 단백질을 해석하였음.
- 이는 물리학 ML방법을 통한 성능 향상 덕분이라고 할 수 있다.

5) 기후 과학

- 몇 십년 후의 지구 기후를 시뮬레이션 하고 기후 변화의 지역적 영향을 예측하고 완화 및 적응을 위한 조치를 취할 수 있는 방법을 찾기 위해서는 매우 큰 규모의 시뮬레이션이 가능해야 한다.
- 지구의 대기, 바다와 하천, 얼음, 땅, 인간 활동 및 요소간의 물리적 상호작용을 모델링해야되기 때문이다.
- 이번 키노트에서는 NVIDIA Modulus라는 물리 ML 모델 개발을 위한 프레임워크를 공개하였다.
- Modulus를 통해서 기존보다 훨씬 빠른 속도로 물리 법칙을 에뮬레이션 하는 것이 가능하다.

Building Virtual Worlds with Omnivers

● Omnivers개요

- 기존의 인터넷을 구성하는 주요 요소들이 2D 정보였다면, 이제는 3D 세계를 만들거나 물리적 세계를 모델링 하는 것이 점점 증가할 것이다.

- NVIDIA는 이러한 세계의 제작자들을 위해 Omnivers를 만들었다. 게임 엔진과는 매우 다르고 데이터센터 규모로 설계되었기 때문에 추후 행성급 규모의 세계를 구성하는 것도 가능해질 것이다.
- Omnivers의 기반이 되는 기술
 - 1) showroom : 그래픽, 물리, 물질, AI등 Omnivers의 기술을 보여주는 데모 및 샘플 앱
 - 2) Farm : 여러 베어메탈 또는 가상화된 시스템, 워크스테이션, 서버에 걸쳐 일괄 작업 처리를 지휘하는 시스템 레이어
 - 3) Omniverse AR : 그래픽을 휴대폰이나 AR 안경으로 스트리밍 함.
 - 4) Omniverse VR : 세계 최초의 풀프레임 인터랙티브 레이 추적 VR
- 가상 공장을 구축하고 가상 로봇을 운영하여 로봇 공장, 창고, 자동차 설계 등을 모니터링 하는 디지털 트윈(물리적인 공장의 디지털 버전)을 생성하여 여러 방면에서 활용이 가능함.
- Omniverse의 활용 사례
 - 1) Bentley
 - Omniverse를 활용한 iTwin이 현재 널리 액세스 단계라고 발표하였음
 - 2) Siemens Energy
 - NVIDIA와 함께 디지털 트윈 플랫폼을 개발하였음.
 - 물리 기반 AI 모델을 트레이닝 하기 위한 포인트 클라우드 데이터를 사용하면 유선으로 표시되는 고충실도 유동 분석을 몇 초 만에 완료할 수 있음.
 - 이를 통해 가동 중단시간을 70% 줄여 연간 17억 달러 절약
 - 3) BMW
 - GTC Spring에서 로겐스버그 공장의 디지털 트윈을 구축하는 방법을 시연한 바 있고 이를 1000만 제곱미터에 달하는 다른 공장 3개로 확대하였음.
 - Omniverse에 구축된 Issac Gym을 사용하여 로봇에게 새로운 기술을 가르치는 중.

4) Ericsson

- 5G 안테나와 라디오 합대를 구성, 작동 및 지속적으로 최적화 하기 위해 도시 전체의 디지털 트윈을 구축하고 있음.

Transforming Industries with AI

- DGL(Deep Graph Library)
 - 기존 딥 러닝 프레임워크 위에 그래프 신경망을 구현하기 위해 제작된 python 라이브러리.
 - 그래프 신경망은 관계를 학습하는 것이 가능함. 예를 들어 분자가 단백질에서 어떻게 서로 연결되는지, 사람들이 소셜 그래프에서 어떻게 연결되는지 등을 그래프를 통해 설명하는 것이 가능함.
 - NVIDIA는 그래프 구성부터 하위 그래프 샘플링, DNN 프레임워크로 그래프 투영에 이르기 까지 개발자가 수십억~수조 개의 엣지로 그래프를 학습시키고 추론할 수 있도록 워크플로우를 가속화시키고 있다.
- GNN
 - 금융 서비스, 신약 개발, 디지털 생물학 및 사이버 보안을 위한 새로운 모델
 - GNN을 통해 PayPal은 공모 사기 탐지를 크게 개선시켰고, Amazon은 Amazon Search를 크게 개선하고 남용과 사기 판매자 및 구매자를 줄였고, Pinterest는 검색 및 추천을 5억 사용자로 확장했다.
- Nemo Megatron
 - NVIDIA가 개발한 수십억 및 수조 개의 매개변수의 언어 모델 학습 전용 프레임워크
 - Nemo Megatron을 사용하면 모든 회사에서 최첨단 대형 언어 모델(LLM)을 교육할 수 있음.

- 대형 언어 모델(LLM)의 활용 사례
 - 스웨덴의 자국 역사 디지털화
 - 삼성의 포르투갈어 사용자를 위한 스마트 스피커 구축
 - VinBrain의 의료용 베트남 LLM
 - JD의 고객 5억명 참여를 유도할 전자 사액래 서비스용 LLM
 - Rakuten의 디지털 서비스를 위한 일본어 LLM
 - ServiceNow의 IT/헬프데스크 챗봇 구축
 - Xiaomi의 AI비서

- AI와 기존 산업 및 서비스
 - 금융업 : 신용카드 구매 사기를 줄일 방법을 모색중
 - 전자상거래 분야 : 제품 및 서비스 추천이 10조 달러 규모 산업으로 성장할 전망
 - 화상회의 : 배경 및 소음 제거, 아이컨택, 언어 번역, 조명 조절 등의 기술들이 개발중

- 추론 도구
 - AI애플리케이션의 응답 시간, 일괄 처리량, 연속 스트리밍 등의 요구사항과 다양한 프레임워크, 머신러닝 플랫폼, 기밀성, 보안, 기능 안전 및 신뢰성 요구 등등의 모든 요구사항의 조합은 막대한 경우의 수를 낳기 때문에 추론이라는 것은 매우 복잡한 기술적 뒷받침이 이뤄진 후에야 가능하다.
 - 1) TensorRT compiler
 - Tensorflow와 Pytorch에 통합되어 있음.
 - 이를 사용하면 ML개발자는 3배의 속도 향상을 얻을 수 있음.
 - 2) NVIDIA Triton
 - 세계 최초의 분산 추론 엔진
 - 여러 GPU와 여러 노드에 걸쳐 분산 처리를 수행할 수 있음.

- Metropolis

- NVIDIA의 영상 처리 분석 플랫폼
- 스트리밍 비디오에서 3D 포즈를 감지, 추적, 계산, 추측하고 향후 전체 3D 장면을 재구성할 수 있음

Omniverse Avatar for Intelligent Virtual Assistants

- Maxine
 - 가상 로봇 플랫폼인 아바타 플랫폼
 - 화상 회의에 사용하거나, 게임 캐릭터를 애니메이션 하거나 로봇에 통합될 수 있음.
 - Maxine은 컴퓨터 비전을 사용하여 사용자의 얼굴을 추적하고 표정을 인식할 수 있음.
- RIVA
 - Maxine의 입출력 담당 음성 AI
 - RIVA는 7개의 언어가 사용 가능하다.
 - RIVA를 통한 Maxine의 음성 출력시엔 Omniverse를 통해 3D 얼굴 애니메이션을 구현해 사용자가 선택하는 버전에 따라 다른 언어 서비스를 제공받을 수 있다.

A Robotic Revolution with NVIDIA Orin

- NVIDIA Clara Holoscan
 - 의료업계의 소프트웨어 정의 이미징 플랫폼
- Issac Sim
 - Omniverse에 내장되어 있는 로봇 시뮬레이터
 - 센서를 모델링하고, 물리학이 시뮬레이션 된다. 시뮬레이션 중인 로봇은 실제 스택을 실행하고 있고 실제 맵에 연결된다.
- Issac Sim Replicator

	<ul style="list-style-type: none"> - 센서를 시뮬레이션하고 자동으로 레이블이 지정된 데이터를 생성하며 도메인 무작위화 엔진을 사용하여 풍부하고 다양한 학습 데이터셋을 만든다. ● Maxine을 통한 자율주행 혁신 <ul style="list-style-type: none"> - 운전기사가 무슨 생각을 하는지를 정확히 보여 주고, 신경 그래픽으로 3D 서라운드 뷰를 재구성해 자율주행에 대한 운전자 신뢰를 높임 - 발렛파킹이 가능해짐 - Orin 로보틱스 칩을 통해 실행 가능함.
과제할 당	없음
특이사항	없음
비고	없음