NVIDIA

최고 수준의 AI

GPU 딥러닝 NVIDIA TensorRT
하이퍼스케일 추론 플랫폼과 함께

인공지능의 폭발적 성장

개인화 서비스에 대한 요구는 복잡도와 양적으로 또한 다양한 인공지능 애플리케이션과 제품에서 급격하게 증가하고 있습니다. 애플리케이션은 AI 추론을 사용해 이미지를 인식하고 음성을 이해하고 추천을 만듭니다. 이를 위해 AI 추론은 빠르고 정확하고 쉽게 배포되어야 합니다.

추론 성능의 이해

추론에서 속도는 성능의 첫 번째에 불과합니다. 추론 성능을 완전하게 이해하려면 프로그램 용이성에서 학습 속도까지 7 가지 요소를 고려해야 합니다

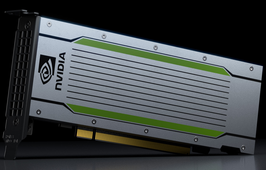
NVIDIA TensorRT 하이퍼스케일 추론 플랫폼은 모든 요소를 제공합니다.
최고의 추론 성능을 규모있게 또한 다양한 기능과 함께 제공합니다. 이를 통해 점점 늘어나는 다양한 최신 신경망들을 다룰 수 있습니다.

NVIDIA TensorRT 하이퍼스케일 추론 플랫폼 내부

강력한 튜링 텐서 코어를 장착한 NVIDIA T4

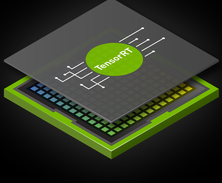
효율적이고 높은 처리량의 추론은 세계 최상급의 플랫폼이 필요합니다. NVIDIA Tesla T4 GPU는 모든 인공지능 추론 작업을 위한 전 세계에서 가장 진보된 가속기입니다. NVIDIA 튜링 텐서 코어를 장착한 T4 는 혁신적인 다중 정밀도 추론 성능을 제공하여 다양한 최신 인공지능 애플리케이션을 빠르게 실행합니다.

다중 정밀도
FP16 | 65 테라플롭스까지
INT8 | 130 톱스까지
TFLOPS = trillion floating-point operations per second
TOPS = trillion operations per second



NVIDIA TENSORRT 의 강력함

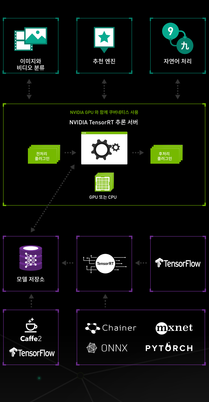
NVIDIA TensorRT™ 는 옵티마이저, 런타임 엔진, 제품 애플리케이션에 배포하기 위한 추론 서버를 포함한 고성능 추론 플랫폼입니다. TensorRT 는 비디오 스트리밍, 추천, 자연어 처리에서 CPU 만 있는 시스템보다 애플리케이션의 속도를 40 배 빠르게 합니다.



- 층과 텐서의 융합
- 가중치와 활성화의 정밀도 조정
- 커널 자동-튜닝
- 동적 텐서 메모리
- 다중 스트림 실행

즉시 제품화가 가능한 데이터 센터 추론

NVIDIA TensorRT 추론 서버는 애플리케이션이 데이터 센터 환경에서 인공지능 모델을 사용할 수 있도록 컨테이너에 담긴 마이크로서비스입니다. GPU 활용을 극대화하고 널리 사용되는 모든 인공지능 프레임워크를 지원하며 쿠버네티스와 도커와 통합되어 있습니다.



애플리케이션 개발자
밑바닥부터 추론 기능을 개발하는 데 시간을 낭비하지 않고 **혁신적인 인공지능 솔루션**을 만드는 것에 집중할 수 있습니다.

데브옵스 엔지니어
추론 서비스를 **여러 애플리케이션**을 손쉽게 배포하고 오**케스트레이션**, **로드 밸런싱**, **오토스케일링**의 장점을 활용할 수 있습니다.

데이터 과학자와 연구자
추론 구현에 대한 걱정없이 널리 사용되는 **모든 인공지능 프레임워크**를 사용하여 모델 설계와 훈련에 집중할 수 있습니다.

최고의 인공지능 플랫폼

www.nvidia.com/data-center-inference

NVIDIA