



AI Infrastructure

AI 인프라



Manish Sainani

Director, Product Management-AI/ML Infrastructure, Google Cloud

#AI Platform #Cloud TPU

Week8. cloud AI

AI Infrastructure

AI 인프라

AI Infrastruture: Why Google Cloud

- ✓ 모든 Use case를 위한 액셀러레이터
비교적 낮은 비용과 고성능 트레이닝까지 모든 use case를 위한 액셀러레이터 제공
- ✓ 비용 및 성능
고성능 GPU 및 Cloud TPU pod으로 빠르게 트레이닝 가능
- ✓ 손쉬운 시작 및 확장
시작하기 쉽고 개발 및 배포를 위한 다양한 서비스 제공

The AI Boom

인터넷 경제의 3원칙인 무어의 법칙은 끝나고 있습니다. 프로세서 트랜지스터는 2년마다 2배 이상 증가하고 있습니다. 그에따라 AI에 대한 수요도 급격하게 증가하고 있습니다. 한 예로 현재 고객들은 AlexNet에서 AlphaGo Zero로 옮길 때 compute 부분이 기존 대비 30만배 증가하고 있으며 ML에 필요한 Petaflops는 3.5 개월마다 두배로 증가하고 있습니다.

구글에서는 현재 딥러닝을 사용하는 Gmail 부터 Translation, Youtube, 자율주행 AlphaGo 등 여러가지 서비스를 제공하고 있으며 이러한 서비스의 경험으로 고객들에게 AI 인프라를 제공해주고 있습니다.

1. 리테일 : 가격 최적화, 추천, 수요예측, 이탈 분석, 비주얼 서치
2. 공공기관 : 스마트 지속 가능한 정부, 스마트 교통, 보안 로그 분석, 재해 복구 및 인도주의 지원, 차량군 관리
3. 금융 : 사기 및 리스크 분석, 금융 범죄 예방, 쿼트 영구, 맞춤형 보험, HFT 분석
4. 미디어 및 엔터테인먼트 : 음성 인식, 자연어 처리, 영상 서치, 번역, 콘텐츠 분류
5. 제조 : 외관 검사, 운영 리포팅, SCM 최적화, 자산 관리, 이상 감지
6. 헬스케어 : 임상 데이터 웨어하우스, 데이터 상호운용성, 의학적 위험 예측, 환자 경험 및 보호, R&D 가속화

구글은 이와같이 여러 전문분야에서 기본 바탕인 인프라 부터 개발환경인 AI Platform, 맞춤형 모델을 만들 수 있는 AutoML, 미리 학습된 모델의 API, 도입을 원활하게 해줄 수 있는 서비스 및 솔루션들을 보유하고 있습니다.

| 액셀러레이터 하드웨어

➤ Cloud TPU

Deep neural networks를 학습하고 실행하기 위한 Google의 맞춤 ASIC입니다. 구글 Cloud에 AI를 위해 만들어 졌으며 성능, 정확성 및 품질에 최적화된 Google 검증 reference 모델들을 제공합니다.

현재 Cloud TPU pods는 GA(General Availability)상태이며 v2, v3 Pod들도 공개되었습니다. 그 외로 PyTorch 1.5 Beta, TF 1.15+/2.2+ GA, GKE, 딥러닝 VM, 등이 출시되고 있습니다.

➤ Cloud GPU

P4 부터 V100까지 N1 VM 제품계열에 해당되는 GPU를 제공해줬으나 이젠 A2 VM계열인 A100을 출시 하였습니다. A100은 더 빠른 인공지능 학습 및 연산처리에 특화된 GPU입니다.

➤ On-Prem과 AI Platform Hybrid

Kubeflow를 사용하고 몇 줄의 코드를 추가하면 Kubeflow Notebook에서 On-Prem 또는 클라우드에서 ML 학습 작업을 실행할 수 있습니다. 학습이 완료된 후 Kubeflow를 사용하여 학습 된 모델을 예측 endpoint로 배포 가능합니다.

➤ Preemptible GPUs

GPU가 포함 된 Preemptible VM을 사용하면 GCE 또는 GKE 용량 풀에서 batch 작업을 실행하고 일반 on-demand 가격의 일부만 지불할 수 있습니다. Preemptible VM을 사용하는 경우 온 디맨드 VM과 비교하여 GPU 에서 70%, vCPU 메모리에서 80%를 절약할 수 있습니다.

| GCP에서 액셀러레이터 활용 방법

1. AI Platform
2. Kubernetes Engine
3. Compute Engine

AI Platform같은 경우 데이터 labeling, 딥러닝 VM 이미지, Notebooks, 내장형 알고리즘, 학습, 예측 서비스들이 포함되어 있습니다. 그리고 BigQuery, Dataflow, Dataprep, Dataproc, Data Studio와 통합이 가능합니다.

| 딥러닝 VM 이미지

Google Cloud Engine에서 ML를 더 쉽고 빠르게 할 수 있습니다.

Fast Prototyping : 딥러닝을 위해 미리 구성된 VM으로 ML 프로젝트를 빠르게 프로토타이핑
CPU, GPU 및 TPU 지원 : 클릭 한 번으로 최신 클라우드 TPU 또는 GPU 를 인스턴스에 추가하고 모델 학습 작업을 가속화

Performance optimized for Google Cloud : 인프라에서 최적의 성능을 얻기 위해 라이브러리 구성

Flexibility : TensorFlow, PyTorch 및 scikit-learn과 같은 다양한 ML 프레임워크 중에서 선택하거나 공통 기본 이미지 위에 직접 설치

베스핀글로벌 인사이트

인공지능은 클라우드 분야에서 가장 많이 논의되고 있는 주제 중 하나입니다. 방대한 양의 데이터를 계산하고 분석할 수 있는 놀라운 잠재력을 갖춘 고급 ML 기술은 복잡한 작업을 더 빠르고 효율적으로 비즈니스에서 사용되고 있습니다.

머신러닝 기반 솔루션은 고객 경험과 ROI를 개선하고 비즈니스에서 경쟁 우위를 확보하기 위해 조직에서 활용되고 있습니다. Google, IBM, Microsoft, Apple 및 Salesforce와 같은 기업들은 이미 ML 혜택을 활용하고 있습니다.

Gartner는 2022년까지 AI 및 ML 기술을 활용하는 새로운 end user 솔루션의 75%가 오픈 소스 플랫폼 대신 구축될 것으로 예측합니다. 그만큼 많은 기업들은 인공지능을 비즈니스에 접목시켜 성장하고 있습니다.

베스핀글로벌은 Google Cloud를 가장 잘 아는 전문가이며, Google Cloud의 프리미어 파트너이자, 국내 최초 Google Cloud의 MSP(Managed Service Provider)입니다. 베스핀글로벌에서는 클라우드 문의나 Google Cloud 관련 무료 컨설팅을 진행하고 있습니다. 아래 문의로 편하게 연락주시기 바랍니다.

문의사항 | 베스핀글로벌 구글사업부 sales.google@bespinglobal.com 070-7931-9600

참고 웹사이트 | <https://cloud.withgoogle.com/next/sf/>