E-BOOK

조금 더 말하고, 조금 더 행동하라

대화형 AI를 위한 데이터 인프라 구축

NetApp

목차

잡담에 싫증이 나셨나요? 더 크게 생각하십시오.

NLP: 자연어 처리라고도 합니다 . 대화형 AI 라고도 합니다 . *말하는 로봇*이라고도 합니다 .

뭐라고 부르든 대화형 AI 시스템은 사람처럼 말하고 , 문맥을 이해하고 , 지능적인 응답을 제공합니다 . 이는 모두 딥 러닝이 엄청나게 발전하여 AI 시스템이 더 자연스럽고 덜 업무적으로 발전했기 때문입니다 .

딥 러닝 덕분에 AI 가 더 사용자 친화적이 되었으며 백 엔드에서 언어학과 규칙 기반 기술에 대한 심층적인 인간적 지식이 필요하지 않게 되었습니다 . 딥 러닝을 통해 매우 복잡한 특정 언어 (예 : 금융 서비스 , 의료 및 생명 과학 , 정부 , 자동차 , 제조 , 철도) 를 사용하는 산업에서 NLP 솔루션을 채택할 수 있는 기회가 열렸습니다 .

데이터는 효율적인 대화를 위한 핵심적인 요소입니다 .

AI 모델은 거대하고 매우 복잡합니다 . 생각의 속도로 움직이는 많은 데이터가 필요합니다 . 효과적인 AI 모델을 구현하려면 NLP 인프라에 다음 기능이 필요합니다 .

1. 대규모 데이터 처리
2. 파이프라인 정리
3. 즉각적인 응답

NLP: 더 이상 챗봇만을 위한 것이 아닙니다 .

스마트 도우미에서 검색 엔진 , 예측 텍스트까지 NLP 는 새로운 글로벌 언어입니다 . 우리 주변 어디에나 있으며 가끔은 생각지도 못한 곳에 있습니다 .

신용도 평가

NLP 를 사용하여 지리적 위치 , 소셜 미디어 활동 , 탐색 동작 , 피어 네트워크 등과 같은 데이터를 기반으로 신용 점수를 생성할 수 있습니다 .

임상 시험 매칭

사람들이 임상 시험이 있다는 사실을 대부분 모르고 있기 때문에 임상 시험에 참석할 환자를 구하는 것이 어려울 수 있습니다 . NLP 를 사용하면 연구원과 제조업체에서 환자를 임상 시험과 자동으로 연결할 수 있습니다 .

사법당국

경찰서에서는 NLP 를 통해 범죄의 동기를 파악하여 국민을 안전하게 보호하고 폭력을 줄이고 , 더 이해심 있고 책임감 있게 치안 활동을 수행합니다 .

차량 유지 관리

NLP 를 사용하면 운전자가 차량을 최상의 상태로 손쉽게 유지할 수 있습니다 . 두꺼운 사용 설명서를 펼치는 대신 소유주가 차량에 " 표시된 경고등이 무엇이니 ?", " 퓨즈를 어떻게 교체해 ?" 와 같이 질문하면 됩니다 .

항공기 수리

NLP 를 사용하면 기계에서 방대한 정비 매뉴얼의 정보를 합성하여 조종사가 보고하는 문제에 대한 이해도를 높일 수 있습니다 .

1. 대규모 데이터 처리

NLP 를 올바르게 수행하려면 엄청난 양의 데이터가 필요합니다 . 그 규모는 사람들이 사용하는 모든 단어의 양이라고 생각하면 됩니다 .

NLP 은 방대한 데이터 라이브러리에 기반하여 입력된 단어를 처리하고 이해 및 추론을 거쳐 몇 밀리초 이내에 지능적인 응답을 생성할 수 있어야 합니다 .

이 요구사항은 규칙과 예외로 가득 찬 인간 언어의 복잡성을 고려하면 특히나 어렵고 , 관용어 , 풍자 , 유머의 미묘한 차이까지 고려하면 훨씬 더 어려워집니다 또한 , 산업별 모델은 특정 도메인 , 회사 또는 제품에 대한 특정 정보가 필요할 수 있습니다 .

따라서 대화형 AI 모델의 크기가 수백만 또는 수십억 개의 매개 변수로 증가합니다 . 일반적으로 데이터가 많을수록 모델이 더 정확해집니다 . 이러한 크기의 모델을 훈련하려면 몇 주의 컴퓨팅 시간이 걸릴 수 있으므로 최상의 머신 러닝 및 딥 러닝 프레임워크가 필요합니다 .

Google Translate

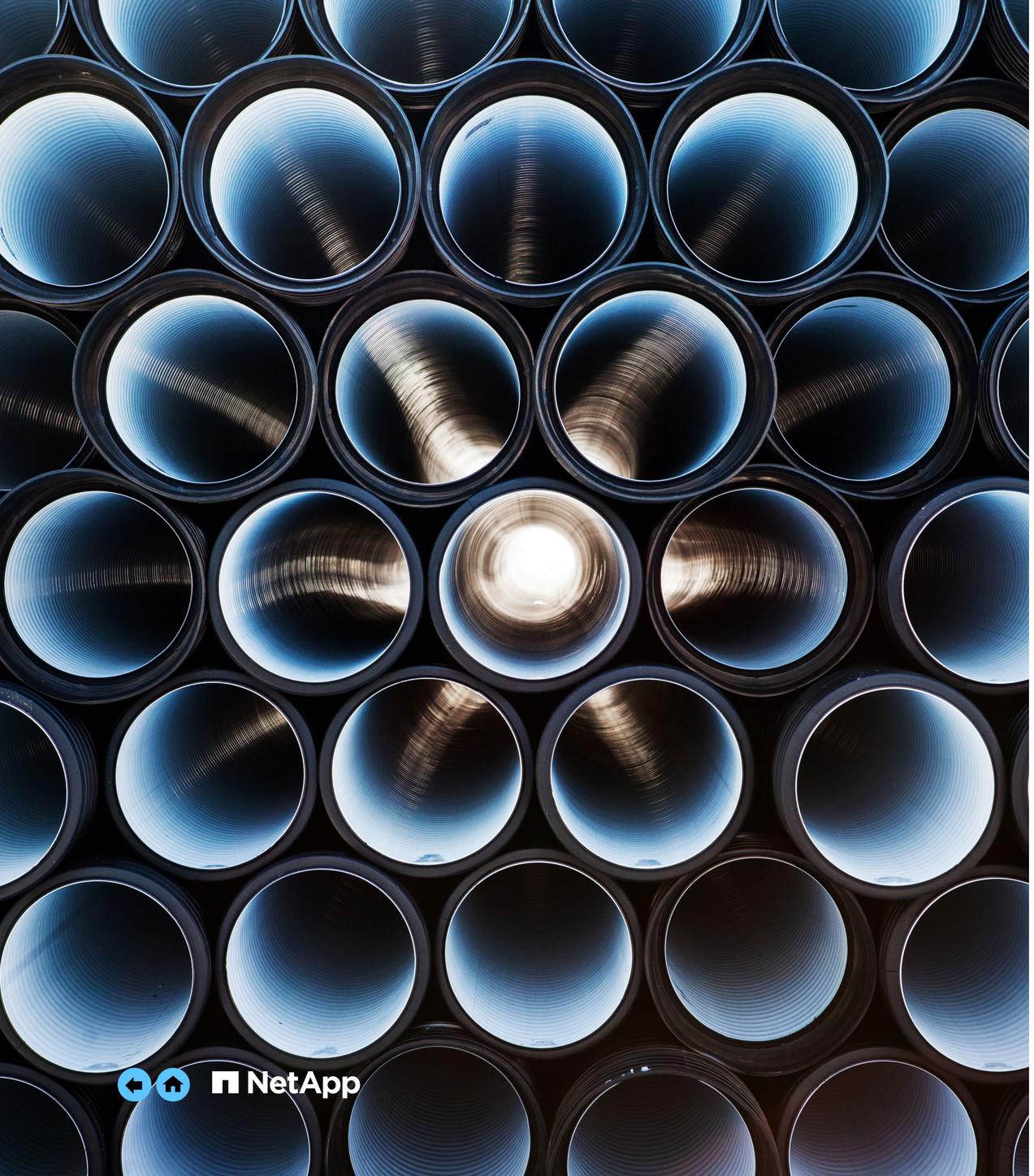
Google Translate 는 100 개 이상의 언어를 지원하며 크라우드 소싱을 통해 제한된 훈련용 말뭉치 (소스 데이터 세트를 위한 고급 단어) 로 언어에 대한 번역과 모델 훈련을 테스트하고 개선할 수 있도록 지원합니다 . Google Translate 는 매일 1,400 억 개 단어를 처리합니다 . 이는 7,000 만 명의 번역사가 작업할 분량입니다 . *매일 .*

Google BERT

Google BERT 는 3 억 4,000 만 개의 매개 변수를 보유한 인기 NLP 모델입니다 . BERT 는 전화 트리 알고리즘과 같은 트랜잭션 음성 인터페이스를 능가하여 진정한 대화를 가능하게 하므로 NLP 의 혁신적인 발전을 나타냅니다 . BERT 는 텍스트를 읽고 질문에 매우 정확하게 답할 수 있습니다 .

BioMegatron

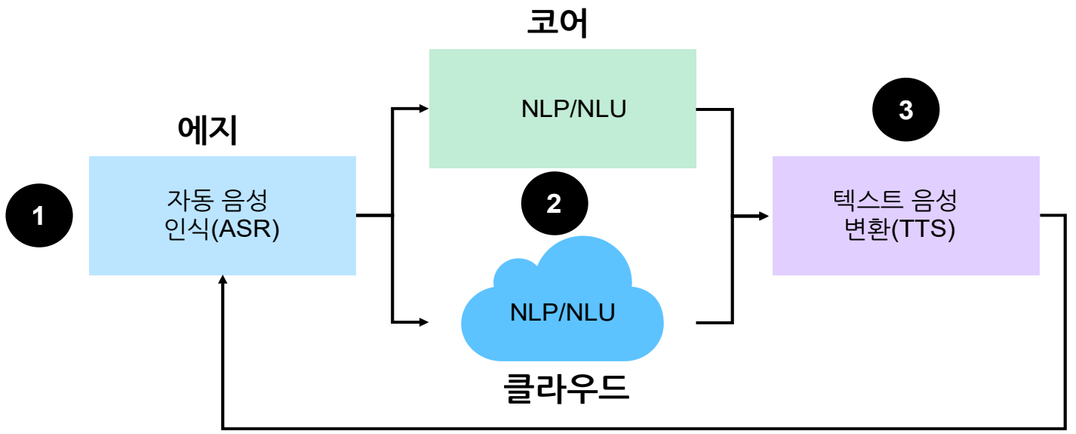
BioMegatron 은 사상 최대 규모의 학습된 생체의학 트랜스포머 기반 언어 모델이며 , 최대 12 억 개의 변형 매개 변수를 보유합니다 . 또한 , 생체의학 주제에 관한 추상화 및 전체 텍스트 저널 문서 저장소인 PubMed 에서 61 억 개 단어를 학습했습니다 .



2. 파이프라인 정리

빠르고 효과적인 NLP 에는 수집 , 인식 , 음성 합성 등 전체 에코시스템을 포괄하는 데이터 파이프라인이 필요합니다 . 데이터가 파이프라인의 각 단계에서 빠르고 자유롭게 이동하여 실시간 언어 처리를 촉진해야 합니다 .

일반적인 NLP 파이프라인은 3 단계로 구성됩니다 .



현대 NLP 인프라에서는 수천 개의 에지 위치에서 매일 테라바이트 단위의 데이터를 수집합니다 . 사일로화된 인프라로 인해 이 데이터에 대한 액세스가 제한될 경우 딥 러닝은 수박 겉핥기에 불과합니다 .

3. 즉각적인 응답

AI 가 인간의 언어를 복제하려면 인간의 뇌가 반응하는 속도보다 더 빠르게 작동해야 합니다 . 모델이 클수록 사용자가 질문한 시간부터 AI 가 응답하는 시간까지 간격이 더 길어집니다 . 자연스럽게 들리기 위해서는 모든 계산이 300 밀리초 이내에 완료되어야 합니다 .

이 프로세스는 다음과 같은 몇 단계로 구성됩니다 .

1. 사용자의 말을 텍스트로 변환
2. 텍스트의 의미 이해
3. 문맥에서 가장 적합한 응답 검색
4. 다시 언어로 응답 제공

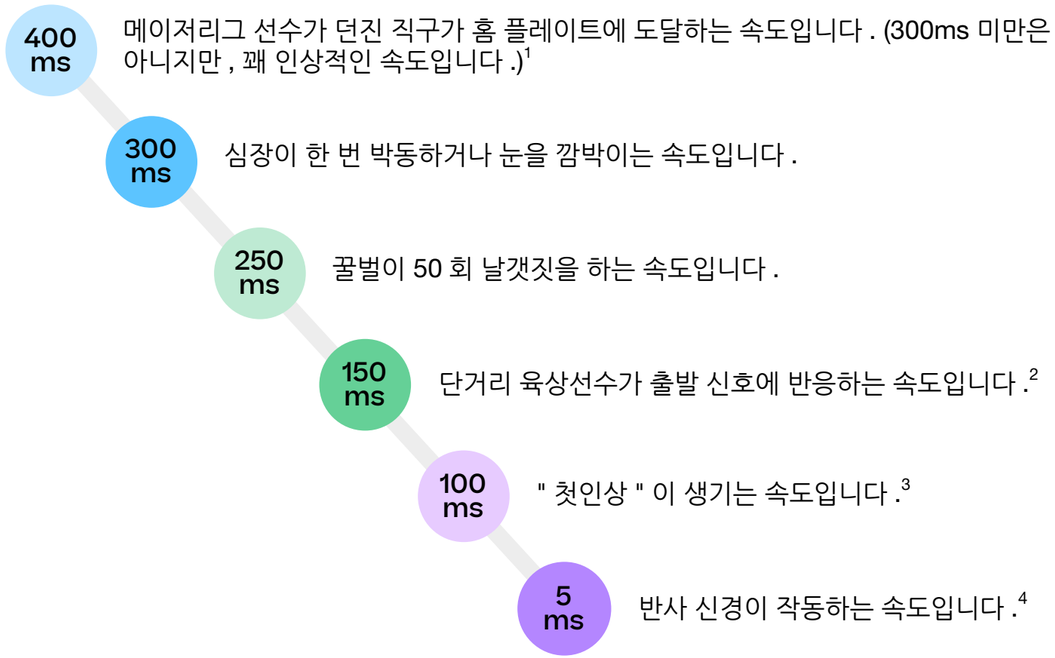
이렇게 엄격한 지연 시간 요구사항으로 인해 대화형 AI 개발자가 타협점을 찾아야 하는 경우가 있습니다 . 복잡한 고품질 모델은 크지 않은 언어 처리 모델에 비해 더 오래 걸릴 수 있습니다 . 작은 모델은 신속하게 결과를 제공하지만 , 미묘한 응답 차이가 부족합니다 .

초조한 구직자처럼 음성 도우미는 대화 중에 " 찾아보겠습니다 ." 라고 말하거나 어색한 침묵을 채우기 위해 삑삑거리는 소리를 내면서 시간을 지연시킬 수 있습니다 . 하지만 이상적인 대화형 AI(NLP 의 성배) 는 사람의 쿼리를 정교하고 정확하게 이해하고 , 원활한 자연어로 신속하게 응답합니다 .

얼마나 빠르게 대화하나요 ?

NLP 는 일반적으로 300 밀리초 (0.3 초) 이내로 실시간 응답을 생성합니다 . 그게 얼마나 빠른 것인가요 ? 매우 빠릅니다 .

300ms 이내에 발생하는 일 :



사용자의 언어로 말하는 NetApp

NVIDIA DGX 시스템 및 NetApp 클라우드 연결 All-Flash 스토리지 시스템 기반의 NetApp® ONTAP® AI 를 사용하면 , 대규모 최신 언어 모델을 훈련 및 최적화하여 신속하게 추론할 수 있습니다 . NetApp 에서 제공하는 Data Fabric 은 에지 , 코어 , 클라우드의 AI 데이터 파이프라인에서 데이터 관리를 간소화합니다 .

- **NetApp AI 솔루션**은 병목 현상을 제거하여 데이터를 효율적으로 수집하고, AI 워크로드를 가속하고, 클라우드를 원활하게 통합할 수 있습니다.
- **NetApp의 통합 데이터 관리 솔루션**은 하이브리드 멀티 클라우드 환경에서 원활하고, 비용 효율적인 데이터 이동을 지원합니다.
- **NetApp의 세계적인 파트너 에코시스템**은 AI 리더, 채널 파트너, 시스템 통합업체, 소프트웨어 및 하드웨어 공급자, 클라우드 파트너를 기술과 완벽하게 통합됩니다. 함께 결합하여 비즈니스 목표를 달성하는 데 도움이 되는 스마트하고 강력하고 신뢰할 수 있는 AI 솔루션을 구축합니다.
- **NetApp 프로페셔널 서비스**는 복잡성을 줄이고 AI 기회와 성공을 넓혀 나가는 데 필요한 전문 지식을 제공합니다.

그리고 NetApp 은 IDC MarketScape 에서 전 세계 스케일아웃 파일 기반 스토리지 부문의 선두업체로 선정되었습니다 .[5] 컴퓨터 비전 워크로드가 스케일아웃 및 파일 기반이므로 이는 매우 중요합니다 .

데이터 과학자를 행복하게 만들기

5배

AI 파이프라인을 통해 5 배 더 많은 데이터를 실행합니다 .

<60 초

몇 시간 또는 며칠이 아닌 몇 초 이내에 데이터 세트 복사

20 분 이내

20 분 이내에 Ansible 과 통합하여 AI 인프라 구성

NetApp 소매 도우미: 성공을 위한 청사진

대화형 AI 서비스 구축을 위한 엔드 투 엔드 프레임워크인 NVIDIA Jarvis 를 사용하여 NetApp 과 NVIDIA 는 언어 또는 텍스트 입력을 허용하고 기상관측소 API, Yelp Fusion API, eBay Python SDK 에 연결하여 날씨 , 관심 포인트 , 재고 가격에 대한 질문에 답하는 가상 소매 도우미를 구축했습니다 . 확인하기

NetApp Retail Assistant(NARA) 는 다음을 기반으로 구축되었습니다 .

- **NVIDIA Jarvis.** Jarvis는 지연 시간을 낮게 유지하도록 최적화된 엔드 투 엔드 딥 러닝 파이프라인을 사용하여 대화형 AI를 위한 GPU 가속 서비스를 제공합니다.
- **NetApp ONTAP AI.** 이 검증된 아키텍처는 NVIDIA DGX 시스템과 NetApp All-Flash 스토리지를 결합합니다. ONTAP AI는 데이터 흐름을 안정적으로 간소화하여 지연 시간 요구사항을 초과하지 않고 복잡한 대화형 모델을 훈련하고 실행할 수 있도록 지원합니다.
- **NVIDIA NeMo.** GPU 가속 대화형 AI 모델을 구축, 훈련, 미세 조정하기 위한 Python 툴킷인 NeMo를 사용하면 실시간 자동 음성 인식(ASR), 자연어 처리(NLP), 텍스트 음성 변환(TTS) 애플리케이션을 비롯한 간편한 API를 통해 모델을 구축할 수 있습니다.

NLP에 반대하시나요?

자 , 그렇다면 숲속 동물과 대화하시겠습니까 ? 다람쥐에게 말을 가르칠 수는 없습니다 . 하지만 여러분에게 NLP 에 적합한 인프라를 구축하는 방법을 알려 드릴 수는 있습니다 .

NetApp AI 솔루션에 관해 자세히 알아보십시오 .

- NetApp AI
- ONTAP AI
- NLP용 NetApp 솔루션

질문이 있습니까 ? NetApp AI 솔루션 전문가가 대기하고 있습니다 .

1. O'Neill, Shane. 실시간 인창: 200밀리초 이내에 발생하는 일 Nanigans.
2. Welsh, Tim. 생각하는 데 걸리는 정확한 시간 The Christian Science Monitor. 2015년 7월 1일
3. Wargo, Eric. 첫인상을 갖는 데 걸리는 시간 심리과학협회. 2006년 7월 1일
4. Wise, Jeff. 생각의 속도는? New York Magazine. 2016년 12월 19일
5. Potnis, Amita. IDC MarketScape: 전 세계 스케일아웃 파일 기반 스토리지 부문 2019 공급업체 평가. IDC. 2019년 12월.

NetApp 정보

평범함으로 가득한 세상에서 NetApp은 특별함을 선사합니다. NetApp은 귀사가 데이터를 최대한 활용할 수 있도록 돕는다는 한 가지 목표에 주력하고 있습니다. NetApp은 귀사에서 사용 중인 엔터프라이즈급 데이터 서비스를 클라우드로 전환하고, 클라우드의 유연성을 데이터 센터에 제공합니다. 업계 최고 수준의 NetApp 솔루션은 다양한 고객 환경과 세계 최대의 퍼블릭 클라우드에서 작동합니다.

클라우드 주도형 데이터 중심 소프트웨어 회사인 NetApp만이 고유한 Data Fabric을 배포하고, 클라우드를 단순화하고 연결하며, 언제 어디서나 원하는 사람에게 원하는 데이터와 서비스, 애플리케이션을 안전하게 제공하도록 지원할 수 있습니다.

자세한 내용은 **www.netapp.com/ko**을 참조하십시오.