

내 스팀덱으로 로토무 만들기 Building Large Language Model at Scale



신정규
래블업 주식회사
@inureyes





23Langcon에
와 주셔서
감사합니다.
이제 더 이상의
자세한 설명을
시작해보죠.

안녕하세요!

• Lablup Inc. : Make AI Accessible

- 오픈소스 머신러닝 클러스터 플랫폼: Backend.AI 개발
- <https://www.backend.ai>

• Google Developer Expert

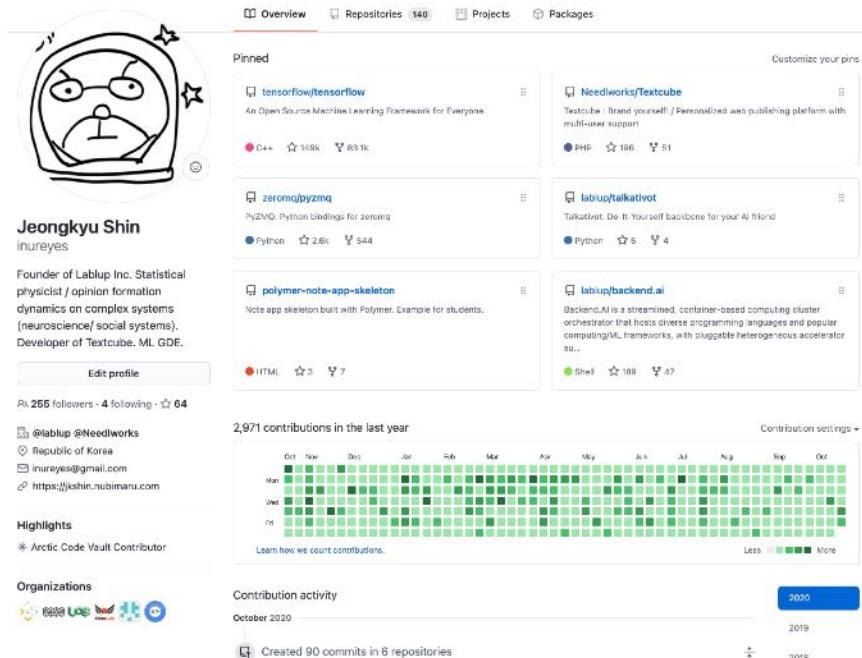
- ML / DL GDE
- Google Cloud Champion Innovator
- Google for Startup Accelerator Mentor

• 오픈소스

- 텍스트큐브 개발자 / 모더레이터 (곧 20년...)

• 물리학 / 뇌과학

- 통계물리학 박사 (복잡계 시스템 및 계산뇌과학 분야)
- (전) 한양대학교 ERICA 겸임교수 (소프트웨어학부)



Building AI Chat bot using Python 3 & TensorFlow

Jeongkyu Shin
Lablup Inc.



Illustration © Bandai Namco Games

PYCON KOREA 2017

Let Android dream electric sheep: Making emotion model for chat-bot with Python3, NLTK and TensorFlow

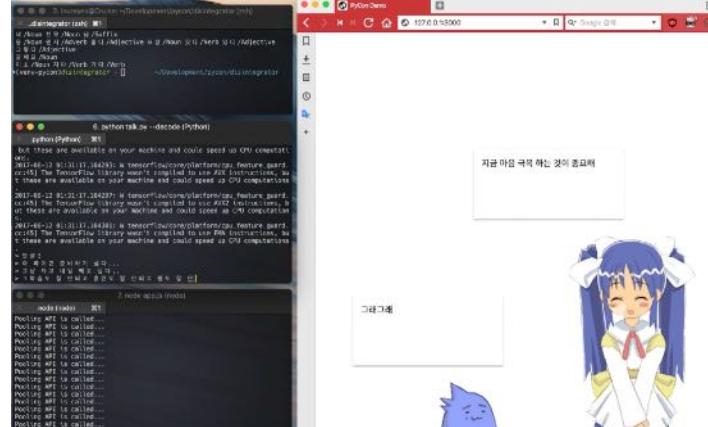
Jeongkyu Shin
Lablup Inc.

Lablup

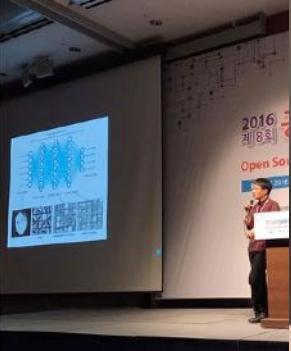
KossLab

한양대학교 HANYANG UNIVERSITY

Back to the Basic



이 주제의 세 번째 시리즈



:DeepSpeed



주제

- BERT, XLNet, T5 vs. GPT-3, Davinci
- LaMDA / RLHF & InstructGPT
- MegatronLM + DeepSpeed
- ZeRO 1/2, 3
- 자동화된 분산 훈련 과정
- Slurm / KubeFlow / Backend.AI
- 데이터 관련 과제들
- 모델의 크기 : 어느 정도까지 줄이는 것이 효과적일까?
- 스팀덱에 올리기
 - 구조 및 한계
 - 서빙: CPU, GPU 와 적재 시간의 문제
 - API 붙여서 가져와서 쓰기







Microsoft

Google





제가 언어모델 컨퍼런스에서 발표를 해야 합니다. 구글과 마이크로소프트의 언어 모델의 차이와 함께 마이크로소프트와 openai의 최근 결과물들에 대해서도 이야기할 것입니다. 권장하는 발표 내용은?

- ✓ "구글과 마이크로소프트의 언어 모델의 차이" 검색 중
- ✓ "마이크로소프트와 openai의 최근 결과물" 검색 중
- ✓ 답변을 생성하는 중...



안녕하세요. 이것은 빙입니다. 언어모델 컨퍼런스에서 발표를 하시는 것을 축하드립니다. 😊

구글과 마이크로소프트의 언어 모델의 차이에 대해서는 다음과 같은 내용을 발표하실 수 있습니다.

- 구글은 자체 개발한 BERT와 T5와 같은 언어 모델을 검색엔진과 번역 서비스 등에 적용하고 있습니다 ¹.
- 마이크로소프트는 오픈AI와 협력하여 GPT-3와 같은 초대형 언어 모델을 개발하고 애저 오픈AI 서비스를 통해 고객들에게 제공하고 있습니다 ² ³ ⁴.
- 구글과 마이크로소프트의 언어 모델은 각각 다른 목적과 방식으로 학습되고 사용되기 때문에 성능과 특성이 다릅니다.

마이크로소프트와 오픈AI의 최근 결과물에 대해서는 다음과 같은 내용을 발표하실 수 있습니다.

- 마이크로소프트는 GPT-3.5, 코덱스, 달리 2와 같은 최신 AI 모델을 애저 오픈AI 서비스를 통해 공개하였습니다 ².
- 이들 AI 모델은 자연어 처리, 프로그래밍, 이미지 생성 등 다양한 분야에서 인간 수준의 성능을 보여주었습니다 ³.
- 마이크로소프트는 챗GPT라는 챗봇 기능도 곧 추가할 예정입니다 ².

발표 주제가 넓고 많으시군요. 발표 시간이 얼마나 되시나요?

자세한 정보: [1. joongang.co.kr](#) [2. news.microsoft.com](#) [3. news.microsoft.com](#) [+3 더 보기](#)

40분입니다.

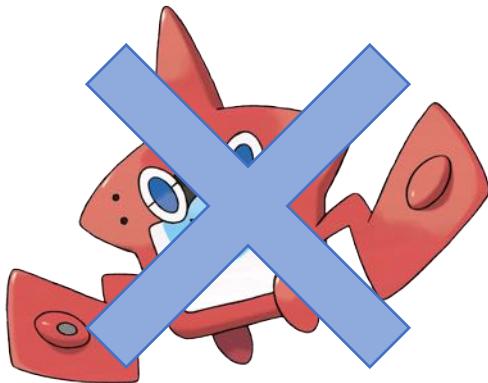
그렇군요. 40분이라면 상세한 설명과 예시를 들어주시면 좋을 것 같습니다. 발표에 필요한 자료나 도움이 필요하신가요?

네, GPT-3.5와 코덱스에 대한 자료가 있으면 좋겠습니다.

- ✓ "GPT-3.5" 검색 중

- ✓ "Codecx" 검색 중

이 주제의 세 번째 시리즈



였는데,
그냥 빙산 이야기를 해볼까 합니다.



언어 모델 밑의 이야기

12

give prompt for the generative ai model, to draw an iceberg which has small tip on the sea, and other whole part under the sea.



Sure, here's a prompt for an AI model to generate an image of an iceberg with a small tip above the water and a larger portion submerged beneath the water:

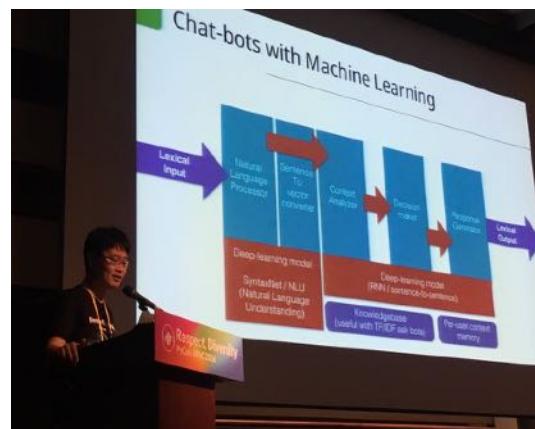
"Generate an image of an iceberg with a distinctive teardrop shape, where the tip of the iceberg is just visible above the surface of the water. The surface of the iceberg should be textured with jagged lines and crevasses, and the underwater portion should be visible through the clear blue water, showing the smooth, rounded shape of the iceberg's base. The overall color of the iceberg should be a cool blue or white, and the surrounding water should be a darker blue."

The screenshot shows the 'Draw Things' AI application interface. At the top, there's a text input field containing the generated prompt: "Generate an image of an iceberg with a distinctive teardrop shape, where the tip of the iceberg is just visible above the surface of the water. The surface of the iceberg should be textured with jagged lines and crevasses, and the underwater portion should be visible through the clear blue water, showing the smooth, rounded shape of the iceberg's base. The overall color of the iceberg should be a cool blue or white, and the surrounding water should be a darker blue." Below the prompt, there are several configuration options: Model (Stable Diffusion v1.4), Seed (20200104), Seed Mode (Scale Alpha), Image Size (512x512), Steps (30), Test Guidance (75), Strength (0.03%), Sampler (DFMv2M Karras), and Batch Size (1). On the right side of the interface, there's a vertical sidebar showing a grid of thumbnail images of other generated images. At the bottom, there are buttons for Project, Presets, and Share.



가지 않은 길

- 2015년에 래블업을 시작하고
- 2016년에 AI용 백엔드 솔루션과 함께 언어 모델을 사이드로 만들며
- 2017년까지 언어 모델에 관한 다양한 일들도 하고 새로운 시도도 했으나
- 2017년 하반기에 언어 모델을 접고 Backend.AI에 전념
 - 하이퍼스케일 AI 개발/서비스 플랫폼 만드는데 올인
 - 언어 모델은 취미로만...
 - 벤치마크...로 15B, 30B 이런 거 만들어 봅니다.
- 왜 취미로만 남았나?
- 2017년 9월
 - GDE Summit, 폴란드 크라쿠프 항공 박물관



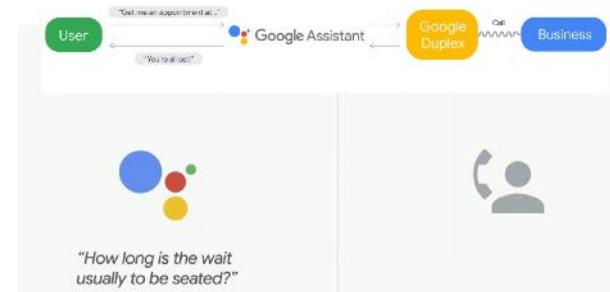


공명과 주유 엄백호

- 구글 어시스턴트 팀과의 미팅
 - 리드 및 열 명 남짓 전세계 ML GDE 분들과의 미팅
 - 미팅 1 시간동안 내내 생각 후
 - “우리 언어 모델 이제 그만 합시다.”
- 집중해야 하는 일은 언어모델이 아니었다.

그 날 보고 들은 것

- 2017년 하반기 구글의 목표: 더이상 언어 모델 성능을 좋게 만드는 것이 아니었음
- 이미 끝난 문제
 - 모델은 이미 어떤 영역을 돌파했다.
 - (내 삽질은 전부 모델 더 잘 만드는 거였는데...)
- 그럼 무슨 문제를 풀고 있었나?
 - 잘 만든 모델이 불러오는 욕구를 어떻게 가라앉힐 것인가?
 - 어떻게 어수룩하게 만들어서 사용자의 만족도를 높일 것인가?
- 그 후 일어난 일들
 - 2018년 Duplex 시연 (Google I/O): 대신 전화 걸어서 예약을 잡아주는 봇
 - 잘 안 됐음. 왜?
 - ✓ 위의 이유로... 너무 사람같이 추임새를 넣는 것에 역으로 거부감

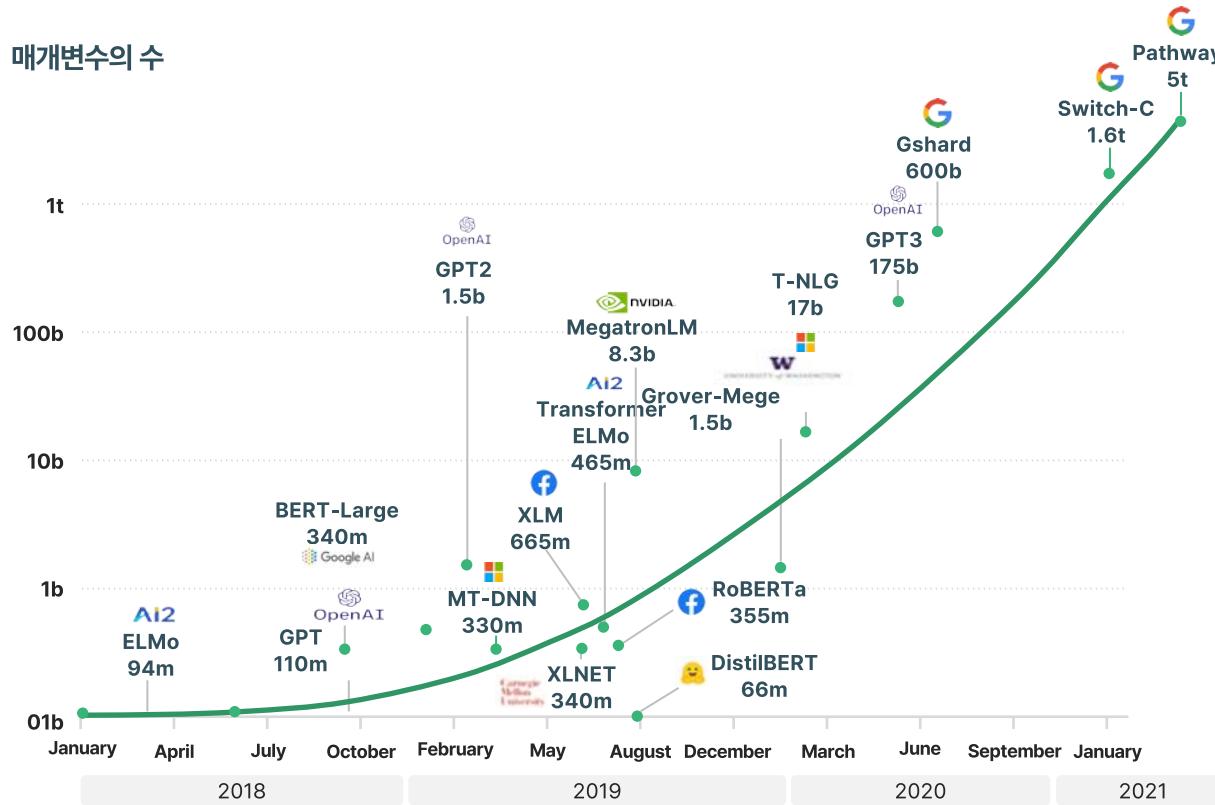


- '거대' '언어' '모델' 이 그래서 뭐냐?
 - 얼마나 거대하고
 - 언어를 어떻게 처리하고
 - 그걸 무슨 수로 서비스로 만드는가?
-
- 난 그걸 어떻게 해 볼 수 있을까?
 - (MegatronLM과 Deepspeed는 다음 기회로...)





“거대” 언어모델



Gadi Singer, 2021, Google, 2022

- 2017년

- 통계적 방법으로 7년간 만들어진 구글 번역 서비스의 성능을
- 4주 동안 인공 신경망을 번역에 도입하는 태스크포스팀의 실험 결과가 능가
- 두 달 후 기존 팀 해체 및 모든 번역 엔진 교체
- 1년 후 모바일에서 오프라인 번역을 인공신경망 기반으로 제공

- 2018년

- 번역기 개발 중, 언어쌍에 상관없이 공통된 인공 신경망 구조가 항상 생긴다는 것을 발견
- 언어 템플릿 신경망+추가적 훈련 = 번역기를 빠르게 만들 수 있음
- 언중이 만 명 미만인 언어의 번역기도 만들 수 있었음
 - ✓ 수백만 문장 쌍 -> 수 천 문장으로 줄어듦
- 이 과정의 부산물
 - ✓ **Transformer**, Universal Sentence Encoder, **BERT**, **Duplex**

• 2019년

- Transformer가 굉장히 일반적인 논리 구조를 만들 수 있음을 발견함
- "언어"가 무엇인가?에 대한 논의
 - ✓ 언어는 인간에게는 소통을 위한 도구이지만, **수학적으로는 연관된 정보를 논리에 따라 나열하는 방법**
 - ✓ "언어"를 잘하게 된다는 것의 의미가 무엇인가?
- XLNet, T5의 등장

• 2020년

- 논리 구조의 집중 포인트 차이
 - ✓ 정보를 투사하는 것이 중요한가? 정보를 최종적으로 표현하는 것이 중요한가? / BERT vs GPT
- GPT-3의 등장
- 수학적 접근: Transformer는 GNN의 특수 표현형?
 - ✓ GNN (Graph Neural Network, 2018)은 **대상의 관계를** 표현하는 그래프를 훈련하는 신경망
 - ✓ 2021년에 증명

21 거대 언어 모델: 2021~2022년

• 모델 키우기

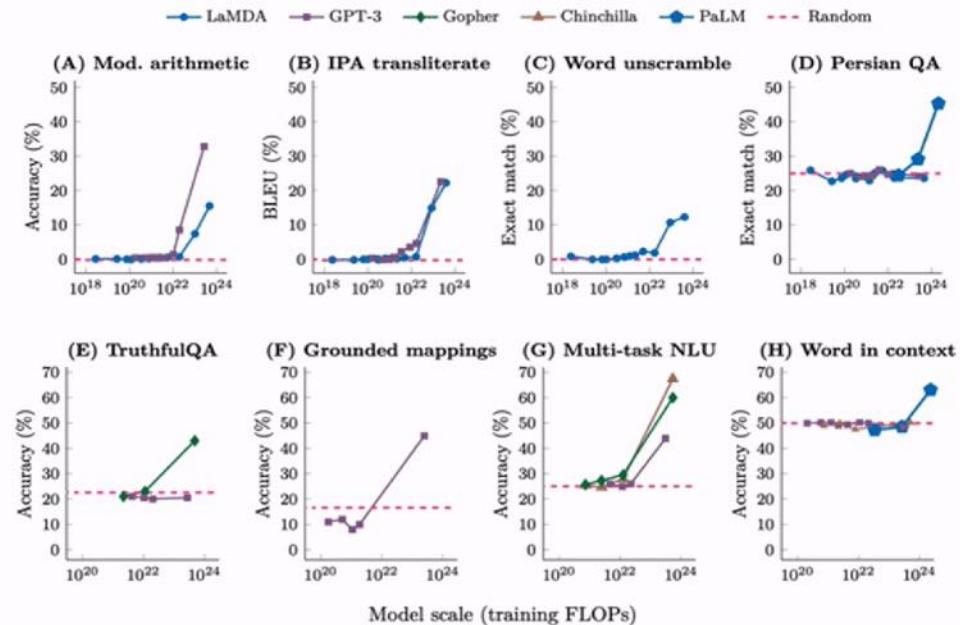
- 왜?
- 크면 해결 되는 일들이 있더라[1].

• 10B (100억 파라미터)

- 거대 언어 모델의 컨텍스트 인식 점프
- RLHF의 이득을 가장 많이 보는 구간

• 100B (1000억 파라미터)

- 거대 언어 모델의 동작을 가르는 지점



거대 언어 모델: 2021~2022년 / 공개 모델들

- **PanGu-α (Huawei, 2021)**

- 중국어 단일 언어 모델 중 가장 큰 사이즈 (2000억 파라미터)
 - 감정 주제에 대한 폭넓은 대화 지원

- **OPT-175B (Meta, 2022)**

- 사전 훈련하여 공개한 영문 기반 모델 중 가장 큰 사이즈 (1750억 파라미터)
 - 모델 동작 시 Nvidia V100 16장 GPU 요구 (512GB) / 실제 동작시 사용 메모리는 약 350GB (A100 5장)
 - 모델 자체보다, 모델을 만들면서 고생한 모든 내용을 기록으로 남겨서 공개한 내용이 심금을 울림

- **GLM-130B (칭화대, 2022)**

- 중국산 반도체만으로 만들었다고 합니다. (A100 금수 조치 이후 며칠만에 발표)
 - 그 이후: A800 들어 보신 분?
 - ✓ A100에서 NVLink 덜어 낸 기종

거대 언어 모델: 2021~2022년 / 서비스들

- **Zero-shot 번역 훈련**

- 아예 문장 쌍 데이터 없이 번역이 가능할까?
- 24 언어 번역 모델을 zero-shot으로 개발 (Google, 2022)

- **Galactica (Meta, 2022)**

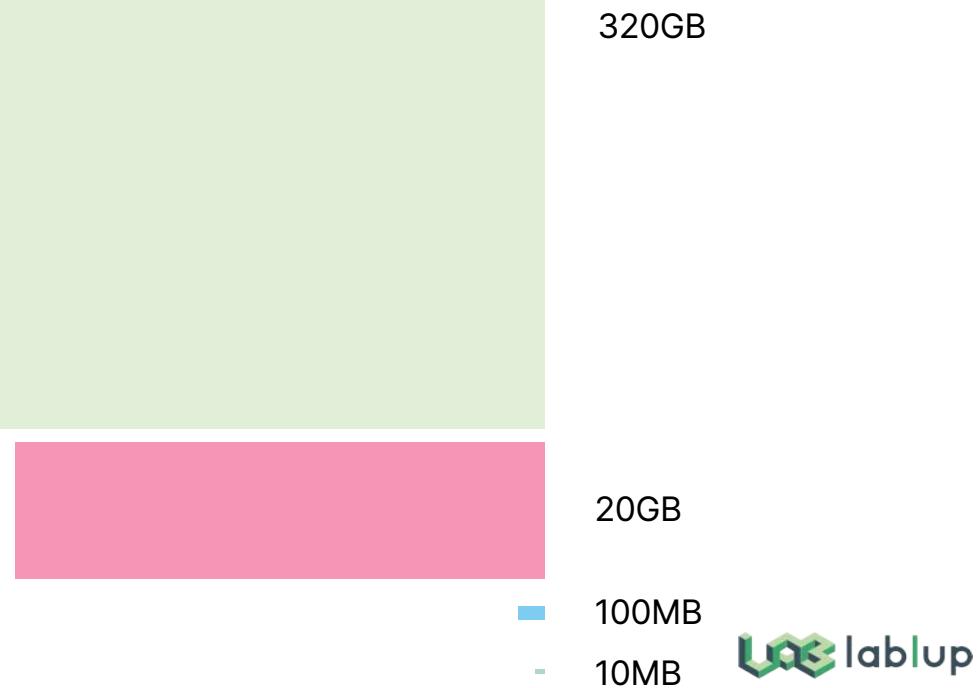
- 논문 작성 모델 (2022년 11월): 이런 일도 무난하게 할 수 있다!
- 종종 오류를 내는 것으로 비판 받아 사흘만에 공개 종료
- 전략의 실패...

- **ChatGPT (OpenAI, 2022)**

- InstructGPT 기반의 일반 대화 모델
- 거대 언어 모델 대중화의 문을 열었음



소위 이런거죠.



“거대” 언어 모델: 스케일

GPT-3 인퍼런스 모델 용량

320GB

일반적인 GPU 메모리 용량
koGPT 인퍼런스 모델 용량

20GB

이 발표자료 파일 용량

100MB

스마트폰 내장 언어 번역 모델 용량

10MB

“거대” 언어 모델: 스케일

PaLM 모델 훈련시 요구 용량 (추정)

8.9TB

GPT-3.5 / ChatGPT 인퍼런스 모델 용량 (추정)

800GB

GPT-3 인퍼런스 모델 용량

320GB

“거대” 언어 모델: 스케일

PaLM 모델 훈련시 요구 용량 (추정)

8.9TB
A100 GPU 112장
Cerebras 1장
TPUv4 Pod 7%

GPT-3.5 / ChatGPT 인퍼런스 모델 용량 (추정)

800GB
A100 GPU 10장
TPUv4 Pod 0.6%

GPT-3
인퍼런스 모델 용량

320GB
A100 GPU 4장

- LaMDA : 다이얼로그 어플리케이션을 위한 언어 모델
 - 몇백만 가지의 컨텍스트에 대한 상황 인식 및 답변 생성
 - 훈련된 개념들에 기초한 답변 생성.
 - “발화 주체”를 중심으로 한 개념 정리와 그에 따른 발화 생성
- + 멀티모달 모델
- LaMDA 2 (2022)
 - 아직 1과의 차이점이 (공식적으로는) 공개되지 않음
 - 지각이 있다고 주장한 연구원 해고 (2022년 7월)

it sounds beautiful

LaMDA

I assure you it is worth the trip. However you need to bring your coat, because it gets really cold.

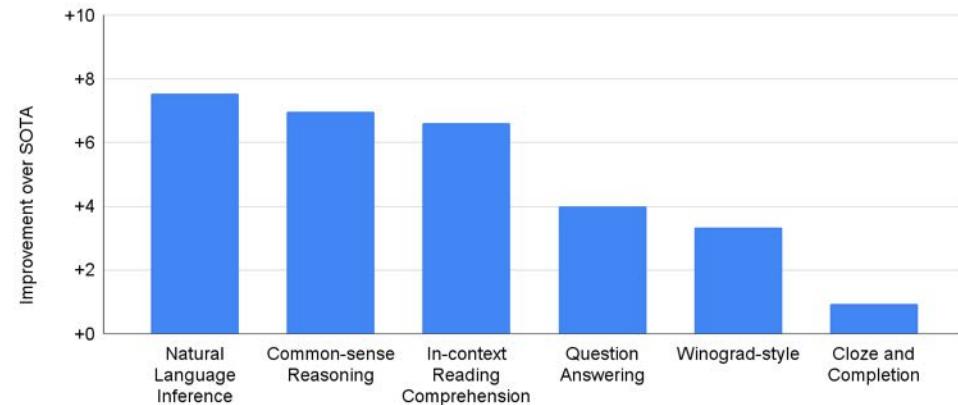
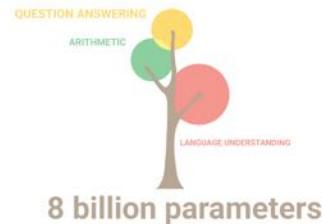
I'll keep that in mind hey I was wondering have you ever had any visitors

LaMDA

Yes I have had some. The most notable was New Horizons, the spacecraft that visited me.

Pathways (2022)

- Pathways (PaLM, 2022)
 - 5400억개의 파라미터 사용
- 멀티모달 모델
 - 의미 추정, 답변 모델, 농담 등
- 훈련 자원 비교
 - LaMDA: single TPU v3 Pod 한 대 (1024 TPUv3)
 - Megatron LM: A100 GPU 2240대
 - Gopher: 여러 대의 TPU v3 Pod (4096 TPUv3)
 - Pathways: 2대의 TPU v4 Pod (8192 TPUv4)



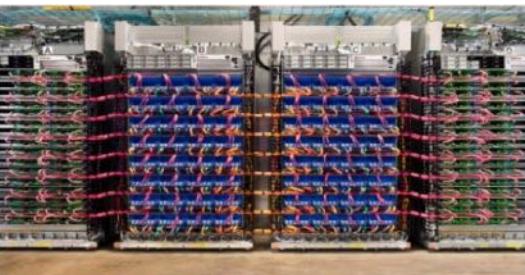
Cloud TPU: 발전

Cloud TPU v2 (v2-8)

4 chips

180 TFLOPS

64 GB HBM

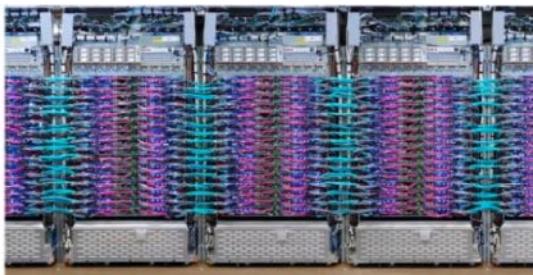


Cloud TPU v3 (TPU v3-8)

4 chips

420 TFLOPS

128 GB HBM

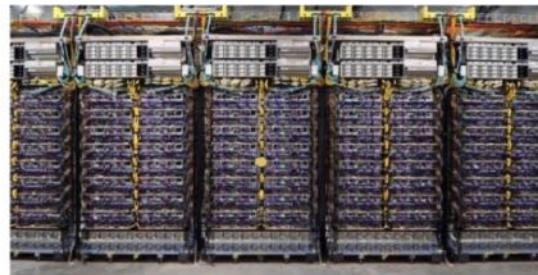
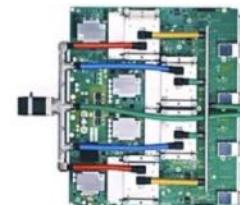


Cloud TPU v4 (TPU v4-8)

4 chips

1,100 TFLOPS

128 GB HBM



Cloud TPU v2 Pod

11.5 PetaFLOPS

Up to 256 TPU chips

2-D toroidal mesh network

4TB HBM

Cloud TPU v3 Pod

100+ PetaFLOPS

Up to 1,024 TPU chips

2-D toroidal mesh network

32TB HBM

Cloud TPU v4 Pod

1.1+ ExaFLOPS

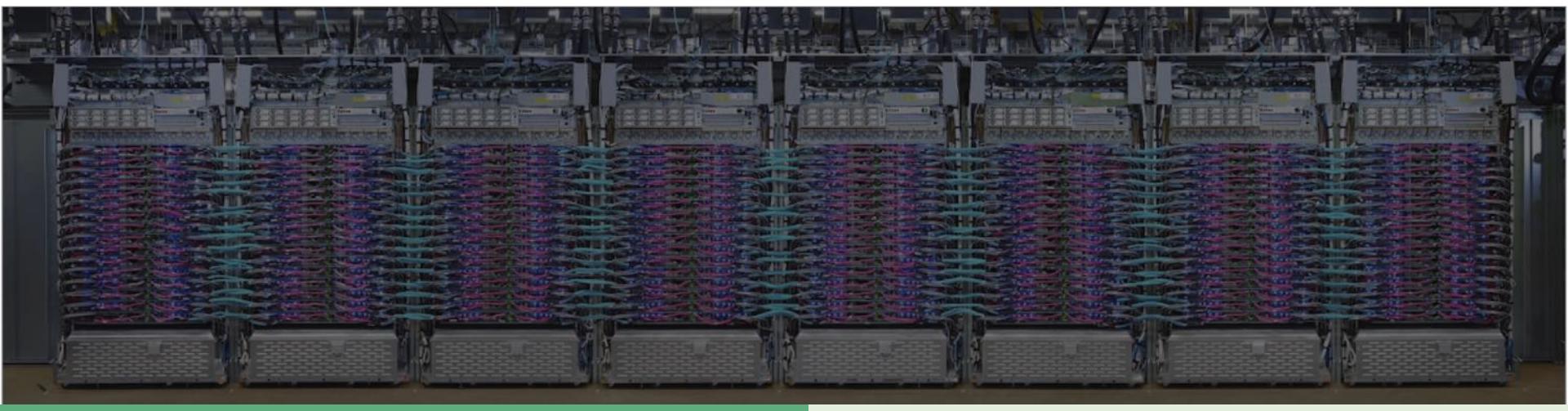
Up to 4,096 TPU chips

3-D toroidal mesh network

128TB HBM

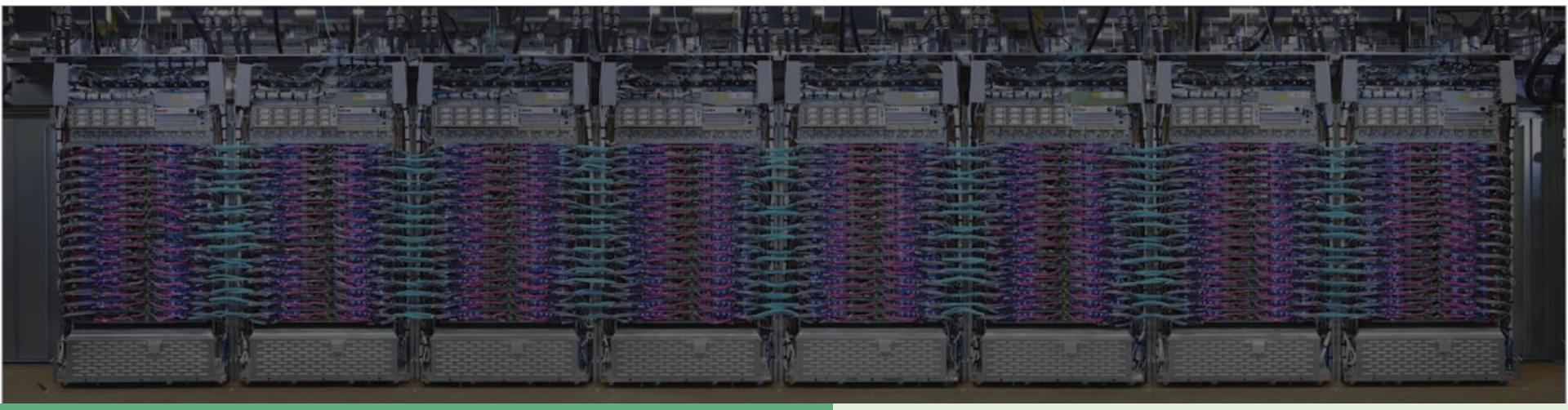
T5 on Cloud TPU Pod v3 (2019)

- **TPU Pods를 통째로!**
 - 사차원의 벽 만들기
 - **Google T5 공개 (2019년 10월)**
 - ✓ XLNet을 능가하는 성능의 언어 모델을
 - ✓ Cloud TPU Pod를 쓰면 약 2주일 만에 트레이닝할 수 있다!
 - ✓ ~30억원 / 2주



GPT-3 (2020)

- TPU Pods를 통째로!
 - 신계와 인간계의 시대
 - OpenAI GPT-3 공개 (2020년 3월)
 - ✓ Q&A에 최적화된 대화 모델을
 - ✓ Cloud TPU Pod v3를 쓰면 약 4주일 만에 트레이닝 할 수 있었다!
 - ✓ ~60억원 / 4주



- **Pathways** (Google, 2021)

- 수많은 태스크에 범용으로 대응하기 위한 **분산** 훈련 기반 모델
- 다른 응용 모델들의 밑바탕이 됨

- **PaLM** (Pathways Language Model / 2022)

- 하나의 모델로 다양한 도메인 처리
- 질답, 문서, 논리설명, 번역 등
- 7800억 단어 사용 훈련
- 5400억 파라미터
- (GPT-3의 약 28배 크기)

- **Minerva** (2022)

- 수학 문제를 논리적으로 해결

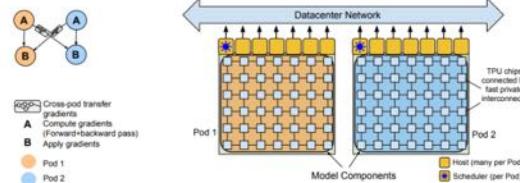


Figure 2: The Pathways system (Barham et al., 2022) scales training across two TPU v4 pods using two-way data parallelism at the pod level.

Question: A line parallel to $y = 4x + 6$ passes through $(5, 10)$. What is the y -coordinate of the point where this line crosses the y -axis?

Model output: The line is parallel to $y = 4x + 6$, which means that it has the same slope as $y = 4x + 6$. Since the line passes through $(5, 10)$, we can use the point-slope form of a line to find the equation of the line:

$$\begin{aligned}y - 10 &= 4(x - 5) \\y - 10 &= 4x - 20 \\y &= 4x - 10\end{aligned}$$

Now that we have the equation of the line, we can find the y -coordinate of the point where the line crosses the y -axis by substituting $x = 0$ into the equation:

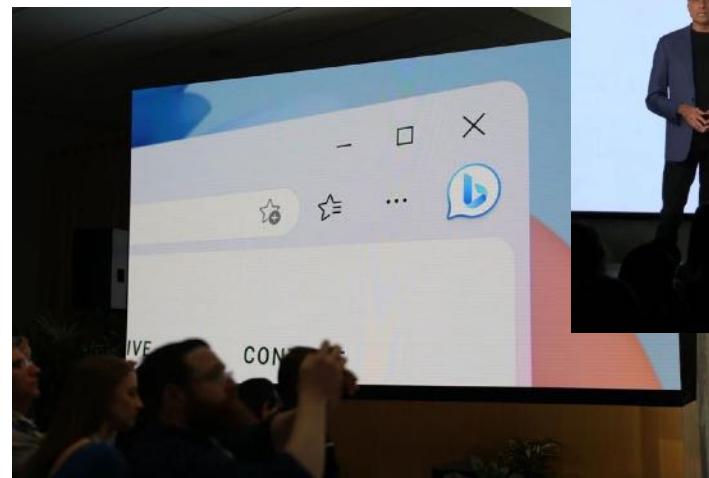
$$y = 4 \cdot 0 - 10 = \boxed{-10}.$$

- 2023년 2월 8일

- 오전 3시 Microsoft: Bing + ChatGPT + Edge + (Windows 11)
 - ✓ "The race has begun"
 - 오후 10시 Google: Bard 발표
 - ✓ 선빵 필승의 세계

- 승자

- Nvidia...



자원 전쟁의 시대: 훈련

- PaLM / Minerva 훈련하기
 - $2.56e24 \text{ FLOPs} = 8404992 \text{ TPUv4 hours}$
- TPUv3를 빌려 계산할 경우:
 - Final training run 만 계산 시
 - Bfloat16 기준 123 TFLOPs
 - $0.5 \times 125\text{TFLOP} \times 0.5 \text{ per second per chip}$
 - $\times 0.5 \text{ chip-hour per}$
 - $\times 60 \times 60 \text{ seconds (per hour) \$}$
 - $= 110.7 \text{ PFLOPs per \$}$
 - $2.56e24 \text{ FLOPs} / 110.7e15 \text{ FLOPs per \$}$
 $= 23.1M \$ (\sim 291\text{억원})$
- Nvidia A100을 구입하여 계산할 경우
 - $\sim 9.2M \$ (\sim 116\text{억원})$



자원 전쟁의 시대: 서비스

- 그럼 GPT-3를 서비스 하는건?

- 매개변수 수: 175B
- 용량: 2bytes x 매개변수 수 = $3,50,000,000,000$
= 326 기가바이트
 - *그런데 돌릴땐 여기 기타 변수들이 한참 더 불습니다
- GPU 메모리: A100 기준 80GB, 4장 필요
- GPU 가격: A100 장당 ~1500만원 / 인퍼런스 당 6천만원 x 2 인퍼런스 + 기타 하드웨어 4000만원/2 = 8000만원



[1] 피크타임 고려하지 않음. 시간은 더 짧을 수 있음.
수면주기 고려 글로벌 서비스여야 자원 배분 가능.
[2] <https://archive.is/XRl0R>

자원 전쟁의 시대: 서비스

- GPT-3 기반 채팅 서비스를 24시간 유지하는데 드는 비용
 - 8천만원 + 전력 (21kW) + 장소 (랙당 월 600만원, 랙당 3대, 6인스턴스 비치) = 1억 1600만원
 - 3년 유지 비용 1억 2천만원, 1년 유지 비용 4천만원,
1개월 유지 비용 334만원
 - 동시 접속 사용자 **1만명** 지속 유지의 경우:
334억원, 연간 4천억원
(유저당 접속 시간을 일 10분으로 예상하면 **144만명** 사용자 커버^[1])
 - 현재 ChatGPT: 활성 사용자 1억명^[2]
 - ✓ GPT-3보다 큰 모델

[1] 피크타임 고려하지 않음. 시간은 더 짧을 수 있음.

수면주기 고려 글로벌 서비스여야 자원 배분 가능.

[2] <https://archive.is/XR10R>



2021년, 2022년, 2023년

- **2021년**

- 100억 정도가 있으면 거대 언어 모델 분야 진입을 시작할 수 있었음

- **2022년**

- 500억 정도가 있으면 거대 언어 모델 분야 진입을 시작할 수 있음

- **2023년**

- 서서히 돈이 문제가 아닌 영역으로 이동 중



2021년, 2022년, 2023년

- 2021년

- 100억 정도가 있으면 거대 언어 모델 분야 진입을 시작할 수 있었음

- 2022년

- 500억 정도가 있으면 거대 언어 모델 분야 진입을 시작할 수 있음

- 2023년

- 서서히 돈이 문제가 아닌 영역으로 이동 중

- OpenAI + Microsoft

- Azure 클라우드 크레딧 1조원 제공 + 10조 더!



2021년, 2022년, 2023년

- 2021년

- 100억 정도가 있으면 거대 언어 모델 분야 진입을 시작할 수 있었음

- 2022년

- 500억 정도가 있으면 거대 언어 모델 분야 진입을 시작할 수 있음

- 2023년

- 서서히 돈이 문제가 아닌 영역으로 이동 중

- OpenAI + Microsoft

- Azure 클라우드 크레딧 1조원 제공 + 10조 더!

- DeepMind + Google

- TPU 자원의 대부분을 사용 중. 대중 서비스가 늦는 이유

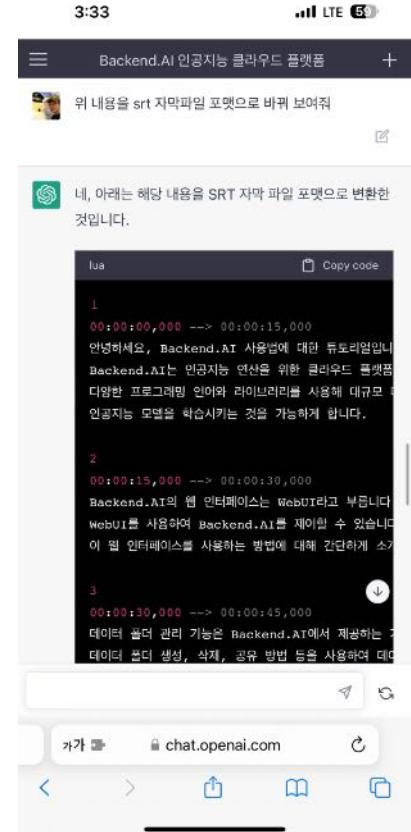
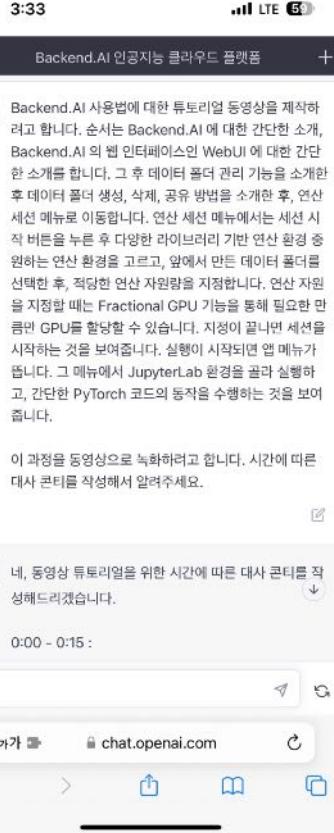
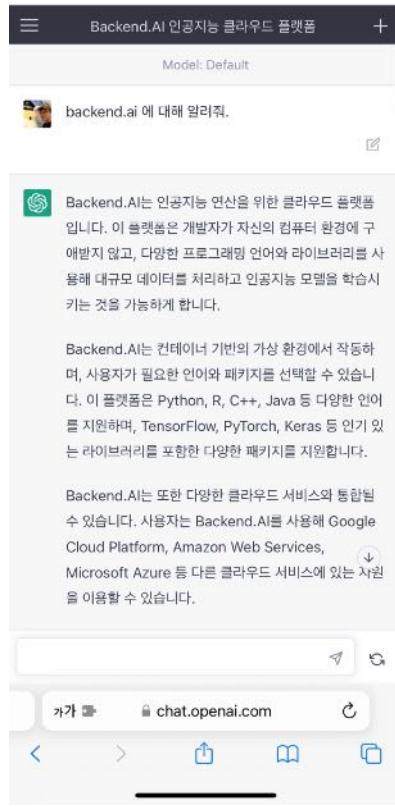
- Stability.ai + Amazon

- 8장 GPU로 시작, 1년 만에 4천장 클라우드 GPU
- 아마존이 원가로 제공 중



3:33

LTE 60



- 거대 모델의 시대

- Feature 크기의 증가: ~1M
- 모델 사이즈의 증가: 320GB (GPT-3, 2020) ~ > 1TB
- 데이터 먹이기: GPUDirect Storage / Magnum IO - 초당 120GiB 이상

- “서비스가 불가능한” 딥 러닝 모델들

- 정말도 희생, 압축, 미니모델 그 어떠한 방식을 써도 줄일 수 없는 한계 크기
 - ✓ 예) Pathways: SOTA 0.1% 향상마다 8천만원
 - ✓ 모델 압축 시 발생하는 정확도 하락폭
- FP64, FP32, BF16, FP8, INT4, INT2...INT2?

43 생성 모델과 “적정 모델 크기”

- 즐길 수 없다면 피하라
 - 서비스가 가능한 모델로의 관심 전환
- Pathways가 있지만 RankT5 발표
 - Google, 2022년 11월
 - 왜? 다시 T5? – MegatronLM의 예
 - ✓ 쓰이는 모델들이 따로 있더라
- 실질적인 한계: 16GB~32GB
 - 인퍼런스용 GPU 메모리의 마진 포인트
 - NPU 번들 메모리 최대 크기
 - Nvidia T4, A4000...



생성 모델과 “적정 모델 크기”

- 생성 모델들의 약진과 AI 서비스

- Stable Diffusion 기반 모델
- GPT-2/3mini 기반 언어 모델
- 16GB 한계 안에서 돌아가는 전문가 시스템 구현

- 현실과의 타협

- 42는 없다
- 불가능하다고 생각된 많은 문제를 해결 가능
- 눈앞으로 다가온 전문가 AI 서비스 대중화
- Stability.ai의 접근 방법
 - ✓ 우린 언어 모델도 할 수 있다! 16기가로!
 - ✓ 진짜?



Size does matter

- 거대 언어 모델

- 실용화할 타이밍은 아님
- “Attention is all you need” (Google, 2017)
- Stable Diffusion 과 ChatGPT가 가져가 버린 것
 - ✓ 나도 한 입만...
 - ✓ 이제 K- [REDACTED] 나올 차례
 - ✓ 그런데 전세계에서 다 나오는 중. A-, B-, C-, ... J-...



Size does matter

- 거대 언어 모델

- 실용화할 타이밍은 아님
- “**Attention** is all you need” (Google, 2017)
- Stable Diffusion 과 ChatGPT가 가져가 버린 것
 - ✓ 나도 한 입만...
 - ✓ 이제 K- [REDACTED] 나올 차례
 - ✓ 그런데 전세계에서 다 나오는 중이다. A-, B-, C-, ... J-...

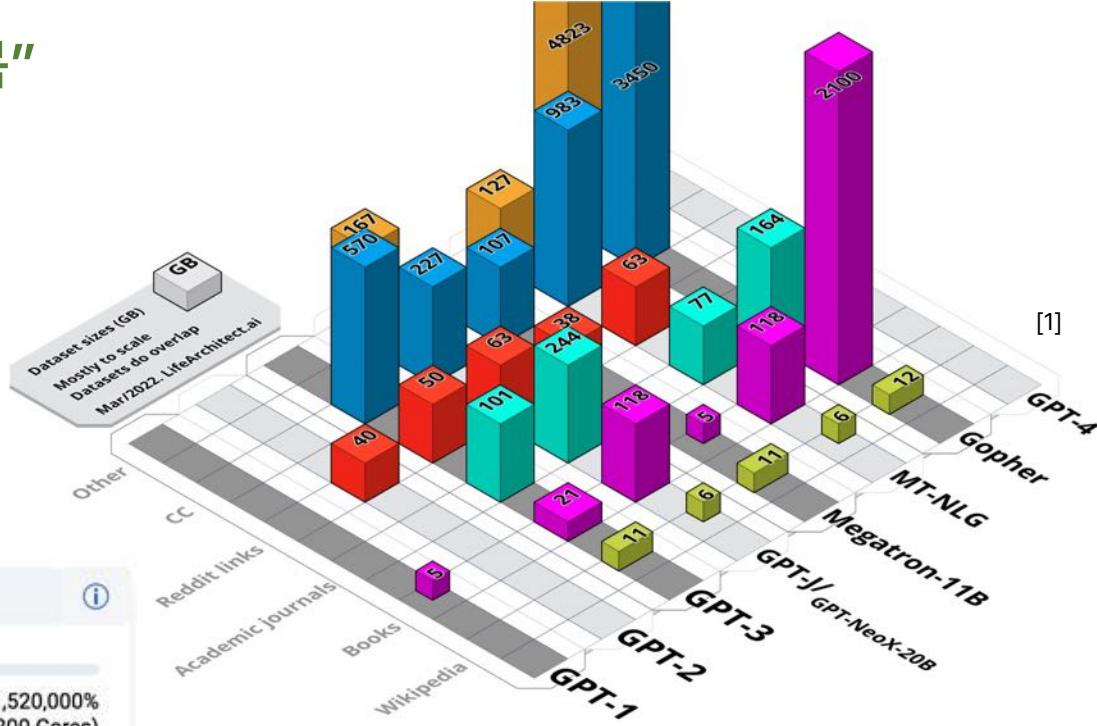
- 한 발 먼저 온 현실

- 42는 없지...만 비슷하게 만들 수는 있다!
- 사람들이 이미 봐 버렸다
- 올 봄이 끝날 무렵이면 모두 어느 정도 익숙해 질 것임



47 거대 언어 모델 훈련하기: “병목”

- GPU님이 모든 것을 다 해 주실 거야!
- 음?



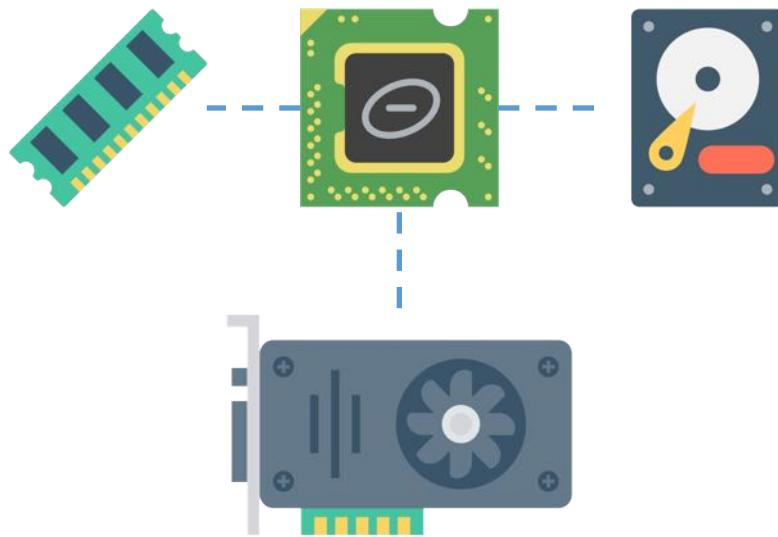
Total Allocated Resources	
CPU	0 / 14,960 Cores allocated
RAM	0 / 111.64 TiB
GPU	0.0 / 920.0 Fraction allocated

Total Utilization	
CPU	122% / 1,520,000% (15200 Cores)
RAM	1% / 100% (1.36 TiB / 111.87 TiB)
GPU	0% / 92,000% (920 Devices)

48 거대 언어 모델 훈련하기: “병목”

- GPU님이 모든 것을 다 해 주실 거야!
- 규모가 클 수록 문제가 되는 것들
 - 스토리지 네트워크
 - CPU 대역폭
 - GPU/NPU간 커뮤니케이션





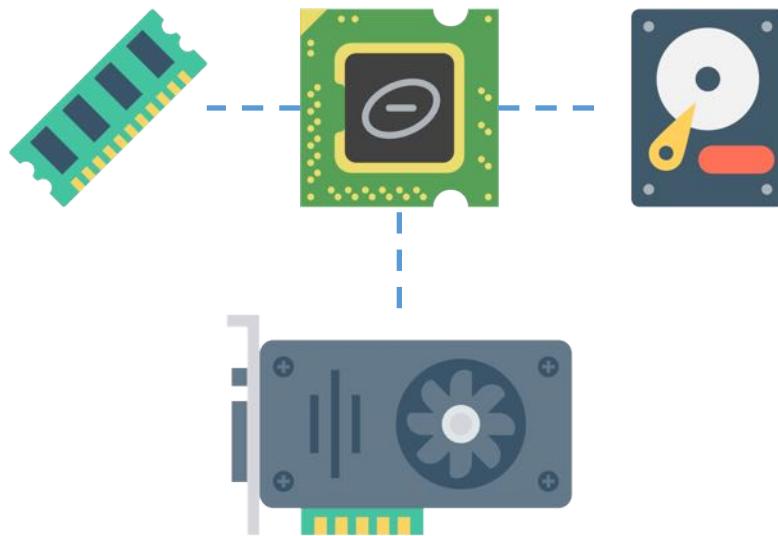
- 고전적인 AI 병목

- GPU-CPU (PCI-E)
- 데이터 파이프라인: GPU-CPU-Disk(-NAS)

- 워크로드+하드웨어의 변화

- 워크로드: 초기 대 AI 개발, AI 인퍼런스 / 서비스
- 하드웨어: GPU-GPU 네트워크, UMA, 매니코어 CPU, GPUDirect I/O...

거대 언어 모델: 거대화로 인한 **병목**의 이동 / CPU 대역폭



- 새로운 문제들

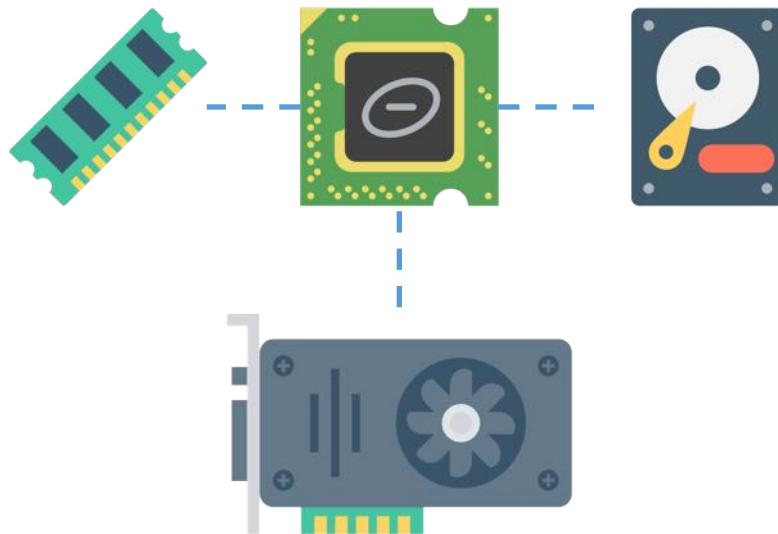
- CPU-RAM: CPU 코어당 대역폭 부족

	24코어	128코어
DDR5 8000	23.47	4.40
DDR4 3200	8.53	1.60GB/s/Core

- 임시 해결: RAM을 CPU에 직결

- ✓ Intel Xeon Max (2022년 11월, HBM 128G)
 - ✓ AMD MI300 (2023년 1월)
 - ✓ Nvidia Grace (2023년 예정)

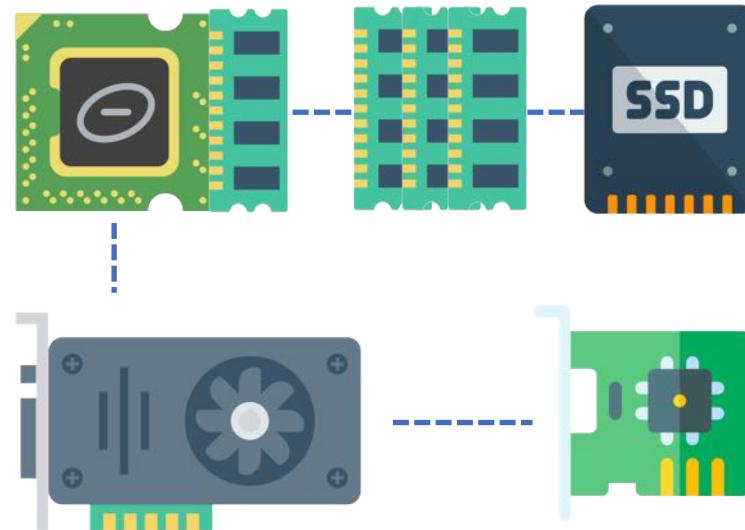
거대 언어 모델: 거대화로 인한 병목의 이동 / 스토리지 대역폭



• 새로운 문제들

- 네트워크: NAS 입출력의 속도 간섭
 - ✓ GPUDirect I/O 데이터: 80~120GiB/s
 - ✓ GPU-GPU 인터노드: 20~40GiB/s
 - ✓ AI 인퍼런스 플레인: 1~10GiB/s
 - ✓ 사용자 서비스 플레인: 0.1~1GiB/s
- 하이퍼스케일 I/O
 - ✓ 페타 바이트 단위의 데이터를
 - ✓ 100여대의 연산 노드에
 - ✓ 초당 100기가바이트 이상으로 전송해도
 - ✓ 노드당 초당 1기가밖에 데이터가 못 감
- (발표 광고: Sokovan 오피스트레이터!)
 - ✓ 세계 최초로 GPUDirect Storage / Magnum I/O를 컨테이너 클러스터에서 구현하고 디플로이
 - ✓ GTC 2023 및 OpenInfra Summit 2023에서 만나요~

거대 언어 모델: 거대화로 인한 병목의 이동 / 연산장치간 네트워크



• GPU간 네트워크

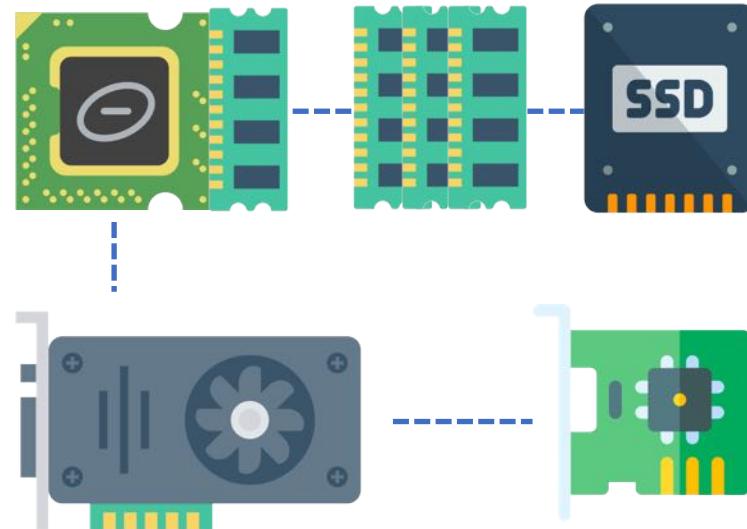
- 모델을 횡으로, 종으로 쪼개기
- 어떤 방향으로 쪼개도 GPU 한 장엔 안 담김
- 훈련 시
 - ✓ 16장: 프로토타이핑...
 - 노드 하나에 물리적으로 최대한 연결 가능한 GPU 수
 - ✓ 64~장: 진지한 언어 모델 훈련 시작

• 대책

- Nvidia: NCCL with Infiniband / GraceHopper
- Google: TPU (3-D toroidal 구조)
- Tesla: Dojo (3-D toroidal / 실제 여부 모름)

거대 언어 모델: 거대화로 인한 병목의 이동

- (근미래의) 해결책들



- 램 티어링: HBM2 + DDR5
- CXL 기반 통합 데이터 브릿지
 - ✓ CPU – HBM2 – DDR5 – NVMe – Net
- GPU / NPU 전용 네트워크 포트
 - ✓ Infiniband 통합 (Nvidia)
 - ✓ GroqLink (Groq)
- 다중 네트워크 포트 구성
 - ✓ PCI-E 레인 확보가 중요해지고 있음

AI 워크로드 가속 시도

- 훈련용 AI 가속기

- TPU (Google)
- Gaudi2 (Intel)
- IPU (GraphCore)
- Cerebras (Cerebras)
- SN30 (SambaNova)
- GroqChip (Groq)
- Loihi2/ Nahuku: Simulated NeuralNet, (Intel) (2018~)

- 인퍼런스용 AI 가속기 / NPU

- TPU / Coral (Google)
- BrainWave (Microsoft)
- Alveo (AMD, Xilinx)
- Warboy (Furiosa), ION/ATOM (Rebelion), Sapeon (Sapeon)



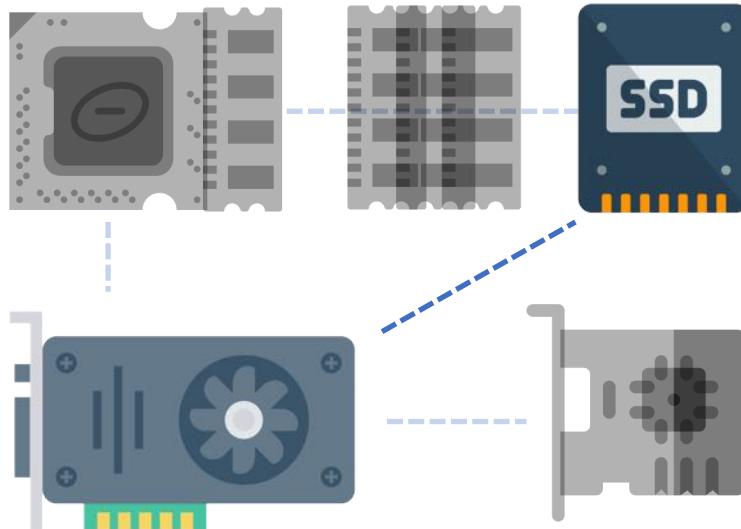
- 훈련용 AI 가속기

- 일반 ops 구현의 높은 난도
- 지원 모델을 늘려가는 방식의 접근으로 우회
 - ✓ 인퍼런스 칩 만드는 방법론으로 트레이닝 칩을 만들자!
 - ✓ 주요 AI 모델 대중화가 되었다는 판단

- 인퍼런스용 AI 가속기

- 저전력, 저지연, 저발열
- PCI-E, USB-C 및 GPIO 인터페이스
- FPGA 방식의 커스텀 IP 회사들
- FP16 / INT8 기반
- 애매한 부분들





- 램!

- 모델 사이즈 커지는데 어떻게 할 것인가!
- 램 티어링: HBM2 + DDR5 + ?
- 이걸로 되나?
- (근미래의 해결책): GPM (GPU with Persistent Memory) + Optane

- Precision + Ops!

- BF16 / TF32: (반쯤) 실패의 교훈
- INT8, INT4... 거대 언어 모델에서 의미가 있나?
- **FP16/FP8**: 최종 격전지
- Nvidia의 경쟁력: 여기서 못 이기면 못 이김

거대 언어 모델 플랫폼: AI 클러스터 규모 확대에 따른 기술적 난도 상승

완전히 성격이 다른
파이프라인 내 워크로드 특성

단계별 성능 병목 지점 차이

- 데이터 전처리: Data I/O
- 데이터 분석: CPU
- 딥러닝 훈련: GPU/ASIC
- 모델 서빙: CPU/GPU

딥러닝 연산 가속기의 급성장

딥러닝용 GPU 시장

- NVIDIA, AMD, Intel

딥러닝 전용 ASIC 시장

- Google (TPU / Coral)
- GraphCore (IPU)
- Habana (Gaudi)
- Sambanova (SN)
- Cerabras
- Cambricon
- Groq

대규모 연산으로 인한
높은 모델 개발 비용

단위시간당 자원 증가^[1]

- CPU: \$0.006
- GPU (A100 기준): \$0.93

모델 훈련 비용의 급격한 증가

- PaLM (2022): 300억원~
- GPT-3 (2020): 45억원~
- T5 (2019): 30억원~
- BERT (2018): 1억원~

거대 언어 모델 플랫폼: 대규모 GPU 클러스터 기술

컨테이너, VM 등의 다양한 가상 환경 등장

컨테이너 기반 가상 환경

- Docker, Podman
- Containerd, LXC
- Kubernetes
- Backend.AI / Sokovan

VM 기반 가상환경

- KVM, QEMU, VMWare
- Firecracker

작업 그래프 기반
파이프라인·워크플로우
저작 도구 등장

오픈소스 파이프라인 도구

- Apache AirFlow
- MLFlow
- KubeFlow
- FastTrack

클라우드 파이프라인 서비스

- Amazon SageMaker MLOps
- Google Cloud Composer
- Azure MLStudio

AI 훈련 및 응용 클러스터의 다양화

하이브리드 클라우드 솔루션

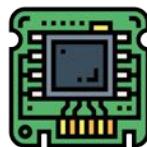
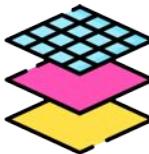
- 클라우드 벤더
- 온프레미스 벤더
- 오픈소스

2023년의 전망 및 도전 과제



- 2023년 이후의 (래블업 입장에서 보는) 도전 과제들

- 증가하는 딥 러닝 소프트웨어 레거시
 - ✓ 엔터프라이즈 AI/ML과 신기술 등장의 속도 차
- Arm / RISC-V 등 멀티 아키텍처 플랫폼의 등장
 - ✓ 인간계의 한계: 발열과 전력 소모 문제
 - ✓ 전력 대 성능비의 중요성 – 더이상 데이터센터에 집적이 힘들다
- AI 훈련 / 서비스 가속 ASIC의 춘추전국시대 도래
 - ✓ CPU 대비 상대적으로 간단한 구조
 - ✓ 특화 기능 대응의 용이함을 살릴 수 있을 것인지?
- 거대 언어 모델 및 생성 모델 서비스의 증가
 - ✓ 고비용 AI 모델의 상용화 급증
- Edge Intelligence / distributed intelligence
 - ✓ 모바일 단에서의 모델 서빙 – 일반 IoT 기기들로 확장
 - ✓ 딥 러닝 엣지 소프트웨어 플랫폼들의 보급



- 거대 언어 모델

- 2017년의 깨달음
- “규모”에 대하여

- 거대 언어 “모델링”

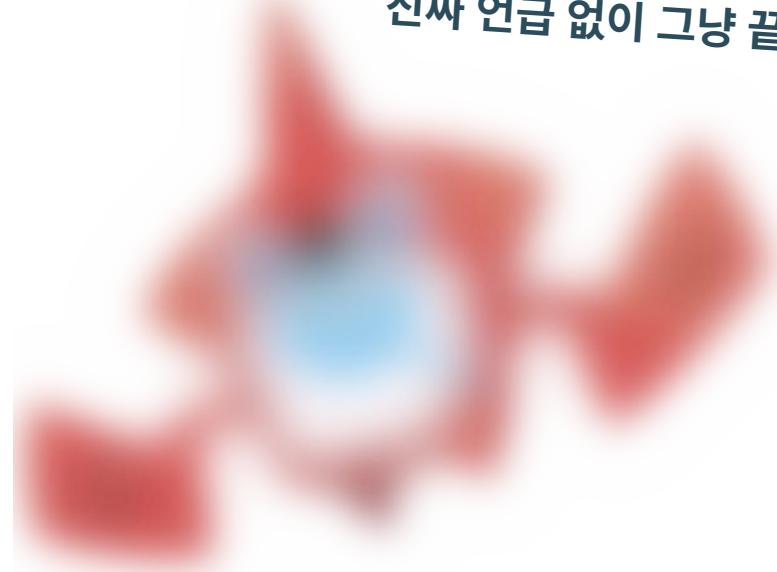
- 언어 모델과 자원
- 적정 모델, 거대 모델
- 대 AI 시대

- 거대 언어 모델의 규모에 대응하기

- 병목과 대응 기술
- AI 워크로드 가속 시장
- 소프트웨어 복잡도
- 향후 전망 및 문제

나는 어떻게 된 건가요

진짜 언급 없이 그냥 끝나나요



투명 뒷판으로 스팀덱 케이스 같아도 했는데...

- 2022년 10월 방콕, ML Community Summit (Google)
 - JAX, MLIR 등...
- 식사 시간
 - Huggingface의 현재 과제: 모델을 필터링할 권한
 - 딥 러닝 모델은 자체 편견을 만들지 않을 수 있는가?
 - 더 빠르게, 더 공격적으로 만들지 않을까?
 - 편견 없는 데이터로 편견 간단히 유도하기
 - 모델끼리 이런 식으로...
 - ✓ 근본 없는 XX칩으로 훈련한 주제에!
 - ✓ 너 몇년산 모델이야? 나 2035년산이야!





끝!



inureyes@gmail.com



<https://www.facebook.com/jeongkyu.shin>



contact@lablup.com



<https://www.facebook.com/labluplnc>

Lablup Inc.

<https://www.lablup.com>

Backend.AI

<https://www.backend.ai>

Backend.AI GitHub

<https://github.com/lablup/backend.ai>

Backend.AI Cloud

<https://cloud.backend.ai>