

자연어처리 개요 및 대용량 스케일에서의 **LLM**

고석현 - Sionic AI

오늘 발표 자료를 도와주신... 송영숙 박사님 죄송합니다..!



소속공제 길벗-모두의 시리즈

모두의 한국어 텍스트 분석 with 파이썬 기초부터 챗GPT까지 처리

박조은, 송영숙 지 | 길벗 | 2023년 05월 29일

★★★★★ 9.6 회원리뷰(10건) | 판매지수 3,186 베스트 OS/데이터베이스 7:

정가 27,000원

판매가 24,300원 (10% 할인)

YES포인트 1,350원 (5% 적립)

5만원이상 구매 시 2천원 추가적립

결제혜택 카드/간편결제 혜택을 확인하세요

배송안내 서울특별시 영등포구 은행로 11(여의도동, 일신빌딩) 지역변경

일반배송 2일 이내(9/26, 화) 출고예정

배송비 1,000원

저는



- Data Lab

컨텐츠, 이미지, 검색, 추천



- Co-Founder

자연어 대화엔진

Naver에 M&A로 exit

Naver



- AI Software Platform

Hyper Clova - LLM

Naver Cloud - LM Platform SaaS



- ML Platform



- Enterprise LLM

저는

챗봇 1만 개의 모델 서빙하기: AI 서비스 어디까지 해봤니

Hyper clova 거대 모델 기반의 ML 서빙과 성능 최적화 (GPT-3)

케어콜 - 코로나로부터 우리를 지키는 믿음직한 AiCall

물 흐르듯 자연스러운 ML 서비스 만들기 - MLOps 를 넘어 서비스로

딥러닝을 딥하지 않게 사용해서 자연어 검색 만들기

오늘은 몇 가지 간단한 질문들로

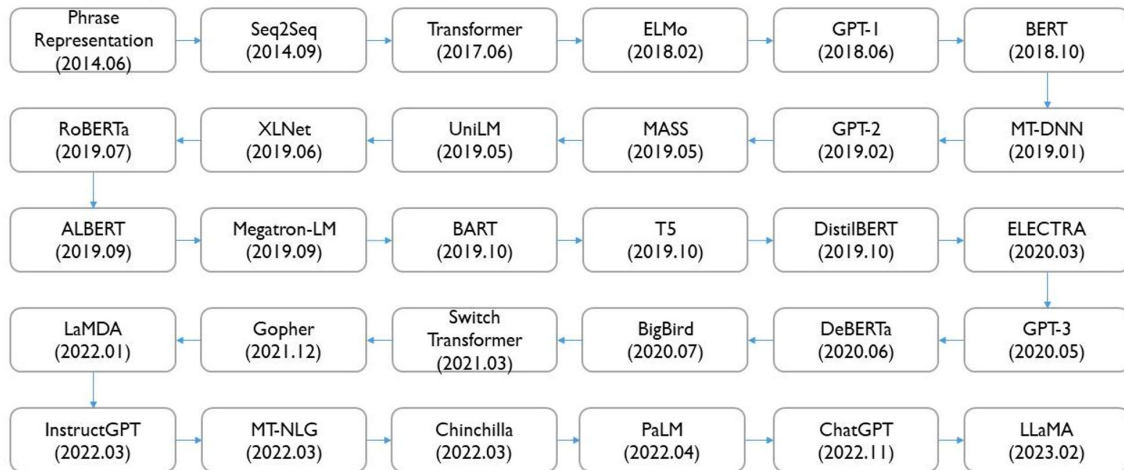
아주 얇고 넓게 LLM을 살펴 보겠습니다.

History of (Large) Language Models

고려대학교 산업경영공학과 DSBA 연구실

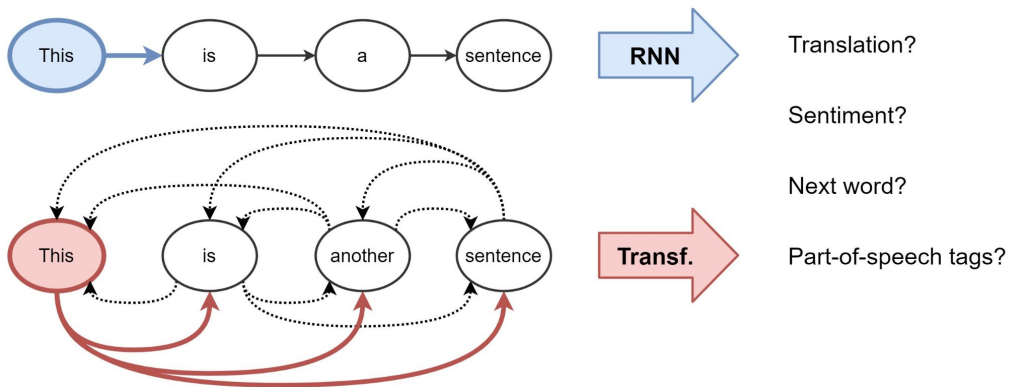
홈페이지: <http://dsba.korea.ac.kr>

유튜브 채널: <https://www.youtube.com/channel/UCFq01cgCcEwhXI7BvcwIQyg>



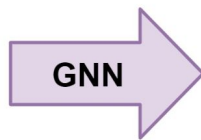
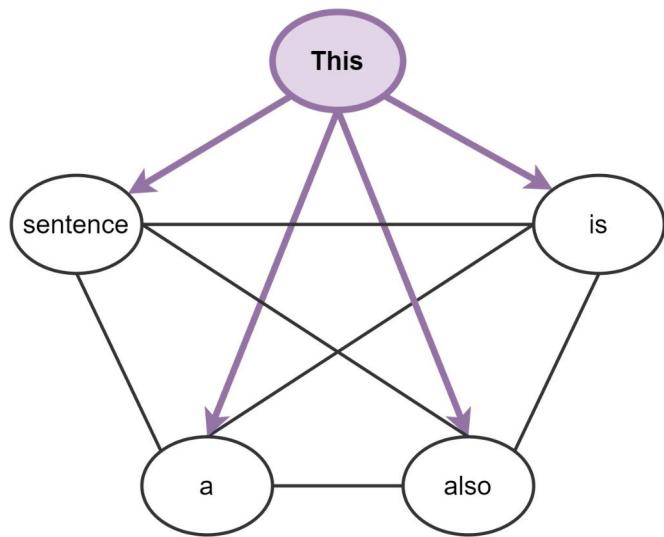
트랜스포머는 무엇일까요

Transformers are a special case of Graph Neural Networks



문장은 완전히 연결된 단어 그래프입니다.

Transformers are a special case of Graph Neural Networks



Translation?

Sentiment?

Next word?

Part-of-speech tags?

딥러닝은 무엇일까요

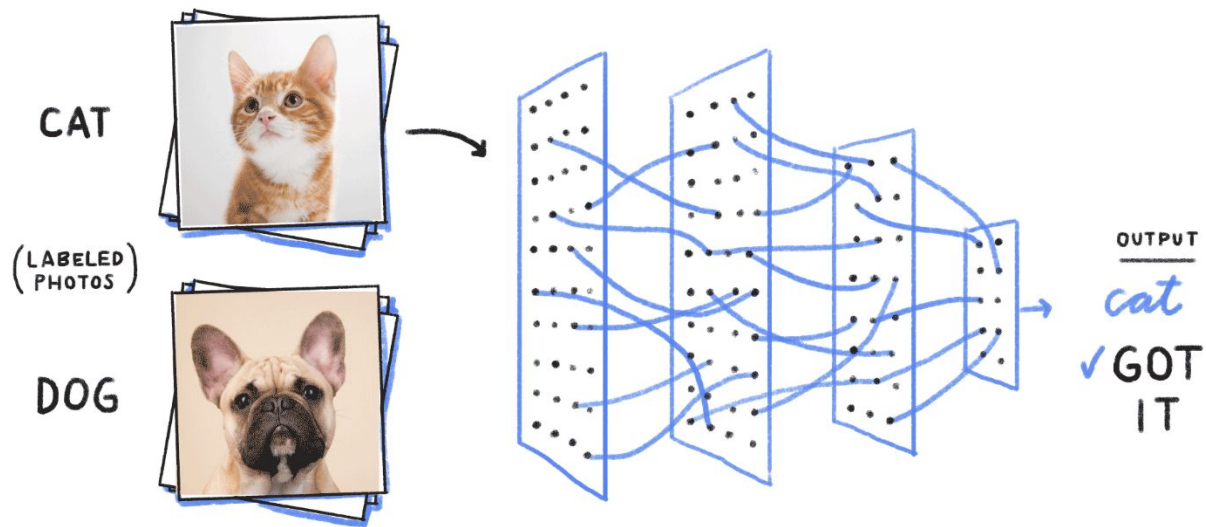
복수의 층으로 구성된 **NN**를 사용하는 모델과 해당하는 방식들

딥러닝은 무엇일까요

비선형 데이터의 집합에 선형성을 부여해주는 **projection matrix**의 학습 방식

딥러닝은 무엇일까요

feature의 정의를 스스로 하는 모델



딥러닝은 무엇일까요

지금은 Open AI로 이동한 전 Tesla Head of AI 의 Andrej Karpathy 표현을 빌려보면

- 데이터로 프로그램의 코드를 일부분 대체 하는 개발 수단



Andrej Karpathy

Nov 12, 2017 · 9 min read · [Listen](#)



Software 2.0

I sometimes see people refer to neural networks as just “another tool in your machine learning toolbox”. They have some pros and cons, they work here or there, and sometimes you can use them to win Kaggle competitions. Unfortunately, this interpretation completely misses the forest for the trees. Neural networks are not just another classifier, they represent the beginning of a fundamental shift in how we develop software. They are Software 2.0.

딥러닝은 무엇일까요



WIKIPEDIA
The Free Encyclopedia

Search Wikipedia



Please join us for the #WikiForHumanRights writing challenge on “Right to a healthy environment”, or join related events hosted by the international Wikimedia community

Main menu [\[hide\]](#)

[Main page](#)

[Contents](#)

[Current events](#)

[Random article](#)

[About Wikipedia](#)

[Contact us](#)

[Donate](#)

[Switch to old look](#)

[Contribute](#)

[Help](#)

[Learn to edit](#)

[Community portal](#)

Universal approximation theorem

8 languages

[Article](#) [Talk](#)

[Read](#) [Edit source](#) [View history](#)

From Wikipedia, the free encyclopedia

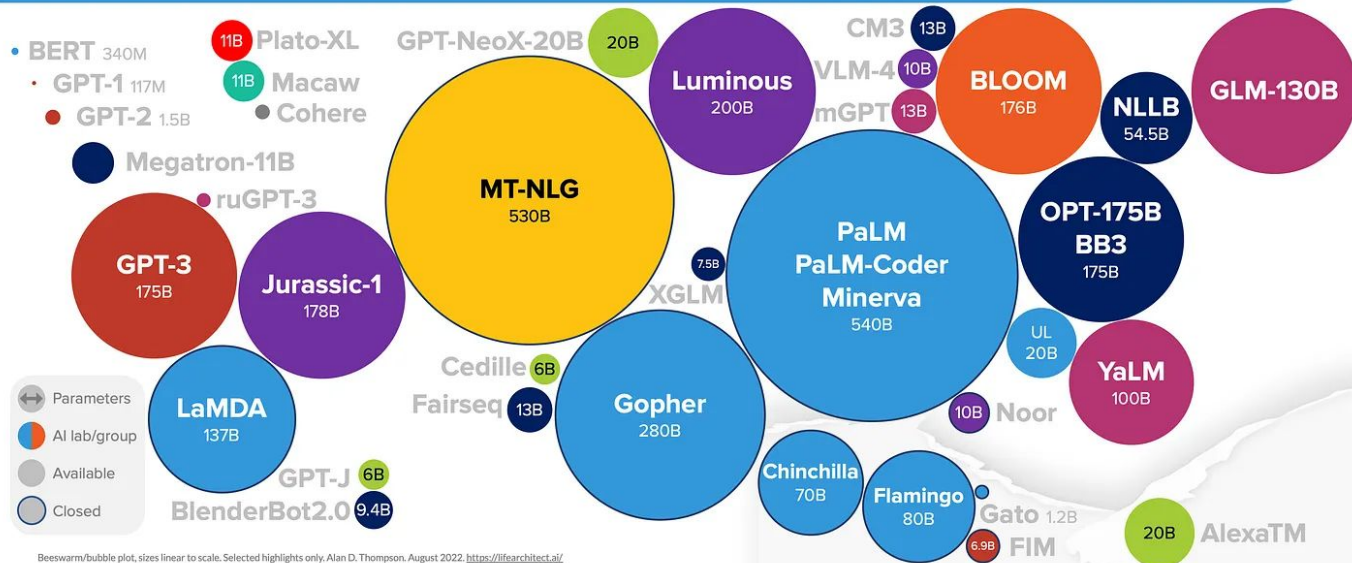
In the [mathematical](#) theory of [artificial neural networks](#), **universal approximation theorems** are results^{[1][2]} that establish the [density](#) of an algorithmically generated class of functions within a given function space of interest. Typically, these results concern the approximation capabilities of the [feedforward architecture](#) on the space of continuous functions between two [Euclidean spaces](#), and the approximation is with respect to the [compact convergence](#) topology.

However, there are also a variety of results between [non-Euclidean spaces](#)^[3] and other commonly used architectures and, more generally, algorithmically generated sets of functions, such as the [convolutional neural network](#) (CNN) architecture,^{[4][5]} radial basis-functions,^[6] or neural networks with specific properties.^{[7][8]} Most universal approximation theorems can be parsed into two classes. The first quantifies the approximation capabilities of neural networks with an arbitrary number of artificial neurons (“*arbitrary width*” case) and the second focuses on the case with an arbitrary number of hidden layers, each containing a limited number of artificial neurons (“*arbitrary depth*” case). In addition to these two classes, there are also universal approximation theorems for neural networks with bounded number of hidden layers and a limited number of neurons in each layer (“*bounded depth and bounded width*” case).

Universal approximation theorems imply that neural networks can *represent* a wide variety of interesting functions when given appropriate weights. On the other hand, they typically do not provide a construction for the weights, but merely state that such a construction is possible.

그럼 LLM 은 무엇일까요

LANGUAGE MODEL SIZES TO AUG/2022



그럼 LLM 은 무엇일까요

Emergence를 빼놓고 이야기 할 수 없습니다.

창발

한글 35개 언어 ▾

문서 토론

읽기 편집 역사 보기 도구 ▾

위키백과, 우리 모두의 백과사전.

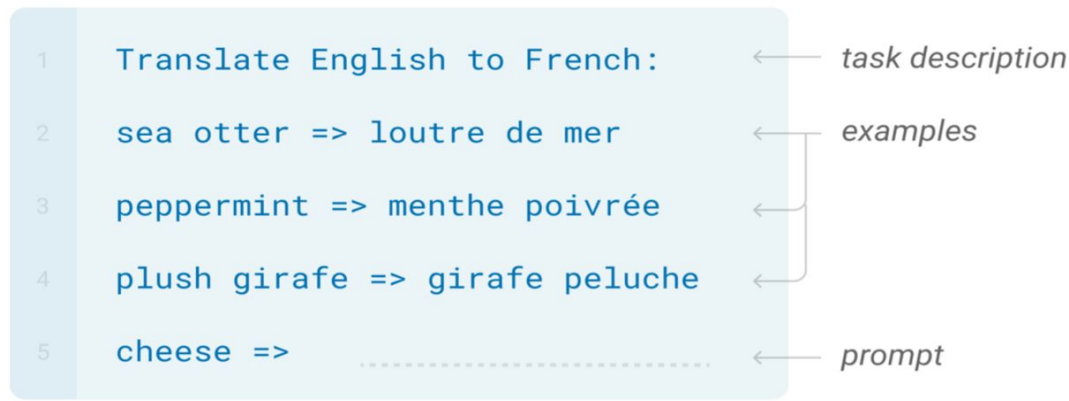
창발(創發)또는 **떠오름 현상**은 하위 계층(구성 요소)에는 없는 특성이나 행동이 상위 계층(전체 구조)에서 자발적으로 돌연히 출현하는 현상이다. 또한 불시에 솟아나는 특성을 창발성(영어: emergent property) 또는 이머전스(영어: emergence)라고도 부른다. 자기조직화 현상, 복잡계 과학과 관련이 깊다.

그럼 LLM 은 무엇일까요

아주 작은 예시로 전혀 새로운 **task**를 할 수 있습니다.

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



The diagram illustrates a prompt structure for a Few-shot LLM task. It consists of five lines of text, each preceded by a number in a light blue box. To the right of the text, arrows point from labels to specific parts of the prompt. The first line, '1 Translate English to French:', is labeled 'task description'. The next three lines, '2 sea otter => loutre de mer', '3 peppermint => menthe poivrée', and '4 plush girafe => girafe peluche', are grouped by a bracket and labeled 'examples'. The final line, '5 cheese =>,', is labeled 'prompt'.

```
1 Translate English to French:
2 sea otter => loutre de mer
3 peppermint => menthe poivrée
4 plush girafe => girafe peluche
5 cheese => .....,
```

task description

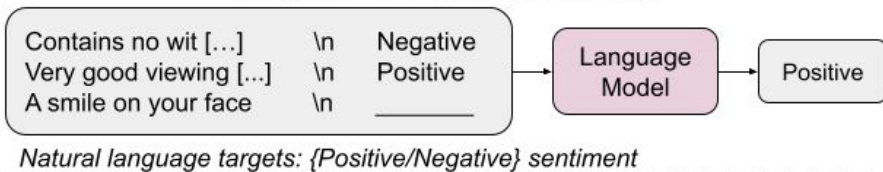
examples

prompt

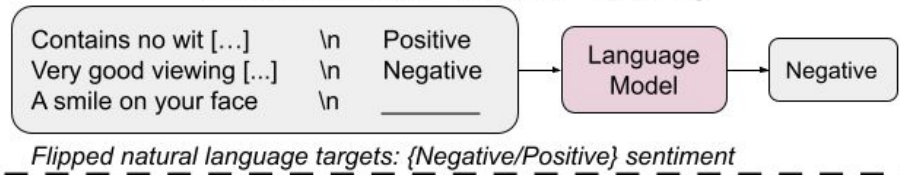
그럼 LLM 은 무엇일까요

ICL 을 통해 사전 학습된 지식과 상반되는 동작도 할 수 있습니다.

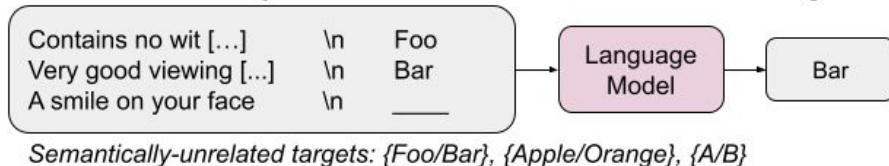
Regular In-Context Learning



Flipped-Label In-Context Learning

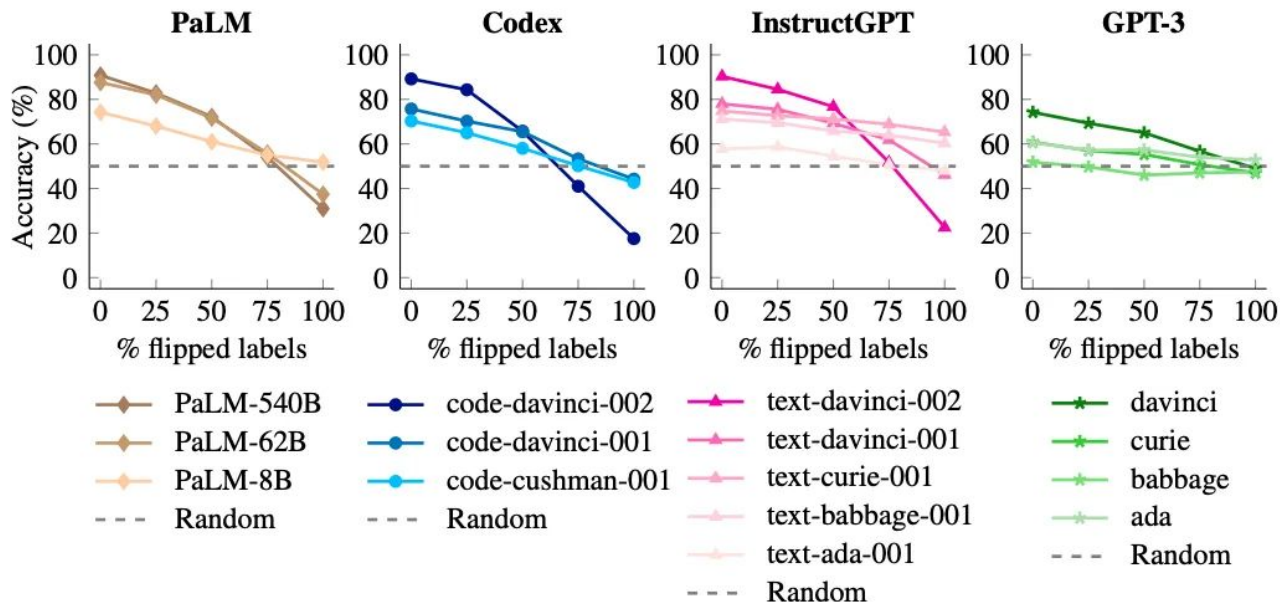


Semantically-Unrelated Label In-Context Learning



그럼 LLM 은 무엇일까요

그리고 이런 추론 능력은 모델의 규모에 따라 증가합니다.





Chain of Thought

GPT-4 Is a *R*easoning Engine

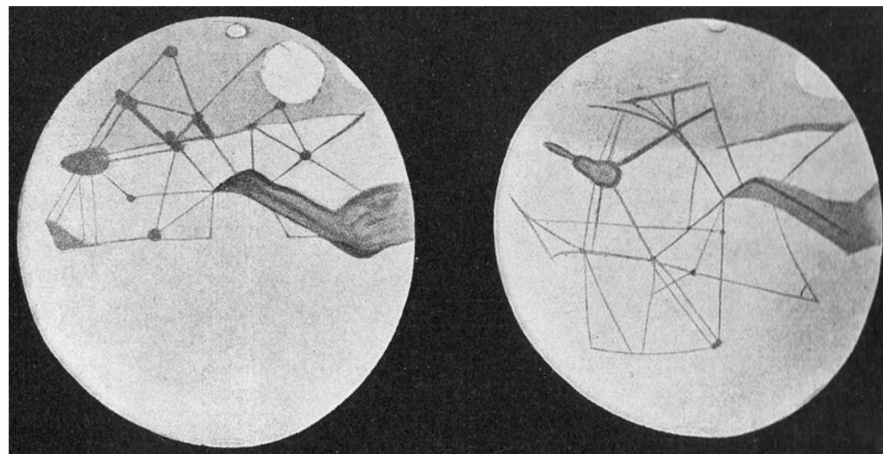
Reason is only as good as the information we give it

BY DAN SHIPPER

MARCH 31, 2023

♥ 291

[Listen](#)



Drawings of canals on Mars by Percival Lowell

GPT-4는 추론 엔진(Reasoning Engine)입니다.

Share



김육영

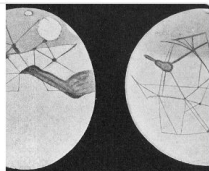
ADMIN

5 days ago

GPT-4 Is a Reasoning Engine

Reason is only as good as the information we give it

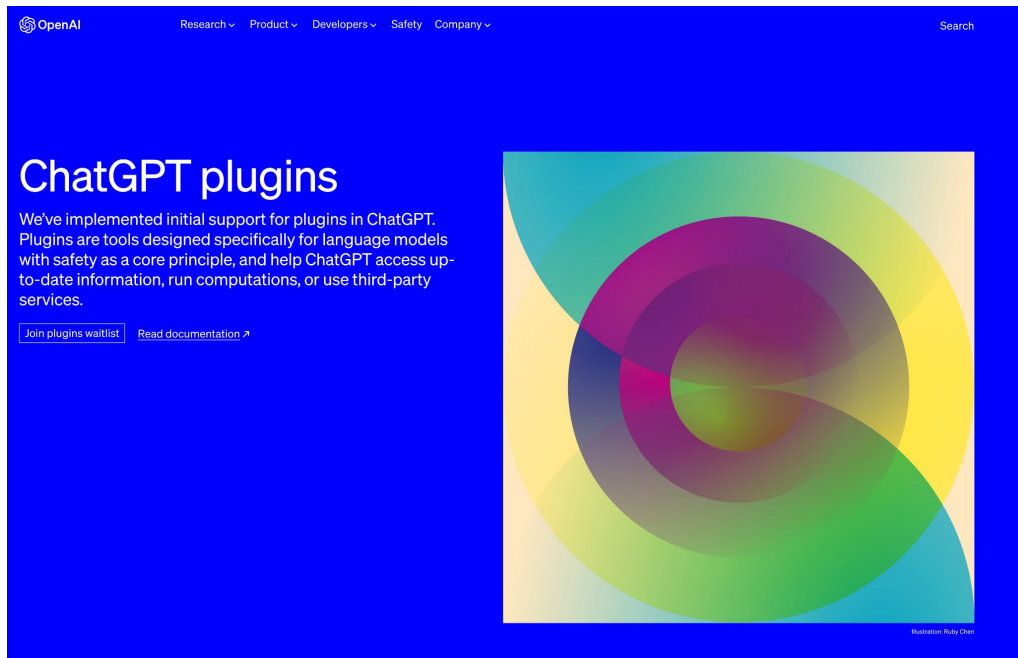
<https://every.to/chain-of-thought/gpt-4-is-a-reasoning-engine>



- GPT 모델은 데이터 베이스가 아니라, 추론 엔진이라는 것이 중요합니다.
- AI 유용성 향상이 추론 능력의 향상 뿐 아니라, 정확한 타이밍(적시)에 유용한 지식을 액세스 하는 능력 향상에서 비롯될 것이라고 (글쓰기가) 예측하기에 중요합니다.
- GPT의 웹 검색 베이스 기반 (GPT Model : Browing)은 웹 조사를 수행하여 자체 지식 기반을 만듭니다. 그런 다음에 수집된 정보를 분석하여 보다 정확한 답변을 추출하는 것입니다.
- GPT-4가 인터넷, 혹은 데이터베이스에 연결되면 모호한 통계적 추론에 의존할 필요가 없습니다. 대신하여 명확한 답변을 생성할 수 있습니다.
- 이런 지식에서 적시에 적합한 지식을 찾는 것, AI에서는 일반적으로 벡터 데이터베이스로 수행 됩니다.
- 실제로 AI를 구축하는 회사의 성공을 전체적으로 나타내는 투자를 하고 싶다면 벡터 데이터베이스 제공업체 또는 이들의 바스켓에 투자하는 것이 현명한 조치입니다(파인콘 etc.)
- 개인 지식 저장소는 매우 가치가 있을 겁니다.
- 우리는 입력의 중요성을 과소 평가하는 경향이 있는 듯 합니다. AI로 정확한 결과를 얻기 위해서는 입력한 정보가 무엇인지가 중요합니다. 그 답은 우리가 분석을 위하여 제공하는 정보에 의해 크게 좌우됩니다.

개인과 조직의 지식은 점점 더 중요해집니다.

이제 우리는 동안 쌓기만 했던 많은 지식을 자동으로 사용할 방법을 알게 됩니다.



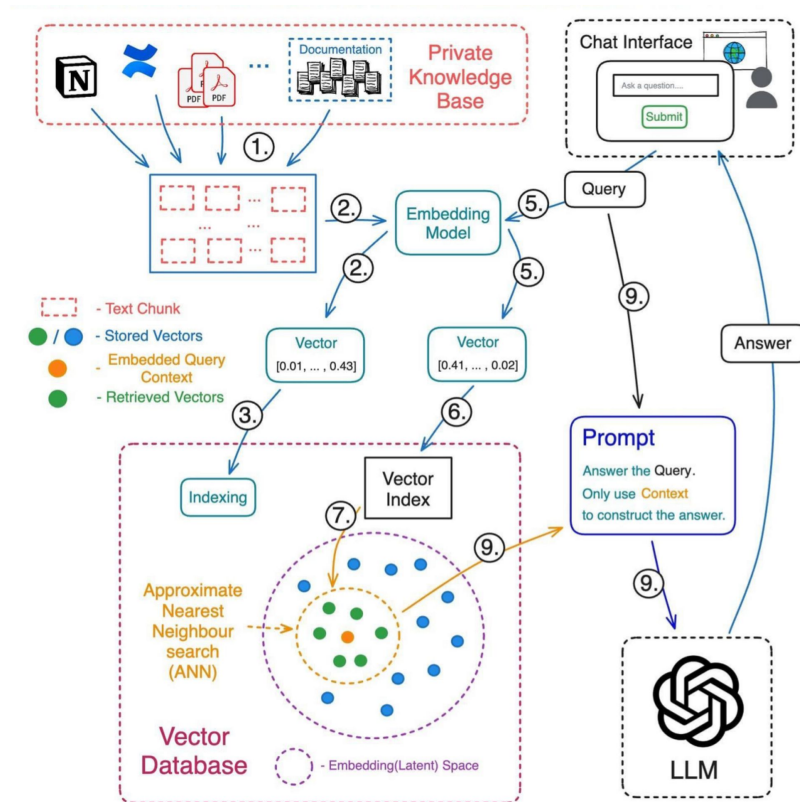
결국 생성 AI의 방향은 초 개인화된 검색 추천과 함께

정보의 품질

정보의 대한 평판과 가시성 확보

생성 모델의 **context** 제어

LLM의 장기 기억을 담당하는 RAG (검색 증강 생성)



하지만 **Vector** 관련 기술적으로 초기 단계

알려진 일반적인 모델링 방법이 없으며

차원의 저주 (**Curse of dimensionality**) 로 인한 직접적인 **ANN** 근사 성능 저하가 더욱 심해진 현상을 해결하지 못하였고

충분히 적은 데이터를 고품질의 **vector search**로 접근 하는 방향의 중요성이 알려지지 않음

vector search를 이루는 기반 기술중 하나를 보면

ANN Benchmarks Home Datasets Algorithms Contact

Info

ANN-Benchmarks is a benchmarking environment for approximate nearest neighbor algorithms search. This website contains the current benchmarking results. Please visit <http://github.com/erikbern/ann-benchmarks/> to get an overview over evaluated data sets and algorithms. Make a pull request on [Github](#) to add your own code or improvements to the benchmarking system.

Benchmarking Results

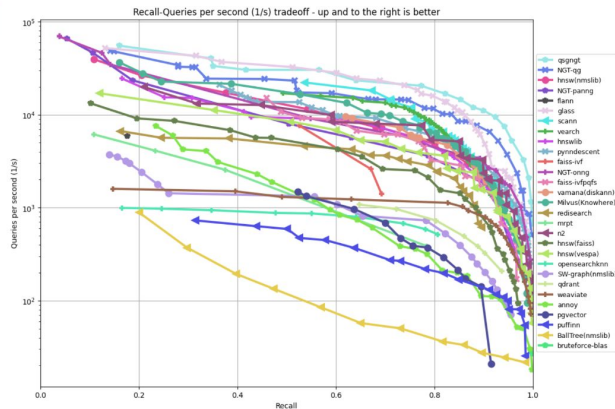
Results are split by distance measure and dataset. In the bottom, you can find an overview of an algorithm's performance on all datasets. Each dataset is annotated by ($k = \dots$), the number of nearest neighbors an algorithm was supposed to return. The plot shown depicts *Recall* (the fraction of true nearest neighbors found, on average over all queries) against *Queries per second*. Clicking on a plot reveals detailed interactive plots, including approximate recall, index size, and build time.

Benchmarks for Single Queries

Results by Dataset

Distance: Angular

nytimes-256-angular ($k = 10$)



몇 가지 시사점

검색은 무엇일까요

(IR) Information Retrieval ?

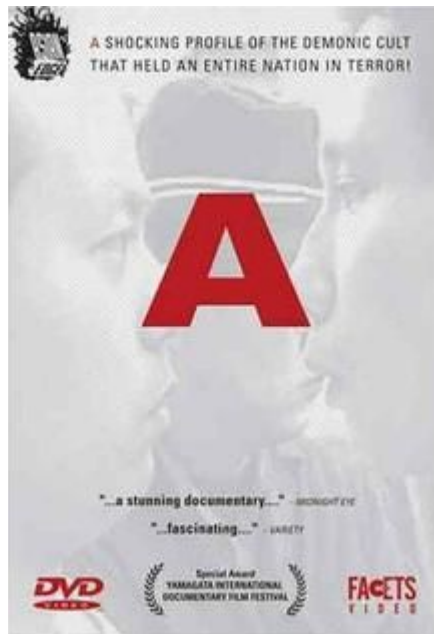
(내 의도와) 적합한 정보를 골라내는것

KR

a 라는 영화 알아?



제가 알아듣기 어려운 질문 같습니다. "a"라는 제목의 영화가 무엇인지 더 자세한 정보를 제공해 주시면 대답해 드릴 수 있을 것 같습니다.



검색은 무엇일까요

정보가 (지나치게 많아서) 오염되거나 유저에게 적합한 결과가 아니라면 좋은 검색이 아닌 것

KR

You know the movie named A?



I'm sorry, but I would need more information to provide an accurate answer. There are many movies with titles that start with the letter A, and without additional context or details, I'm not sure which movie you are referring to. Could you please provide more information or clarify your question?



KR

movie name is "A"



I'm still unsure which movie you are referring to, as there are many movies with a one-word title "A." Could you provide any additional information about the movie such as the director, main actors, or a brief plot summary? This would help me to better understand which movie you are asking about.



검색 잘 하는것은 충분히 어려우며 동시에 방대한 분야

항상 유저들은 업계 최고 수준의 엔진과 포탈의 검색 결과를 익숙하게 사용

즉 제품에 대한 검색에 대한 기대치는 매우 높음

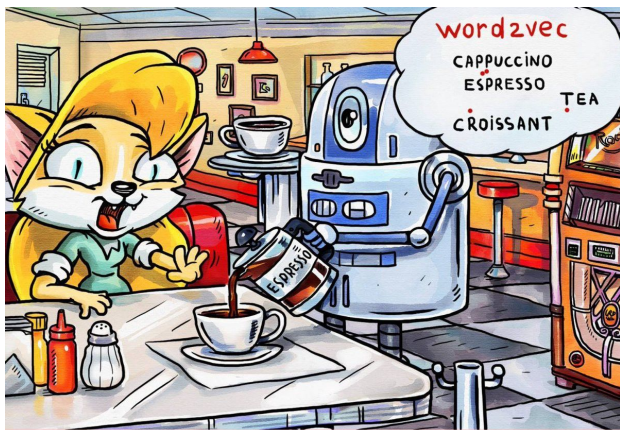
대규모의 인프라와 검색 팀을 가지고 있지 않다면 좋은 검색 만들고 운영하는건
어려운 과제

하지만 LM과 LLM은 소규모 검색을 쉽고 빠르고 만들수 있는 강력한 도구

Architecture

Sparse Embedding 에서 기대하는 Lexical 성능의 저조

DPR의 구현과 같은 bm25 결과를 일부 활용하거나 적절한 거리 함수 기반 보정이 필요



- Espresso? But I ordered a cappuccino!
- Don't worry, the cosine distance between them is so small that they are almost the same thing.

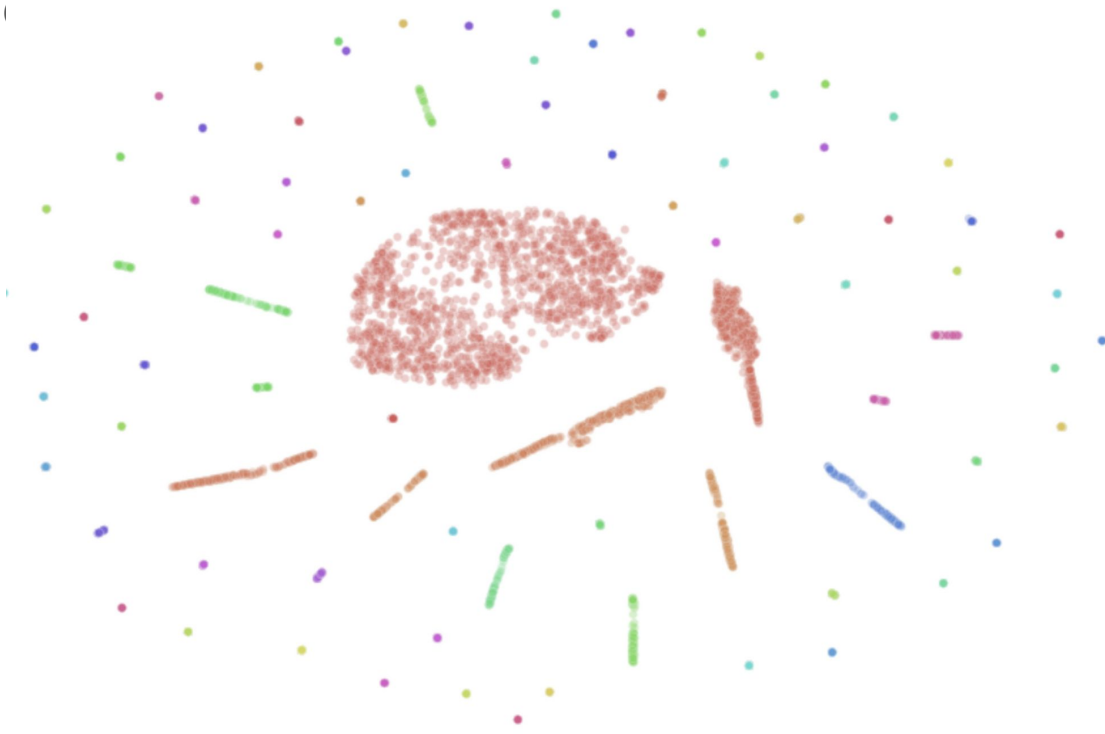
Architecture

다음과 같이 시각화된 embedding이 있을 때



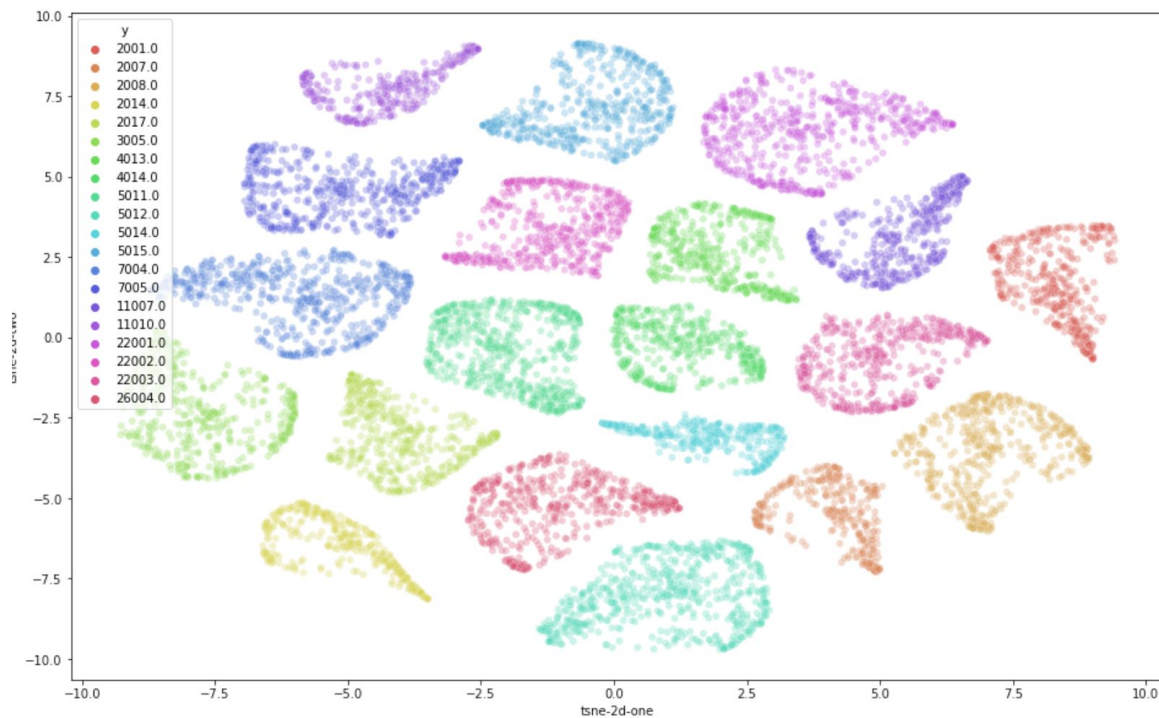
Architecture

metric learning 이 적용되지 않은 일반적인 classification 의 모델 embedding 의
대한 결과 manifold



Architecture

metric learning 이 적용된 manifold



Architecture - Engineering 문제점

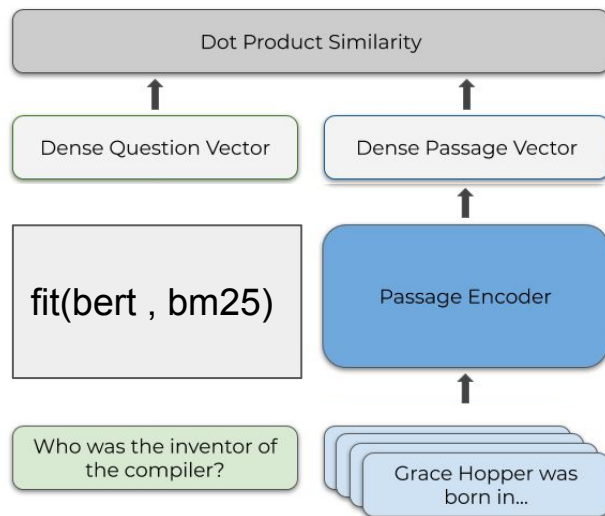
- BERT 110M 수준의 모델은 연산 비용이 높음 (하지만 GPT와 비교할수 없게 가볍다.)
 - GPU 를 사용한 단순한 구조의 real time inference로 약 50~100 QPS (query per second) 성능
- ANN등 벡터 유사도 연산에 다수의 서빙과 인프라가 필요
 - 주기적으로 indexing 처리와 별도 서빙 서버등 필요
 - 관련된 인프라가 많고 복잡할 경우 MSA의 저주에 걸릴 수 있다.
- ANN의 활용 효율이 많이 떨어짐
 - ANN은 효율적인 벡터의 숫자가 존재 약 적어도 10만 이상의 타겟 벡터에서 의미가 있음
 - 효과적으로 근사되는 차원의 범위가 존재 예를 들면 20~ 256 정도가 일반적
 - 768 ~ 2048 혹은 30000 차원 수준의 dense vector 검색에서는 recall 성능 감소가 아쉬움

Architecture - Model

Query -> Query' 를 찾는 문제는 충분히 robust 하며 적은 데이터로 동작 가능

하지만 결국은 Query -> Documents 를 찾는 QA 문제를 잘 풀어야 Contents 검색이 가능

Query -> Documents'



Architecture - Encoding

“query: how much protein should a female eat”

->

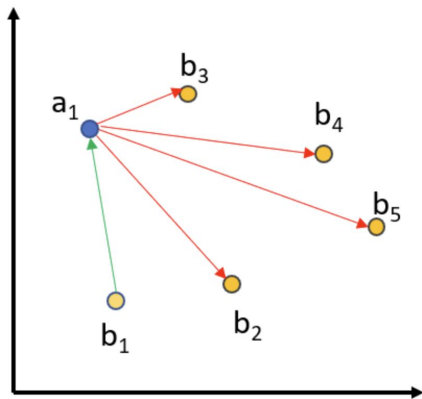
"passage: As a general guideline, the CDC's average requirement of protein for women ages 19 to 70 is 46 grams per day. But, as you can see from this chart, you'll need to increase that if you're expecting or training for a marathon. Check out the chart below to see how much protein you should be eating each day.

Architecture - Model

LM을 MultipleNegativesRankingLoss 를 통해 bm25 결과로 초기 세팅하고

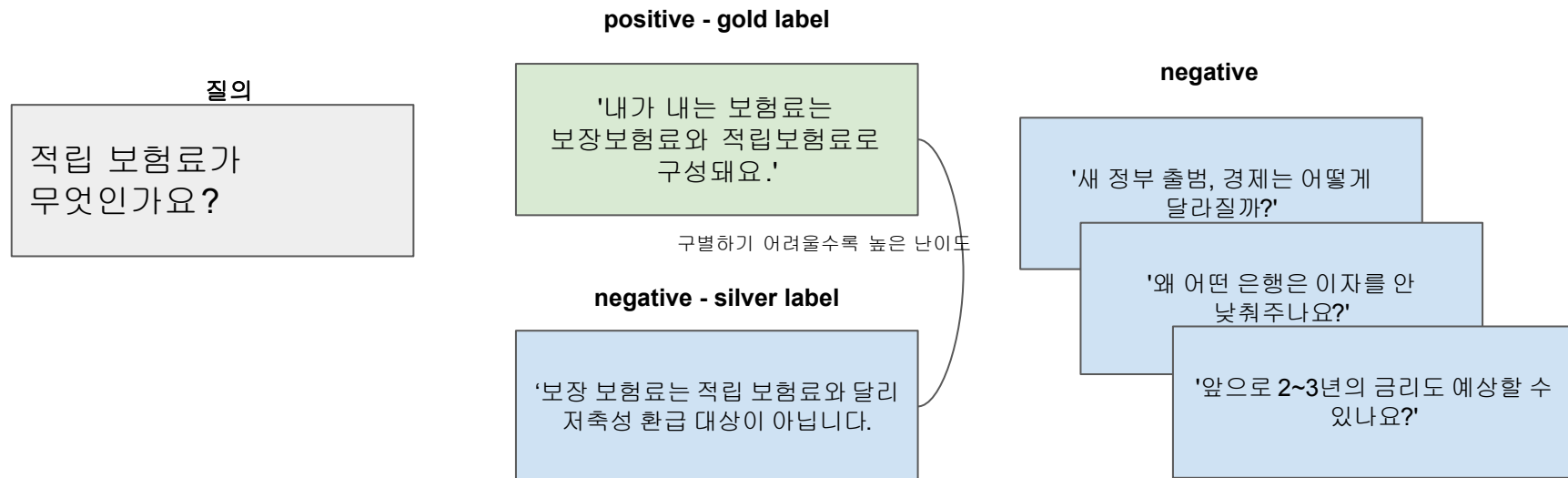
사내 테스트와 AB 테스트로 지속적으로 implicit feedback 을 수집하는 추천 태스크 간주

Efficient Natural Language Response Suggestion for Smart Reply <https://arxiv.org/abs/1705.00652> pdf



Architecture - Model

추가로 LLM을 통한 데이터 augmentation, BM25를 통한 negative sample 생성
gold - silver 의 차이를 최소화하는 hard negative를 얼마나 잘 세팅하는지가 중요



Architecture - Model

arXiv > cs > arXiv:2308.14963

Search...

Help | Advance

Computer Science > Information Retrieval

[Submitted on 29 Aug 2023]

Vector Search with OpenAI Embeddings: Lucene Is All You Need

Jimmy Lin, Ronak Pradeep, Tommaso Teofili, Jasper Xian

We provide a reproducible, end-to-end demonstration of vector search with OpenAI embeddings using Lucene on the popular MS MARCO passage ranking test collection. The main goal of our work is to challenge the prevailing narrative that a dedicated vector store is necessary to take advantage of recent advances in deep neural networks as applied to search. Quite the contrary, we show that hierarchical navigable small-world network (HNSW) indexes in Lucene are adequate to provide vector search capabilities in a standard bi-encoder architecture. This suggests that, from a simple cost-benefit analysis, there does not appear to be a compelling reason to introduce a dedicated vector store into a modern "AI stack" for search, since such applications have already received substantial investments in existing, widely deployed infrastructure.

Subjects: **Information Retrieval (cs.IR)**

Cite as: [arXiv:2308.14963](https://arxiv.org/abs/2308.14963) [cs.IR]

(or [arXiv:2308.14963v1](https://arxiv.org/abs/2308.14963v1) [cs.IR] for this version)

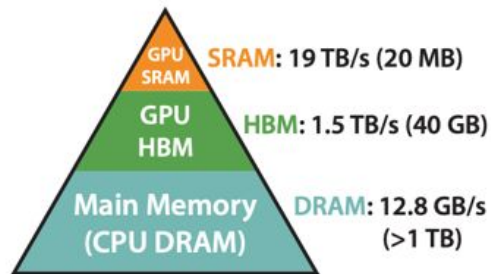
<https://doi.org/10.48550/arXiv.2308.14963> 

Submission history

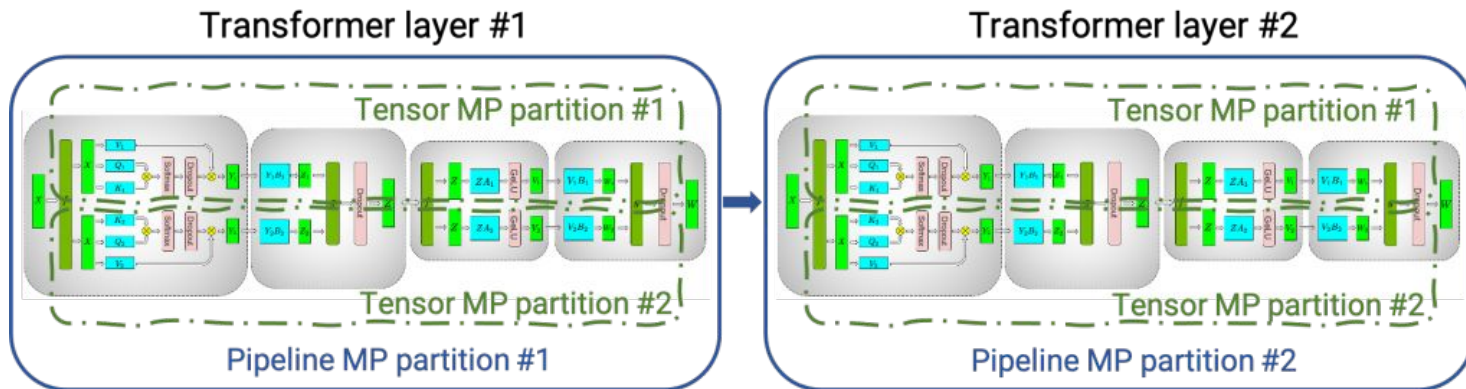
From: Jimmy Lin [[view email](#)]

[v1] Tue, 29 Aug 2023 01:30:23 UTC (123 KB)

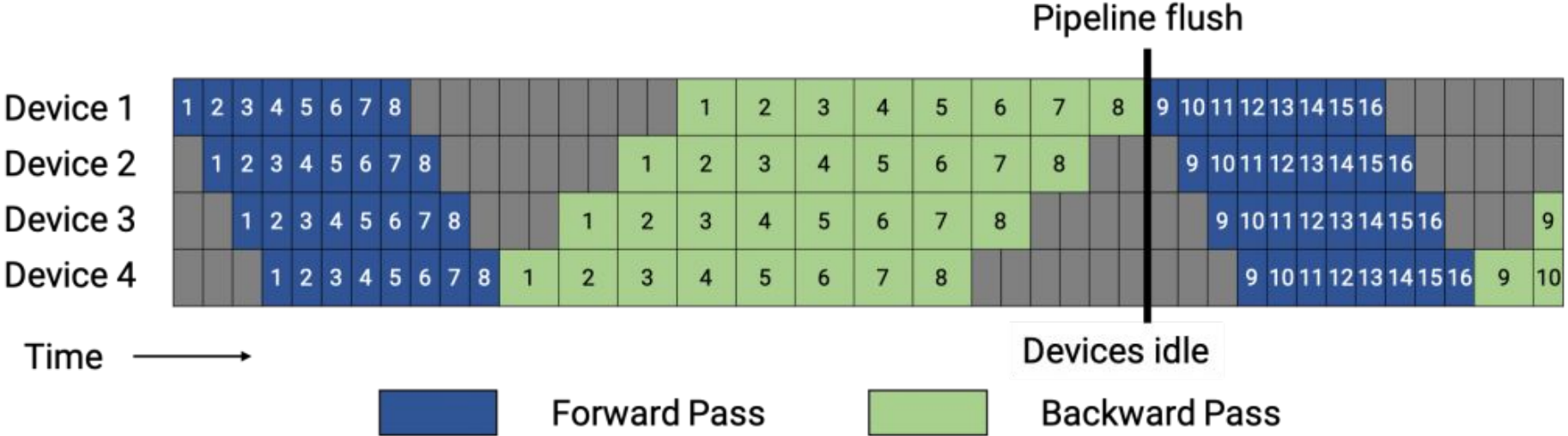
이제 대용량 LLM



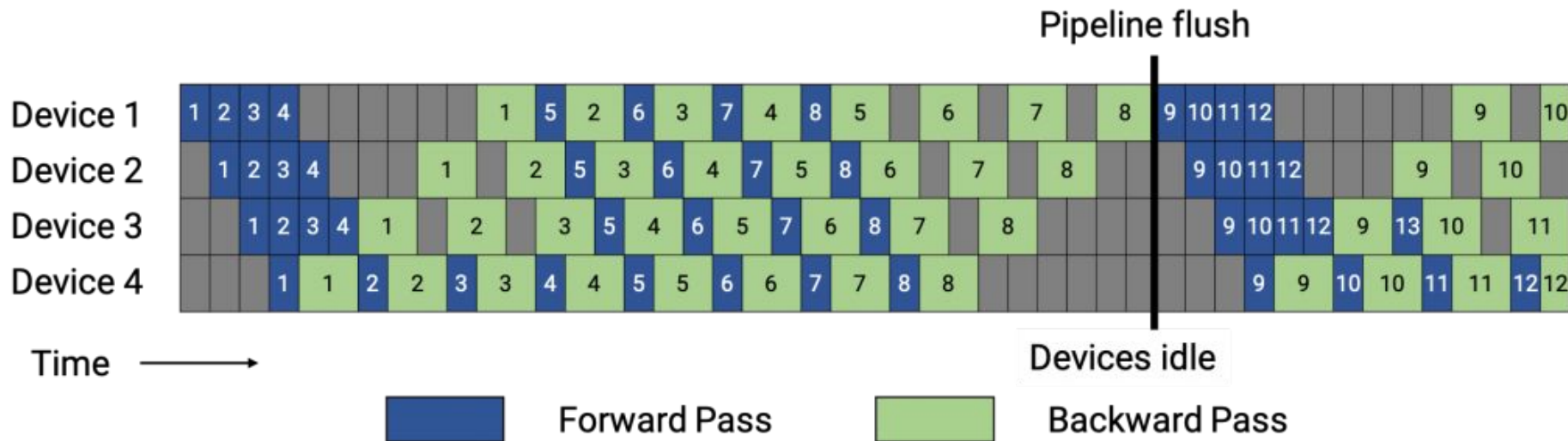
Memory Hierarchy with Bandwidth & Memory Size



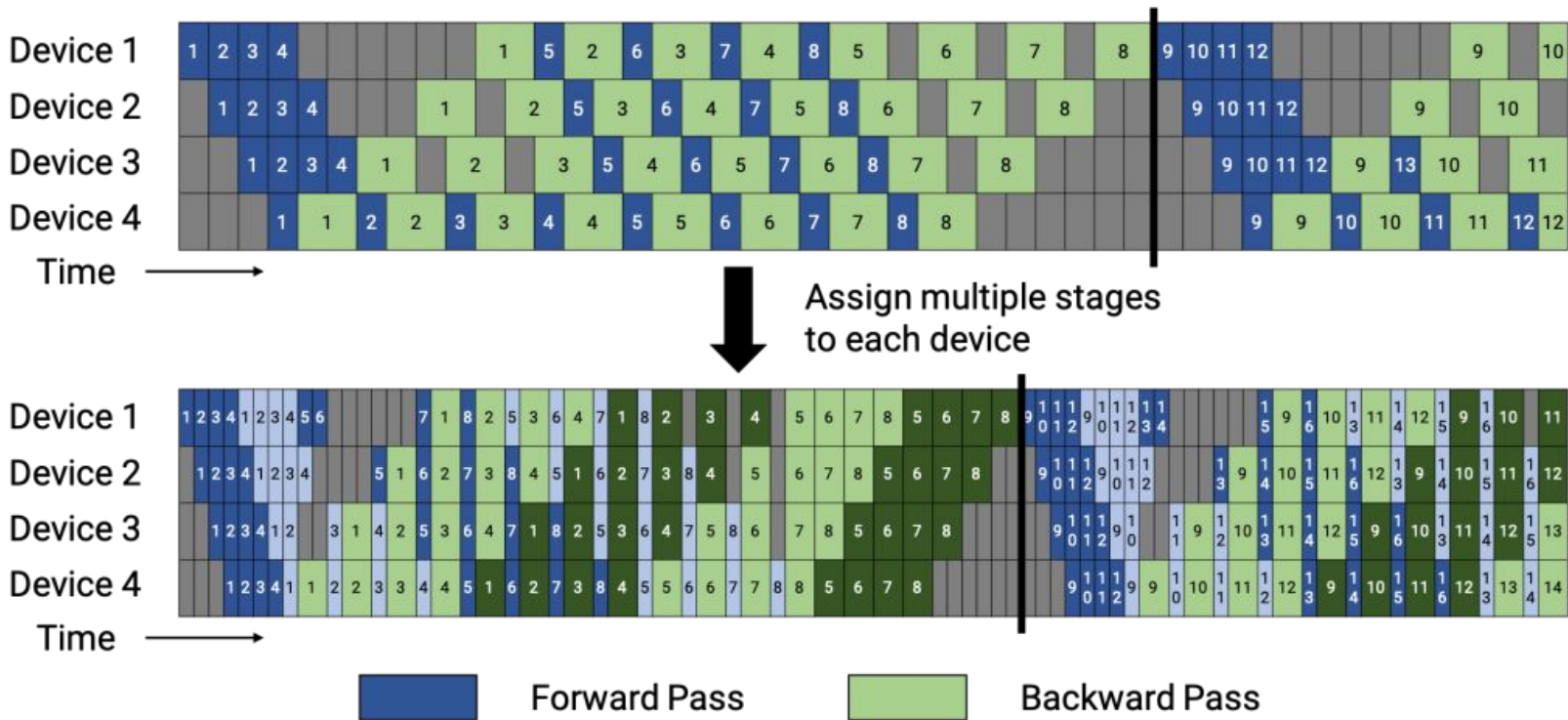
이제 대용량 LLM



이제 대용량 LLM

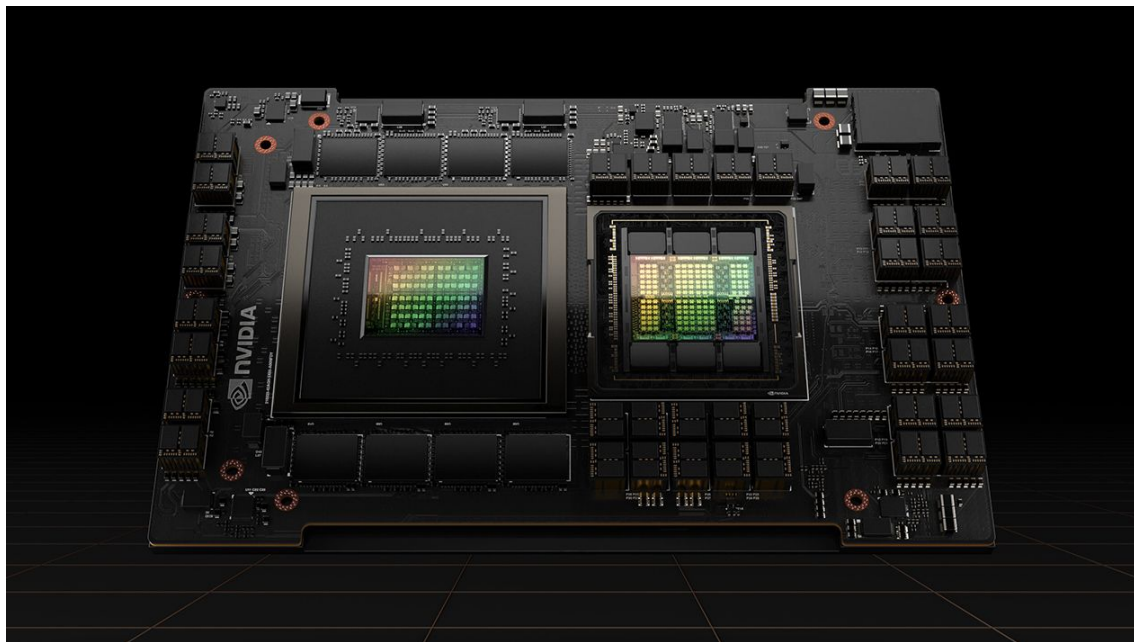


이제 대용량 LLM



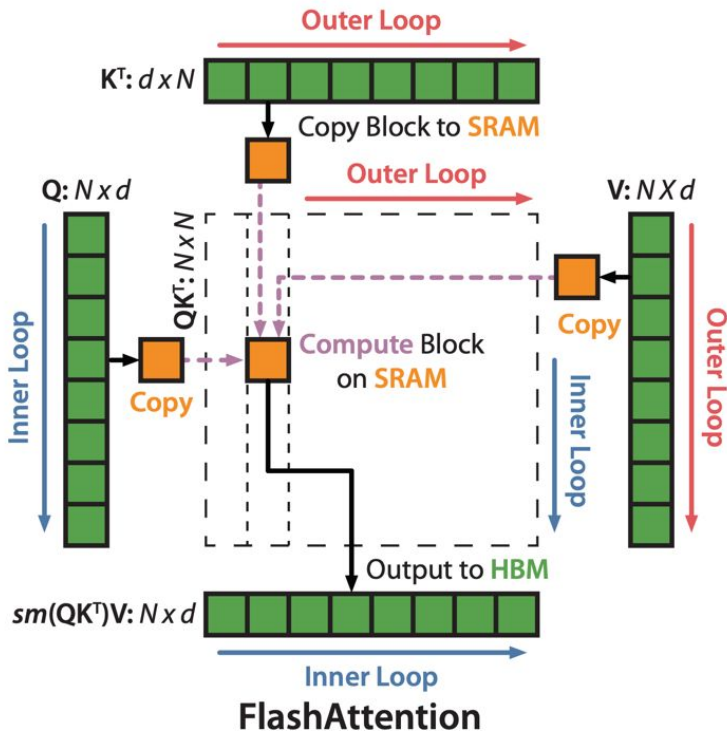
그래서... 최대한 카드에서 처리하는것이 중요

H100, A100, RTX 6000 ada



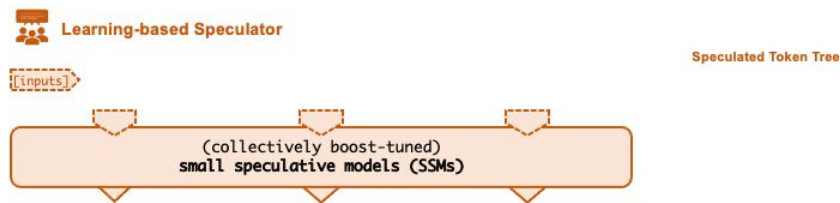
Flash Attention

타일링(Tiling), 재계산 (Recomputation) 두개만 기억하세요.



Speculative Inference

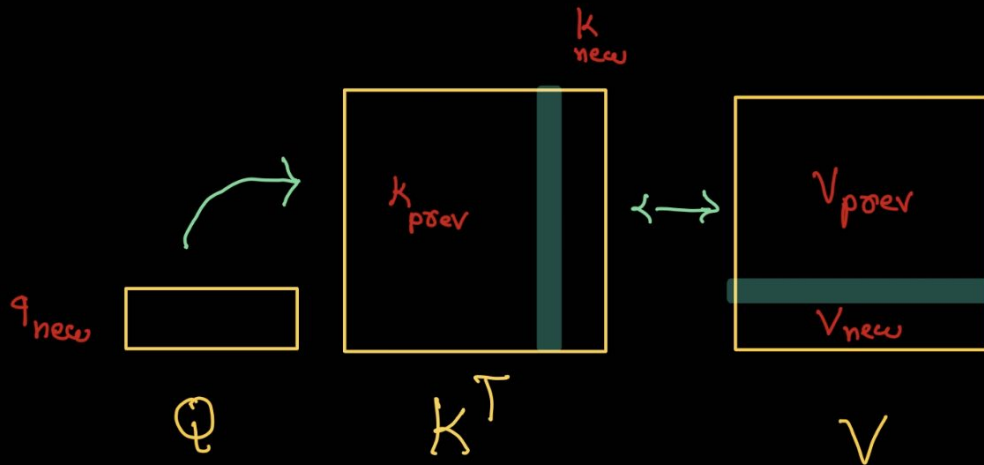
추측적 병렬 추론



K V cache

중간 결과 저장

```
def attention(q, k, v, mask): # [n_q, d_k], [n_k, d_k], [n_k, d_v], [n_q, n_k] -> [n_q, d_v]
    return softmax(q @ k.T / np.sqrt(q.shape[-1]) + mask) @ v
```



k_{prev} and v_{prev} are cached

Apple

How many tflops is the M2 Ultra GPU?

27.2 TFLOPS

In total, the M2 Max GPU contains up to 608 execution units maximum floating point (FP32) performance of 13.6 TFLOPS. The M2 Ultra features a 36-core GPU with up to 9728 ALUs and **27.2 TFLOPS** of FP32 performance.

NVIDIA A100 for
NVIDIA HGX™

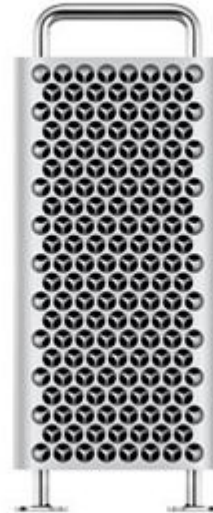
NVIDIA A100 for
PCIe

NVIDIA Ampere

FP64: 9.7 TFLOPS

FP64 Tensor Core: 19.5 TFLOPS

FP32: 19.5 TFLOPS



Nvidia

	NVIDIA A100 for NVIDIA HGX™	NVIDIA A100 for PCIe
GPU Architecture	NVIDIA Ampere	
Double-Precision Performance	FP64: 9.7 TFLOPS FP64 Tensor Core: 19.5 TFLOPS	
Single-Precision Performance	FP32: 19.5 TFLOPS Tensor Float 32 (TF32): 156 TFLOPS 312 TFLOPS*	
Half-Precision Performance	312 TFLOPS 624 TFLOPS*	
Bfloat16	312 TFLOPS 624 TFLOPS*	
Integer Performance	INT8: 624 TOPS 1,248 TOPS* INT4: 1,248 TOPS 2,496 TOPS*	
GPU Memory	40 GB HBM2	
Memory Bandwidth	1.6 TB/sec	

