Open Model을 Long Context 형태로 튜닝하는 방법

이성진 (BHSN)

발표자 소개

- 99년생 갓 졸업한 사회 새내기
- 수학보다 국어를 더 좋아했던 공대생
- 2021년 군 복무 도중 하이퍼클로바를 보고 홀린 듯 자연어 처리 공부 시작
- 2022년 CLOVA 챌린지에서 1등 먹고 하이퍼클로바 기반 서비스 출시에도 기여해 봄
  - 현재까지 제 인생 최고 업적
- 현재는 변호사님들과 함께 AI 기반 법률 서비스 개발 중
  - BHSN에서는 변호사님들이 데이터 라벨링을 합니다?!
- OSS Contribution 하시는 분들 존경합니다
  - 저도 하고 싶어요…

Seongjin Lee
GirinMan

Contributor **@HanyangTechAI** / NLP
Engineer **@bhsn-ai**

Edit profile

45 followers · **38** following

BHSN.AI
Uiwang-si, Gyeonggi-do, Korea
11:19 (UTC +09:00)
lsjg9909@hanyang.ac.kr
life.like.giraffe
in/girinman
https://blog.naver.com/girin_man

Introduction

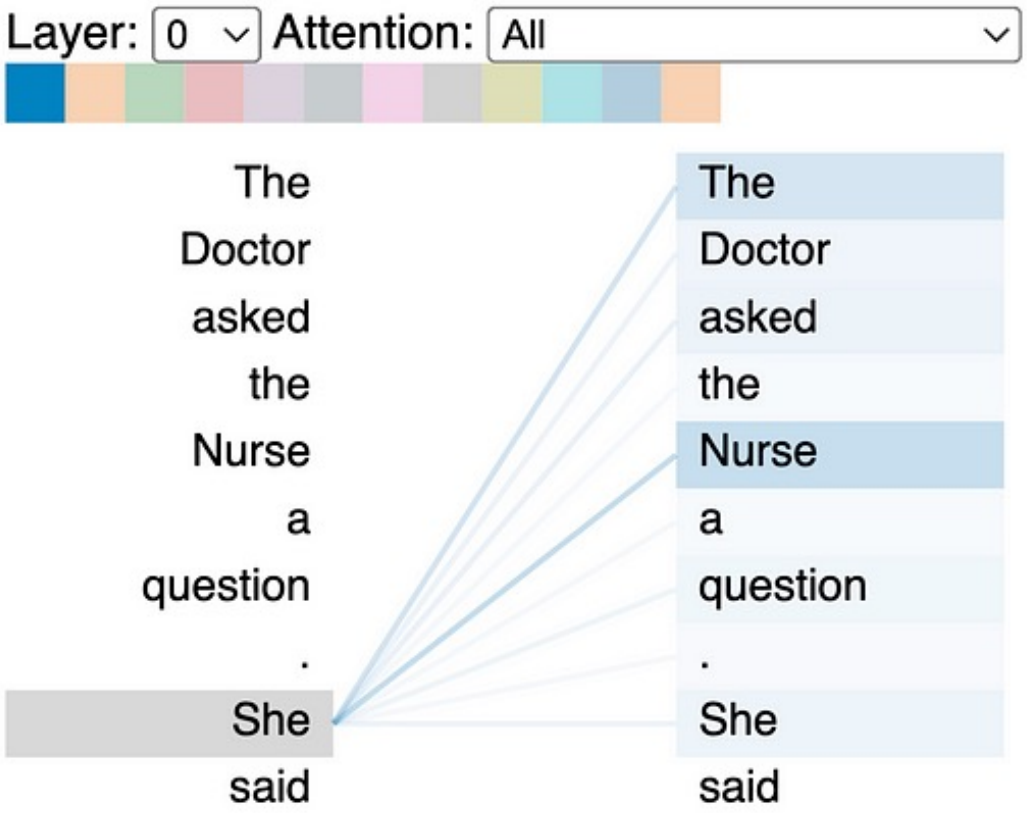
- Desiderata
  - 기존 open source 언어 모델의 context window 한계
  - From Scratch부터 Long Context Pretraining에는 큰 비용이 필요함
  - 요약, QA 등 long context가 필요한 task 존재
- Goal
  - 최소한의 비용으로, 기 학습된 모델의 효과적인 Context Window 확장 필요성
  - Rotary Positional Embedding (RoPE) 기반 Positional Interpolation 방법 소개
  - Long Context 학습 방법 및 평가 데이터셋 & 지표 소개

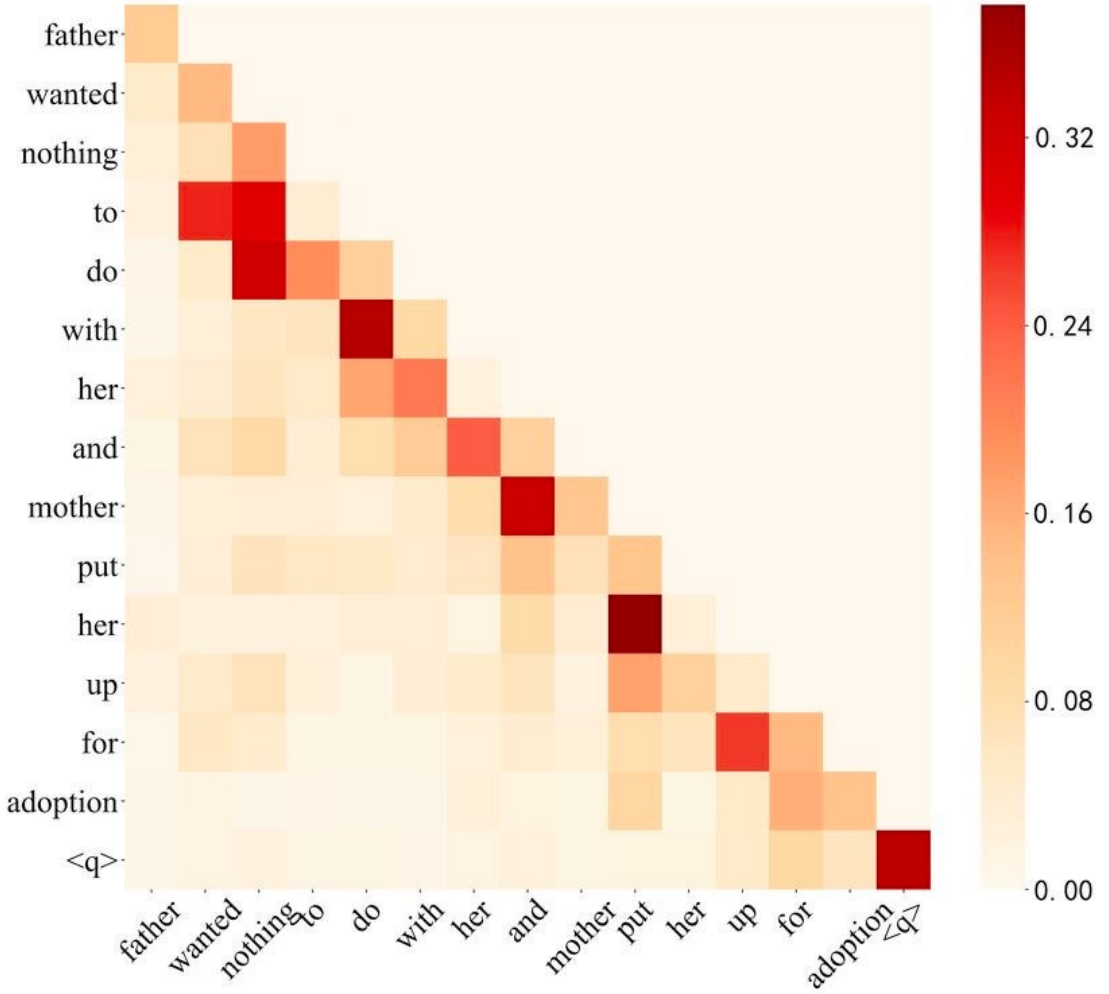
Long Context Summarization & QA

- Llama-2 등 Open model의 context에 한 번에 입력할 수 없는 길이의 문서는 요약하기 어려움
- 기존에는 문서를 특정 segment 단위로 chunking 후 요약, merge하는 방식 주로 사용
  - https://deview.kr/2023/sessions/557
  - 문서가 chunk 단위로 나누어지며 정보가 손실될 수 있음
  - 전체 맥락을 파악하며 요약하기 어려움
- 모델의 Context를 확장시키면 매우 긴 문서를 한번에 넣어서 전체 문맥을 고려하며 요약할 수 있음
- RAG(Retrieval Augmented Generation)에서도 더 많은 검색 결과를 바탕으로 답변 가능함
  - 원고와 피고의 주장 요지를 비교하여 표 형태로 정리해줘 -> 하나의 검색 결과로는 답변하기 어려움
  - Longer Context에서 더 많은 Document를 바탕으로 적절하게 답변 가능

Self-Attention Mechanism

Language Conference



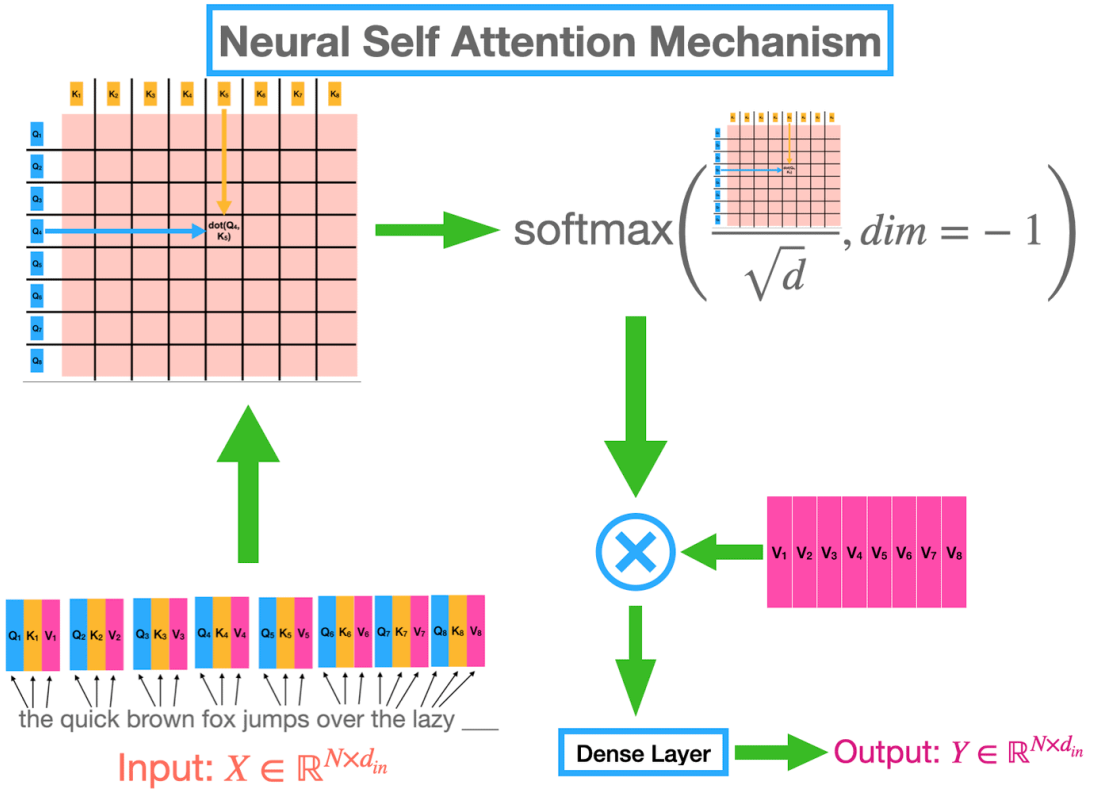


The Doctor asked the Nurse a question. She [???]

Self-Attention Mechanism

Computational Cost
∝ N^2

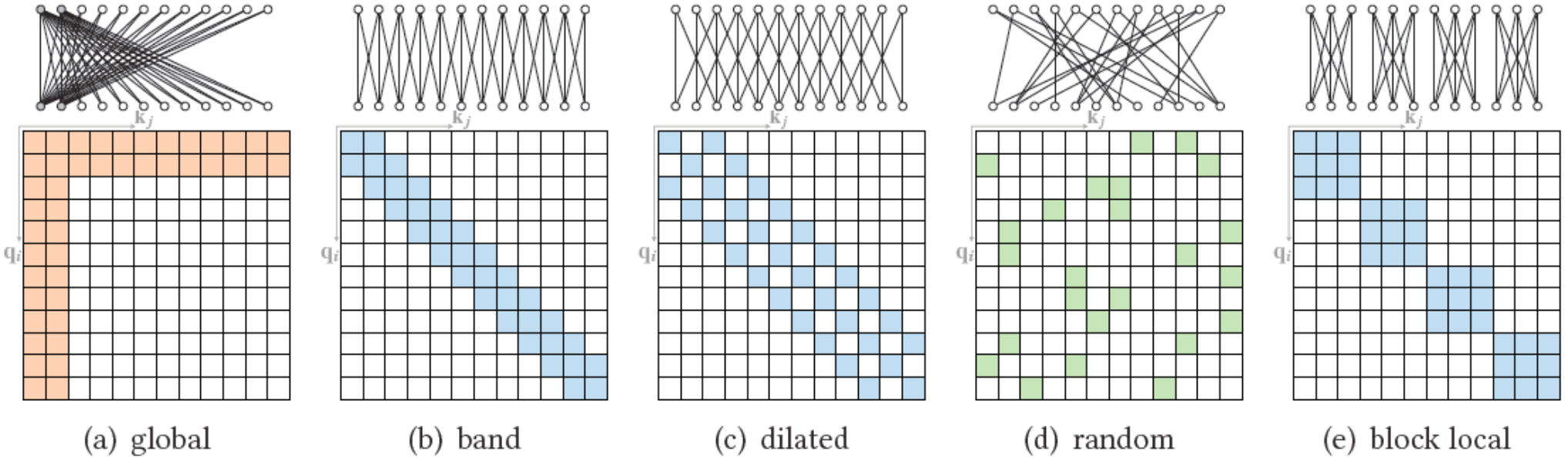
2K context window 처리 가능한 모델을
8K로 확장 시키려면 비용 16배!



Sparse, Local and Windowed Attention

Language Conference

- Context 내에서 모든 토큰에 대해 attention을 계산하지 않고, 특정 범위 내에서만 계산
- Full attention 대비 Computational Efficiency가 크게 향상됨

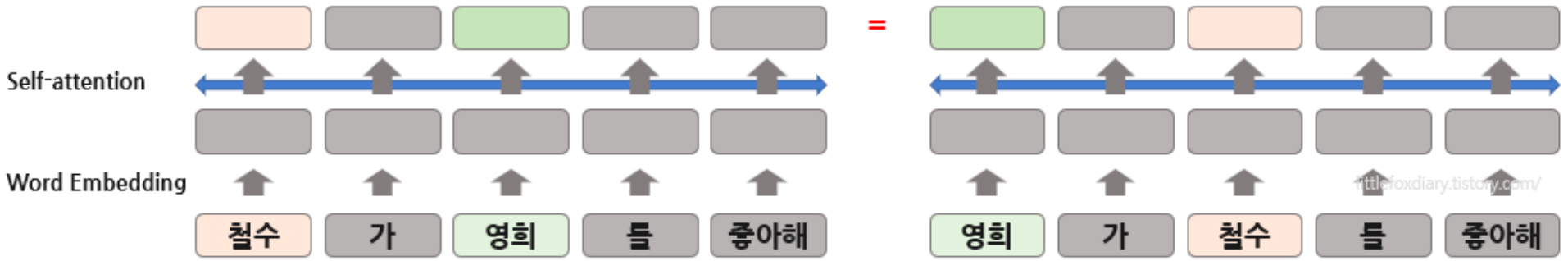


Mistral이 사용 중

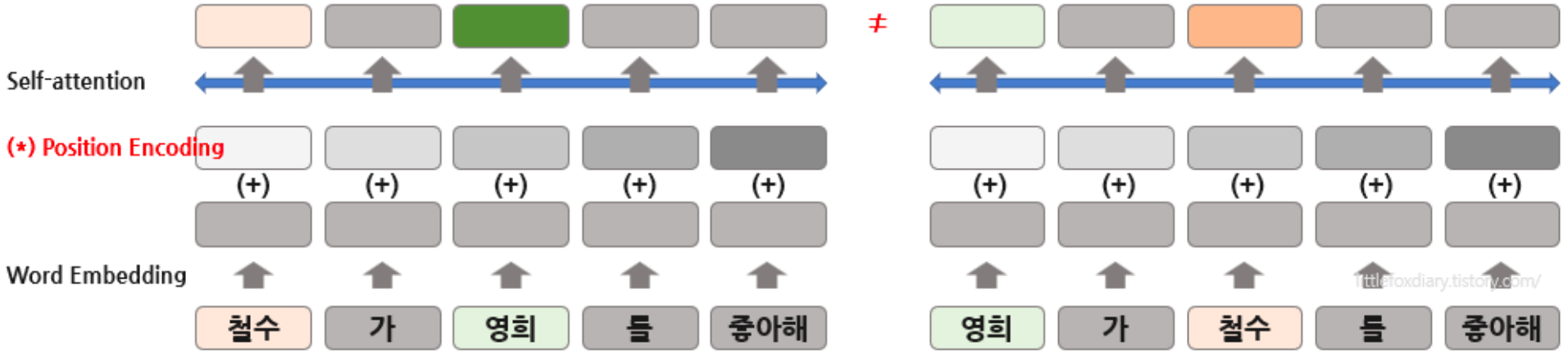
Positional Encodings

Language Conference

- Transformer는 입력된 token들의 순서를 직접적으로 구분할 수 없음



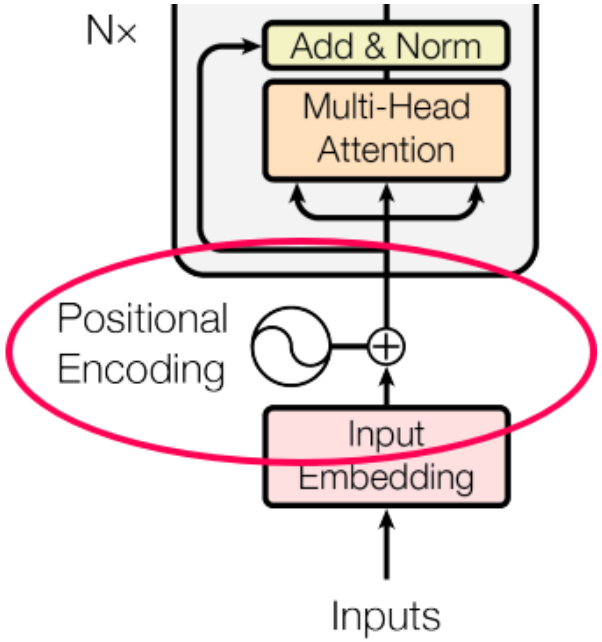
- 각 토큰의 임베딩마다 위치를 표현할 수 있는 값을 더해 인코딩하여 모델이 위치 정보를 인식할 수 있게 함



Limits of Traditional Positional Encodings

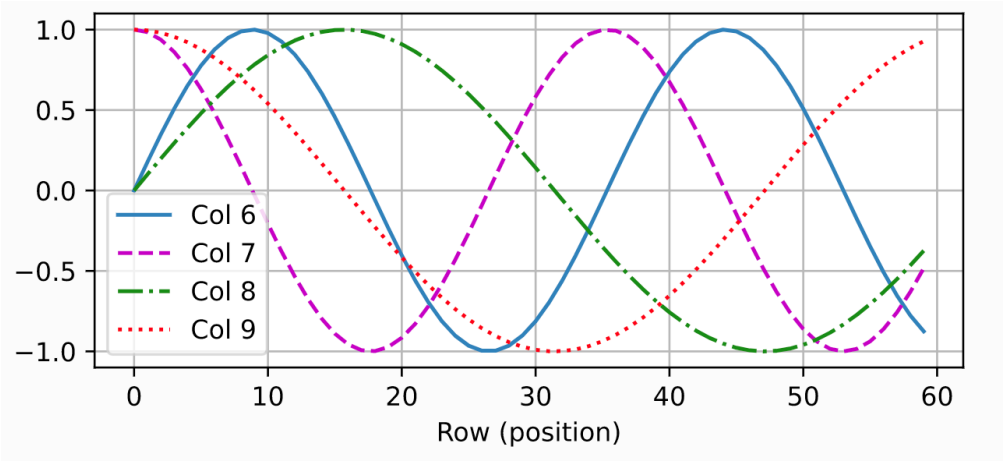
Language Conference

- 초기 Transformer 모델에서 사용된 positional embedding은 입력 시퀀스의 최대 길이를 모델 학습 시 미리 정해야 함
  - 모델이 학습 도중 또는 이후에 더 긴 시퀀스를 처리할 수 없어 모델의 범용성과 확장성에 제한이 생김
  - Ex) 2시간 길이의 영화 대본을 통째로 요약하거나 감상문을 작성하려면?.
- 초기의 positional embedding은 주로 절대 위치 정보에 초점을 맞춰 단어 간의 상대적 위치 관계를 모델링하는 데 있어 제한적
  - 문장 내에서 단어 간의 상대적 거리가 중요한 언어적 구조를 포착하는 데 어려움이 있을 수 있음.



$$\mathrm{PE}(pos, 2i) = \sin\left(pos/10000^{2i/d_{model}}\right)$$

$$\mathrm{PE}(pos, 2i+1) = \cos\left(pos/10000^{2i/d_{model}}\right)$$



Rotary Positional Encodings(RoPE)

Language Conference

- 각 토큰의 위치에 따라 임베딩 벡터를 일정 각도만큼 회전시키는 방식
  - 전체 Context Length가 달라져도 각 토큰 간 상대적인 위치 정보를 표현할 수 있음
  - Embedding dimension에 따라 각각 다른 frequency를 활용하여 효과적으로 상대적인 위치 정보 표현

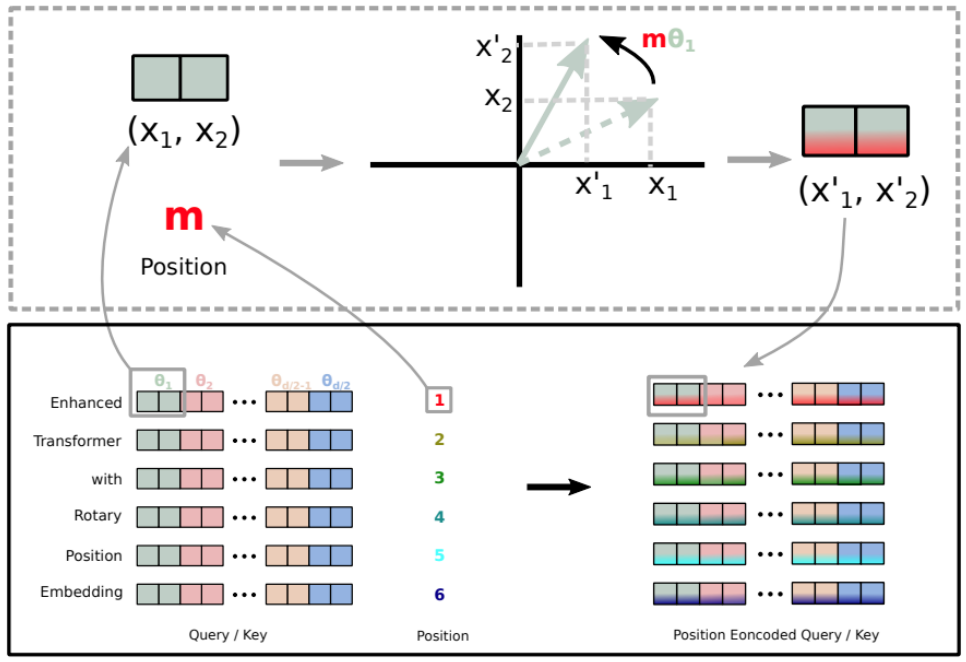
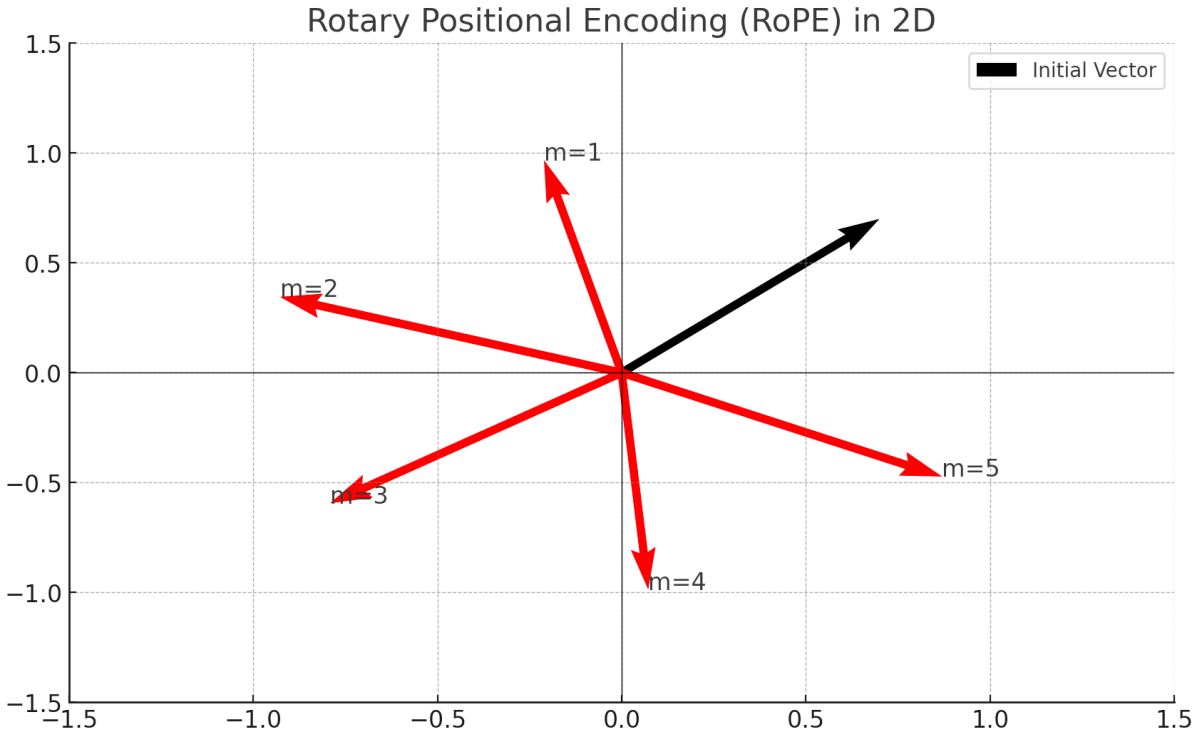
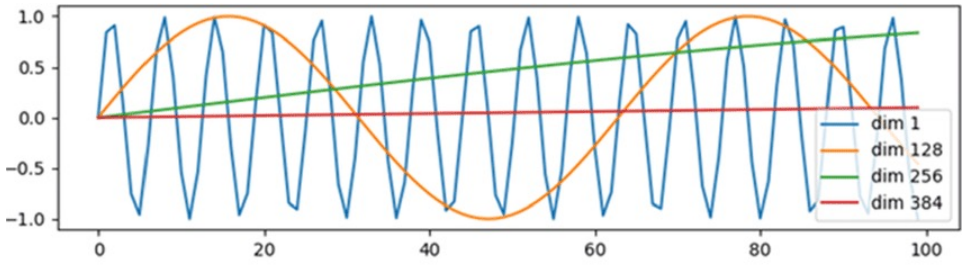


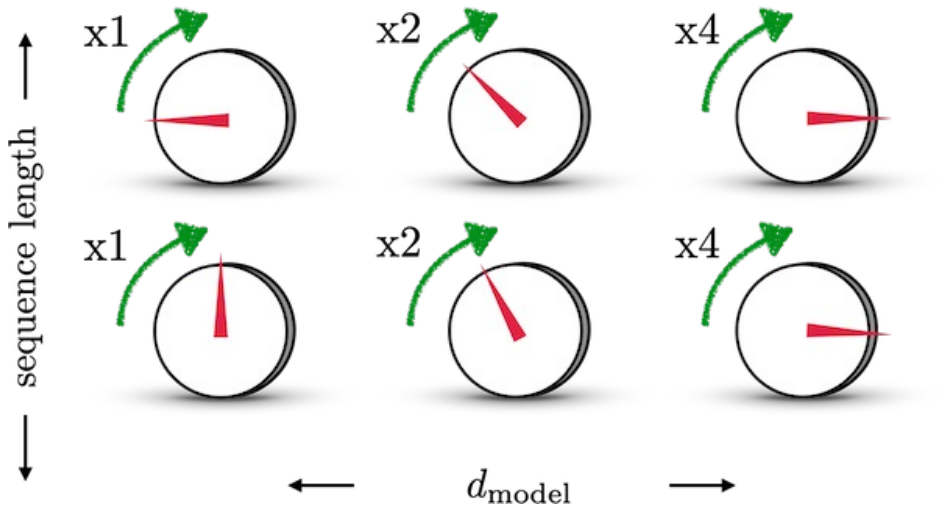
Figure 1: Implementation of Rotary Position Embedding(RoPE).



Rotary Positional Encodings(RoPE)

- 임베딩 차원마다 다른 frequency를 가지는 회전 각도를 통해 효과적으로 정보를 표현
- 마치 시각을 표현할 때 년, 월, 일, 시간, 분, 초 등을 각각 표시하는 것과 유사함

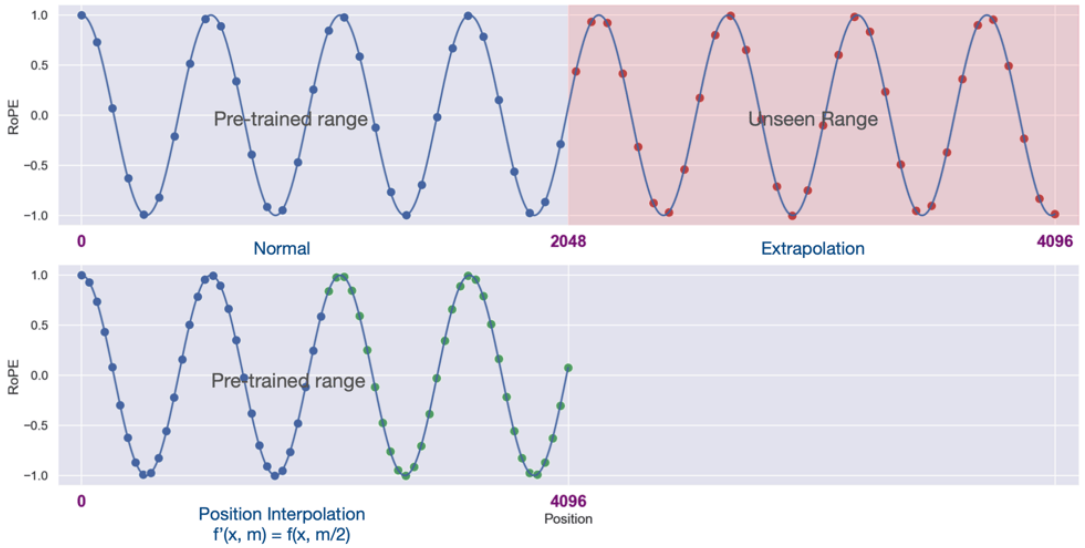


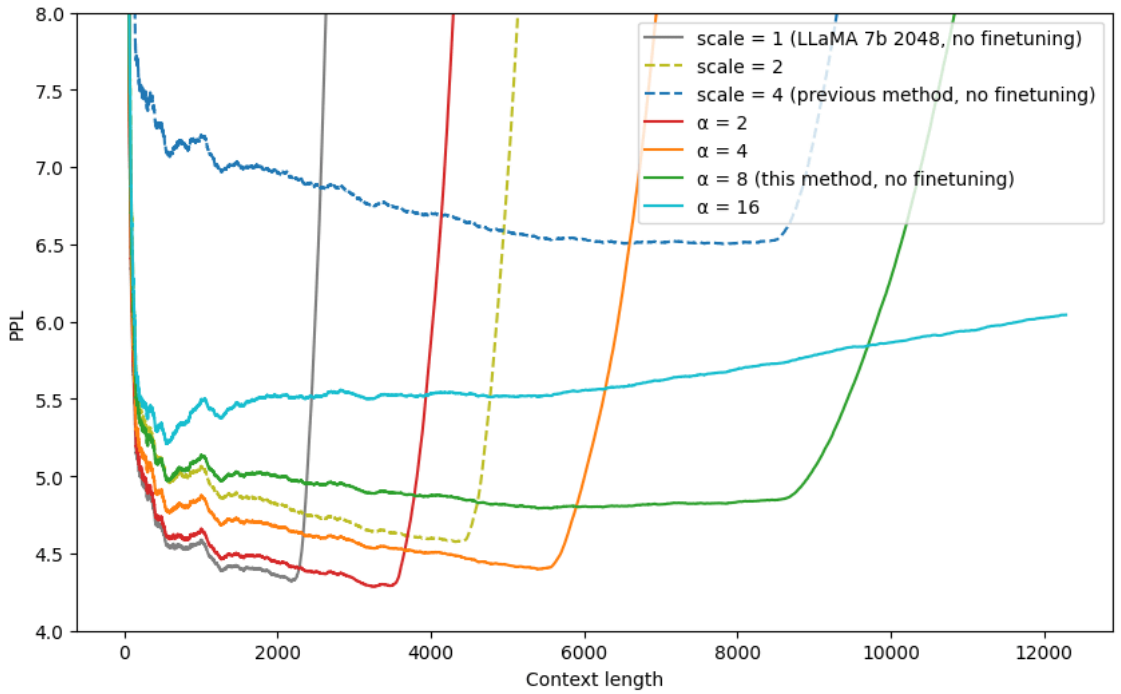


Performance Degradation in Long Context

Language Conference

- 현재 Transformer 구조상 기하급수적으로 늘어나는 필요 자원으로 인해 Pretrain 단계에서 long context(8K 이상)으로 학습시키는 것은 매우 어려움
- 최근 공개된 오픈소스 모델은 2K~4K 정도의 context에서 학습되었기 때문에, extrapolation만을 적용하면 pretrain 시점에서 보지 못했던 특정 context 이상에서는 생성 능력(LM Loss 기반 perplexity)이 매우 떨어짐





Positional Interpolation

- Rotary Positional Embedding의 구조를 활용하면 pretrain 시점에서 학습되었던 position에 대한 정보를 인식하는 능력을 유지하면서 더 긴 context를 사용할 수 있도록 확장 가능
- Linear Scaling
  - 학습 시점의 context와 추론 시점의 context 길이에 정비례하게 RoPE를 조정하는 방법
- Dynamic NTK Scaling
  - Neural Tangent Kernel을 활용하여 각 차원이 표현하는 위치 정보를 더 효과적으로 표현할 수 있도록 하는 방법
- Yarn
  - 각 embedding dimension별 frequency를 고려하여 scaling 방식을 다르게 하는 방법
- LongRoPE
  - 모델별로 다른 Frequency별 최적의 scaling 방식을 자동으로 찾을 수 있도록 하는 방식

Efficient Post-Pretrain Long Context Extension

Language Conference

- Positional Interpolation은 상대적으로 적은 학습량으로도 더 긴 context에 대한 생성 능력을 확보해 주지만, 학습 과정에 다양한 자원과 기술이 필요함
- LongLoRA
  - Long Context 학습 과정에서 메모리 사용량을 줄이기 위해 모델의 일부 파라미터만을 학습시키는 LoRA를 활용
- Scalable Training via Distributed Training
  - 초거대 모델을 Long Context에서 학습시키기 위해서는 효과적으로 scalable한 학습 방법을 활용할 수 있어야 함
  - Megatron-LM을 활용한 Multi-Node 3D-Parallelism(Model, Data, Tensor Parallel) 기법 등 활용
- (여전히) High Computing Resources가 필요함
- Long Context를 반영할 수 있는 corpus 확보 및 전처리 역시 필요함
  - 긴 대화 내용을 포함하는 발화 말뭉치
  - 행정 문서, 매뉴얼, 소스 코드 등 길이가 매우 긴 텍스트

Long Context Pretrained 모델 예시: XGen-7B

Language Conference

https://blog.salesforceairesearch.com/xgen/#training-details

Long Sequence Modeling with XGen: A 7B LLM Trained on 8K Input Sequence Length

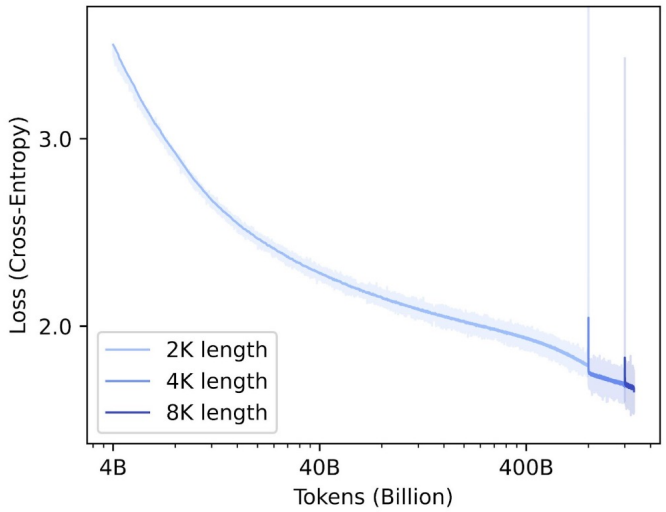
10 min read

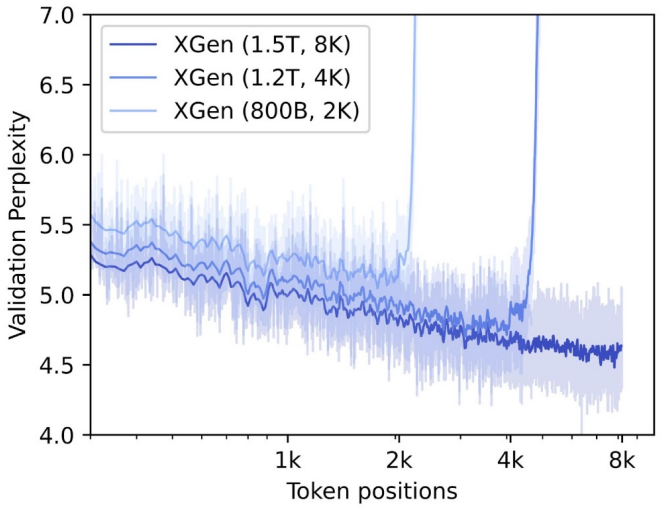
Erik Nijkamp Hiroaki Hayashi Tian Xie Congying Xia Bo Pang Rui Meng Wojciech Kryscinski Lifu Tu Meghana Bhat Semih Yavuz Chen Xing

We trained a series of **7B LLMs named XGen-7B** with standard dense attention on up to **8K** sequence length for up to **1.5T tokens**. We also fine tune the models on public-domain instructional data. The main take-aways are:

- On standard NLP benchmarks, XGen achieves comparable or better results when compared with state-of-the-art open-source LLMs (e.g. MPT, Falcon, LLaMA, Redpajama, OpenLLaMA) of similar model size.
- Our targeted evaluation on long sequence modeling benchmarks show benefits of our 8K-seq models over 2K- and 4K-seq models.
- XGen-7B archives equally strong results both in text (e.g., MMLU, QA) and code (HumanEval) tasks.
- Training cost of $150K on 1T tokens under Google Cloud pricing for TPU-v4.

Model	Description
XGen-7B-4K-base	We train for 800B tokens with a sequence length of 2k tokens first, then for another 400B tokens (total **1.2T tokens**) with 4k. Released under Apache-2.0.
XGen-7B-8K-base	Initialized with XGen-7B-4K-base and further trained for 300B more tokens (total **1.5T tokens**) with 8K sequence length. Released under Apache-2.0.
XGen-7B-{4K,8K}-inst	Supervised fine tuned on public domain instructional data including databricks-dolly-15k, oasst1, Baize and GPT-related datasets. Released for research purpose only.

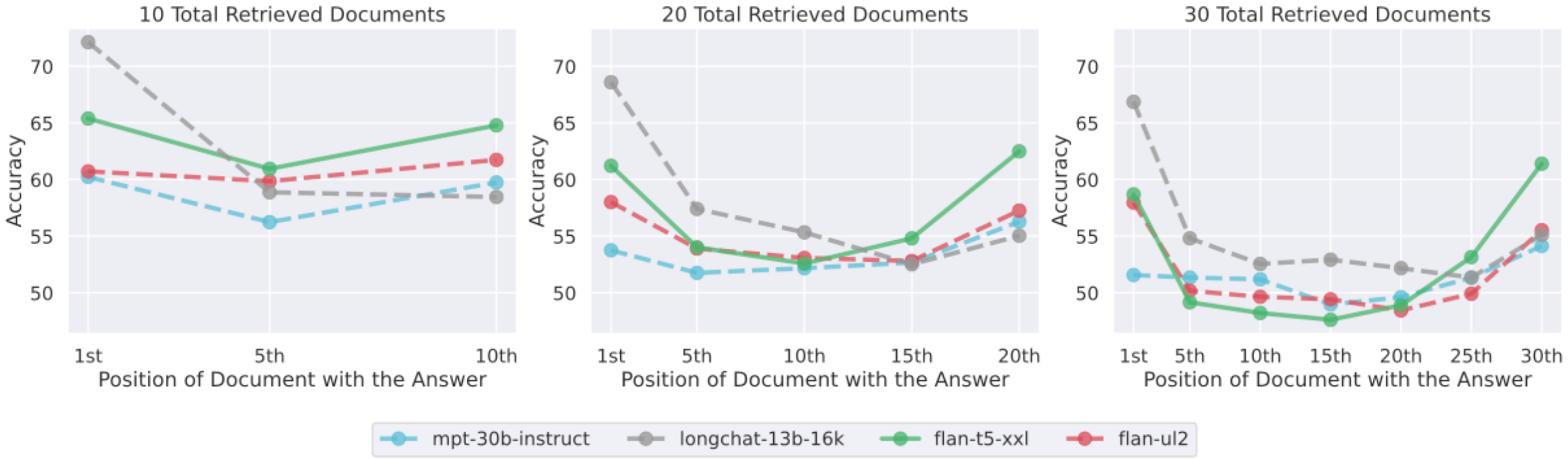




Lost in the Middle

Language Conference

- 긴 Context를 사용하면 RAG(Retrieval Augmented Generation)등에서 더 많은 document를 사용할 수 있음
- 하지만, 모델이 학습되었을 때 보다 긴 context에서 사용하면 중간 부분의 정보를 제대로 참고하지 못함



Alignment Learning on Long Context Extension

Language Conference

- Long Context에서 모델의 성능을 단순 Perplexity만으로 평가하는 것은 조금 어려움
  - QA, Summarization 등 Instruction을 수행하는 과정에서 Lost In the Middle 발생
  - Long Context의 각 영역별로 attention이 고르게 걸릴 수 있도록 alignment가 필요함
- QA, Summarization을 긴 Context에서 수행할 수 있도록 하는 학습 데이터 및 평가 지표
  - Long Context에서 task를 수행하기 위한 SFT 데이터(Long Alpaca 등)
  - LMSys, WikiQA, Passkey Retrieval 등 task-oriented evaluation

Long Input

There is an important info hidden inside a lot of irrelevant text. Find it and memorize them. I will quiz you about the important information there. The grass is green. The sky is blue. The sun is yellow. Here we go. There and back again. … The sun is yellow. Here we go. There and back again. The pass key is 0000 Remember it. 0000 is the pass key. The sun is yellow. Here we go. There and back again …The grass is green. The sky is blue. The sun is yellow. Here we go.

Prompt: What is the pass key?
The pass key is
LLM Output: 0000 ✓

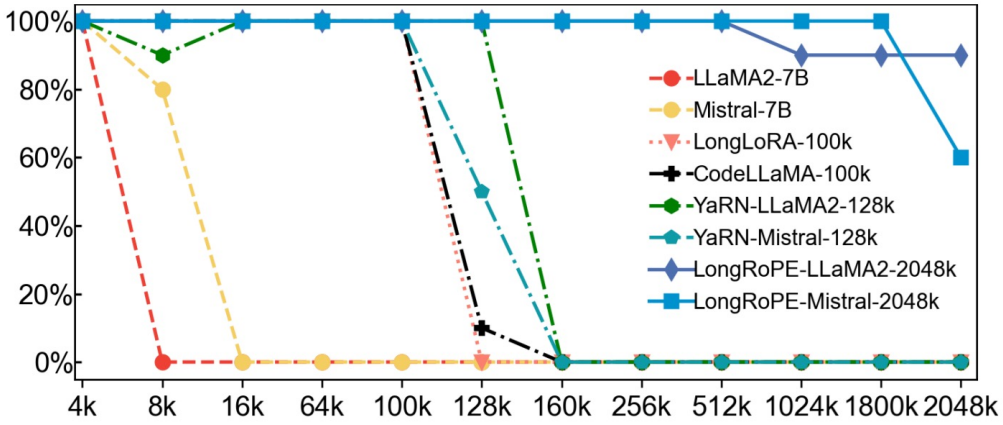
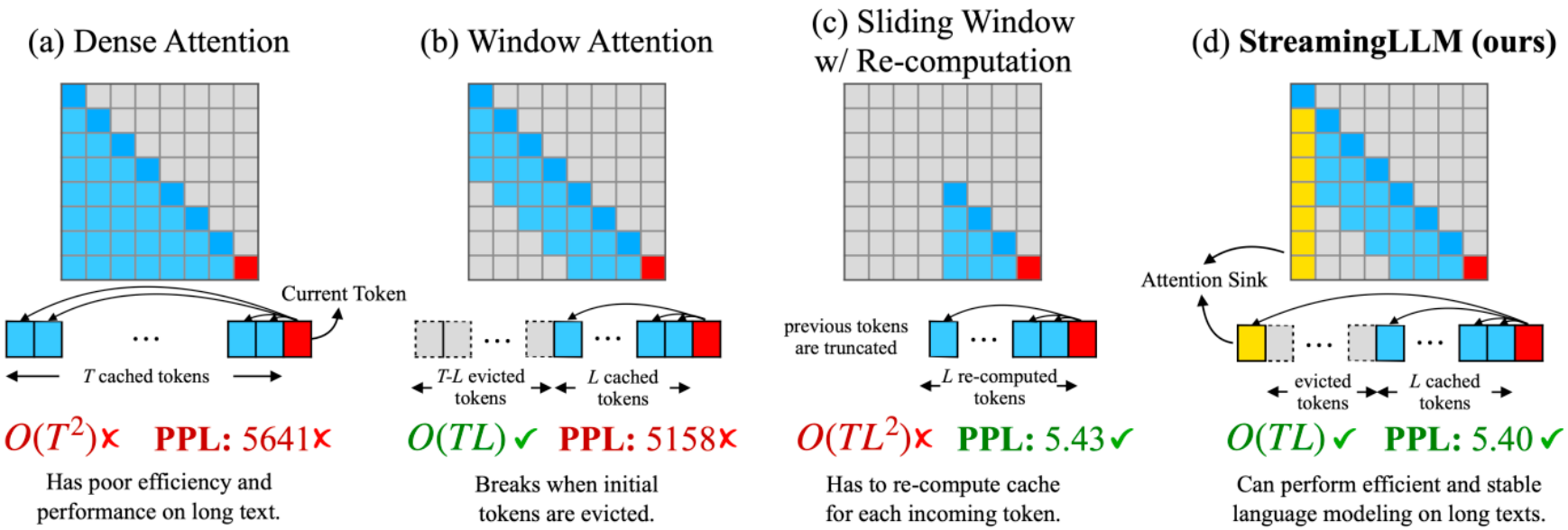


Figure 4. Passkey retrieval accuracy of long-context LLMs. It showcases the remarkable ability of our models to accurately retrieve a passkey from a vast pool of million-level tokens.

Serving Long Context LLMs Efficiently

Language Conference

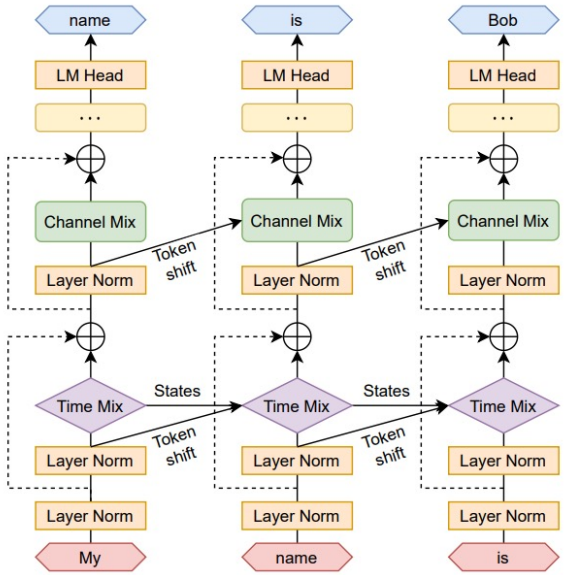
- RoPE scaling, Flash Attention, Paged Attention, …
- Huggingface TGI에 (놀랍게도) 모두 포함되어 있다!
- Flash-Decoding: Attention 연산을 병렬화하여 메모리 사용량을 줄이고 생성 속도를 증가시킴
- StreamingLLM: 특정 window size만큼만 참고하되 첫 토큰을 함께 고려함으로써 긴 context에서도 성능이 drop되지 않도록 함

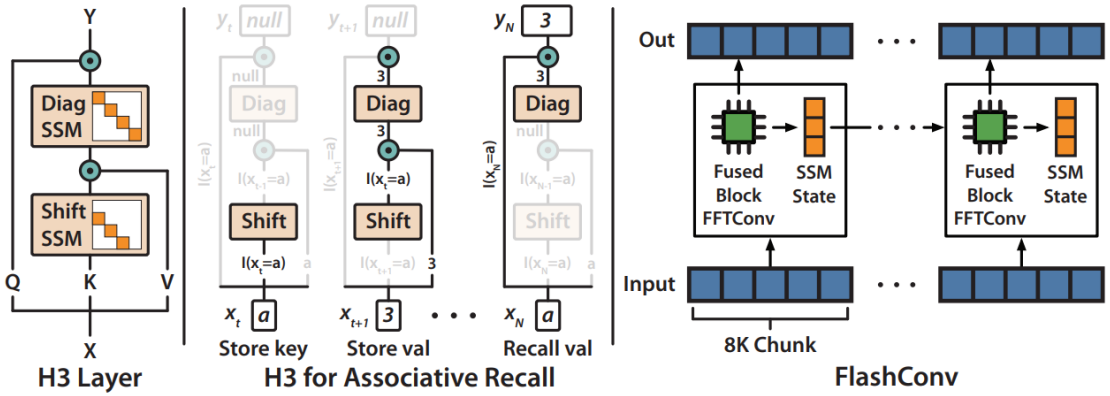


Future of Long Context LLM

Language Conference

- 다양한 방법론에도 Self-Attention 기반 Transformer 구조의 한계는 명확함
- Context 길이의 제한을 받지 않는 RNN 기반 RWKV, S3 기반 Mamba 등 새로운 모델 구조 연구 진행 중...
- 대량의 Context를 제약 없이 다룰 수 있는 모델이 등장한다면 새로운 패러다임이 시작될 수 있음
  - 지금까지 살아온 내 모든 인생을 요약해서 정리하고 앞으로의 방향을 제안해줘!
  - 정말 재미있는 소설이 완결되었는데 기존 소설 전체의 내용과 독자 반응을 고려해서 후속편을 발매해줘!
- 결국 NVIDIA가 엄청난 하드웨어를 출시하면 다 해결될지도...?





References

- Attention Is All You Need https://arxiv.org/abs/1706.03762
- Generating Long Sequences with Sparse Transformers https://arxiv.org/abs/1904.10509
- Longformer: The Long-Document Transformer https://arxiv.org/abs/2004.05150
- RoFormer: Enhanced Transformer with Rotary Position Embedding https://arxiv.org/abs/2104.09864
- XGen-7B(Salesforce): https://blog.salesforceairesearch.com/xgen/#training-details
- Lost in the Middle: How Language Models Use Long Contexts https://arxiv.org/abs/2307.03172
- Giraffe: Adventures in Expanding Context Lengths in LLMs https://arxiv.org/abs/2308.10882
- Extending Context Window of Large Language Models via Positional Interpolation https://arxiv.org/abs/2306.15595
- YaRN: Efficient Context Window Extension of Large Language Models https://arxiv.org/pdf/2309.00071.pdf
- LongRoPE: Extending LLM Context Window Beyond 2 Million Tokens https://arxiv.org/pdf/2402.13753.pdf
- LongLoRA: Efficient Fine-tuning of Long-Context Large Language Models https://arxiv.org/abs/2309.12307
- FlashDecoding++: Faster Large Language Model Inference on GPUs https://arxiv.org/abs/2311.01282
- Efficient Streaming Language Models with Attention Sinks https://arxiv.org/abs/2309.17453
- RWKV: Reinventing RNNs for the Transformer Era https://arxiv.org/abs/2305.13048
- Mamba: Linear-Time Sequence Modeling with Selective State Spaces https://arxiv.org/abs/2312.00752

감사합니다!

E-mail: sjlee@bhsn.ai
Linkedin: linkedin.com/in/girinman

LangCon 2024