2024.11.18

RETRO

Improving Language Models by Retrieving from Trillions of Tokens
(Sebastian Borgeaud et al. 2022)

Pseudo-Lab | 신서현

Background

Parametric

모델 파라미터에 모든 정보 저장
학습 데이터 → 모델 내부화된 기억

장점: 모델이 정보 내재화
단점: 메모리 사용량 증가

Non-parametric

외부 데이터베이스 사용 (e.g., 키-값 쌍)
필요 시 정보 검색

장점: 실시간 데이터 업데이트 가능
단점: 검색 속도 중요

Background

Self-attention

동일한 입력 내에서 정보를 상호 참조

Cross-attention

두 개 이상의 서로 다른 소스 간 정보를 연결

RETRO : Abstract

Introduction

autoregressive language modeling

previous method more parameters , more memory

increasing the number of parameters in Transformer models
⇨ tremendous increase in training **energy cost**
⇨ resulted in a generation of **dense LLMs** with 100+ billion parameters

alternate path **RETRO;** Retrieval Enhanced TRansfOrmers

↳ semi-parametric approach
: 모델이 **외부 지식 DB**에 접근해 필요한 지식을 활용

RETRO：2.1. Training dataset

Introduction

for both training and retrieval data
use a multi-lingual version of **MassiveText** (Rae et al., 2021)
：multiple sources, multi-lingual

tokenize the dataset by SentencePiece
(128,000 tokens)

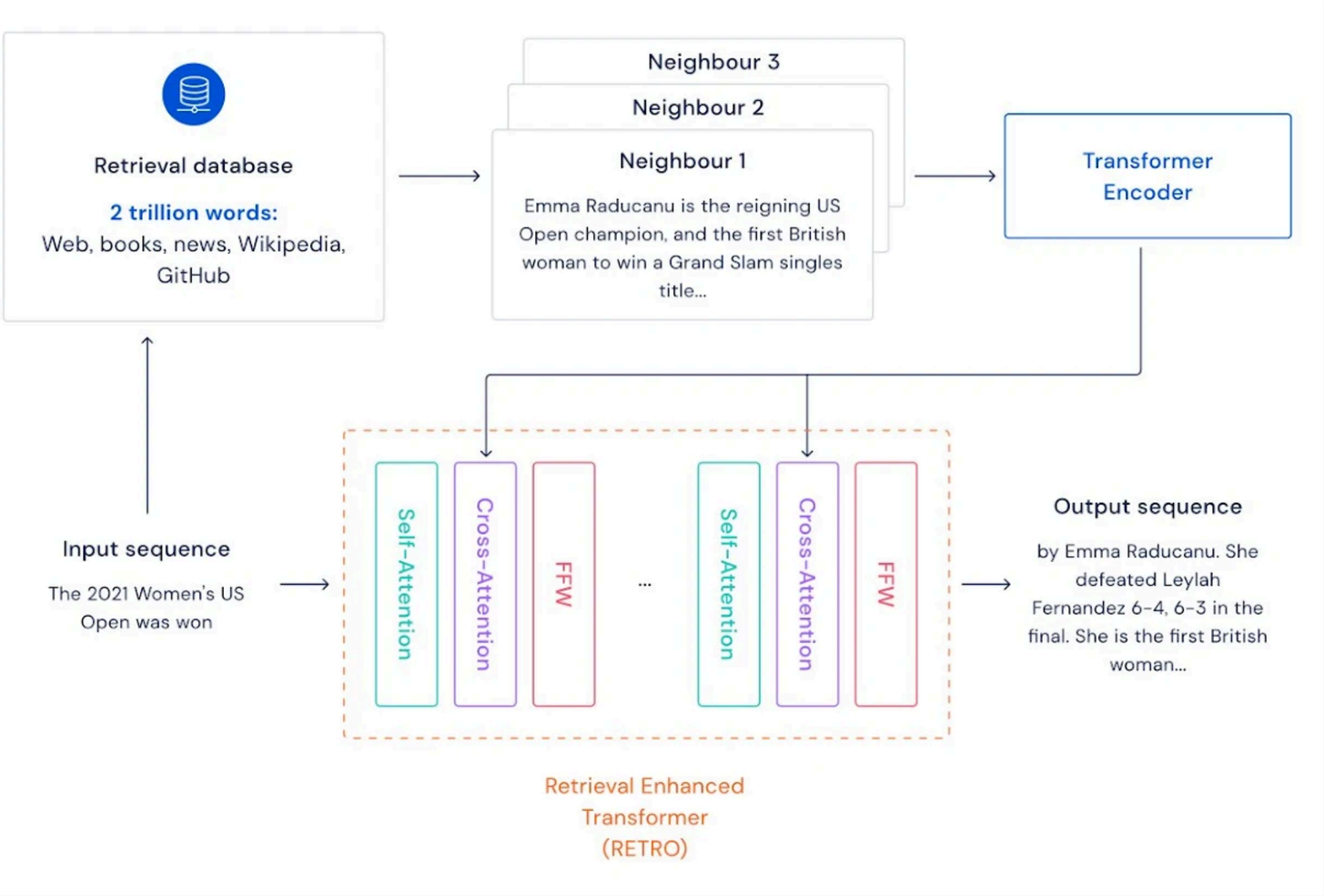
training retrieval：600B tokens
evaluation retrieval：1.75T tokens

To limit test set leakage - **MinHash**
compute the 13-gram Jaccard similarity
(train-test documents) using the MinHash scheme
remove all training documents with high similarity (0.8 or higher)
to a validation or test set document

Source	Language	Token count (M)	Documents	Sampling weight
Web	En	483,002	604,938,816	0.314
	Ru	103,954	93,004,882	0.033
	Es	95,762	126,893,286	0.033
	Zh	95,152	121,813,451	0.033
	Fr	59,450	76,612,205	0.033
	De	57,546	77,242,640	0.033
	Pt	44,561	62,524,362	0.033
	It	35,255	42,565,093	0.033
	Sw	2,246	1,971,234	0.0044
	Ur	631	455,429	0.0011
Books	En	3,423,740	20,472,632	0.25
News	En	236,918	397,852,713	0.1
Wikipedia	En	3,977	6,267,214	0.0285
	De	2,155	3,307,818	0.003
	Fr	1,783	2,310,040	0.003
	Ru	1,411	2,767,039	0.003
	Es	1,270	2,885,013	0.003
	It	1,071	2,014,291	0.003
	Zh	927	1,654,772	0.003
	Pt	614	1,423,335	0.003
	Ur	61	344,811	0.0001
	Sw	15	58,090	0.0004
Github	-	374,952	142,881,832	0.05
Total	-	5,026,463	1,792,260,998	1

deepmind blog：Figure 1

Introduction



https://deepmind.google/discover/blog/improving-language-models-by-retrieving-from-trillions-of-tokens

Method

retrieval-enhanced autoregressive language model

use a **chunked cross-attention module** to incorporate the retrieved text
time complexity linear in the amount of retrieved data

n-token-long example X into a sequence of l chunks of size m = n/l
n=2048, m=64

chunked Trillion 단위의 database에서 retrieval
저장 공간과 계산을 효율화하기 위해 chunked token을 기본단위로 사용

$$C_1 \triangleq (x_1, \ldots, x_m), \ldots, C_l \triangleq (x_{n-m+1}, \ldots, x_n) \in \mathbb{V}^m$$

cross-attention 각 청크 C에 대해 Retrieval 데이터를 검색하고, 이를 Cross-Attention으로 통합

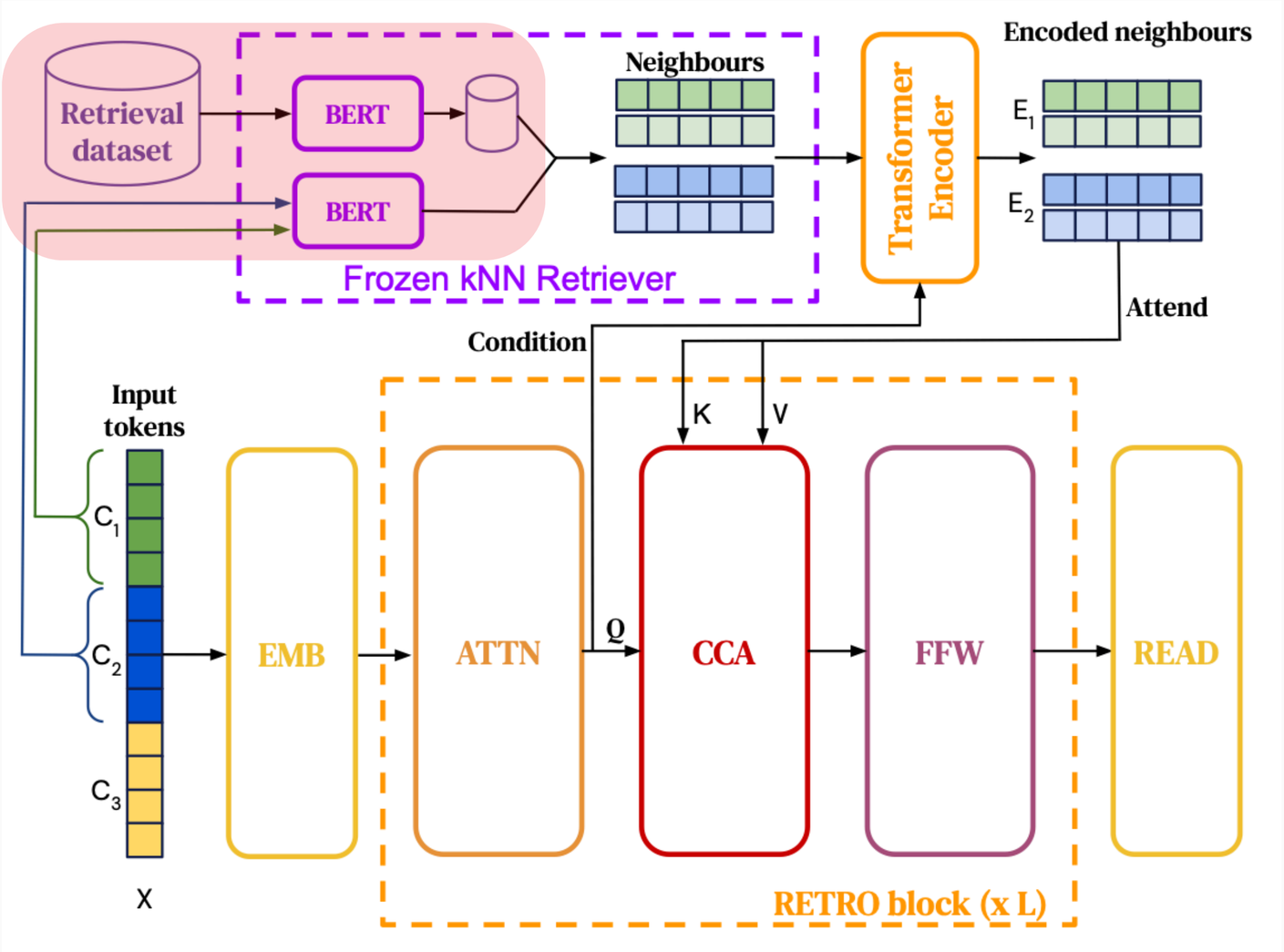
Method

retrieval database

key : frozen BERT embeddings
value : raw chunks of text tokens

key	value
key 1	key에 해당하는 raw text
	next text chunk

고정된 모델(frozen BERT) 사용
→ 학습 중 전체 데이터베이스에 대해 임베딩을 다시 계산할 필요 없음.



RETRO：2.3. Nearest neighbour retrieval

Method

input chunk

↓

L2 distance

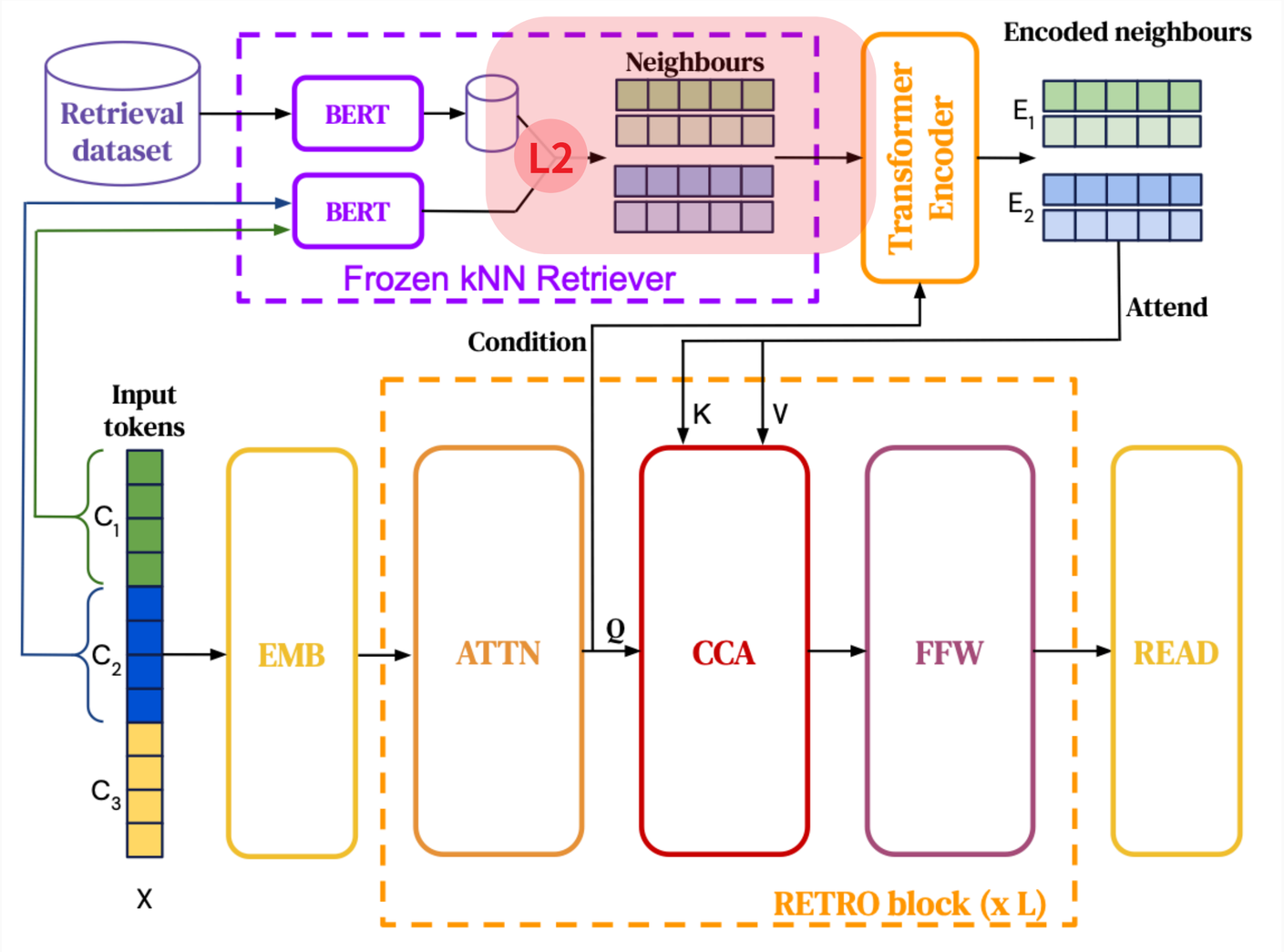
$$d(C, N) = ||\textsc{Bert}(C) - \textsc{Bert}(N)||_2^2$$

↓

nearest neighbours

↓

Transformer encoder



Method

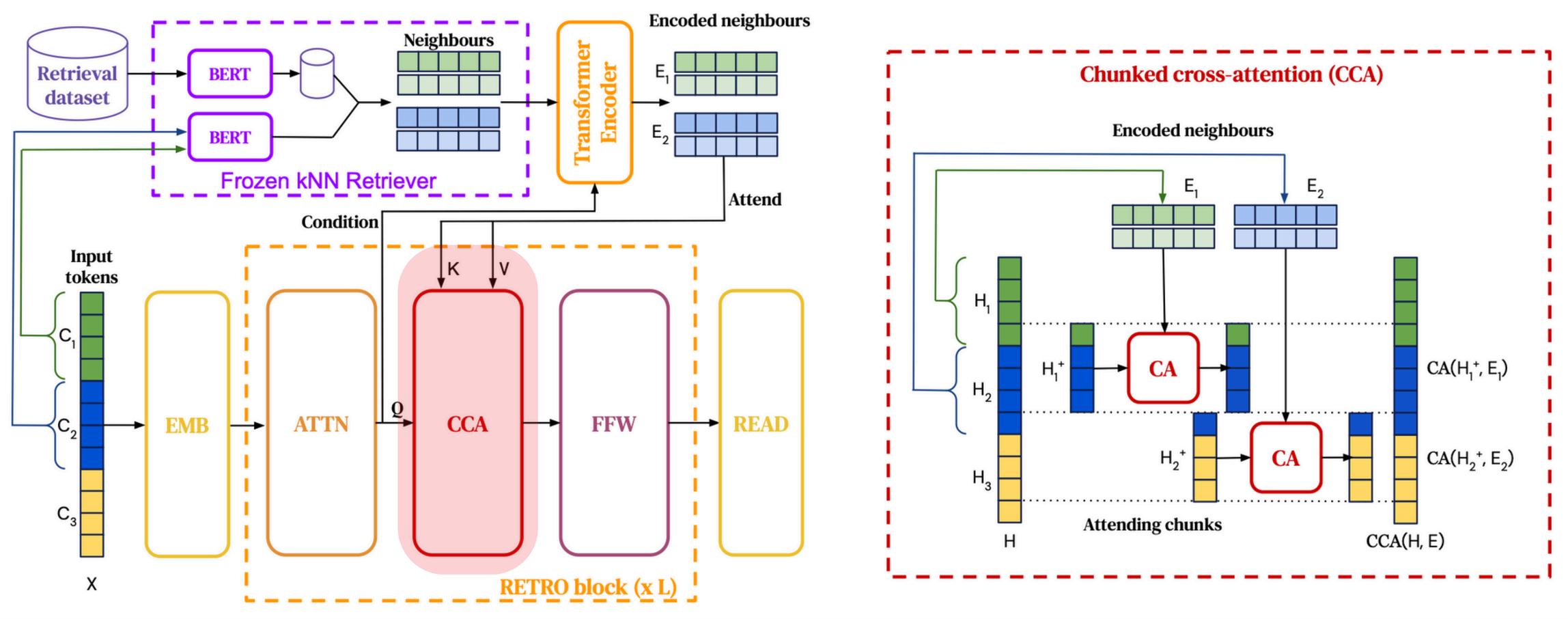
LM architecture

$$\textsc{Lm}(H) \triangleq \textsc{Ffw}(\textsc{Attn}(H))$$

RETRO model architecture

$$\textsc{Retro}(H, E) \triangleq \textsc{Ffw}(\textsc{Cca}(\textsc{Attn}(H), E))$$

Method


Retrieval dataset
BERT
BERT
Neighbours
Frozen kNN Retriever
Transformer Encoder
Encoded neighbours
E_1
E_2
Attend
Condition
Input tokens
C_1
C_2
C_3
X
EMB
ATTN
Q
K
V
CCA
FFW
READ
RETRO block (x L)
Chunked cross-attention (CCA)
Encoded neighbours
E_1
E_2
H_1
H_2
H_3
H
H_1^+
CA
H_2^+
CA
$CA(H_1^+, E_1)$
$CA(H_2^+, E_2)$
Attending chunks
CCA(H, E)

Method

Split decoder input into l - 1 attending chunks

maintain autoregressivity

(dependency propagation)

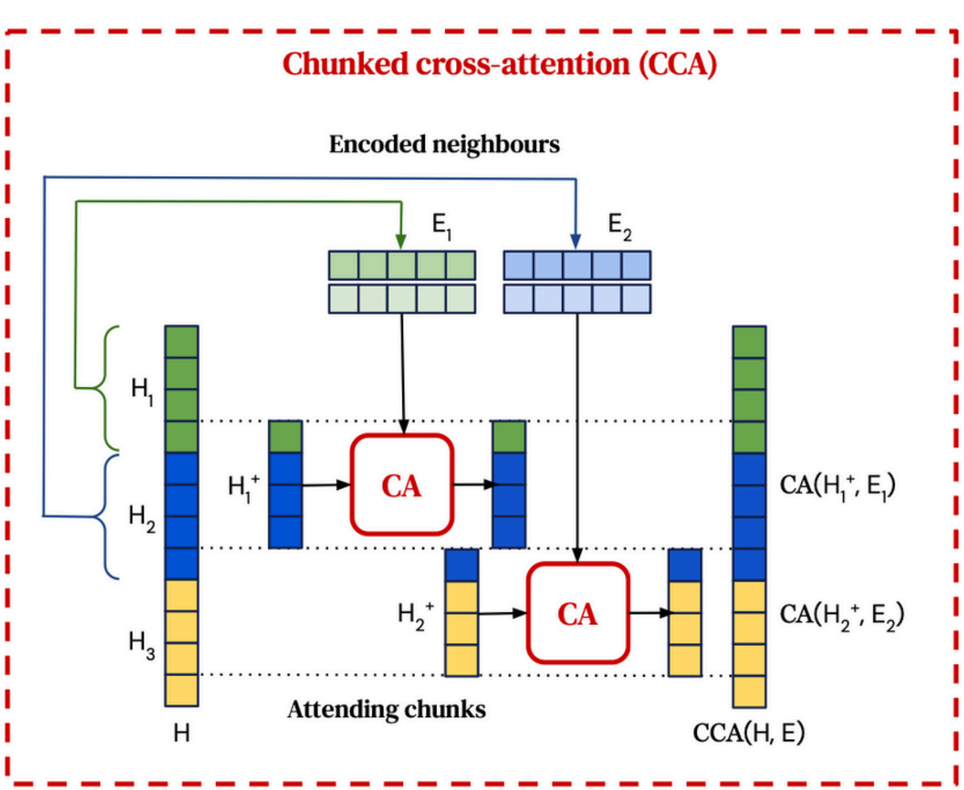
Chunked Cross-Attention

$$\left(H_u^+ \triangleq (h_{u\,m+i-1})_{i\in[1,m]} \in \mathbb{R}^{m\times d}\right)_{u\in[1,l-1]}$$

최종 출력 활성화 함수

$$\textsc{Cca}(H, E)_{u\,m+i-1} \triangleq \textsc{Ca}(h_{u\,m+i-1}, E_u)$$

$$\textsc{Ca}(h, Y) \triangleq \text{softmax}(YKQ^T h)YV,$$



H = input 활성화
E = retrieved neighbor
Q,K,V는 각각 쿼리(query), 키(key), 값(value)
Y · K = 검색된 데이터 Y의 키를 계산

Results

Datasets evaluated on

- C4
- Wikitext103
- Curation Corpus
- Lambada,
- The Pile

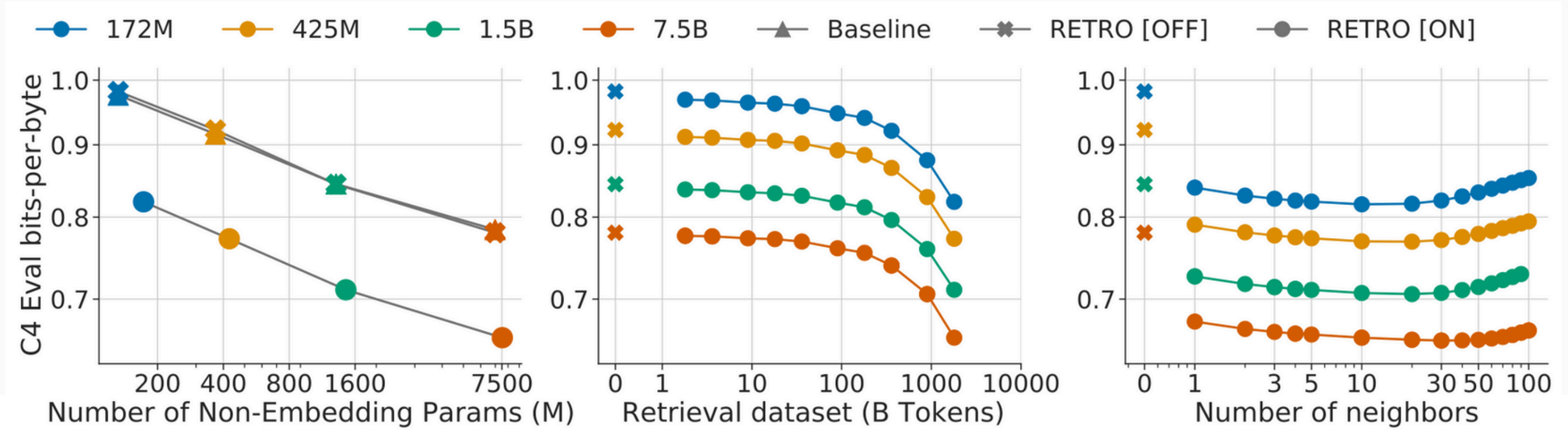
Metric : bits-per-byte(bpb)

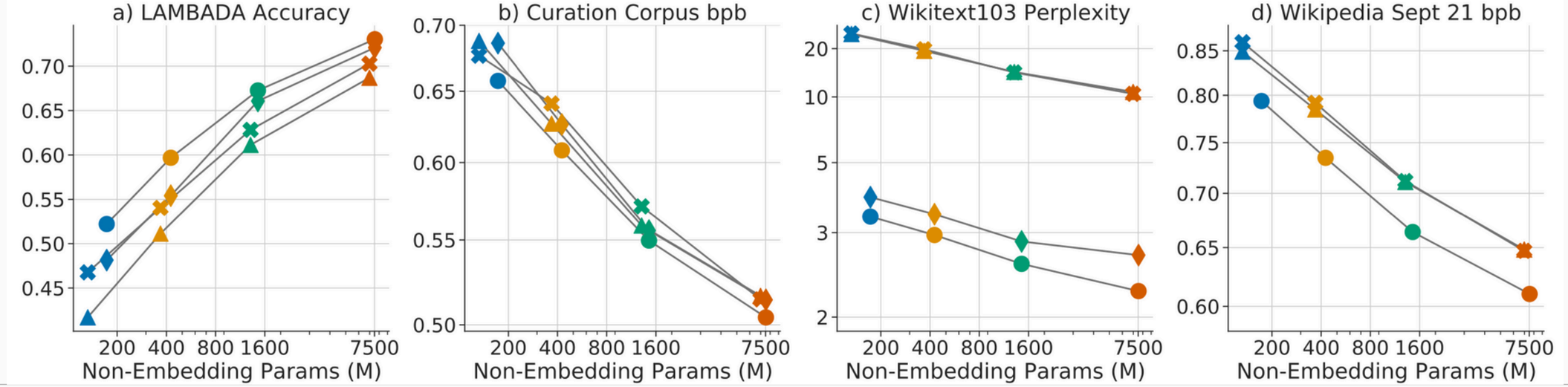
$$\text{perplexity} = 2^{\frac{|bytes|}{|words|} \cdot bpb}$$

tokenizer와 무관하게 모델의 언어 모델링 성능을 측정

BPB가 낮을수록 모델 성능이 좋음

Results

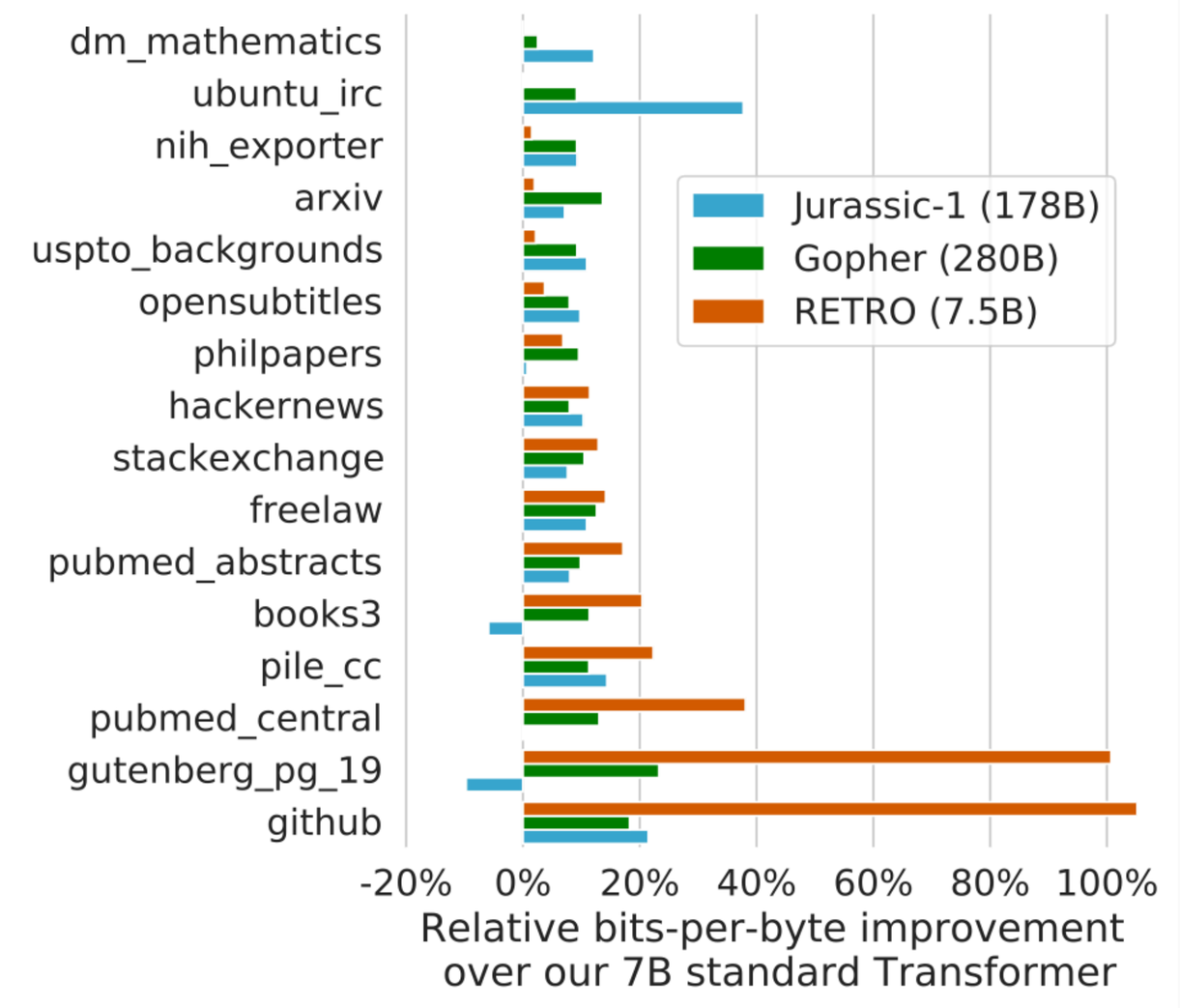

172M
425M
1.5B
7.5B
Baseline
RETRO [OFF]
RETRO [ON]
C4 Eval bits-per-byte
Number of Non-Embedding Params (M)
Retrieval dataset (B Tokens)
Number of neighbors


a) LAMBADA Accuracy
b) Curation Corpus bpb
c) Wikitext103 Perplexity
d) Wikipedia Sept 21 bpb
Non-Embedding Params (M)

Results

The Pile dataset

Jurrasic-1：178B
Gopher：280B



Results

Table 16. **Daniel Radcliffe**, from Wikitext103Valid, retrieval data from c4. The chunks C_2 and C_3 are almost entirely retrieved from neighbours $[N_1, F_1]$ and $[N_2, F_2]$ respectively, up to formatting differences, which dramatically reduces the loss for these tokens. This example illustrates that when training data leaks into evaluation sets despite deduplication, our RETRO model can directly exploit this leakage.

C_u colored by loss difference $L_{\text{RETRO[OFF]}} - L_{\text{RETRO}}$ $\leqslant -0.5$, $= 0$, $\geqslant 0.5$	C_u colored by LCP with $\text{RET}(C_u - 1)$ LCP $= 0, 1, 2, 3, 4, \geqslant 5$	$[N_u^1, F_u^1]$ colored by LCP with C_{u+1} LCP $= 0, 1, 2, 3, 4, \geqslant 5$	$[N_u^2, F_u^2]$ colored by LCP with C_{u+1} LCP $= 0, 1, 2, 3, 4, \geqslant 5$
= Daniel Radcliffe =Daniel Jacob R adcliffe (born 23 July 1989) is an English actor who rose to prominenc e as the title character in the Harr y Potter film series. He made his ac ting debut at 10 years of age in BBC One's 1999 television film David Co pperfield, followed by his cinematic debut	= Daniel Radcliffe = Daniel Jacob R adcliffe (born 23 July 1989) is an English actor who rose to prominenc e as the title character in the Harr y Potter film series. He made his ac ting debut at 10 years of age in BBC One's 1999 television film David Co pperfield, followed by his cinematic debut	Daniel Jacob Radcliffe (born 23 July 1989) is an English actor who rose to prominence as the title character in the Harry Potter film series. He made his acting debut at 10 years o f age in BBC One's 1999 television f ilm David Copperfield, followed by h is cinematic debut in 2001's The Tai lor of Panama. At age 11, he was cas t as Harry Potter in the first Harry Potter film, and starred in the ser ies for 10 years until the release o f the eighth and final film in 2011. Radcliffe began to branch out to s tage acting in 2007, starring in the London and New York productions of Equus, and	Daniel Jacob Radcliffe (born 23 July 1989) is an English actor who rose to prominence as the title character in the Harry Potter film series. He made his acting debut at 10 years o f age in BBC One's 1999 television m ovie David Copperfield, followed by his film debut in 2001's The Tailor of Panama. At age 11, he was cast as Harry Potter in the first Harry Pot ter film, and starred in the series for 10 years until the release of th e eighth and final film in 2011. Rad cliffe began to branch out to stage acting in 2007, starring in the Lond on and New York productions of Equus , and in the
in 2001's The Tailor of Panama. At age 11, he was cast as Harry Potter in the first Harry Potter film, and starred in the series for 10 years u ntil the release of the eighth and f inal film in 2011.Radcliffe began to branch out to stage acting in 200 7, starring in the London and New	in 2001's The Tailor of Panama. At age 11, he was cast as Harry Potter in the first Harry Potter film, and starred in the series for 10 years u ntil the release of the eighth and f inal film in 2011.Radcliffe began to branch out to stage acting in 200 7, starring in the London and New	in 2001's The Tailor of Panama. At age 11, he was cast as Harry Potter in the first Harry Potter film, and starred in the series for 10 years u ntil the release of the eighth and f inal film in 2011.Radcliffe began to branch out to stage acting in 200 7, starring in the London and New Yo rk productions of Equus, and in the 2011 Broadway revival of the musical How to Succeed in Business Without Really Trying. He starred in the 201 2 horror film The Woman in Black, an d played beat poet Allen Ginsberg in the 2013 independent film Kill Your Darlings.He has contributed to ma ny charities	of Panama. At age 11, he was cast a s Harry Potter in the first Harry Po tter film, and starred in the series for 10 years until the release of t he eighth and final film in 2011.R adcliffe began to branch out to stag e acting in 2007, starring in the Lo ndon and New York productions of Equ us, and in the 2011 Broadway revival of the musical How to Succeed in Bu siness Without Really Trying. He sta rred in the 2012 horror film The Wom an in Black, and played beat poet Al len Ginsberg in the 2013 independent film Kill Your Darlings. He has con tributed to many charities, includin g Demelza House Children's
York productions of Equus, and in t he 2011 Broadway revival of the musi cal How to Succeed in Business Witho ut Really Trying. He starred in the 2012 horror film The Woman in Black, and played beat poet Allen Ginsberg in the 2013 independent film Kill Y our ¡unk¿.He has contributed to ma ny charities,	York productions of Equus, and in t he 2011 Broadway revival of the musi cal How to Succeed in Business Witho ut Really Trying. He starred in the 2012 horror film The Woman in Black, and played beat poet Allen Ginsberg in the 2013 independent film Kill Y our ¡unk¿.He has contributed to ma ny charities,	York productions of Equus, and in t he 2011 Broadway revival of the musi cal How to Succeed in Business Witho ut Really Trying. He starred in the 2012 horror film The Woman in Black, and played beat poet Allen Ginsberg in the 2013 independent film Kill Y our Darlings.He has contributed to many charities, including Demelza H ouse Children's Hospice and The Trev or Project. He also made public serv ice announcements for the latter. In 2011, he was awarded the Trevor Pro ject's "Hero Award."Sources disagr ee about Radcliffe's personal wealth ; he was reported to have earned £1 million for the first Harry Potter	in the 2011 Broadway revival of the musical How to Succeed in Business Without Really Trying. He starred in the 2012 horror film The Woman in B lack, and played beat poet Allen Gin sberg in the 2013 independent film K ill Your Darlings. He has contribute d to many charities, including Demel za House Children's Hospice and The Trevor Project. He also made public service announcements for the latter . In 2011, he was awarded the Trevor Project's "Hero Award."
including ¡unk¿ Hospice Care for Ch ildren, and The Trevor Project for s uicide prevention among LGBTQ youth, which gave him its Hero Award in 20 11.= = Early life = =Radcliffe w as born in West London, England. He is the only child of Alan George Rad cliffe, a literary agent, and	including ¡unk¿ Hospice Care for Ch ildren, and The Trevor Project for s uicide prevention among LGBTQ youth, which gave him its Hero Award in 20 11.= = Early life = =Radcliffe w as born in West London, England. He is the only child of Alan George Rad cliffe, a literary agent, and		

2024.11.18

감사합니다

Pseudo-Lab | 신서현