LongEmbed: Mở rộng mô hình embedding cho truy xuất ngữ cảnh dài

Hứa Mạnh Tân, Nguyễn Tấn Tài, Nguyễn Nhựt Thành

Trường Đại học Công nghệ Thông tin - ĐHQGTPHCM

GVHD: TS. Nguyễn Thị Quý

Mục lục

Giới thiệu: Lý do dùng embedding cho ngữ cảnh dài?

- Nhiều ứng dụng thực tế (wikipedia dài, biên bản cuộc họp, báo cáo hành chính) cần truy xuất thông tin trong văn bản dài.
- Hầu hết embedding model hiện tại chỉ hỗ trợ 512 tokens (hoặc vài nghìn tokens) $\Rightarrow$ giới hạn khả năng truy xuất.
- Huấn luyện mô hình embedding dài từ đầu tốn kém (ví dụ để train BGE-M3 chạy 8k context cần 96 GPU A100).

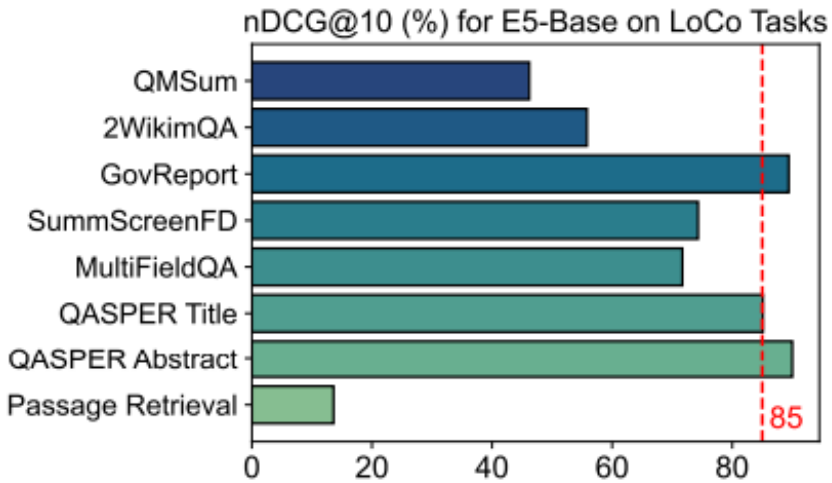
$\Rightarrow$ **Ý tưởng**: Mở rộng ngữ cảnh cho mô hình embedding hiện có mà không huấn luyện lại toàn bộ.

Giới thiệu: Đóng góp chính của bài báo

- **Xây dựng Benchmark LongEmbed:** Gồm 2 tác vụ tổng hợp và 4 tác vụ thực tế để đánh giá toàn diện khả năng truy xuất ngữ cảnh dài.
- **Khảo sát chiến lược mở rộng:** Đánh giá hiệu quả của các phương pháp không cần huấn luyện (training-free) và tinh chỉnh (fine-tuning).
- **Phân tích APE vs. RoPE:** Phát hiện *Rotary Position Embedding (RoPE)* có khả năng ngoại suy tốt hơn hẳn so với *Absolute Position Embedding (APE)*.
- **Đề xuất mô hình mới:** Giới thiệu các mô hình cải tiến (*E5-Base-4k*, *E5-RoPE-Base*) khắc phục hạn chế của các mô hình cũ.

LONGEMBED: Hạn chế của các benchmark hiện tại

- **BEIR**: độ dài văn bản ngắn (trung bình dưới 300 từ) ⇒ không phù hợp
- **LoCo**: có xu hướng thiên lệch, thông tin quan trọng nằm ở đầu văn bản ⇒ không phù hợp



Hình 1: Kết quả của E5-Base trên 8 tác vụ LoCo

LONGEMBED

Dataset	Domain	# Queries	# Docs	Avg. Query Words	Avg. Doc Words
		Real Tasks			
NarrativeQA	Literature, Film	10,449	355	9	50,474
QMSum	Meeting	1,527	197	71	10,058
2WikiMultihopQA	Wikipedia	300	300	12	6,132
SummScreenFD	ScreenWriting	336	336	102	5,582
		Synthetic Tasks			
Passkey	Synthetic	400	800	11	†
Needle	Synthetic	400	800	7	†

Hình 2: Tổng quát benchmark LONGEMBED

Phương pháp luận: APE vs RoPE

1. APE(Absolute Position Embedding)

- **Định nghĩa:** Chiến lược mã hóa vị trí chiếm ưu thế trong các mô hình embedding (theo kiến trúc BERT).
- **Cơ chế**: Chuyển đổi các ID vị trí thành các vector vị trí, rồi cộng vào vector nhúng, rồi đưa vào chuỗi các lớp Transformer

2. RoPE(Rotary Position Embedding)

- **Định nghĩa:** Chiến lược mã hóa phổ biến nhất trong kỷ nguyên LLM (LLaMA, QWen, Mistral...).
- **Cơ chế:** Mã hóa thông tin bằng ma trận xoay

Phương pháp luận: RoPE

- RoPE được áp dụng lên các vector Query (q) và Key (k) tại mỗi lớp của mạng Transformer.
- Với một vector ẩn h có chiều d và chỉ số vị trí m, RoPE thực hiện phép xoay trong không gian phức:

$$f(h, m) = [(h_0 + ih_1)e^{im\theta_0}, (h_2 + ih_3)e^{im\theta_1}, \dots]$$

  Trong đó θ_j là tần số góc quy định tốc độ xoay.
- Điểm chú ý (attention score) giữa một query tại vị trí m và key tại vị trí n phụ thuộc vào khoảng cách tương đối $(m - n)$ thông qua hàm:

$$a(q, k) = g(q, k, (m - n)\theta)$$

Phương pháp luận: Mở rộng APE-based model

- **Parallel Context Windows (PCW):**
  - **Cơ chế:** Chia văn bản dài (D) thành nhiều đoạn nhỏ (chunks), mã hóa các đoạn nhỏ song song, rồi tổng hợp kết quả embedding của các đoạn tạo thành embedding cho toàn văn bản
  - **Ưu điểm:** đơn giản, dễ cài đặt
  - **Nhược điểm:** mất đi sự tương tác ngữ nghĩa giữa các token nằm ở các đoạn khác nhau

Phương pháp luận: Mở rộng APE-based model

- **Position Reorganization (GP and RP):**
  - **Nguyên lý:** Tái sử dụng các ID vị trí ban đầu để gán cho văn bản dài hơn.
  - **Grouped Positions (GP):** Nhóm các vị trí lại với nhau. Công thức: $f_{gp}(pid) \rightarrow \lfloor pid/s \rfloor$. Ví dụ: Nếu $s = 2$, các vị trí 0 và 1 đều được gán ID là 0.
  - **Recurrent Positions (RP):** Tái sử dụng lặp lại các ID vị trí. Cho phép mô hình xử lý ngữ cảnh dài bằng cách "xoay vòng"các embeddings vị trí đã học.

Phương pháp luận: Mở rộng APE-based model

- **Linear Position Interpolation (PI):**
  - **Nguyên lý:** Tạo ra các embedding vị trí mới bằng cách nội suy tuyến tính từ các embedding có sẵn.
  - **Cách thực hiện:** Ánh xạ vị trí: $f_{pi}(pid) \rightarrow pid/s$. Mở rộng ma trận embedding vị trí ban đầu E_o sang E_t.
  - **Cách nội suy:**
    - Tại các điểm nguyên: $E_t[i \cdot s] = E_o[i]$.
    - Tại các điểm không nguyên: Sử dụng trung bình có trọng số giữa hai embedding lân cận.

Phương pháp luận: Mở rộng RoPE-based model

‘

1. **Self Extend (SE):**
   - Đặc điểm RoPE: Hoạt động trên các vector Query và Key tại mỗi lớp để mã hóa vị trí tương đối $\rightarrow$ Linh hoạt trong việc tổ chức lại vị trí.
   - Thay vì gán vị trí nhóm cho tất cả các token như APE , Self Extend giữ lại vị trí tương đối cho các token trong cửa sổ lân cận gần nhất (w).
2. **NTK-Aware Interpolation (NTK):**
   - **Vấn đề của PI**: PI co giãn đều mọi tần số, làm mất thông tin ở các tần số cao $\rightarrow$ Mô hình mất khả năng phân biệt thứ tự của các từ đứng gần nhau
   - **Nguyên lý:** : Thay vì chia tỷ lệ trực tiếp, NTK thực hiện biến đổi cơ số của tần số góc θ. Tần số cao thì giảm ít và ngược lại $\Rightarrow$ bảo toàn chính xác cục bộ, mở rộng khả năng nhìn thấy ngữ cảnh xa
   - **Mục đích:** Phân tán áp lực nội suy trên nhiều chiều dữ liệu.

Thực nghiệm

Setup:

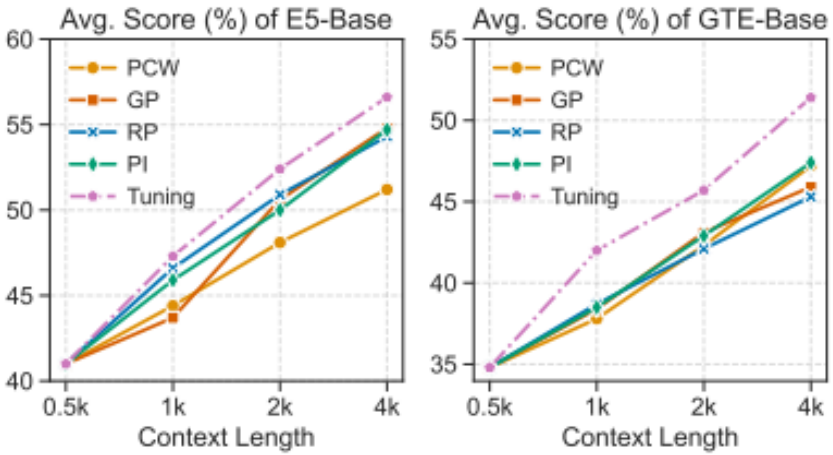
- Bộ dữ liệu: LongEmbed
- Mô hình được đánh giá: E5-Base, GTE-Base, BGE-Base, Contriever, E5-Mistral, Jina-V2, v.v.
- Các mô hình được dùng mở rộng ngữ cảnh:
  - Nhóm APE: E5-Base, GTE-Base.
  - Nhóm RoPE: E5-Mistral (4k context), E5-RoPE Base.

Thực nghiệm

Model	Param.	CTX Len.	Synthetic (Acc@1)		Real (nDCG@10)				Avg.
			Passkey	Needle	NQA	QMS	SFD	WQA	
512 Context Models									
$E5_{Base}$ (Wang et al., 2022)	110M	512	38.0	28.5	25.3	23.8	74.7	55.8	**41.0**
E5-$RoPE_{Base}$	110M	512	38.5	31.5	24.6	23.2	66.6	58.8	40.5
GTE_{Base} (Li et al., 2023)	110M	512	31.0	24.5	28.6	21.8	55.8	47.3	34.8
BGE_{Base} (Xiao et al., 2023)	110M	512	18.0	25.3	25.6	22.4	60.3	51.7	33.9
Contriever (Izacard et al., 2021)	110M	512	38.5	29.0	26.7	25.5	73.5	47.3	40.1
GTR_{Base} (Ni et al., 2022)	110M	512	38.5	26.3	26.5	18.3	63.7	52.2	36.5
$\geq$ 4k Context Models									
E5-Mistral (Wang et al., 2023b)	7B	4,096	71.0	48.3	44.6	43.6	96.8	82.0	**64.4**
Jina-V2 (Günther et al., 2023)	137M	8,192	50.3	54.5	37.9	38.9	93.5	74.0	58.2
Nomic-V1(Nussbaum et al., 2024)	137M	8,192	60.7	39.5	41.2	36.7	93.0	73.8	57.5
BGE-M3 (Chen et al., 2024)	568M	8,192	59.3	40.5	45.8	35.5	94.0	78.0	58.9
OpenAI-Ada-002	-	-	50.8	36.8	41.1	40.0	91.8	80.1	56.8
Our Extended Models									
$E5_{Base}$ + Tuning (4k)	110M	4,096	67.3	41.5	30.4	35.7	95.2	69.2	56.6
E5-$RoPE_{Base}$ + SelfExtend (4k)	110M	4,096	73.5	53.5	32.3	39.1	91.9	74.6	60.8
E5-Mistral + NTK (32k)	7B	32,768	**93.8**	**66.8**	**49.8**	**49.2**	**97.1**	**95.2**	**75.3**

Hình 3: Kết quả của các mô hình trên bộ LongEmbed

Thực nghiệm

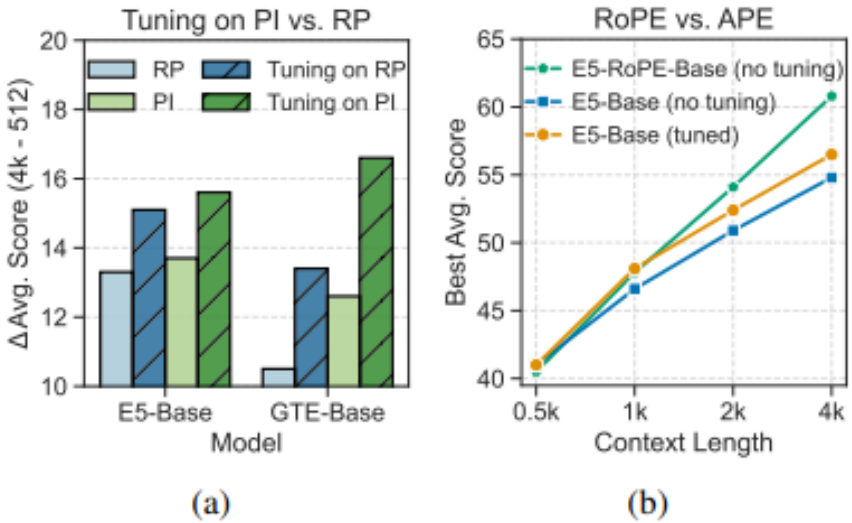


Hình 4: APE-based model

Model	Synthetic		Real				Avg.
	P	N	NQA	QMS	SFD	WQA	
E5-RoPE$_{Base}$	*38.5*	*31.5*	*24.6*	*23.2*	*66.6*	*58.8*	*40.5*
+PCW (4k)	42.5	50.8	25.1	34.9	**94.9**	69.3	52.9
+GP (4k)	68.0	38.8	25.9	30.9	85.8	65.8	52.5
+PI (4k)	68.3	36.0	25.9	30.8	84.9	65.3	51.9
+SE (4k)	**73.5**	**53.5**	**32.3**	**39.1**	91.9	**74.6**	**60.8**
+NTK (4k)	66.3	46.5	25.5	35.8	90.8	71.7	56.1
E5-Mistral	*71.0*	*48.3*	*44.6*	*43.6*	*96.8*	*82.0*	*64.4*
+PCW (32k)	63.5	49.5	**59.3**	51.3	**97.3**	91.2	68.7
+GP (32k)	81.0	48.8	37.0	42.9	90.6	88.1	64.7
+PI (32k)	89.8	48.5	37.8	40.4	76.8	63.0	59.4
+SE (32k)	90.8	52	49.3	48.7	97.2	**96.4**	72.4
+NTK (32k)	**93.8**	**66.8**	49.8	**49.2**	97.1	95.2	**75.3**

Hình 5: RoPE-based model

Thực nghiệm



Hình 6

Kết luận

Kết luận:

- Các chiến lược mở rộng không cần huấn luyện (training-free) có hiệu quả cao, tăng độ dài đầu vào lên nhiều lần trên benchmark LONGEMBED.
- Mô hình dựa trên RoPE vượt trội hơn hẳn so với APE trong việc mở rộng ngữ cảnh.

Hạn chế:

- Nghiên cứu hiện tại chủ yếu tập trung vào các phương pháp training-free.

Tài liệu tham khảo I


Dawei Zhu, Liang Wang, Nan Yang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li.
Longembed: Extending embedding models for long context retrieval.
arXiv preprint arXiv:2404.12096, 2024.