Báo cáo Demo: Extending Embedding Models for Long Context Retrieval

Hứa Mạnh Tân - 23521396
Nguyễn Tấn Tài - 23521376
Nguyễn Nhựt Thành - 23521451
Khoa khoa học máy tính

December 25, 2025

Bài báo cáo trình bày một thử nghiệm demo về bài toán **retrieval** sử dụng **Long Document Embeddings (Long Embed)**. Chúng tôi sử dụng bộ dữ liệu LEMBNeedleRetrieval, LEMBPasskeyRetrieval và đánh giá các mô hình embedding bằng các metric phổ biến như NDCG, MAP, Recall, Precision và MRR.

1 Giới thiệu

Trong Natural Language Processing (NLP), bài toán *Extending Embedding Models for Long Context Retrieval* đóng vai trò quan trọng khi tìm kiếm thông tin từ các tài liệu dài. Long Embed là phương pháp embedding các tài liệu dài, cho phép truy xuất chính xác các đoạn văn liên quan tới câu hỏi.

2 Bộ dữ liệu

Bộ dữ liệu sử dụng trong demo là **LEMBNeedleRetrieval**:

- **Corpus**: tập các document tổng hợp (synthetic) với thông tin hiếm..
- **Queries**: câu hỏi cần tìm thông tin cụ thể trong các document.
- **Qrels**:chỉ định document chứa thông tin đúng.

Bộ dữ liệu sử dụng trong demo là **LEMBPasskeyRetrieval**:

- **Corpus**:tập document tổng hợp (synthetic) có chứa "passkey" – thông tin mục tiêu
- **Queries**: các câu hỏi nhằm truy xuất passkey
- **Qrels**:chỉ định document chứa passkey đúng.

3 Mô hình

Thử nghiệm với các mô hình embedding:

- E5-Base-4k
- E5-RoPE-Base (không được đánh giá do giới hạn tài nguyên tính toán)

4 Đánh giá

Các metric dùng để đánh giá:

- **Precision@k (P@k)**: tỷ lệ các tài liệu truy xuất trong top-k là đúng.

$$P@k = \frac{|\text{relevant documents in top-}k|}{k}$$

- **Recall@k (R@k)**: tỷ lệ các tài liệu đúng được tìm thấy trong top-k.

$$R@k = \frac{|\text{relevant documents in top-}k|}{|\text{total relevant documents}|}$$

- **Mean Average Precision@k (MAP@k)**: trung bình của Precision tại các vị trí có tài liệu đúng.

$$MAP@k = \frac{1}{|Q|} \sum_{q \in Q} AP@k(q), \quad AP@k(q) = \frac{1}{m_q} \sum_{i=1}^{k} P(i) \cdot rel(i)$$

  Trong đó m_q là số tài liệu đúng cho query q, $rel(i) = 1$ nếu tài liệu thứ i đúng, ngược lại 0.

- **Normalized Discounted Cumulative Gain@k (NDCG@k)**: đo chất lượng xếp hạng, ưu tiên các kết quả đúng ở vị trí cao.

$$NDCG@k = \frac{DCG@k}{IDCG@k}, \quad DCG@k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

- **Mean Reciprocal Rank@k (MRR@k)**: trung bình nghịch đảo của vị trí xuất hiện đầu tiên của kết quả đúng.

$$MRR = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{\text{rank of first relevant doc for } q}$$

5 Kết quả demo

- Trong các thí nghiệm sau, nhóm cố gắng giữ cố định kiến trúc mô hình và chỉ thay đổi cách xử lý positional embedding nhằm đánh giá tác động của long embedding đến hiệu quả retrieval.

- Đánh giá mô hình E5-Base-4k trên đầu vào dài 1024 token, so với thiết lập 1024 token của model với tập dữ liệu LEMBNeedleRetrieval

Metric	@1	@5	@10	@100	@1000
NDCG	0.66	0.80	0.82	0.83	0.83
MAP	0.66	0.77	0.78	0.78	0.78
Recall	0.66	0.92	0.98	1.00	1.00
MRR	0.66	0.77	0.78	0.78	0.78
Precision	0.66	0.18	0.09	0.01	0.001

Table 1: Chạy bình thường

Metric	@1	@5	@10	@100	@1000
NDCG	0.70	0.82	0.84	0.85	0.85
MAP	0.70	0.79	0.80	0.80	0.80
Recall	0.70	0.94	0.98	1.00	1.00
MRR	0.70	0.79	0.80	0.80	0.80
Precision	0.70	0.18	0.09	0.01	0.001

Table 2: mở rộng embedding

- Đánh giá mô hình E5-Base-4k trên đầu vào dài 512 token, so với thiết lập 1024 token của model với tập dữ liệu LEMBPasskeyRetrieval

Metric	@1	@5	@10	@100	@1000
NDCG	0.70	0.85	0.86	0.86	0.86
MAP	0.70	0.81	0.82	0.82	0.82
Recall	0.70	0.96	1.00	1.00	1.00
MRR	0.70	0.81	0.82	0.82	0.82
Precision	0.70	0.20	0.10	0.01	0.001

Table 3: Chạy bình thường

Metric	@1	@5	@10	@100	@1000
NDCG	1.00	1.00	1.00	1.00	1.00
MAP	1.00	1.00	1.00	1.00	1.00
Recall	1.00	1.00	1.00	1.00	1.00
MRR	1.00	1.00	1.00	1.00	1.00
Precision	1.00	0.20	0.10	0.01	0.001

Table 4: mở rộng embedding

6 Kết luận

Qua các thí nghiệm trên hai tập dữ liệu LEMBNeedleRetrieval và LEMB-PasskeyRetrieval, có thể nhận thấy rằng việc mở rộng positional embedding giúp mô hình E5-Base-4k cải thiện đáng kể hiệu quả truy xuất thông tin trong các kịch bản xử lý văn bản dài.

Cụ thể, khi độ dài ngữ cảnh đầu vào tăng lên (512 và 1024 token), phương pháp mở rộng embedding cho kết quả tốt hơn so với thiết lập positional encoding gốc, thể hiện qua sự cải thiện nhất quán ở các chỉ số NDCG, MAP và MRR trên nhiều mức top-k. Điều này cho thấy mô hình duy trì khả năng biểu diễn và tập trung vào thông tin quan trọng tốt hơn khi xử lý các ngữ cảnh dài.

Ngoài ra, kết quả trên tập LEMBPasskeyRetrieval cho thấy long embedding đặc biệt hiệu quả trong các bài toán yêu cầu truy xuất chính xác một

thông tin quan trọng (passkey) nằm sâu trong văn bản dài. Trong khi đó, trên LEMBNeedleRetrieval, cải thiện diễn ra ở mức độ vừa phải nhưng vẫn ổn định, phản ánh khả năng tổng quát hóa tốt hơn của mô hình khi mở rộng positional embedding.

Tổng thể, demo này khẳng định rằng long embedding không làm thay đổi kiến trúc hay năng lực học ngôn ngữ cốt lõi của mô hình, mà đóng vai trò quan trọng trong việc duy trì tính nhất quán của biểu diễn embedding, từ đó nâng cao hiệu quả truy xuất trong các kịch bản ngữ cảnh dài.