NEAR^2: A Nested Embedding Approach to Efficient Product Retrieval and Ranking

Shenbin Qian[1], Diptesh Kanojia[1], Samarth Agrawal[2], Hadeel Saadany[3], Swapnil Bhosale[1], Constantin Orasan[1], Zhe Wu[2]

[1]University of Surrey & [2]eBay Inc & [3]Birmingham City University

{s.qian, d.kanojia, s.bhosale, c.orasan}@surrey.ac.uk, {samagrawal, zwu1}@ebay.com, hadeel,saadany@bcu.ac.uk

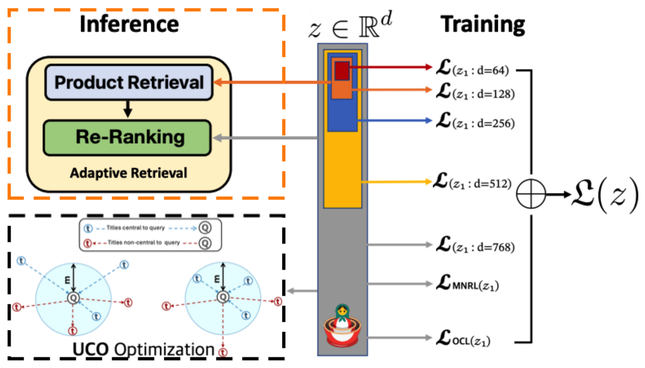
Introduction

E-commerce information retrieval (IR) systems struggle to **simultaneously achieve high accuracy in interpreting complex user queries and maintain efficient processing of vast product catalogs**. The **dual challenge** lies in precisely matching user intent with relevant products while managing the computational demands of real-time search across massive inventories. We propose a **N**ested **E**mbedding **A**pproach to product **R**etrieval and **R**anking (NEAR^2) approach, which can achieve **efficient product retrieval and ranking** using **much smaller embedding sizes** of encoder-based Transformer models. NEAR^2 can improve model performance on these challenging datasets using significantly smaller embedding dimension sizes. Our contributions can be summarized as follows:

- We propose NEAR^2, a nested embedding approach, which can achieve up to 12$\times$ efficiency in embedding size and 100$\times$ smaller in memory usage during inference while introducing no extra cost in training.
- We evaluate NEAR^2 on four different test sets that contains various types challenging queries. Evaluation results show that our approach achieves an improved performance using a much smaller embedding dimension compared to any existing models.
- We conduct ablative experiments on different encoder-based models fine-tuned using different IR loss functions. We find that NEAR^2 is robust to different IR losses or loss combinations for continued fine-tuning.
- We perform a qualitative analysis on retrieved product titles using challenging queries. Our analysis re-affirms the superior performance of our approach and reveals that the similarity scores from NEAR^2 models are more reliable than those of baseline models.

Methodology

Matryoshka Representation Learning MRL develops representations with diverse capacities within the same higher-dimensional vector by explicitly optimizing sets of lower-dimensional vectors in a nested manner, as illustrated in Figure 1.



The initial $m-$dimensions of the Matryoshka representation, where $m \in M$, the set of nested representation sizes, form a compact and information-dense vector that matches the accuracy of a separately trained $m-$dimensional representation, but requires no extra training effort. The MRL loss is formally defined in Equation 1, where L_{task} is the loss for downstream tasks such as the cross-entropy loss for classification tasks. $f_m(x)$ is the output of the m-th nested embedding representation, and c_m is the importance weight for the m-th embedding representation. For our product retrieval and ranking task, we set the multiple negative ranking loss (MNRL) as our L_{task}.

$$L_{MRL} = \sum_{m \in M} c_m L_{task}(f_m(x), y) \quad (1)$$

Multiple Negative Ranking Loss MNRL measures the difference between relevant (positive) and irrelevant (negative) examples associated with a given query. This technique ensures a clear separation by reducing the distance between the query and positive samples while increasing the distance from negative samples. Using multiple negative examples enhances the model's ability to discern varying levels of irrelevance, refining its optimization. The MNRL objective function is formulated as follows:

$$MNRL = \sum_{i=1}^{P} \sum_{j=1}^{N} max(0, f(q, p_i) - f(q, n_j) + margin) \quad (2)$$

Pre-trained Language Models We initially leveraged **BERT**, a publicly available pre-trained encoder Transformer model. For our specific use case in e-commerce, we also employed **eBERT**, a proprietary multilingual language model pre-trained internally at eBay. This custom model was pre-trained on a corpus of approximately three billion product titles, supplemented by data from general domain sources like Wikipedia and RefinedWeb. Expanding our experimental approach, we also incorporated **eBERT-siam**, a fine-tuned variant of eBERT using a Siamese network architecture. This model aims to generate semantically aligned embeddings for item titles, making it particularly effective for similarity-based search and retrieval tasks. Consistent across all models, we maintained a uniform architectural design of 12 layers with a dimension size of 768.

User-intent Centrality Optimized (UCO) Models Saadany et al. (2024) show how current IR systems have problems in achieving user-centric product retrieval and ranking due to implicit or alphanumeric queries. They curated a dataset with user-intent centrality scores and proposed a few models optimized for user-intent using an MNRL loss for retrieval and ranking, and an online contrastive loss (OCL) for user-intent centrality.
They applied the two losses in a transfer learning setup for eBERT and eBERT-siam models, and performed fine-tuning for centrality classification. Their results indicate that the UCO models achieve an improved performance for retrieval and ranking. To improve model efficiency and meanwhile leverage optimized performance of the UCO models, we continued training them using NEAR^2 for both **eBERT-UCO** and **eBERT-siam-UCO** models.

Data

We utilized eBay's internal graded relevance (IGR) datasets to train our nested embedding representation. These datasets comprise user search queries alongside the product titles retrieved on the platform. They are annotated by humans following specific guidelines to generate two types of buyer-focused relevance labels.

- The first is a relevance ranking scheme, where query-title pairs are assigned a rank from (1) Bad, (2) Fair, (3) Good, (4) Excellent, to (5) Perfect. A "Perfect" rating signifies an exact match between the query and title, indicating high confidence that the user's needs are fully met, whereas a "Bad" rating indicates no alignment between the query and the product title. This ranking methodology aligns with previous studies.
- The second annotation type is a binary centrality score, derived through majority voting among multiple annotators, indicating whether a product aligns with the user's expressed query intent. Centrality scoring differs from relevance ranking in that it assesses whether an item is an outlier or unexpected in the retrieval set versus being a core match to user expectations.
- To compare the results of our approach with those reported in Saadany et al. (2024), we utilized the Common Queries (**CQ**), CQ Balanced (**CQ-balanced**), CQ Common String (**CQ-common-str**), and CQ Alphanumeric (**CQ-alphanum**) test sets proposed in their paper. The CQ test set was constructed using queries with both positive (relevancy ¿ 3) and negative (relevancy ¡ 3) titles, resulting in a dataset skewed toward positive pairs due to the nature of e-commerce data collection. To address this imbalance, a new version, CQ-balanced, was created with approximately equal numbers of positive and negative query-title pairs. The CQ-common-str set was derived by selecting queries where the exact query string appeared in both positive and negative titles, ensuring a strong correlation between relevance scores (both graded relevance and binary centrality). Finally, CQ-alphanum was created to include only query-title pairs containing alphanumeric characters, allowing for a more focused evaluation.

Results

Delta in precision, recall, NDCG, and MRR at k on all the test sets for different models fine-tuned using **NEAR^2 at 64 dimensions** of the entire embedding size (768)

Model	Precision@k			Recall@k			NDCG@k			MRR@k
	3	5	10	3	5	10	3	5	10	10
CQ test										
eBERT-siam	+11.80%	+11.79%	+11.49%	+9.99%	+9.72%	+9.07%	+11.50%	+11.23%	+10.65%	+9.06%
eBERT-UCO	+2.98%	+3.28%	+3.90%	+3.12%	+2.99%	+3.16%	+3.27%	+3.34%	+3.47%	+3.03%
eBERT-siam UCO	+2.82%	+2.75%	+3.16%	+2.72%	+2.45%	+2.50%	+2.91%	+2.77%	+2.80%	+2.58%
CQ-balanced test										
eBERT-siam	+8.85%	+8.45%	+7.31%	+8.85%	+8.43%	+7.28%	+10.28%	+10.03%	+9.56%	+10.48%
eBERT-UCO	+3.19%	+2.87%	+2.42%	+3.15%	+2.81%	+2.41%	+3.36%	+3.19%	+3.03%	+3.25%
eBERT-siam UCO	+2.77%	+2.45%	+2.09%	+2.75%	+2.48%	+2.05%	+3.06%	+2.93%	+2.77%	+3.01%
CQ-common-str test										
eBERT-siam	+6.62%	+4.90%	+3.00%	+6.59%	+4.84%	+3.01%	+8.57%	+7.70%	+6.99%	+8.51%
eBERT-UCO	+1.69%	+1.53%	+0.81%	+1.68%	+1.51%	+0.86%	+1.56%	+1.48%	+1.27%	+1.38%
eBERT-siam UCO	+1.49%	+1.22%	+0.81%	+1.48%	+1.18%	+0.83%	+1.86%	+1.72%	+1.59%	+1.85%
CQ-alphanum test										
eBERT-siam	+5.82%	+5.84%	+6.15%	+4.70%	+4.59%	+5.01%	+5.52%	+5.40%	+5.35%	+4.41%
eBERT-UCO	+3.64%	+3.75%	+3.92%	+3.61%	+3.55%	+3.60%	+3.30%	+3.33%	+3.40%	+2.57%
eBERT-siam UCO	+2.32%	+2.13%	+2.68%	+2.15%	+1.87%	+2.36%	+2.33%	+2.13%	+2.38%	+2.28%

- Comparing results upon using NEAR^2 *vs* existing models, we find that our approach remarkably improves performance on all test sets for all models in Methodology, even using embeddings with a dimension size of 64, which is 12$\times$ smaller in size and more than 100$\times$ smaller in memory usage than the full model (see the following table).

Embedding Size	Memory Usage (MB)
768	398.03
512	2.77
256	4.09
128	0.55
64	1.56

- When comparing results of different dimension sizes from the largest (768) to the smallest (64), as shown in the following table for the **CQ test** set, we discover that the drop in performance is not significant. Embeddings of some smaller dimensions are even slightly better than larger ones. For example, the performance of the eBERT-siam model using NEAR^2 at dimension 512 is slightly better than 768 for precision, NDCG and MRR. This is also true for other models such as BERT, eBERT and eBERT-UCO, which further indicates the effectiveness of our approach for product retrieval and ranking.

Delta in precision, recall, NDCG, and MRR at k on **CQ test** set for different models fine-tuned using **NEAR^2** for all dimension sizes

Model	Dimension	Precision@5	Recall@5	NDCG@5	MRR@10
eBERT-siam	768	+13.33%	+11.77%	+13.10%	+10.20%
	512	+13.35%	+11.87%	+13.16%	+10.30%
	256	+13.26%	+11.68%	+13.05%	+10.19%
	128	+13.10%	+11.37%	+12.80%	+10.16%
	64	+11.79%	+9.72%	+11.23%	+9.06%
eBERT-UCO	768	+4.25%	+4.04%	+4.34%	+3.50%
	512	+4.27%	+3.97%	+4.37%	+3.57%
	256	+4.18%	+3.83%	+4.23%	+3.49%
	128	+3.86%	+3.52%	+3.97%	+3.42%
	64	+3.28%	+2.99%	+3.34%	+3.03%
eBERT-siam-UCO	768	+3.85%	+3.75%	+3.82%	+3.05%
	512	+3.85%	+3.72%	+3.81%	+3.00%
	256	+3.62%	+3.47%	+3.61%	+2.96%
	128	+3.46%	+3.27%	+3.46%	+2.96%
	64	+2.75%	+2.45%	+2.77%	+2.58%

Conclusion

This paper propose a nested embedding approach for efficient product retrieval and ranking, NEAR^2, achieving up to 12$\times$ efficiency in embedding size and 100$\times$ smaller in memory usage during inference, without any increase in pre-training costs. Tested across diverse datasets, our method outperforms existing models with smaller embedding dimensions, demonstrating its robustness across challenging evaluation sets, and with efficiency.

References H. Saadany, S. Bhosale, S. Agrawal, D. Kanojia, C. Orasan, and Z. Wu. 2024. Centrality-aware Product Retrieval and Ranking. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 215–224, Miami, Florida, US. Association for Computational Linguistics.