On the Theoretical Advantages of Bilinear Similarities in Dense Retrieval

Shubham Chatterjee[1][0000−0002−6729−1346]

Missouri University of Science and Technology, Rolla, MO, USA
shubham.chatterjee@mst.edu

Abstract. We present a theoretical and empirical study of bilinear similarity functions in neural IR, showing that they are strictly more expressive than dot-product and weighted dot-product (WDP) models under fixed embeddings. We prove this separation formally and illustrate it with the Structured Agreement Ranking Task, where a simple rank-2 bilinear model achieves 100% accuracy while all WDP models fail. This highlights the importance of modeling feature interactions for conditional relevance. On MS MARCO, low-rank bilinear models significantly outperform dot-product baselines: a rank-32 model triples performance (MRR@10: 0.090 vs. 0.031), and rank-128 approaches a 4× gain. These results offer a principled and practical case for using low-rank bilinear models in dense retrieval. **Code:** https://github.com/shubham526/bilinear-projection-theory

1 Introduction

Dense retrieval maps queries and documents into a shared embedding space and assesses relevance via a dot product. This setup, fundamental to dual encoder architectures like DPR [10], enables efficient approximate nearest-neighbor (ANN) search [20, 14, 13, 6]. Dot product-based models such as E5 [19] and Instructor [18] have become the de facto standard, particularly in RAG pipelines.

Despite widespread adoption, the inherent limitations of dot product similarity have become increasingly apparent. The dot product assumes a simplistic linear interaction between embeddings, thus restricting relevance modeling to effectively rank-1 interactions [4]. This "low-rank bottleneck" significantly constrains the capacity to capture complex semantic relationships, a problem that is especially pronounced when using general-purpose pre-trained embeddings not specifically optimized for dot-product interactions [4, 2, 21].

Recognizing these limitations, researchers have explored bilinear similarity functions of the form $s(q, d) = q^\top W d$, which offer greater expressive power through a learnable interaction matrix W [9, 16]. While the principle that higher-rank models are more expressive is well-known, its direct consequences for dense retrieval have not been fully explored. Key questions remain for practitioners: when does this additional expressiveness lead to meaningful gains, how can we formalize the limitations of dot-product models in a retrieval context, and how can we manage the complexity of bilinear models in a principled way?

Contributions. This paper addresses these questions by connecting established theory with rigorous empirical validation, providing a practical framework for leveraging bilinear similarities in dense retrieval.

1. **A Concrete Illustration of Expressiveness Limits.** While it is known that bilinear models are more expressive, we provide a clear, intuitive demonstration of this gap. We introduce the *Structured Agreement Ranking Task*, a synthetic task designed to model conditional relevance. On this task, a simple rank-2 bilinear model achieves perfect accuracy, whereas all weighted dot-product models fail universally (100% vs. 0%). This provides a concrete failure case that highlights a qualitative, not just quantitative, advantage.
2. **Validating Theoretical Gains with Real-World Performance.** We demonstrate that this theoretical advantage translates into substantial practical improvements. On the MS MARCO passage ranking benchmark using fixed BERT embeddings, low-rank bilinear models yield large gains over dot-product: a rank-32 model achieves a nearly 3x higher MRR@10 (0.090 vs. 0.031), and a rank-128 model approaches a 4x improvement (0.127 vs. 0.031).
3. **Principled, Theory-Guided Rank Selection.** We show how to make bilinear models efficient without resorting to heuristic guesswork. By applying established low-rank approximation theory (i.e., the Eckart-Young-Mirsky theorem), we connect the model's approximation error directly to its singular value spectrum. Our experiments validate this, showing a strong correlation between retrieval performance and discarded singular values. This provides a practical, theory-guided method for choosing a rank that balances performance and efficiency.

2 Related Work

The reliance on dot-product similarity in dense retrieval has faced increasing scrutiny as empirical evidence reveals its limitations. Early work in cross-modal retrieval demonstrated the benefits of bilinear formulations [9, 16], while recent theoretical analysis by Ding et al., 2025 [4] formalized the "low-rank bottleneck," proving that dot-product similarity restricts systems to rank-1 interactions. This limitation is particularly pronounced with general purpose embeddings, paralleling similar constraints identified in attention mechanisms [2]. Researchers have explored various solutions: Multi-vector models [8, 11] use multiple query representations, and the Mixture-of-Logits framework [4] demonstrates that weighted combinations of dot products can represent arbitrarily high-rank interactions, though at significant computational cost.

Bilinear similarities offer a structured middle ground between simple dot products and computationally expensive cross-attention mechanisms. The concept has deep roots in machine learning, from Factorization Machines in recommendation systems [17] to bilinear pooling in multimodal tasks [12]. Neural IR models such as DRMM [5] and MatchPyramid [15] have long used interaction matrices at the term level, but applying these concepts to global dense embeddings remains underexplored. The computational challenge of full bilinear mod-

els requiring storage of $O(n^2)$ has driven interest in low-rank approximations, inspired by the successes in LSI [3] and LoRA [7]. However, theoretical understanding of expressiveness hierarchies under fixed embeddings, concrete failure cases for linear similarity, and principled design methods for bilinear models remained gaps that our work addresses through rigorous theoretical analysis and empirical validation.

3 Expressiveness: Bilinear vs. Dot-Product Similarities

The core is: *with fixed query and document embeddings, what ranking patterns can different similarity functions express?* We begin by formalizing the widely understood principle that bilinear similarities are more expressive than standard dot-product similarity. This setting reflects the common practice of using frozen pre-trained embeddings, where the similarity function becomes the primary tool for task adaptation. Understanding the precise limits of these functions is essential for principled retrieval system design and provides the foundation for the more complex, retrieval-specific scenarios we analyze in later sections.

Dot-product similarity measures how much a query and document align in their feature space, much like comparing two lists by counting overlapping items. However, relevance often depends not just on which features are present, but on how they interact. For example, consider the query "machine learning healthcare applications." A document focused heavily on machine learning but barely mentioning healthcare may score similarly to one that emphasizes healthcare but says little about machine learning. In contrast, a third document that meaningfully connects both topics, such as the use of machine learning in clinical settings, would be intuitively more relevant. Dot-product similarity may not distinguish this, as it assigns scores based solely on total feature overlap. Formally, it computes $q^\top d$, projecting one vector onto another in the original embedding space. In contrast, the bilinear similarity is computed as $q^\top W d$, where the document vector is first transformed by the matrix W before projection. This allows the model to highlight relevant feature interactions and ignore irrelevant ones, tailoring similarity to the task.

To formalize this analysis, we introduce precise definitions for the fixed-embedding retrieval setting.

Definition 1 (Problem Setting). *We assume fixed embedding functions $\phi_q : Q \to \mathbb{R}^n$ and $\phi_d : D \to \mathbb{R}^n$, where Q is the set of queries, D is the set of documents, and n is the embedding dimension. For a weight matrix $W \in \mathbb{R}^{n\times n}$, the bilinear similarity is defined as: $s_W(q,d) := \phi_q(q)^\top W \phi_d(d)$. The standard dot-product similarity is the special case where W is the identity matrix I_n, i.e., $s_{dot}(q,d) := s_{I_n}(q,d)$. A weighted dot-product (WDP) similarity, which we use as a key comparator, is the special case where W is a diagonal matrix, i.e., $s_v(q,d) := \phi_q(q)^\top diag(v)\phi_d(d)$, for some weight vector $v \in \mathbb{R}^n$.*

Definition 2 (Ranking Equivalence). *For a fixed query $q \in Q$ and a finite set of candidates $\mathcal{D}_q \subseteq D$, two similarity functions* ***induce the same ranking*** *if they impose identical total orders on $\mathcal{D}_q$.*

The set of functions representable by the dot product is a subset of those representable by a bilinear model. To show that this inclusion is *strict*, we can construct a simple case where the dot product fails. Consider 2D embeddings with query $\phi_q(q) = [1, 1]^\top$ and two documents $\phi_d(d_1) = [1, 0]^\top$ and $\phi_d(d_2) = [0, 1]^\top$. The dot product yields a tie: $s_{\text{dot}}(q, d_1) = 1$ and $s_{\text{dot}}(q, d_2) = 1$. However, a simple bilinear model using a diagonal weight matrix $W_* = \text{diag}(2, 1)$ breaks this tie: $s_{W_*}(q, d_1) = 2$ while $s_{W_*}(q, d_2) = 1$. This illustrates that even a simple WDP model can express feature preferences that the standard dot product cannot. While this example is straightforward, it establishes the foundation for asking a deeper question: are there ranking tasks that *all* diagonal models fail, which a non-diagonal bilinear model can solve? We address this in Section 4.

This expressiveness advantage has several important practical consequences:

1. Any dot-product system can be viewed as an un-trained bilinear model with $W = I$, providing a natural upgrade path by simply making W learnable.
2. The constructive example suggests that even simple (e.g., diagonal or low-rank) weight matrices can yield significant expressive power over the standard dot product.
3. With frozen pre-trained embeddings becoming standard practice, learning a similarity function like W offers a parameter-efficient method for task adaptation without expensive embedding re-training.
4. Although a full rank W has $O(n^2)$ complexity, in Section 5 we show that low-rank approximations are highly effective, making bilinear models computationally manageable with modern hardware.

However, the fixed embedding assumption is crucial. When embeddings can be learned end-to-end, the advantage of a learnable similarity function may be less pronounced, as the embedding function itself can learn to compensate for the limitations of the dot product.

4 When Bilinear Strictly Outperforms Linear

To demonstrate a fundamental architectural limitation of weighted dot-product (WDP) models, we introduce the **Structured Agreement Ranking Task**. This synthetic task is designed to show that no *single, universal* WDP model (with one fixed weight vector v) can generalize to handle tasks requiring conditional relevance, even though a specialist WDP model could be tuned for any single instance. We show that the bilinear formulation, in contrast, is sufficiently expressive to solve the general task, highlighting a clear qualitative separation between the model classes. The task uses hypercube embeddings in $\{-1, +1\}^n$, where agreement is simple multiplication ($q_k d_k = \pm 1$). For any query q and a pair of "critical" indices $I_0 = \{i_1, i_2\}$, we create four documents: two that agree with q on the features in I_0 (d_1, d_2) and two that disagree (d_3, d_4). The goal is to produce the ranking $\{d_1, d_2\} \succ \{d_3, d_4\}$.

Theorem 1 (Bilinear vs. Weighted Dot-Product Separation). *For the Structured Agreement Ranking Task:*

(i) ***Bilinear suffices:*** *For any critical feature set $I_0 = \{i_1, i_2\}$, a simple rank-2 bilinear model solves the task perfectly for all queries q.*
(ii) ***WDP fails universally:*** *For dimension $n \geq 3$, no single, universal weight vector v can solve the task for all possible choices of the critical pair I_0.*

Proof (Proof Sketch). **(i) Bilinear suffices:** A rank-2 matrix W_{I_0} that only considers the two critical indices gives a score of $+2$ to agreeing documents and -2 to disagreeing ones, achieving perfect separation.

(ii) WDP fails universally: A proof by contradiction shows that for a universal weight vector v to exist, the sum of weights for any critical pair must exceed the sum of all other weights. By considering two different but overlapping critical pairs (e.g., $\{j_1, j_2\}$ and $\{j_1, j_3\}$), this requirement leads to the contradictory conditions that $v_{j_2} > v_{j_3}$ and, by symmetry, $v_{j_3} > v_{j_2}$. Thus, no such universal vector can exist. The full, corrected proof is in the appendix.

Experimental Validation. We validate these findings experimentally. On the synthetic task, the bilinear model achieves 100% accuracy, while a trained WDP model scores 0% on unseen critical pairs, confirming the architectural failure. This theoretical advantage translates to significant real-world gains: on the MS MARCO passage ranking task [1], a rank-128 bilinear model improves MRR@10 by over 4x compared to the dot-product baseline when using fixed `bert-base-uncased` embeddings (0.127 vs 0.031). However, when using embeddings already fine-tuned for dot-product (`all-mpnet-base-v2`), the bilinear model offers no benefit, confirming its value is greatest in frozen-embedding scenarios.

Takeaway. WDP models are architecturally incapable of modeling the conditional feature relevance required by many retrieval tasks. The bilinear formulation resolves this, and its theoretical advantage leads to substantial performance gains on real-world benchmarks. This makes low-rank bilinear similarity a powerful and principled upgrade for retrieval systems, especially when using general-purpose or frozen embeddings.

5 Analysis of Low-Rank Bilinear Approximations

Although bilinear models offer superior expressiveness, their $O(n^2)$ complexity from the full interaction matrix W limits practical deployment. To address this, we apply low-rank approximations grounded in the Eckart–Young–Mirsky theorem, which provides optimal compression via singular value decomposition (SVD). By connecting this theory with dense retrieval, we translate abstract approximation bounds into concrete performance gains.

Theorem 2 (Pointwise Error Bound for Low-Rank Approximation). *For any matrix $W \in \mathbb{R}^{n \times n}$ and its optimal rank-r approximation W_r, the error in the bilinear similarity score is bounded by the first discarded singular value:*

$$|s_W(q, d) - s_{W_r}(q, d)| \leq \|q\|_2 \|d\|_2 \sigma_{r+1}(W)$$

where $\sigma_{r+1}(W)$ is the $(r+1)$-th largest singular value of W. This bound is tight.

Proof (Proof Sketch). The error in the similarity score is $q^\top(W - W_r)d$. By substituting the SVD of the residual matrix $W - W_r$, the error can be expressed as a sum involving the products of singular values and vector projections. Applying the triangle and Cauchy-Schwarz inequalities bounds this sum by the largest discarded singular value, $\sigma_{r+1}(W)$, multiplied by the norms of the query and document vectors. The bound's tightness is shown by selecting the corresponding singular vectors for q and d. The full proof is in the appendix.

Our experiments on MS MARCO empirically validate the effectiveness of low-rank bilinear approximation. Retrieval performance (MRR@10) improves consistently with increasing rank r, although with diminishing returns beyond rank 96. In particular, MRR@10 is strongly correlated with the magnitude of the first discarded singular value, σ_{r+1}, confirming that our theoretical bound is a strong predictor of real-world performance. The singular value spectrum of the learned interaction matrix W^* decays rapidly, indicating that the matrix is naturally low-rank and well-suited for approximation. This analysis reveals clear efficiency "sweet spots," allowing principled trade-offs between performance and cost without expensive hyperparameter tuning. For example, a rank-32 model achieves 50.8% of full performance with 24× compression, a rank-64 model delivers 68.9% with 12× compression, and a rank-96 model captures 86.1% with 8× compression.

Takeaway. Low-rank approximations make bilinear retrieval efficient without sacrificing expressiveness. The singular value spectrum of the interaction matrix offers a principled way to balance accuracy and compression, enabling theory-driven rank selection with minimal tuning. This provides three benefits: (1) **Predictable performance** through error bounds, (2) **Flexible deployment** across efficiency levels, and (3) **Theoretical grounding** validated by the strong correlation between singular value decay and retrieval quality.

6 Conclusion

We present a comprehensive case for adopting bilinear similarities in dense retrieval, both theoretically and empirically. Standard dot-product models, especially those with fixed embeddings, face structural limitations. Our work addresses this through three main contributions.

First, we illustrate the expressiveness gap with a concrete task in which bilinear models succeed while weighted dot-product models fail entirely. Second, we show that this theoretical edge leads to real-world gains, achieving up to a 4× improvement in MRR@10 in MS MARCO. Third, we address bilinear models' computational cost by leveraging low-rank approximation theory, linking performance directly to the singular value spectrum. This enables principled, efficient rank selection without costly tuning. Our findings reframe similarity function design as a critical, tunable component of retrieval, showing that expressiveness and efficiency can coexist. Low-rank bilinear models thus offer a practical and principled upgrade to standard dual encoder architectures.

References

1. Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., Nguyen, T., Rosenberg, M., Song, X., Stoica, A., Tiwary, S., Wang, T.: Ms marco: A human generated machine reading comprehension dataset (2018), https://arxiv.org/abs/1611.09268
2. Bhojanapalli, S., Yun, C., Rawat, A.S., Reddi, S., Kumar, S.: Low-rank bottleneck in multi-head attention models. In: Proceedings of the 37th International Conference on Machine Learning. ICML'20, JMLR.org (2020)
3. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American society for information science **41**(6), 391–407 (1990)
4. Ding, B., Zhai, J.: Retrieval with learned similarities. In: Proceedings of the ACM on Web Conference 2025. p. 1626–1637. WWW '25, ACM (Apr 2025). https://doi.org/10.1145/3696410.3714822, http://dx.doi.org/10.1145/3696410.3714822
5. Guo, J., Fan, Y., Ai, Q., Croft, W.B.: A deep relevance matching model for ad-hoc retrieval. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. p. 55–64. CIKM'16, ACM (Oct 2016). https://doi.org/10.1145/2983323.2983769, http://dx.doi.org/10.1145/2983323.2983769
6. Hofstätter, S., Lin, S.C., Yang, J.H., Lin, J., Hanbury, A.: Efficiently teaching an effective dense retriever with balanced topic aware sampling. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 113–122. SIGIR '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3404835.3462891, https://doi.org/10.1145/3404835.3462891
7. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: Proceedings of the 10th International Conference on Learning Representations (ICLR) (2022)
8. Humeau, S., Shuster, K., Lachaux, M.A., Weston, J.: Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. In: Proceedings of the 8th International Conference on Learning Representations (ICLR) (2020)
9. Kang, C., Liao, S., He, Y., Wang, J., Niu, W., Xiang, S., Pan, C.: Cross-modal similarity learning: A low rank bilinear formulation. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. p. 1251–1260. CIKM '15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2806416.2806469, https://doi.org/10.1145/2806416.2806469
10. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., Yih, W.t.: Dense passage retrieval for open-domain question answering. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 6769–6781. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.550, https://aclanthology.org/2020.emnlp-main.550/
11. Khattab, O., Zaharia, M.: Colbert: Efficient and effective passage search via contextualized late interaction over bert. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 39–48. SIGIR '20, Association for Computing Machinery, New York,

NY, USA (2020). https://doi.org/10.1145/3397271.3401075, https://doi.org/10.1145/3397271.3401075
12. Kim, J.H., On, K.W., Lim, W., Kim, J., Ha, J.W., Zhang, B.T.: Hadamard product for low-rank bilinear pooling. In: Proceedings of the 5th International Conference on Learning Representations (ICLR) (2017)
13. Liu, Z., Xiao, S., Shao, Y., Cao, Z.: RetroMAE-2: Duplex masked auto-encoder for pre-training retrieval-oriented language models. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2635–2648. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.acl-long.148, https://aclanthology.org/2023.acl-long.148/
14. Ni, J., Qu, C., Lu, J., Dai, Z., Hernandez Abrego, G., Ma, J., Zhao, V., Luan, Y., Hall, K., Chang, M.W., Yang, Y.: Large dual encoders are generalizable retrievers. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 9844–9855. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022). https://doi.org/10.18653/v1/2022.emnlp-main.669, https://aclanthology.org/2022.emnlp-main.669/
15. Pang, L., Lan, Y., Guo, J., Xu, J., Wan, S., Cheng, X.: Text matching as image recognition (2016), https://arxiv.org/abs/1602.06359
16. Qian, Q., Baytas, I.M., Jin, R., Jain, A., Zhu, S.: Similarity learning via adaptive regression and its application to image retrieval (2015), https://arxiv.org/abs/1512.01728
17. Rendle, S.: Factorization machines. In: 2010 IEEE International Conference on Data Mining. pp. 995–1000 (2010). https://doi.org/10.1109/ICDM.2010.127
18. Su, H., Shi, W., Kasai, J., Wang, Y., Hu, Y., Ostendorf, M., Yih, W.t., Smith, N.A., Zettlemoyer, L., Yu, T.: One embedder, any task: Instruction-finetuned text embeddings. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Findings of the Association for Computational Linguistics: ACL 2023. pp. 1102–1121. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.findings-acl.71, https://aclanthology.org/2023.findings-acl.71/
19. Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., Wei, F.: Text embeddings by weakly-supervised contrastive pre-training (2024), https://arxiv.org/abs/2212.03533
20. Xiong, L., Xiong, C., Li, Y., Tang, K.F., Liu, J., Bennett, P., Ahmed, J., Overwijk, A.: Approximate nearest neighbor negative contrastive learning for dense text retrieval. In: Proceedings of the 9th International Conference on Learning Representations (ICLR) (2021)
21. Zhai, J., Gong, Z., Wang, Y., Sun, X., Yan, Z., Li, F., Liu, X.: Revisiting neural retrieval on accelerators. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. p. 5520–5531. KDD '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3580305.3599897, https://doi.org/10.1145/3580305.3599897