When Reducing Representations Improves Performance

Andrea Pasin[1][0009−0007−5193−0741], Guglielmo Faggioli[1][0000−0002−5070−2049], Nicola Ferro[1][0000−0001−9219−6239], Raffaele Perego[2][0000−0001−7189−4724], and Nicola Tonellotto[3][0000−0002−7427−1001]

[1] University of Padua, Italy
[2] ISTI-CNR, Italy
[3] University of Pisa, Italy

Abstract. Neural models have transformed Information Retrieval (IR) by enabling semantic search, representing queries and documents as dense embeddings in latent spaces. However, recent works indicate the contribution of single dimensions in these representations to ranking quality is uneven: some dimensions are essential, while others may even degrade performance. Dimension IMportance Estimators (DIMEs) are heuristics to guide the search for the subsets of dimensions that induce an optimal subaspace where retrieval is more effective. To explore these subspaces, DIMEs rely on two simplifying assumptions: the linearity of subspaces and the independence of dimensions. In this paper, we move a step forward by relaxing the independence assumption and employing genetic algorithms to select the optimal set of dimensions. We show that selecting optimal dimensions for individual queries can achieve up to 0.981 nDCG@10 and 0.831 AP using state-of-the-art dense retrieval models on the considered datasets. Additionally, we identify subsets of dimensions that improve ranking quality across multiple queries simultaneously. Finally, we show that a dataset-specific subset of dimensions enables dense retrieval models to generalize across other datasets without loss of performance.

Keywords: Dense Representations · Ranking · Optimization · Genetic Algorithms · Effectiveness · Information Retrieval

1 Introduction

Recent advances in Large Language Models (LLMs) have enabled Information Retrieval (IR) models to shift from keyword-based search to semantic search. This transition has improved ranking effectiveness, particularly in handling language variations, synonyms, and contextual ambiguities in queries and documents [1]. Modern retrieval systems leverage these models to capture the semantic meaning of documents and queries. In particular, dense IR systems rely on contextualized embeddings [34] to represent queries and documents as dense vectors in a shared, lower-dimensional latent space. Each dimension encodes a latent feature, with its value reflecting the importance of that feature in the text.

The improved effectiveness of dense representations comes with reduced interpretability compared to lexical representations. We do not know the meaning of the dimensions, the reason for their value, or if they are needed [33]. Indeed, recent work reveals that not all dimensions in these dense representations equally contribute to ranking quality: some are useful, while others harm performance [9, 11]. Building upon this empirical observation, Faggioli et al. [11] introduced the *manifold clustering hypothesis*, showing that retrieval effectiveness can be improved by projecting documents and queries onto a query-specific sub-space of the original latent space. To find such a subspace, Faggioli et al. developed heuristics called Dimension IMportance Estimators (DIMEs). To make the problem computationally tractable and avoid searching through all possible solutions, DIMEs introduce two simplifying assumptions: (i) the search is restricted to *linear subspaces*, and (ii) *dimension independence* is assumed. While these assumptions reduce computational complexity, they provide limited insight into the structure of embedding spaces. In particular, the independence assumption ignores potential interactions between dimensions, which may be crucial for accurately characterizing the embedding space.

This paper further explores the manifold clustering hypothesis, aiming to determine whether more effective query-specific DIMEs can be discovered by relaxing the assumption of independence among dimensions. Furthermore, we investigate if the improvements due to the DIME-induced shift of reference spaces are preserved across queries and there are subspaces that consistently perform well across entire query sets. Finally, we explore the potential for transferring dimensionality reduction strategies across different collections. We formalize our analysis into the following research questions:

RQ1 **Effective dimensionality reduction**. To what extent can query-specific dimension selection improve performance?
RQ2 **A subspace for all queries in a dataset**. Are there subsets of dimensions that consistently improve model performance across multiple queries?
RQ3 **Generalization on multiple datasets**. Can we learn a robust subset of dimensions that enables a model to generalize across different datasets?

This paper is primarily theoretical in nature. However, given the analytical intractability of studying the manifold clustering hypothesis, we adopt an empirical heuristic approach. We run extensive experiments on TREC Deep Learning 2019/2020 [5, 6] and TREC Robust 2004 [32], evaluating dense retrieval models such as TAS-B [13], Contriever [14], and Dragon [21].

To explore subspaces without assuming independent dimensions, we use genetic algorithms (GAs) to approximate optimal subsets of dimensions. These are guided by cost functions defined using relevance judgments from ground truth pools, placing our study in an oracle setting with ideal, noise-free feedback. We recognize that relying on ground truth judgments limits real-world applicability, as such data is rarely available at scale. However, we aim to provide a conceptual and empirical foundation for understanding the potential of dimension selection and to encourage future work on developing practical, signal-based approximations of these oracle-informed strategies.

Our genetic approach, applied on a query-per-query basis, selects subspaces leading to substantial performance gains, improving nDCG@10 up to +113% and AP up to +191%. Secondly, we find that certain dimension subsets consistently improve performance across multiple queries. Finally, we demonstrate that dimension subsets learnt on one dataset can be transferred to other datasets without degrading effectiveness, indicating promising generalization capabilities.

2 Background and Related Work

This section discusses related work regarding Dense Retrieval, representation reduction techniques, and GAs.

Dense Retrieval. Dense retrieval has emerged as a powerful paradigm in IR, particularly for tasks requiring semantic understanding, such as question answering, ad hoc search, and recommender systems [34]. Modern dense IR models can be broadly categorized into three types: cross-encoders, bi-encoders, and late-interaction models. Cross-encoders, such as BERT [10], RoBERTa [22], and ELECTRA [3], jointly encode query-document pairs, computing their interactions at query time. While effective, they are computationally expensive. Late-interaction models, such as ColBERT [19], compare contextual representations of individual terms rather than reducing queries and documents to single embeddings [12]. While offering improved retrieval accuracy, they introduce overhead in time and space. Bi-encoders, in contrast, independently encode queries and documents using neural networks, allowing for offline document pre-computation and indexing. These models can be symmetric, i.e., using the same encoder for queries and documents, as in TAS-B [13] and Contriever [14], or asymmetric, as in Dragon [21]. The bi-encoder architecture enables scalable retrieval using efficient indexing tools, such as FAISS [16], and data structures, such as HNSW [23], which speed up nearest-neighbour search over large vector embedding datasets.

In this work, we focus on bi-encoders and investigate to what extent their encoding of queries and documents in the latent space is redundant and noisy.

Representation Reduction Techniques. Dimension IMportance Estimators (DIMEs) have recently been proposed as query-dependent methods to improve the ranking performance of bi-encoder models by exploiting correlations across dimensions between the query representation and those of relevant or irrelevant documents [2, 8, 11]. They leverage the observation that not all representation dimensions contribute equally to retrieval effectiveness, and aim to identify through Pseudo Relevance Feedback or by exploiting dense representations of Large Language Model-generated answers query-dependent subsets of dimensions to enhance ranking quality. However, a key limitation of previous approaches is the assumption of independence across dimensions: each is scored in isolation, ignoring interactions that may influence their combined impact on ranking. This assumption was primarily introduced to ensure computational tractability, since modeling all possible dependencies among dimensions quickly becomes intractable at scale. In this work, we address this limitation by framing the task as an instance of Best Subset Selection, a well-known NP-hard problem [35]. To tackle the combinatorial complexity, we resort to GAs, which provide an efficient heuristic for exploring large search spaces and identifying high-quality suboptimal subsets of dimensions. Unlike previous DIMEs, our approach captures dependencies among dimensions, enabling a richer and more effective exploration of the representation space.

Genetic algorithms. GAs are optimization techniques inspired by the process of natural selection [18, 28]. They work by simulating evolution through operations such as selection, crossover, and mutation [7]. In GAs, potential solutions to a problem are represented as individuals in a population. Over successive generations, the fittest individuals (i.e., the ones representing better solutions to the considered problem) are selected to reproduce, combining their "genetic" information to create new solutions [17]. This iterative process allows the algorithm to explore a solution space efficiently and converge toward optimal or near-optimal solutions.

GAs have already been successfully employed in IR. For example, Martín-Bautista and Miranda [24] described the use of GAs for feature selection, preserving a limited number of relevant features associated with documents to enhance efficiency in classification problems. Kraft et al. [20] employed GAs for weighted boolean query reformulation in order to achieve higher effectiveness, considering both precision and recall. Özel [27] leveraged GAs for web-page classification during crawling processes. Similar to these works, we leverage GAs to identify effective subsets of dimensions that can be used to improve the ranking quality of dense retrieval models.

3 Methodology

3.1 Background on Dimension Importance Estimators

A bi-encoder dense retrieval model projects a query q and the corpus of documents $\mathcal{C} = \{d_1, ..., d_n\}$ into a latent h-dimensional space $\mathbb{R}^h$. At retrieval time, given a query q and its representation $\mathbf{q} \in \mathbb{R}^h$, the retrieval score for a document d represented in the latent space as $\mathbf{d} \in \mathbb{R}^h$ is the dot product $\mathbf{q} \cdot \mathbf{d}$. Thus, we can define a ranked list of documents $\mathcal{R}_q = \{d_1, ..., d_k\}$, containing the top-$k$ documents sorted according to their retrieval score. Finally, if a set of relevance labels $\mathcal{L}$ is available, we can compute an effectiveness measure $M(\mathcal{R}_q; \mathcal{L})$ that quantifies how effective $\mathcal{R}_q$ is in satisfying the information need represented by q. Faggioli et al. [11] observed that given a representation space in $\mathbb{R}^h$, there exists a linear subspace[4] $\mathbb{R}^{|\delta|}$ that insists on the dimensions $\delta \subset \{1, ..., h\}$, with $|\delta| < h$, where the retrieval is more effective. In other terms, called $\mathcal{R}_{q,\delta}$ the ranked list

[4] Faggioli et al. postulate that the subspace can be an arbitrary manifold, but focus only on independent linear subspaces, i.e., subspaces constructed by removing some dimensions.

of documents obtained by carrying out the retrieval in $\mathbb{R}^{|\delta|}$, we observe that $M(\mathcal{R}_q;\mathcal{L}) < M(\mathcal{R}_{q,\delta};\mathcal{L})$. We refer to $(\mathbf{q}\cdot\mathbf{d})_\delta$ as the dot product between the query and document representations in $\mathbb{R}^{|\delta|}$. The problem of finding the optimal set of dimensions δ where the retrieval is most effective is intractable: there are 2^h possible linear subspaces of $\mathbb{R}^h$. Hence, Faggioli et al. propose a class of models, called Dimension IMportance Estimators (DIMEs) $u : (\{1,...,h\};\mathbf{q},\theta) \mapsto \mathbb{R}$ that, given the index of a dimension, the query representation $\mathbf{q}$, and possibly a set of additional inputs θ, outputs a score measuring how important the dimension is.

3.2 DIMEs and Genetic Algorithms

The Dimension IMportance Estimator (DIME) task can be formulated as the optimization problem of finding the set of dimensions that maximizes effectiveness. In this paper, we employ GAs to find suboptimal sets of dimensions for single or groups of queries. Differently from the heuristic approaches proposed in [2, 8, 11], we account for dependencies between dimensions by designing an evolutionary framework designed to balance exploration and exploitation of our solution space. The individuals in the population of our GA are subsets of representation dimensions, represented as binary strings where 1 in the i-th position means that dimension i is retained in the query vector representation, 0 otherwise. The population is initialized with a random sample of individuals. At each generation, we employ a fitness-based strategy that guides the selection of individuals by favouring the subsets of dimensions that provide the best ranking quality while preserving diversity. This increases the likelihood of generating high-quality individuals over time. New individuals are generated by recombining components from selected parents in the current population. For example, given two parent solutions encoded as bit strings 10110011 and 01101100 in a toy 8-dimension model, a new individual is formed by randomly selecting each bit from one of the parent, creating, for example, an offspring like 11111100 corresponding to a DIME selecting the first six dimensions in the representation. This crossover mechanism promotes diversity and facilitates exploration of different regions in the solution space. To further maintain variation and prevent premature convergence, the algorithm introduces small random changes, called mutations, in the current population of solutions. Mutations involve flipping a bit of an individual with low probability, ensuring that new genetic material is occasionally introduced into the population. Our approach also includes elitism, a mechanism that guarantees the best solution found in the current generation is carried forward unchanged to the next generation, thereby ensuring that progress is not lost.

3.3 RQ1: Effective Dimensionality Reduction

RQ1 focuses on determining the largest improvement achievable by a query-specific DIME when the dimension-independence assumption is relaxed. To approximate the optimal subset, we define the following objective function, which guides the GA for each query q using the relevance judgments $\mathcal{L}$:

$$\max_{\delta_q} \frac{1}{|\mathcal{M}|} \sum_{M \in \mathcal{M}} M(\mathcal{R}_{q,\delta_q}; \mathcal{L}) - \frac{|\delta_q|}{h} \tag{1}$$

where δ_q represents the subset of dimensions selected for query q (i.e., an individual in the GA) and $\mathcal{M}$ is the set of evaluation measures to maximize. This objective function seeks to identify the subset of dimensions that maximizes the evaluation measures while simultaneously minimizing the number of dimensions considered. In our experiments we set $\mathcal{M} = \{\text{nDCG@10}, \text{AP}\}$. Notably, this approach does not require explicitly specifying the number of retained dimensions, as the GA automatically selects it for each query according to the objective function.

3.4 RQ2: A Subspace for all Queries in a Dataset

We turn here our attention to a broader question: is there a subset of dimensions capable of enhancing the model performance consistently across all queries in a given dataset? We aim to identify a single subset of dimensions that yields an overall performance boost when considering the entire dataset. This approach prioritizes global effectiveness, as it seeks to balance and optimize performance across all queries rather than focusing on localized improvements.

To address this challenge, we employ a methodology similar to the one described earlier in Section 3.3, again leveraging a GA to explore and refine the selection of dimensions, guided by the following optimization function:

$$\max_{\delta} \frac{1}{|Q|} \sum_{q \in Q} \left(\frac{1}{|\mathcal{M}|} \sum_{M \in \mathcal{M}} M(\mathcal{R}_{q,\delta}; \mathcal{L}) \right) \tag{2}$$

where δ represents the subset of dimensions selected for all queries and Q is the set containing all the queries in a dataset. Similarly to Equation 1, we employ nDCG@10 and AP as the evaluation measures to maximize. This allows us to systematically search for subsets of dimensions that maximize the overall effectiveness on all queries of a dataset.

3.5 RQ3: Generalization on Multiple Datasets

Finally, we aim to identify a robust subset of dimensions that preserves retrieval performance advantages across datasets without relying on dataset-specific relevance labels. In practice, this means selecting a subset of dimensions from one dataset and evaluating whether the performance boost is maintained when applied to other datasets. If successful, this approach could indicate the presence of systematic noise in the embedding space, with certain dimensions consistently degrading retrieval performance.

We fix one dataset as the training set and determine the subset of dimensions to retain based on the output of our optimization problem. We then evaluate the effectiveness of these retained dimensions on other datasets used as test sets. Compared to the previous experiments, this setup employs proper training and test splits, making it an analysis of the latent dimension content and a methodology suitable for real-world application.

Let $\mathcal{L}_q^+$ and $\mathcal{L}_q^-$ denote the sets of relevant and non-relevant documents for query q in the training dataset. We then proceed to define the following optimization function:

$$\max_\delta \sum_{q \in Q} \left(\frac{1}{|\mathcal{L}_q^+|} \sum_{d^+ \in \mathcal{L}_q^+} (\mathbf{q} \cdot \mathbf{d}^+)_\delta - \frac{1}{|\mathcal{L}_q^-|} \sum_{d^- \in \mathcal{L}_q^-} (\mathbf{q} \cdot \mathbf{d}^-)_\delta \right) \tag{3}$$

This function identifies the subset of dimensions that maximizes the dot-product scores for relevant documents while simultaneously minimizing them for non-relevant ones. Unlike the previous functions, it does not directly optimize evaluation measures such as nDCG@10 or AP. Instead, it seeks a subset that enhances the discriminative power between relevant and non-relevant documents. Therefore, this new objective function aims at finding a subset of dimensions with high generalization capabilities, even across different datasets.

4 Experimental Assessment

We present here the experimental settings and a comprehensive analysis of the experiments conducted to answer our three research questions.

4.1 Experimental Settings

We empirically validate our research questions on three dense retrieval models operating in 768-dimensional latent spaces: Contriever [14], TAS-B [13], and Dragon [21]. The model weights were fine-tuned on the MS-MARCO collection and are publicly available through the HuggingFace repository. In terms of datasets, we consider three experimental collections: TREC Deep Learning '19 (DL 19) [6], TREC Deep Learning '20 (DL 20) [5], and TREC Robust '04 (RB 04) [32]. The first two focus on ad-hoc passage retrieval, with 43 and 54 annotated queries, respectively, and are based on the MS-MARCO [26] passages collection. On the other hand, RB 04 contains 249 queries and is derived from the TIPSTER disks. For computational reasons, we consider 100,000 retrieved documents for every query—approximately, 19% of the corpus. As all the considered dense IR systems have been fine-tuned on the MS-MARCO passage collection, they are in-domain IR systems for DL 19 and DL 20, whereas RB 04 represents a zero-shot application of these models.

Our GA employs uniform crossover mechanism with probability $P_{crossover} = 0.5$ to set a dimension based on one of the two parents. Mutation follows a bit-flip strategy with probability $P_{mutation} = \frac{2}{768}$ per bit. Selection is tournament-based, with size 3, ensuring that the fittest individual from each group is chosen for crossover while preserving diversity. retain the best solution across generations.

Table 1. Answering RQ1: retrieval performance using the best query-specific subset of dimensions found by the GA compared to the baseline using all dimensions. Symbol * denotes statistically significant differences with respect to the baseline.

		Genetic algorithm			Baseline		
Dataset	Model	nDCG@10	AP	Avg N° dims.	nDCG@10	AP	Avg N° dims.
DL 19	TAS-B	**0.975***	**0.813***	55	0.717	0.481	768
	Contriever	**0.964***	**0.811***	58	0.674	0.493	
	Dragon	**0.981***	**0.831***	56	0.740	0.522	
DL 20	TAS-B	**0.972***	**0.800***	57	0.684	0.478	768
	Contriever	**0.972***	**0.803***	55	0.672	0.482	
	Dragon	**0.976***	**0.824***	59	0.718	0.509	
RB 04	TAS-B	**0.953***	**0.619***	55	0.447	0.213	768
	Contriever	**0.960***	**0.659***	55	0.465	0.239	
	Dragon	**0.948***	**0.633***	58	0.461	0.228	

The population size is fixed at 100 and initialized randomly. The algorithm terminates upon convergence or after reaching a maximum number of generations. For the experiments in Sections 4.2 and 4.3, the maximum number of generations was set to 500, while for Section 4.4 it was set to 50, balancing effectiveness with computational feasibility under practical timing constraints.

To test the statistical significance of the improvement over the baseline, we use two-way Analysis of Variance (ANOVA) [29] and Tukey's Honestly Significant Differences (HSD) post-hoc test [31], with significance level of 0.05. The experiments are carried out on a machine with 4 Intel Xeon CPUs (72 cores in total) and 1.5 TB of RAM. The code is publicly available.[5]

4.2 RQ1: Effective Dimensionality Reduction

Our first investigation concerns the maximum performance achievable through the selection performed by the GA of an effective query-specific subset of dimensions using the optimization function defined in Equation 1. With our settings the GA examines for each query at most $500 \cdot 100 = 50,000$ possible subsets of dimensions in the $2^{768} \approx 1.55 \cdot 10^{231}$ possible subsets.able 1 reports the results of this oracle experiment. Despite exploring only a limited portion of the solution space, the subsets of dimensions identified by the GA yield a substantial improvement in retrieval performance. On average, selecting about 8% of the total dimensions per query leads to gains of up to 113% in nDCG@10 and 191% in AP on RB 04. These findings highlight the effectiveness of dimensionality reduction in enhancing ranking quality. Selecting with our GA a set of dimensions yields nearly perfect nDCG@10 values, while AP reaches up to ≈ 0.8, which is remarkably high. This leads to the following observations:

[5] To be released upon acceptance

Table 2. Answering RQ2: retrieval performance using the best subset of dimensions found by the GA for all the queries of the test data compared to the baseline using all dimensions. Symbol * denotes statistically significant differences with respect to the baseline.

		Genetic algorithm			Baseline		
Dataset	Model	nDCG@10	AP	N° dims.	nDCG@10	AP	N° dims.
	TAS-B	**0.784***	**0.526***	404	0.717	0.481	
DL 19	Contriever	**0.793***	**0.541***	367	0.674	0.493	768
	Dragon	**0.836***	**0.549***	410	0.740	0.522	
	TAS-B	**0.775***	**0.511***	414	0.684	0.478	
DL 20	Contriever	**0.770***	**0.522***	381	0.672	0.482	768
	Dragon	**0.802***	**0.540***	412	0.718	0.509	
	TAS-B	**0.486***	**0.221***	512	0.447	0.213	
RB 04	Contriever	**0.528***	**0.256***	439	0.465	0.239	768
	Dragon	**0.511***	**0.241***	507	0.461	0.228	

- Fewer than 8% of the dimensions contribute positively to ranking quality, while the remaining 92% are either irrelevant or detrimental. This finding calls for a reconsideration of current training strategies to prevent such a large amount of noise from being injected into query representations.
- Our setting is oracular, and this optimization problem cannot be solved directly in real-world scenarios. Nonetheless, the margins of improvement are substantial: in this optimal setting, we more than double effectiveness on the RB 04 collection. This result opens the door to novel strategies aimed at translating theoretical gains into practical deployments. Future research could focus on identifying high-impact dimension subsets without relying on curated ground truth, for instance, by leveraging proxy signals or noisy relevance judgments. Such advancements would make this approach more practical and applicable in production environments.

4.3 RQ2: A Subspace for all Queries in a Dataset

We now investigate whether there exist subsets of dimensions that improve the model performance consistently across all queries in a given dataset. To answer this question, we employ our genetic approach guided by the optimization function defined in Equation 2. Table 2 presents the effectiveness results obtained by selecting a single subset of dimensions for each dataset and dense retrieval model under consideration.

As a first observation, there exist subsets of dimensions that yield significant improvements in the effectiveness of dense retrieval models across different evaluation measures, with gains of up to +16% in nDCG@10 and +11% in AP. This indicates that certain dimensions can substantially enhance performance, whereas others contribute little or may even degrade it. Figure 1 further illus-

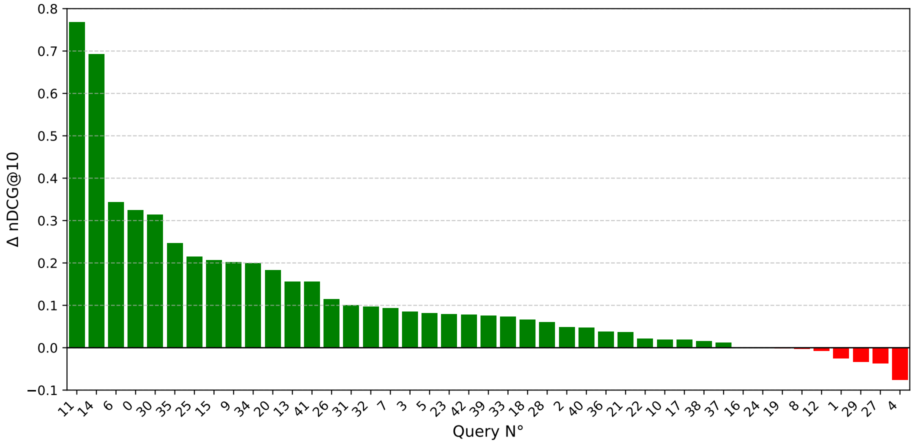

Fig. 1. Difference in nDCG@10 between Contriever using the GA-selected subsets of dimensions and using the full set of dimensions for each query in the DL 19 dataset. For the majority of the queries, we observe a robust performance improvement. Only 6 queries are penalized, but the loss is small.

trates this point: the subsets of dimensions selected by the GA enable Contriever to outperform its full-dimension baseline on nearly all queries, as measured by nDCG@10. Comparable trends are observed for the other models, though results are omitted here due to space constraints.

Despite being empirical bounds, our findings here show a novel insight over previous works [2, 8, 11] that observe a performance improvement only when operating at a query-level: this analysis suggests that it is possible to find a dimension set that optimizes multiple queries at once.

Secondly, we notice that, compared to Table 1, the dimension subsets that optimize the retrieval effectiveness for a set of queries at once are larger compared to those needed for a single query. This is a reasonable behaviour: being able to answer more information needs calls for larger representations. This is further highlighted by the fact that for DL 19 and DL 20 (approximately 50 topics), using around 53% of the full set of dimensions provides the optimal result, while for RB 04 (249 topics), we need 65% of the dimensions.

We highlight again that these outcomes represent an estimate of the maximal improvement that can be achieved by reducing the number of dimensions. Nevertheless, if there exists a set of dimensions that optimizes a group of queries at once, we can directly prune the documents' representation to reduce its memory occupation, and improve efficiency[6] and effectiveness. It remains an open challenge and a relevant future investigation path to identify clusters of queries that benefit from the same dimensionality reduction.

[6] While we do not investigate this empirically for space reasons, the theoretical efficiency gain is proportional to the representation reduction.

Table 3. Answering RQ3: retrieval performance on the target test dataset using the best subset of dimensions found by the GA using DL 19 as training dataset compared to the baseline using all dimensions. Symbol * denotes statistically significant differences with respect to the baseline.

		Genetic algorithm			Baseline		
Model	Dataset	nDCG@10	AP	N° dims.	nDCG@10	AP	N° dims.
TAS-B	DL 19 (Train)	**0.734**	**0.507***	674	0.717	0.481	768
	DL 20 (Test)	0.682	**0.479**		**0.684**	0.478	
	RB 04 (Test)	0.439	0.209*		**0.447**	**0.213**	
Contriever	DL 19 (Train)	**0.683**	**0.501**	668	0.674	0.493	768
	DL 20 (Test)	0.671	0.477		**0.672**	**0.482**	
	RB 04 (Test)	0.462	0.236*		**0.465**	**0.239**	
Dragon	DL 19 (Train)	**0.748**	0.517	671	0.740	**0.522**	768
	DL 20 (Test)	**0.735**	0.508		0.718	**0.509**	
	RB 04 (Test)	0.454	0.223*		**0.461**	**0.228**	

4.4 RQ3: Generalization on Multiple Datasets

Previous experiments showed the effectiveness of the dimension selection at a query and dataset level. We now turn to generalizing across datasets, moving away from the oracle setting and considering a practical machine learning setting, with separate training and test sets. In this experiment, we consider DL 19 as the training dataset to find with our GA and the optimization function provided in Equation 3 a single good subset of dimensions. Then, we test whether the identified subset of dimensions generalizes well across the remaining datasets and, thus, maintains a performance advantage across them.

Table 3 shows the results of this experiment. The selected subset of dimensions enables dense retrieval models to generalize effectively across different datasets. By learning the set of dimensions on DL 19, this clearly induces a performance improvement. The number of dimensions that maximizes the optimization function described in equation 3 ranges from 668 (Contriever) to 674 (TAS-B). This indicates that, by reducing the dimensions by around 13% to 12%, we still maintain a good separation between the scores assigned to relevant and non-relevant documents.

If we use the same set of dimensions to project the representation space also for DL 20 and RB 04, we notice a minor deflation of effectiveness. This decrease occurs, as an order of magnitude, in the third decimal place, indicating a negligible change in terms of ranking. In other terms, while we have saved more than 12% of the resources—both in terms of space on the disk and in the number of required operations to compute the dot product—it is highly unlikely that the user will perceive this change when looking at the ranked list of documents. Importantly, this change is seldom statistically significant, except for AP for RB 04 collection. We stress that, despite the change being significant, Cohen's d [4, 15], which indicates the actual strength of the change, is always below 0.03.[7]

Future research could explore whether the discarded dimensions primarily contain redundant information or noise that does not contribute to model effectiveness. Investigating the role of these dimensions could lead to further optimizations in embedding design, potentially improving both efficiency and interpretability in dense retrieval models.

5 Conclusion and Future Work

In this paper, we presented an in-depth investigation into how strategically selecting dimensions from dense retrieval models' representations can enhance ranking performance. We provided novel empirical evidence supporting the manifold clustering hypothesis by identifying and leveraging only effective subsets of dimensions for each query, using DIMEs in combination with GAs. Our simulation experiments demonstrated substantial improvements in both nDCG@10 and AP across various models and datasets. Notably, we show that our methodology can be extended to entire query sets: the same subset of dimensions can improve retrieval performance even when applied uniformly to all queries in a dataset. Furthermore, we demonstrate that it is possible to learn robust subsets of dimensions that allow a dense retrieval model to generalize effectively to unseen datasets without compromising effectiveness.

Although the insights from this study are primarily theoretical since they mostly rely on an oracle setting, they reveal that dense representations often contain noise or redundancy that can negatively impact ranking quality. By selecting a reduced yet informative subset of dimensions, we not only preserve effectiveness but also achieve significant performance gains.

Future research could explore innovative approaches for selecting effective subsets of dimensions without relying on relevance judgments. This could include unsupervised methods that automatically identify the most important dimensions for a given query, thereby enhancing ranking quality in a more practical way. Moreover, considering other data modalities, such as image-based datasets [25], could help assess how well our findings generalize to different downstream tasks. Additionally, further investigation could focus on understanding the contribution of each model dimension, identifying those associated with noise or redundancy, thus trying to delve deeper into the interpretability of these models. Lastly, more ambitious research could aim to simplify model complexity, which could translate into the generation of lower-dimensional text representations. By extending dimensionality reduction techniques to account for both the quality and interpretability of selected representations, future research directions could further advance our understanding of the impacts of specific models' dimensions on ranking, especially in the context of large-scale real-world applications. By delving deeper into these areas, future research has the potential to

[7] As a reference, a Cohen's d between 0.2 and 0.5 is considered small, while below 0.2 the effect is deemed negligible [30].

build more effective and efficient IR systems capable of better solving complex and specialized retrieval tasks.

Bibliography

[1] Berger, A.L., Caruana, R., Cohn, D., Freitag, D., Mittal, V.O.: Bridging the lexical chasm: statistical approaches to answer-finding. In: Yannakoudakis, E.J., Belkin, N.J., Ingwersen, P., Leong, M. (eds.) SIGIR 2000: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 24-28, 2000, Athens, Greece, pp. 192–199, ACM (2000), https://doi.org/10.1145/345508.345576, URL https://doi.org/10.1145/345508.345576

[2] Campagnano, C., Mallia, A., Silvestri, F.: Unveiling dime: Reproducibility, generalizability, and formal analysis of dimension importance estimation for dense retrieval. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, p. 3367–3376, SIGIR '25, Association for Computing Machinery, New York, NY, USA (2025), ISBN 9798400715921, https://doi.org/10.1145/3726302.3730318, URL https://doi.org/10.1145/3726302.3730318

[3] Clark, K., Luong, M., Le, Q.V., Manning, C.D.: ELECTRA: pre-training text encoders as discriminators rather than generators. CoRR **abs/2003.10555** (2020), URL https://arxiv.org/abs/2003.10555

[4] Cohen, J.: Statistical power analysis for the behavioral sciences. routledge (2013)

[5] Craswell, N., Mitra, B., Yilmaz, E., Campos, D.: Overview of the TREC 2020 deep learning track. CoRR **abs/2102.07662** (2021), URL https://arxiv.org/abs/2102.07662

[6] Craswell, N., Mitra, B., Yilmaz, E., Campos, D., Voorhees, E.M.: Overview of the TREC 2019 deep learning track. CoRR **abs/2003.07820** (2020), URL https://arxiv.org/abs/2003.07820

[7] Deb, K.: An introduction to genetic algorithms. Sadhana **24**, 293–315 (1999)

[8] D'Erasmo, G., Trappolini, G., Silvestri, F., Tonellotto, N.: Eclipse: Contrastive dimension importance estimation with pseudo-irrelevance feedback for dense retrieval. In: Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR), p. 147–154, ICTIR '25, Association for Computing Machinery, New York, NY, USA (2025), ISBN 9798400718618, https://doi.org/10.1145/3731120.3744579, URL https://doi.org/10.1145/3731120.3744579

[9] D'Erasmo, G., Trappolini, G., Tonellotto, N., Silvestri, F.: Eclipse: Contrastive dimension importance estimation with pseudo-irrelevance feedback for dense retrieval. arXiv preprint arXiv:2412.14967 (2024)

[10] Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human

Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pp. 4171–4186, Association for Computational Linguistics (2019), https://doi.org/10.18653/V1/N19-1423, URL https://doi.org/10.18653/v1/n19-1423

[11] Faggioli, G., Ferro, N., Perego, R., Tonellotto, N.: Dimension importance estimation for dense information retrieval. In: Yang, G.H., Wang, H., Han, S., Hauff, C., Zuccon, G., Zhang, Y. (eds.) Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024, pp. 1318–1328, ACM (2024), https://doi.org/10.1145/3626772.3657691, URL https://doi.org/10.1145/3626772.3657691

[12] Formal, T., Clinchant, S., Déjean, H., Lassance, C.: SPLATE: sparse late interaction retrieval. In: Yang, G.H., Wang, H., Han, S., Hauff, C., Zuccon, G., Zhang, Y. (eds.) Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024, pp. 2635–2640, ACM (2024), https://doi.org/10.1145/3626772.3657968, URL https://doi.org/10.1145/3626772.3657968

[13] Hofstätter, S., Lin, S., Yang, J., Lin, J., Hanbury, A.: Efficiently teaching an effective dense retriever with balanced topic aware sampling. In: Diaz, F., Shah, C., Suel, T., Castells, P., Jones, R., Sakai, T. (eds.) SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021, pp. 113–122, ACM (2021), https://doi.org/10.1145/3404835.3462891, URL https://doi.org/10.1145/3404835.3462891

[14] Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., Grave, E.: Towards unsupervised dense information retrieval with contrastive learning. CoRR **abs/2112.09118** (2021), URL https://arxiv.org/abs/2112.09118

[15] Jacob, C.: A power primer. Psychological bulletin **112**(1), 155–159 (1992)

[16] Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with gpus. CoRR **abs/1702.08734** (2017), URL http://arxiv.org/abs/1702.08734

[17] Jong, K.A.D.: Learning with genetic algorithms: An overview. Mach. Learn. **3**, 121–138 (1988), https://doi.org/10.1007/BF00113894, URL https://doi.org/10.1007/BF00113894

[18] Katoch, S., Chauhan, S.S., Kumar, V.: A review on genetic algorithm: past, present, and future. Multim. Tools Appl. **80**(5), 8091–8126 (2021), https://doi.org/10.1007/S11042-020-10139-6, URL https://doi.org/10.1007/s11042-020-10139-6

[19] Khattab, O., Zaharia, M.: Colbert: Efficient and effective passage search via contextualized late interaction over BERT. In: Huang, J.X., Chang, Y., Cheng, X., Kamps, J., Murdock, V., Wen, J., Liu, Y. (eds.) Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020, pp. 39–48, ACM (2020), https://doi.org/10.1145/3397271.3401075, URL https://doi.org/10.1145/3397271.3401075

[20] Kraft, D.H., Petry, F.E., Buckles, B.P., Sadasivan, T.: Applying genetic algorithms to information retrieval systems via relevance feedback. In: Fuzziness in database management systems, pp. 330–344, Springer (1995)

[21] Lin, S., Asai, A., Li, M., Oguz, B., Lin, J., Mehdad, Y., Yih, W., Chen, X.: How to train your DRAGON: diverse augmentation towards generalizable dense retrieval. CoRR **abs/2302.07452** (2023), https://doi.org/10.48550/ARXIV.2302.07452, URL https://doi.org/10.48550/arXiv.2302.07452

[22] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. CoRR **abs/1907.11692** (2019), URL http://arxiv.org/abs/1907.11692

[23] Malkov, Y.A., Yashunin, D.A.: Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(4), 824–836 (2020), https://doi.org/10.1109/TPAMI.2018.2889473

[24] Martín-Bautista, M.J., Miranda, M.A.V.: Applying genetic algorithms to the feature selection problem in information retrieval. In: Andreasen, T., Christiansen, H., Larsen, H.L. (eds.) Flexible Query Answering Systems, Third International Conference, FQAS'98, Roskilde, Denmark, May 13-15, 1998, Proceedings, Lecture Notes in Computer Science, vol. 1495, pp. 272–281, Springer (1998), https://doi.org/10.1007/BFB0056008, URL https://doi.org/10.1007/BFb0056008

[25] Müller, H., Kalpathy-Cramer, J., Jr., C.E.K., Hatt, W., Bedrick, S., Hersh, W.R.: Overview of the imageclefmed 2008 medical image retrieval task. In: Peters, C., Deselaers, T., Ferro, N., Gonzalo, J., Jones, G.J.F., Kurimo, M., Mandl, T., Peñas, A., Petras, V. (eds.) Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers, Lecture Notes in Computer Science, vol. 5706, pp. 512–522, Springer (2008), https://doi.org/10.1007/978-3-642-04447-2“˙63, URL https://doi.org/10.1007/978-3-642-04447-2_63

[26] Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: MS MARCO: A human generated machine reading comprehension dataset. CoRR **abs/1611.09268** (2016), URL http://arxiv.org/abs/1611.09268

[27] Özel, S.A.: A web page classification system based on a genetic algorithm using tagged-terms as features. Expert Syst. Appl. **38**(4), 3407–3415 (2011), https://doi.org/10.1016/J.ESWA.2010.08.126, URL https://doi.org/10.1016/j.eswa.2010.08.126

[28] Srinivas, M., Patnaik, L.M.: Genetic algorithms: A survey. Computer **27**(6), 17–26 (1994), https://doi.org/10.1109/2.294849, URL https://doi.org/10.1109/2.294849

[29] St, L., Wold, S., et al.: Analysis of variance (anova). Chemometrics and intelligent laboratory systems **6**(4), 259–272 (1989)

[30] Sullivan, G.M., Feinn, R.: Using effect size—or why the p value is not enough. Journal of Graduate Medical Education **4**(3), 279–282 (09 2012), ISSN 1949-8349, https://doi.org/10.4300/JGME-D-12-00156.1, URL https://doi.org/10.4300/JGME-D-12-00156.1

[31] Tukey, J.W.: Comparing individual means in the analysis of variance. Biometrics pp. 99–114 (1949)

[32] Voorhees, E.M.: Overview of the TREC 2004 robust track. In: Voorhees, E.M., Buckland, L.P. (eds.) Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004, NIST Special Publication, vol. 500-261, National Institute of Standards and Technology (NIST) (2004), URL http://trec.nist.gov/pubs/trec13/papers/ROBUST.OVERVIEW.pdf

[33] Zhao, T., Lu, X., Lee, K.: SPARTA: efficient open-domain question answering via sparse transformer matching retrieval. In: Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tür, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., Zhou, Y. (eds.) Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, pp. 565–575, Association for Computational Linguistics (2021), https://doi.org/10.18653/V1/2021.NAACL-MAIN.47, URL https://doi.org/10.18653/v1/2021.naacl-main.47

[34] Zhao, W.X., Liu, J., Ren, R., Wen, J.: Dense text retrieval based on pretrained language models: A survey. ACM Trans. Inf. Syst. **42**(4), 89:1–89:60 (2024), https://doi.org/10.1145/3637870, URL https://doi.org/10.1145/3637870

[35] Zhu, J., Zhu, J., Tang, B., Chen, X., Lin, H., Wang, X.: Best-subset selection in generalized linear models: A fast and consistent algorithm via splicing technique. CoRR **abs/2308.00251** (2023), https://doi.org/10.48550/ARXIV.2308.00251, URL https://doi.org/10.48550/arXiv.2308.00251