Dimension Importance Estimation for Dense Information Retrieval

Guglielmo Faggioli
University of Padua, Italy

Nicola Ferro
University of Padua, Italy

Raffaele Perego
CNR, Pisa, Italy

Nicola Tonellotto
University of Pisa, Italy

ABSTRACT

Recent advances in Information Retrieval have shown the effectiveness of embedding queries and documents in a latent high-dimensional space to compute their similarity. While operating on such high-dimensional spaces is effective, in this paper, we hypothesize that we can improve the retrieval performance by adequately moving to a query-dependent subspace. More in detail, we formulate the Manifold Clustering (MC) Hypothesis: projecting queries and documents onto a subspace of the original representation space can improve retrieval effectiveness. To empirically validate our hypothesis, we define a novel class of Dimension IMportance Estimators (DIME). Such models aim to determine how much each dimension of a high-dimensional representation contributes to the quality of the final ranking and provide an empirical method to select a subset of dimensions where to project the query and the documents. To support our hypothesis, we propose an oracle DIME, capable of effectively selecting dimensions and almost doubling the retrieval performance. To show the practical applicability of our approach, we then propose a set of DIMEs that do not require any oracular piece of information to estimate the importance of dimensions. These estimators allow us to carry out a dimensionality selection that enables performance improvements of up to +11.5% (moving from 0.675 to 0.752 nDCG@10) compared to the baseline methods using all dimensions. Finally, we show that, with simple and realistic active feedback, such as the user's interaction with a single relevant document, we can design a highly effective DIME, allowing us to outperform the baseline by up to +0.224 nDCG@10 points (+58.6%, moving from 0.384 to 0.608).

ACM Reference Format:
Guglielmo Faggioli, Nicola Ferro, Raffaele Perego, and Nicola Tonellotto. 2024. Dimension Importance Estimation for Dense Information Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3626772.3657691

1 INTRODUCTION

Information Retrieval (IR) systems have benefited from the emergence of pretrained *Large Language Models (LLMs)*, leading to the development of new systems with improved retrieval effectiveness over the previous state-of-the-art IR systems [10]. These new IR systems leverage neural networks to acquire a comprehensive understanding of documents and queries [23]. Among them, the dense IR systems rely on learning semantic representations for queries and documents, called contextualised word embeddings. These repesentations aim at better encoding the relevance of documents to queries. In dense IR systems, both query and document texts are embedded into the same latent representation space, characterised by a lower dimensionality yet denser representation than traditional IR systems. Dense IR systems differ significantly from traditional IR approaches. Unlike traditional methods like BM25 and query language models, which rely on lexical matching – where query terms in a document indicate relevance. On the other hand, dense IR systems utilize signals derived from semantic similarities in the latent space. This departure allows them to effectively address challenges related to synonymity and polysemy thanks to the underlying LLM [16, 17, 43, 46]. In a dense IR system where queries and documents are encoded as multidimensional vectors, the different dimensions of the embeddings represent features that the model has learned to be important for representing the textual content in the latent space. Each dimension of the vector may correspond to a specific aspect; for example, a given dimension could capture the semantic meaning, the syntactic structure, or other linguistic features of the encoded text. The values along those dimensions measure the importance or presence of those features in a given query or document. Ad hoc retrieval in this setting requires identifying the document embeddings nearest to the query one in the latent space and subsequently ranking them according to the specified similarity measure, e.g., dot product, in line with the *clustering hypothesis* [41], which posits that documents with similar meanings tend to be relevant to the same queries,

Bengio et al. [3] formulated the well-known *manifold hypothesis* on the latent representation spaces for images, stating that high-dimensional data of interest often lives in an unknown lower-dimensional manifold embedded in the representation space. There is strong evidence supporting this hypothesis in image representations [30], and recently, several works in the natural language processing and computational linguistics areas have found that contextualised word embeddings from LLMs lie in low-dimensional linear subspaces [15, 25] or nonlinear manifolds [5, 6].

We conjecture that both the *clustering* and the *manifold hypotheses* hold at the same time for IR and that it is possible to find a subspace of the original latent space that best represents the query and the associated relevant documents. However, instead of assuming a *single* low-dimensional subspace for all the queries and documents, we assume that each query has its low-dimensional

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07.
https://doi.org/10.1145/3626772.3657691

subspace, i.e., we have *multiple* low-dimensional subspaces, one per query, where documents can be projected as well. This aligns with the clustering hypothesis since we speculate that the subspace best representing a query topic and its relevant documents depends on the query itself. Putting everything together, we formulate the following *Manifold Clustering hypothesis (MC hypothesis)* for dense IR systems:

> High-dimensional representations of queries and documents relevant to them often lie in a query-dependent lower-dimensional manifold of the representation space.

If our MC hypothesis holds, there is a query-dependent, low-dimensional manifold in the latent space where retrieval is more effective since the query and its relevant documents are closer than in the original latent space.

While it is possible to imagine both linear and nonlinear subspaces of a given latent space, we further conjecture that not all the dimensions in the latent space of a dense representation are equally crucial for determining the relevance of a document to a specific query. In other terms, we assume it is possible to devise a subset of the dimensions of the latent space, optimal to represent the query and the documents and discard the other ones. This assumption corresponds to restricting ourselves to seek for *linear subspaces* of the original latent space, one for each query. As we will see in the following, this additional assumption, despite possibly being a first approximation, is highly effective and allows us to formulate several efficient heuristics to determine such linear subspaces. Such an assumption implies that not all the dimensions of the latent space are needed to optimize retrieval for a given query but some of them constitute a kind of noise. This is unsurprising when we consider that the number of dimensions is fixed ahead independently from queries. During training, learning algorithms globally optimize the disposal of the embeddings in the latent space. Therefore, they try to exploit the full dimensionality anyway, making it extremely unlikely to completely zero-out some dimensions, which, instead, would produce the best linear subspace for a given query.

To provide the first evidence in support of our hypothesis, we put ourselves in an ideal case, and we assume that relevant documents are known beforehand. In this context, we focus on different state-of-the-art dense IR systems [16, 17, 43] and, by relying on several TREC collections (Deep Learning 2019, 2020, DL HARD 2021, and Robust 2004), we show that there exist query-dependent linear subspaces, i.e., a specific type of manifold where dimensions are zeroed, where dense IR system performance considerably improve, moving from 0.140 to 0.308 (+119.6%) in terms of AP and from 0.360 to 0.677 (+87.8%) in terms of nDCG@10. This made us confident that our hypothesis offers ample room for improving performance.

Then, since known relevant documents are rarely available ahead in operational settings, we devise several methods to estimate which dimensions to retain and which ones to discard, and we call them *Dimension IMportance Estimators (DIMEs)*. We consider DIMEs leveraging statistical heuristics to estimate the importance of each dimension and an estimator that considers some lightweight form of user feedback to identify which dimensions are more effective for ranking. Thorough experimentation of the proposed DIMEs with state-of-the-art dense IR systems on various TREC collections show impressive performance improvements: up to +0.126 (+52.8%, moving from 0.238 to 0.364) in AP and +0.224 (+58.6%, moving from 0.384 to 0.608) in nDCG@10.

The paper is organised as follows: Section 2 summarizes the related work; Section 3 formalizes our methodology and introduces the different DIMEs used in this paper; Section 4 reports the results of our extensive and reproducible evaluation; finally, Section 5 draws some conclusions and outlines future work.

2 RELATED WORK

Classical IR systems are mostly based on exact matching: the presence of a query term within a document is considered an indicator of relevance. This approach is particularly affected by the semantic gap: a concept can be expressed using different synonyms and the same term might be polysemous, impairing the effectiveness of exact matching. With the advent of Neural IR and LLMs the focus shifted from exact term matching to semantic matching. The systems based on this novel paradigm take a piece of text, i.e., a query or a document, and project it into a latent space using a neural network. This novel representation can be either sparse, i.e., it contains as many dimensions as the terms in the vocabulary of the corpus, as in the case of SPLADE [13], or dense. In this paper, we focus on IR systems relying on dense representations. Such representations are typically smaller compared to the vocabulary size, i.e., a few hundred dimensions, but they are also much denser compared to sparse ones. Dense IR approaches project first the documents on the latent space using a projection function called "encoder". Such documents are stored efficiently in a specialized metric index, such as the one offered by FAISS toolkit [18]. At query time, the query is projected on the latent space as well. The encoder used for the query can either be the same as the one used for the documents, or a different one. In this paper, we focus on dense models that use the same encoder and project the query and the documents on the same latent space. Indeed, Izacard et al. [17] empirically observed that using the same encoder improves the robustness when the model is applied in a zero-shot scenario, i.e., the model is trained and tested on two different collections, and leaves unchanged the performances otherwise. Specifically, we focus on three state-of-the-art dense IR models: ANCE [43], Contriever [17], TAS-B [16]. ANCE is a seminal approach that relies on contrastive learning with hard negatives: given a training query, the model is trained by asking it to guess the relevant document between two documents, a relevant one and a non-relevant one ranked high by BM25 (i.e., a hard negative). Contriever is also based on contrastive learning and differs from ANCE mainly based on how positive and negative examples are chosen. Topic Aware Sampling Balanced (TAS-B) is a distillation method based on dual-teacher supervision (teacher models are BERT Cross-encoder [27] and ColBERT [20]). Furthermore, when constructing batches, it relies on Topic Aware Sampling so that batches contain queries on similar topics.

A neighbouring area with our proposal concerns feature selection for machine learning [19, 22, 34]. The objective of feature selection is to isolate a subset of all available features to improve a model's effectiveness while reducing the computational cost. There are several approaches to the feature selection task. Such approaches include the usage of ANalisys Of the VAriance (ANOVA) or the chi-squared statistics to determine the importance of each

feature [4, 11, 38] and approaches based on correlation or mutual information to determine if some features overlap in terms of provided information [29, 39, 45]. Furthermore, feature selection approaches have been successfully applied to the Learning-to-Rank task [9, 14, 31, 32]. While our DIMEs can be categorised as examples of feature selection algorithms, the major difference is that in our case selected features change on a query-per-query basis. Major feature selection approaches identify instead a set of features, regardless of the instance on which to apply the machine learning model [34]. A second difference is that, in the classical Learning-to-Rank task, features may have a semantic meaning and such meaning can be exploited to drive the selection procedure. In our case, no dimension has an active meaning: the latent semantic meaning is provided by the learning procedure and cannot be interpreted. Deciding which dimensions to preserve or remove depends on the underlying representation model and cannot be done before test time when the query is available. Determining if and how current feature selection approaches can applied for the dimension importance estimation task is left as a future work.

Another line of research relevant to our work is related to *Pseudo-Relevance Feedback (PRF)* models. These methods have a widely established and rich literature, starting with the Rocchio approach [33]. As a general pattern, PRF approaches operate by introducing additional terms to the query. Such terms can be chosen either by considering statistics of the terms in pseudo-relevant documents and the corpus [1, 2, 33], or considering the similarity between the query and the terms in a non-contextualized word-embedding space [12, 21, 35, 36, 44]. Most of the PRF approaches can be interpreted under a geometric framework. When we introduce new words in the query, we are implicitly applying a linear transformation to its representation, to "move" it where it is more likely that relevant documents are. Similarly, our DIMEs apply a spatial transformation, i.e., a projection to a linear subspace where relevant documents might be closer to the query. Nevertheless, there are two major differences: i) PRF relies on linear combinations of vectors (i.e., scaling and translations in a representation space), while the MC hypothesis conjectures that projections are the most effective transformations. ii) PRF operates only on the query representation; MC hypothesis is designed to operate the projection both on queries and documents. At the same time, as shown in Section 4, PRF and the MC hypothesis can also synergize. For example, pseudo-relevant documents can be used to instantiate a DIME that operates under the MC hypothesis.

3 METHODOLOGY

Relying on our MC hypothesis and on the assumption that finding linear subspaces is an effective strategy, in Section 3.1 we formalise the dimension importance estimation framework and introduce our *Dimension IMportance Estimators (DIMEs)*, i.e., efficient methods to assign a query-dependent importance score to the dimensions in the latent representation. These DIMEs allow us to sort the dimensions in decreasing order of estimated importance and to select the most important ones, identifying the query-dependent linear subspace at the basis of our assumption. Specifically, in Section 3.2, we define an oracle DIME to provide experimental evidence in support of the MC hypothesis in an ideal scenario where relevance judgments are known. In Section 3.3, we discuss instead several DIME methods for practical use, i.e., when relevance judgments are not known.

3.1 The Dimension Importance Estimation Framework

Let $\mathbf{q}$ and $\{\mathbf{d}1, ..., \}$ denote a query and a corpus of documents represented in the latent space $\mathbb{R}^d$ by a bi-encoder of a dense neural model. The IR system $\langle \mathbf{q}, \{\mathbf{d}1, ..., \}\rangle$ takes as input the representations of the query and the documents and produces a ranked list of documents as output. Let $\mathcal{M}\big(\langle \mathbf{q}, \{\mathbf{d}1, ..., \}\rangle\big)$, be an evaluation measure which assess the performance of the IR system for the query $\mathbf{q}$. Let W denote a subspace of $\mathbb{R}^d$. Furthermore, let π_W be the projection operator that projects a vector from $\mathbb{R}^d$ to W.

Our MC hypothesis implies that we can look for a query-dependent subspace W that maximizes the retrieval effectiveness:

$$\underset{W \subseteq \mathbb{R}^d}{\operatorname{argmax}}\ \mathcal{M}\big(\langle \pi_W(\mathbf{q}), \{\pi_W(\mathbf{d}1), ..., \}\rangle\big), \tag{1}$$

where $\mathcal{M}\big(\langle \pi_W(\mathbf{q}), \{\pi_W(\mathbf{d}1), ..., \}\rangle\big)$ denotes the evaluation measure when both the query $\mathbf{q}$ and the documents are projected from $\mathbb{R}^d$ onto the subspace W by the corresponding projection operator π_W. If eq. (1) finds a subspace W where retrieval performance improves over the full latent space in $\mathbb{R}^d$, then our MC hypothesis holds (at least for the query represented by $\mathbf{q}$).

Exploring any possible linear or nonlinear subspace W is not feasible, since the solution space would be infinite; thus, we need a way to determine a proper subspace by construction. Therefore, we assume that a linear subspace is a suitable simplification and that, among all the linear subspaces, we can restrict ourselves to those obtained by zeroing out one or more dimensions of the representations in $\mathbb{R}^d$. As discussed in Section 1, this assumption is a reasonable first approximation, which might lead to a slightly suboptimal solution, but at the great benefit of a very clear and straightforward constructive way to determine W, as we will further discuss in the following.

Therefore, we specialise the projection operator π_W to π_δ, which removes the components of a vector in $\mathbb{R}^d$ not included in a set of dimensions $\delta \subseteq \{1, \ldots, d\}$, and we rewrite eq. (1) as

$$\underset{\delta \subseteq \{1,...,d\}}{\operatorname{argmax}}\ \mathcal{M}\big(\langle \pi_\delta(\mathbf{q}), \{\pi_\delta(\mathbf{d}1), ..., \}\rangle\big). \tag{2}$$

Even if eq. (2) restricts the infinite solution space of eq. (1) to the finite solution space of finding the best subset of dimensions δ which maximizes $\mathcal{M}$, this is still a huge solution space, corresponding to the power set of the d dimensions having cardinality 2^d.

To make the problem computationally tractable, we introduce an additional assumption by considering the contribution of each dimension to retrieval effectiveness independent from each other. Such an assumption allows us to independently choose the dimensions in δ based on *Dimension IMportance Estimators (DIMEs)*, i.e., functions $u_q : 1, ..., d \rightarrow \mathbb{R}$ that associate to each representation dimension a score estimating its *importance* for query $\mathbf{q}$. In other words, we assume that given two dimensions i and j, $u_q(i) > u_q(j)$ implies that for two sets δ_i and δ_j differing only for the presence of dimension i in δ_i and j in δ_j, $\mathcal{M}\big(\langle \pi_{\delta_i}(\mathbf{q}), \{\pi_{\delta_i}(\mathbf{d}1), ..., \}\rangle\big) > \mathcal{M}\big(\langle \pi_{\delta_j}(\mathbf{q}), \{\pi_{\delta_j}(\mathbf{d}1), ..., \}\rangle\big)$.

Given the previous assumption, to address the problem in Eq. 2, we can rely on a DIME to score the d dimensions for query $\mathbf{q}$ and simply search the solution among the prefixes of the list of dimensions ordered by decreasing DIME score. To formulate its importance estimation, a DIME can rely on several possible sources of information, including the query and document representations.

Using DIMEs has two significant advantages: i) it relaxes the task, making it practical; ii) it lets us explore the behaviour of the proposed approaches for a varying number of subspace dimensions. It is worth noting that our DIMEs are query-dependent. Our objective is not to find a global ordering of the dimensions that optimizes the effectiveness performance on all queries but to find a query-dependent ordering in line with the MC hypothesis.

3.2 Oracle DIME

To assess the impact of the MC hypothesis, we propose an estimator that shows the superior retrieval effectiveness achievable by removing some of the dimensions. This oracle DIME employs all documents annotated as relevant or not relevant and thus cannot be used in real scenarios of practical interest. On the other hand, it is effective in demonstrating that: i) different dimensions have a diverse degree of importance, i.e., the MC hypothesis is grounded on the reality; and ii) there is a large margin of improvement in the effectiveness performance that can be obtained by properly selecting the correct dimensions in dense IR representations.

Let $\mathcal{R}$ be the list of annotated documents for a given query q. A relevance label r, represented as an integer, is associated with each document in $\mathcal{R}$. Depending on the collection, the integer label can be a binary value or a graded relevance assessment. Moreover, let $R \in \mathbb{R}^{d \times |\mathcal{R}|}$ be a matrix s.t. $R_{i,j} = \mathbf{q}_i \cdot \mathcal{R}_i^j$. In other terms, the element in the i-th row, j-th column of R is the product between the i-th component of the query representation $\mathbf{q}$ and the i-th component of the representation of the j-th document in $\mathcal{R}$. To assess if a dimension "correlates" positively with the relevance of documents in $\mathcal{R}$, we build the *relevance vector* $\mathbf{r} \in \mathbb{R}^{|\mathcal{R}|}$, where the j-th element is the relevance label of the j-th document in $\mathcal{R}$. For each dimension i, our oracle estimator u_q^{or} measures the Pearson's correlation ρ between the i-th column of R, $R_{:,i}$, and the relevance vector:

$$u_q^{or}(i) = \rho(\mathbf{r}, R_{:,i}). \tag{3}$$

The oracle DIME associates the maximum importance to the dimension whose corresponding column in R correlates the most with the relevance labels. Thus, the better a dimension ranks the documents according to their relevance, the more important it is.

3.3 DIMEs in Practice

Since the relevance annotations are not available in practice, we now introduce DIMEs that, differently from oracle DIME of Eq. (3), do not rely on such information.

Magnitude DIME. In this case, we assume that the information that allows us to determine the importance of each dimension is already available from the query representation $\mathbf{q}$ itself. Specifically, we hypothesize that the magnitude of each dimension of the query describes how important the dimension is for producing a good ranking. If a dimension is particularly large, it is likely associated with a latent facet that is of great importance to understanding the query. On the other hand, dimensions with small magnitudes are likely to be associated with noise and irrelevant aspects for the query and, therefore, can be neglected. Our magnitude-based DIME heuristic u_q^{mag} for dimension i is thus simply defined as:

$$u_q^{mag}(i) = |\mathbf{q}_i|, \tag{4}$$

where $\mathbf{q}_i$ denotes the i-th component of $\mathbf{q}$. Notice that we consider the absolute value of each element: saying that a query is particularly idiosyncratic toward a specific dimension – even in negative terms – should be of great importance to describe the query. A filter based on this heuristic will retain particularly large-magnitude dimensions and discard small-magnitude dimensions.

PRF DIME. This DIME operates under the assumption that the first top k_f documents retrieved are relevant, and the interaction between such documents and the query can provide effective insights on how to identify the most effective dimensions.

More in detail, given the representations $\mathbf{d}_1, ..., \mathbf{d}_{k_f}$ of the top k_f documents retrieved for the query q, which are assumed to be pseudo-relevant, we construct the representation $\mathbf{p}$ of a generic pseudo-relevant document as the centroid $\mathbf{p}$ of the representations of the retrieved documents. This allows us to instantiate our PRF DIME heuristic u_q^{PRF} for the importance of dimension i as follows:

$$u_q^{PRF}@k_f(i) = \mathbf{q}_i \cdot \mathbf{p}_i. \tag{5}$$

PRF DIME approximates the importance of dimension i as the product between the i-th dimensions of the query and the centroid of the representations of top k_f pseudo-relevant documents. We assume that if the alignment between the query and the archetypal pseudo-relevant document is particularly prominent on a certain dimension, then it is more likely that such dimension is effective for retrieval and therefore should be retained.

LLM DIME. LLMs are the current state of the art for generating documents. Therefore, given a query q, we harness their power to generate an artificial document that can be used to determine which dimensions of $\mathbf{q}$ are the most important. In more detail, we employ a state-of-the-art LLM to generate an answer in response to the query. We are not interested in investigating if the answer returned is correct, as it will not be presented to the user but used only for computing the DIME. To avoid introducing any form of bias, we do not perform any prompt engineering: we directly input the verbatim query to the LLM, without any form of preprocessing, granting the highest possible reproducibility. Once the text in response to the query has been generated by the LLM, we compute its representation $\mathbf{a}$ in the latent space. Then, the DIME based on LLM feedback u_q^{LLM} is defined as follows:

$$u_q^{LLM}(i) = \mathbf{q}_i \cdot \mathbf{a}_i. \tag{6}$$

The dimension importance is given by the product of the i-th dimension of the representations of the query and the LLM-generated answer.

Active-Feedback DIME. This DIME constructs upon the LLM DIME, by replacing the document generated by the LLM, with an actual, human-assessed, relevant document. This importance estimator cannot be a suitable option in an offline scenario, as it requires knowing, for each query, at least one relevant document. Nevertheless,

it can be particularly effective when it comes to online situations. Consider for example the case in which the user has issued a query to a search engine and has retrieved a set of documents, under the form of a *Search Engine Result Page (SERP)*. After inspecting it, the user clicks on a link corresponding to a document they consider relevant. Such a document can be then used to instantiate a DIME, reorganizing the SERP according to the active feedback provided by the user. Let thus us assume to have access to a relevant document in response to a query and let $\mathbf{s}$ be its representation in the latent space. The DIME based on Active-Feedback is defined as follows:

$$u_q^{rel}(i) = \mathbf{q}_i \cdot \mathbf{s}_i. \tag{7}$$

In other terms, the weight of each dimension is the product of the i-th dimension of the relevant document representation and the i-th dimension of the query representation.

While this DIME has a specific area of application, i.e., real-time retrieval, it is also effective in showing the power of DIMEs in identifying the optimal dimensions. In turn, it represents a sort of middle solution between the superior performance of the oracle DIME and the performance of the other, more practical DIMEs.

4 EXPERIMENTAL RESULTS

4.1 Operationalizing DIMEs

Our DIMEs produce for each dimension a score that estimates its expected relevance. Therefore, the higher the DIME score, the more likely it is that the dimension is relevant and effective in producing a good-ranked list of documents. We thus use each DIME to rank dimensions and perform ranked retrieval by considering a lower number of dimensions, studying the effect that such dimensionality reduction has on retrieval performance. In this paper, we are interested in showing that reducing the number of dimensions improves retrieval performance; we leave the task of determining the optimal number of dimensions to retain as future work.

Given a generic DIME u, we compute the projection of the query on the top k dimensions. In practical terms, this corresponds to setting to 0 the $d - k$ dimensions that are not among the top k ones according to the DIME scores. Finally, we use the novel representation of the query to rank the documents, by leaving unaltered the original representations of documents. The zeroing of the query components assures in fact that only the retained dimensions will contribute to the final query/document similarity score. This kind of operationalization of DIMEs allows for their seamless integration in already deployed retrieval pipelines: there is no need for re-indexing the collection, but it is sufficient to operate on the query representations only. Furthermore, it is possible to imagine future operationalizations where computations on ignored dimensions are skipped, increasing also retrieval efficiency.

4.2 Experimental Setup

In our experimental analysis[1], we examine three dense retrieval models: ANCE[2] [43], Contriever[3] [17], and TAS-B[4] [16]. We utilize model weights that were fine-tuned on the MS-MARCO collection and are publicly accessible on the Huggingface repository. All these models operate in 768-dimensional latent spaces. In terms of datasets, we consider four experimental collections: TREC Deep Learning '19 (DL '19) [8], TREC Deep Learning '20 (DL '20) [7], Deep Learning Hard (DL HD) [24], and TREC Robust '04 (RB '04) [42]. The first three focus on ad-hoc passage retrieval, with 43, 54, and 50 annotated queries, respectively, based on the MS MARCO passages collection [26]. RB '04 contains 249 queries and is based on the TIPSTER disks 4 and 5, minus the congressional records, corpus. It is important to note that all the dense IR systems have been fine-tuned on the MS-MARCO passages collection, making them behave as in-domain IR systems for DL '19, DL '20, and DL HD. However, for RB '04, it represents a zero-shot application of the models. To instantiate the DIME based on LLMs, we used GPT4 [28]. To assess whether improvements over the baseline are statistically significant, we use *ANalysis Of the VAriance (ANOVA)* [37] and Tukey's *Honestly Significant Differences (HSD)* post-hoc test [40] with significance level $\alpha = 0.05$.

4.3 Determining if Using Fewer Dimensions is Beneficial for Ranking

To obtain an upper bound of what could be the improvement in terms of performance if we were able to perfectly select the optimal dimensions, we report in Figure 1 the performance of the oracle DIME, presented in Subsection 3.2. For each possible configuration of collection/measure, we compute the performance of the dense IR system when considering only the first k dimensions sorted according to the DIME, ranging from 10% to 100% of the total, with a step of 10% (i.e., 10 different cutoffs). For example, given a representation in $\mathbb{R}^{768}$, we start with the top 77 dimensions identified by the DIME, and continue adding 10% of the dimensions at each step. Of course, using 100% corresponds exactly to using the model without any form of dimension importance estimation. We notice that both TAS-B and Contriever have comparable behaviour, while ANCE presents a widely different pattern across scenarios.

Contriever and TAS-B. For both these systems, the performance shows a convex pattern, starting low, reaching the maximum when more dimensions are included, and then decreasing. The oracle DIME exhibits impressive performance improvement even if only 10% of the dimensions are retained in all scenarios. The only exception is TAS-B on DL '19 with *Average Precision (AP)* as the evaluation measure: in this case, the performances are almost identical using either 10% or 100% of the dimensions. Moving beyond 10% of the dimensions retained always improves the performance. This indicates that the subsequent dimensions provide the IR system with additional relevance signals useful for increasing ranking quality. Then, the pattern observed depends on the scenario. For example, for the RB '04 and DL '20, the performance of both Contriever and TAS-B does not improve by adding subsequent dimensions beyond the first 20%. In other cases, such as DL '19 and DL HD when using AP as the evaluation measure, the improvement continues until 50% to 60% of the dimensions are retained. Let us now provide some evidence of the impressive performance boost achieved with our method based on the oracle DIME. The improvement in AP is up to +0.234 (+95.8%) for Contriever on RB '04 with 40% dimensions.

[1] source code available at: https://github.com/guglielmof/DIME-SIGIR-2024
[2] https://huggingface.co/sentence-transformers/msmarco-roberta-base-ance-firstp
[3] https://huggingface.co/facebook/contriever
[4] https://huggingface.co/sentence-transformers/msmarco-distilbert-base-tas-b

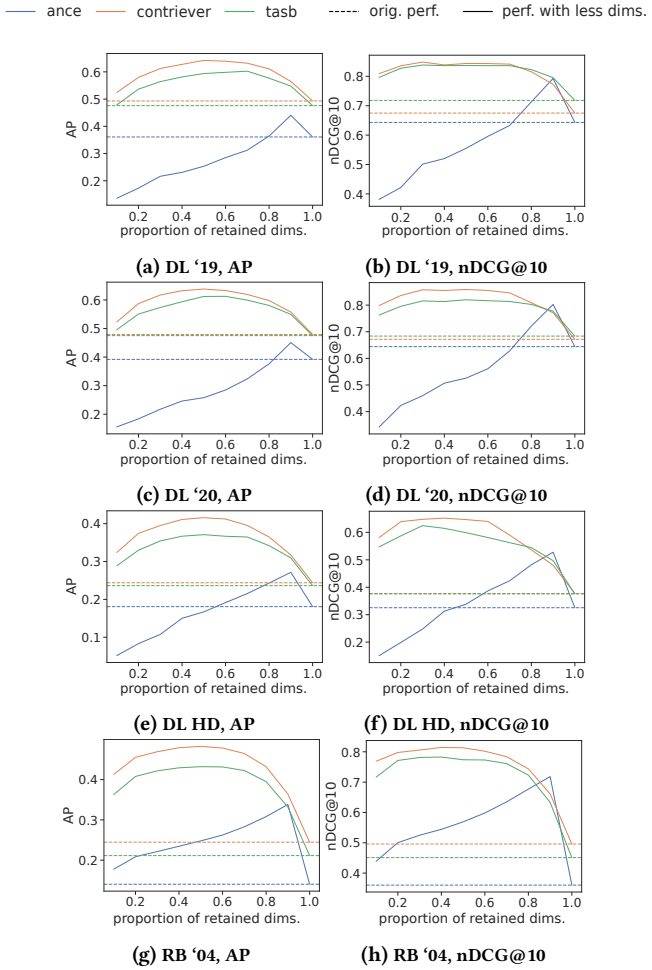


Figure 1: Retrieval performance using our oracle DIME when varying the fraction of retained dimensions. Horizontal dashed lines correspond to the performance of baseline models that use all representation dimensions.

Similarly, for nDCG@10, the improvement induced by dimension pruning can be as big as +0.332 (+73.4%) when using TAS-B for RB '04 queries with 40% dimensions. Moving forward, when more than 60-70% of the dimensions are retained, we notice a decrease in performance that continues until the absolute minimum is reached when 100% of the dimensions are considered (i.e., no DIME method is used). This indicates not only dimensions are not equally useful ranking signals but also that approximately 20-30% of the dimensions are even harmful! If we could recognize such dimensions, we could significantly improve the ranking performance.

ANCE. Interestingly, ANCE exhibits a completely different pattern. In almost all scenarios (with the exclusion of RB '04), if we consider 10% of the dimensions sorted by importance, the achieved performance is much lower than the baseline given by the original model using all dimensions. For both DL '19 and DL '20, the performance remains below the baseline until 70-80% of the dimensions are considered. This indicates that, in general, for ANCE, the information about relevance is distributed across different dimensions. After that, we observe a peak in performance in correspondence of 90% of the dimensions. Then, the last 10% of the dimensions are extremely harmful to ANCE, with a severe drop in performance. Without any dimension importance estimation, such dimensions remain diluted within the other dimensions as a form of noise, which impairs the quality of the model. Despite the different behavior, even with ANCE the performance improvement of oracle DIME is astonishing: +0.156 (+48.1%) in nDCG@10 on DL HD; +0.316 (+87.8%) in nDCG@10 on RB '04.

DIME on out-of-domain collections. An interesting pattern that can be observed is the difference between the behaviour of the models under dimension importance pruning when applied to in-domain and out-of-domain collections. On the out-of-domain collection, i.e., the RB '04, we observe a bigger performance improvement. Vice-versa for the in-domain collections, DL '19, DL '20, and DL HD, the improvement is variable. The most evident case is with ANCE: on RB '04, ANCE improves its performance over the baseline even when 10% of the dimensions are considered. When a dense IR model is applied in a zero-shot fashion on an out-of-domain collection the least important dimensions (the last 30% for Contriever and TAS-B and 10% for ANCE), are extremely harmful. While this analysis was carried out on an oracular DIME that cannot be applied in reality, it provides a good measure of the phenomenon. By properly selecting the dimensions, we can observe an improvement as high as +87.8% (in the case of ANCE and RB '04). There is no need to train additional models, reindex the collections, or change the pipeline: it is enough to identify and set some dimensions to zero to obtain such an improvement.

4.4 Assessing DIMEs Requiring or not User Feedback

Automatic DIMEs not requiring user feedback. Table 1 reports the performance achieved using u^{mag} to identify the most important dimensions. We report the results that we achieve in terms of AP and nDCG@10 when considering a variable number of retained dimensions, from 20% to 100% (original performance) with a step of 20%. We do not notice any relevant improvement over the baselines with this DIME. When there is an improvement, it occurs on the third decimal digit and is not statistically significant. These results suggest that the magnitude of the dimension is not the most prominent element in determining which dimensions are the most important. In other words, dimensions might have been weighted high by the representation model but be not particularly relevant for the ranking of documents. On the other hand, there might be dimensions that received a low weight by the encoder but that are important for the ranking. However, even with this basic DIME we obtain effectiveness figures comparable with those of the baseline by using about 60-40% of the representation dimensions. This allows the IR system to directly skip multiplications and additions necessary for computing the query/document similarity function on dimensions estimated as unimportant by the DIME. This, in turn, results in a reduction of 20-40% of the computational cost.

Nevertheless, it is likely that additional external information is needed beyond the representation of the query to estimate dimension importance. Such information can be provided either by the list of retrieved documents, as in u^{PRF}, or by using a pseudo-relevant document generated by a LLM. To demonstrate this, Table 2 re-

Table 1: Retrieval performance of the considered IR models, when u^{mag} is used to identify the most informative dimensions. While in some cases we observe a slight improvement over the baseline, such an improvement is never statistically significant.

	AP					nDCG@10					AP					nDCG@10				
Retained	0.2	0.4	0.6	0.8	1	0.2	0.4	0.6	0.8	1	0.2	0.4	0.6	0.8	1	0.2	0.4	0.6	0.8	1
	DL '19										DL '20									
ANCE	.047	.241	.319	.354	**.361**	.143	.533	.620	**.644**	.643	.106	.281	.357	.389	**.392**	.217	.520	.604	.638	**.644**
Contriever	.463	.486	.492	**.494**	.493	.648	.672	.676	**.678**	.675	.463	.478	**.479**	**.479**	**.479**	.660	.663	.664	.671	**.672**
TAS-B	.450	.466	.472	**.476**	**.476**	.693	.710	.714	**.719**	.718	.440	.465	.472	**.475**	**.475**	.658	.678	**.688**	.685	.684
	DL HD										RB '04									
ANCE	.021	.126	.164	.180	**.181**	.068	.284	.322	**.326**	.325	.027	.086	.124	.138	**.141**	.093	.260	.326	.351	**.362**
Contriever	.222	.238	.239	**.246**	.244	.362	.373	.376	**.379**	.377	.223	.238	.243	**.245**	**.245**	.462	.483	.494	.498	**.499**
TAS-B	.211	.228	.232	**.236**	**.236**	.335	**.377**	.375	.374	.376	.189	.206	.211	**.212**	**.212**	.418	.444	.450	**.454**	.453

Table 2: Performance of the DIMEs that do not require explicit user feedback. In bold, the best performance observed for each triple IR system, test collection, and evaluation measure. Values marked with * are a statistically significant improvement over the baseline using all the dimensions (corresponding to Retained = 1).

		AP					nDCG@10					AP					nDCG@10				
	Retained	0.2	0.4	0.6	0.8	1	0.2	0.4	0.6	0.8	1	0.2	0.4	0.6	0.8	1	0.2	0.4	0.6	0.8	1
		DL '19										DL '20									
ANCE	u^{PRF}@1	.033	.255	.342	**.372**		.082	.559	.644	.658		.083	.288	.366	.390		.175	.549	.616	.648	
	u^{PRF}@2	.036	.257	.339	.370		.095	.567	.637	.652		.083	.287	.366	.389		.176	.542	.612	.647	
	u^{PRF}@5	.034	.257	.344	.370	.361	.088	.568	.633	.647	.643	.077	.290	.364	.391	.392	.155	.545	.613	.645	.644
	u^{LLM}	.032	.260	.351	.370		.081	.569	.651	**.663**		.084	.284	.374	**.397**		.171	.537	.629	**.655**	
Contriever	u^{PRF}@1	.483	.503	.507	.507		.676	.685	.686	.689		.488	.498	.497	.495		.711*	.703*	.701*	.692	
	u^{PRF}@2	.493	.503	.508	.507		.672	.675	.679	.685		.478	.488	.494	.495		.682	.685	.687	.685	
	u^{PRF}@5	.491	.503	.511	.509	.493	.646	.664	.680	.681	.675	.490	.495	.497*	.495	.479	.698*	.687	.690	.686	.672
	u^{LLM}	.516	.528	**.534***	.527		.720	.742*	**.752***	.750*		.503*	**.512***	.511*	.504*		.719*	.722*	**.725***	.710*	
TAS-B	u^{PRF}@1	.487	.506*	.505*	.503		.719	.731	.733	.729		.468	.486	.491	.491		.697	.699	.709	.703	
	u^{PRF}@2	.491	.507*	.508*	.503*		.718	.733	.731	.726		.466	.481	.488	.487		.684	.698	.710	.707	
	u^{PRF}@5	.495	.501	.503*	.502*	.476	.709	.721	.719	.721	.718	.465	.478	.486	.487	.475	.683	.687	.693	.695	.684
	u^{LLM}	.512*	**.529***	.527*	.521*		.747	.749	**.760***	.755*		.483	.498	**.501***	.500*		.708	.706	.710	**.712**	
		DL HD										RB '04									
ANCE	u^{PRF}@1	.015	.126	.170	.183		.042	.266	.326	.332		.020	.096	.134	.143		.074	.284	.343	.357	
	u^{PRF}@2	.014	.125	.169	.182		.051	.274	.325	.328		.019	.092	.133	.143		.066	.273	.341	.356	
	u^{PRF}@5	.019	.125	.174	.183	.181	.054	.274	.330	.330	.325	.017	.090	.131	.144	.141	.058	.263	.334	.359	.362
	u^{LLM}	.012	.129	.175	**.186**		.042	.284	**.339**	.348		.020	.092	.134	**.146**		.078	.280	.354	**.371**	
Contriever	u^{PRF}@1	.248	.255	.254	.254		.396	.395	.387	.389		.254*	.267*	**.269***	**.269***		.512*	.522*	.527*	.523*	
	u^{PRF}@2	.253	.261	.261	.264		.395	.391	.394	.399		.257*	.266*	.268*	.267*		.500	.513*	.517*	.515*	
	u^{PRF}@5	.247	.255	.253	.252	.244	.379	.385	.383	.387	.377	.257*	.267*	.266*	.265*	.245	.504	.513*	.511*	.512*	.499
	u^{LLM}	.259	.267	**.270***	**.270***		.392	.409	**.414***	.412*		.257*	.267*	**.269***	.265*		.527*	**.539***	**.539***	.530*	
TAS-B	u^{PRF}@1	.223	.234	.234	.237		.349	.376	.374	.375		.221*	.231*	.232*	.230*		.458	.475*	.475*	.471*	
	u^{PRF}@2	.226	.239	.242	.243		.359	.377	.382	.391		.227*	.233*	**.234***	.231*		.465	.474*	.476*	.470*	
	u^{PRF}@5	.238	.239	.247	.248	.236	.364	.371	.384	.381	.376	.226*	.230*	.230*	.229*	.212	.462	.460	.462	.464	.453
	u^{LLM}	.243	.254	**.258**	.250		.385	.397	**.401**	.397		.217	.233*	.232*	.231*		.462	.487*	**.488***	.485*	

ports the results in terms of AP and nDCG@10 when filtering the dimensions using the DIMEs described in Subsections 3.3, namely the approaches based on PRF and a pseudo-relevant document generated using a LLM. The patterns follow what was observed for the oracular DIME. First of all, we notice that regardless of the setup – the triple IR model, collection, and measure considered – it is always possible to find at least one DIME that for some fraction of dimensions retained can remarkably outperform the baseline using 100% dimensions. The most effective approach is u^{LLM}, the DIME that exploits a pseudo-relevant document generated using an LLM. The overall improvement depends on multiple factors, such as which collection is considered, which measure, and which IR system is used. The improvement for ANCE is generally lower than for the other systems. With ANCE it is important to use a large fraction of the dimensions to provide an effective ranking. At the same time, there are a few dimensions, approximately 20%, which are harmful to the system: including such dimensions severely damage the ranking. Therefore, the DIME task with ANCE is particularly challenging as it is necessary to identify exactly the fraction of dimensions to retain. Specifically, we observe improvements over the baseline only when 80% of the dimensions are retained and these improvements are never statistically significant. On the contrary, for both Contriever and TAS-B, we can observe an impressive

improvement over the baseline. Indeed, the improvement for Contriever is between +0.023 (+9.55%) (AP for RB '04) up to +0.077 (+11.5%) in the case of nDCG@10 for DL '19. For TAS-B on the other hand the improvement is between 0.021 (+8.96%) in the case of AP for DL HD, to 0.053 (+11.2%) for DL '19. If we look at the results in terms of DIMEs, u^{PRF} is typically ineffective when we consider ANCE. We hypothesize that the top retrieved documents are not sufficiently relevant to represent effective pseudo-relevant documents. Vice-versa, in the case of both Contriever and TAS-B, u^{PRF} provides an improvement on average, although not always significant. As a general trend, when it comes to u^{PRF}, small values of k_f are more effective than large ones. The improvement provided by u^{LLM} on the other hand is always statistically significant for Contriever and TAS-B, except for TAS-B on the DL HD collection. As for the oracle DIME, the analysis highlights the large impact of using DIMEs for zero-shot application of IR models: when it comes to the RB '04 collection, in almost all scenarios there is a significant improvement over the baseline for both Contriever and TAS-B.

A Specific Use case: When a User Feedback is Available. The previous experiments show that DIMEs, especially u^{LLM}, are – to various extents – effective in identifying the most important dimensions and thus improve the retrieval in a completely automatic way. Nevertheless, we can easily imagine a hybrid scenario in which a user provides some feedback. Consider for example a user clicking on a relevant document of a SERP. In this case, we receive active feedback making us aware of a document that the user considers relevant to the query. Can we exploit such information to improve the retrieval by following our DIME approach? To this end, we use the active-feedback DIME u^{rel}. In particular, for each query, we assume that the user provides us with feedback on a single document that is highly relevant to the query. To simulate such feedback, for each query, we randomly pick a document with maximum relevance among those annotated for the query. We leave it as future work determining what happens when partially relevant or non-relevant documents are used as feedback. For DL '19, DL '20, and DL HD we randomly pick a document annotated with relevance "3" – the maximum – for queries having them, else "2". For RB '04 we sample among documents annotated with either "2" or "1", depending on the maximum relevance of the documents annotated for the query. Once we have a relevant document for each query, we instantiate u^{rel} and use the products of the weights in each dimension of the representations of the query and the relevant document to sort the dimensions in order of importance. Table 3 shows the performance achieved if, based on such active-feedback DIME, we retain a varying fraction of the representation dimensions. To simulate a real-life scenario, Table 3 reports the results when considering a single relevant document returned as feedback. First of all, it is interesting to notice that in all scenarios there is an improvement over the baseline. In particular, in the case of Contriever and TAS-B, the improvement is significant (and very large), regardless of the collection or evaluation measure considered. The maximum improvement is observed on the DL HD, where Contriever and TAS-B reach an impressive improvement in nDCG@10 of +0.220 (+57.7%) and +0.225 (+58.6%), respectively. ANCE, on the other hand, remains the most challenging model, with improvements that are not significant, although they are quite large in some cases (e.g., +0.056 of nDCG@10 with DL HD). Table 3 assumes a single relevant document as active feedback, to obtain comparable results. Nevertheless, we can imagine that different users might click on different documents. We are thus interested in determining if, when using different documents as feedback, the user will observe widely different performances. To this end, we repeat the experiment mentioned above 1,000 times: for each query, we pick a random highly relevant document and use it to instantiate u^{rel}. In this setting, we carry out retrieval and measure the average performance over the test queries of DL '19. Figure 2 shows the results of this experiment. Specifically, the plots report the distribution of nDCG@10 scores measured by randomly selecting 1,000 times the relevant document used to instantiate u^{rel} as a function of the fraction of dimensions retained. In line with Table 3 (but also with the oracle DIME used in Figure 1) for both Contriever and TAS-B (Figures 2b and 2c), all the fractions of retained dimensions allows improving the performance over the baseline (dashed red line). In the case of Contriever, we see that the settings using 0.4 or 0.2 of the dimensions obtain the best performance, while we achieve slightly lower nDCG@10 scores with 0.6 and 0.8 even if, also in these cases, we strongly outperform the baseline. Similarly, for TAS-B, 0.6, 0.4, and 0.2 achieve almost identical top performance, while 0.8 performs slightly lower even if always better than the baseline. On average, the choice of the relevant document instantiating the DIME has a limited impact as the performances are distributed in an interval of ±0.025 around the mean. For ANCE (Fig. 2a), in line with previous analyses, the improvement is observed only when 60%/80% of the dimensions are retained. Even in this case, the improvement is observed independently of the relevant document considered.

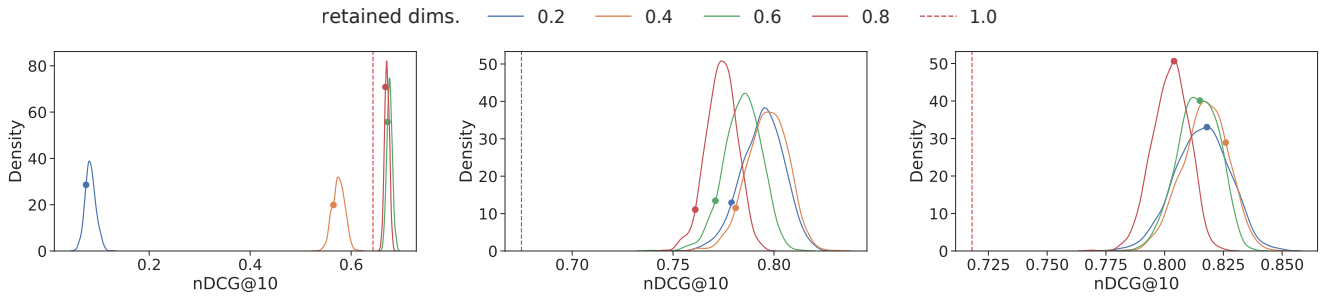
To conclude the analysis of our methodology, it is worth noting that the u^{mag}, u^{PRF}, and u^{LLM} were entirely automatic – therefore, they represent a full-fledged improvement over the current state of the art. On the contrary, the results of the u^{rel} DIME cannot be compared with purely automatic ranking strategies, as it requires some active feedback from the user. Nevertheless, its application is simple as it requires a single relevant document – we can rely for example a click of the user. Thus, it can be used online to reduce the dimensions and retrieve more precise new documents or re-rank those already retrieved. Finally, it provides a clear view of what are the achievable improvements using proper DIME techniques.

5 CONCLUSION AND FUTURE WORK

This paper introduces the MC hypothesis for the latent space learned by dense IR neural models: "high-dimensional representations of queries and documents relevant to them often lie in a query dependent lower-dimensional manifold of the representation space". According to this hypothesis, for a given query there is a subspace of the learned representation space where the representations of relevant documents tend to cluster closer around the query representation. To ground our hypothesis in empirical reality, we focus on searching such subspaces by assuming dimension independence and restricting our search to linear subspaces, i.e., subspaces of the original space where some dimensions are zeroed. To address this task practically, we define the problem of Dimension Importance Estimation. Given a dense IR model and a query, it consists of determining which dimensions of the high dimensional space are the

Table 3: Performance of the Active-Feedback DIME. Contriever and TAS-B show a significant improvement, regardless of the proportion of retained dimensions.ANCE improves when 60-80% dimensions are retained. Values marked with * are a statistically significant improvement over the baseline using all the dimensions (corresponding to Retained = 1).

	AP					nDCG@10					AP					nDCG@10				
Retained	0.2	0.4	0.6	0.8	1	0.2	0.4	0.6	0.8	1	0.2	0.4	0.6	0.8	1	0.2	0.4	0.6	0.8	1
	DL '19										DL '20									
ANCE	.034	.271	.363	**.381**	.361	.075	.565	**.672**	.668	.643	.059	.279	.378	**.393**	.392	.134	.571	.645	**.668**	.644
Contriever	.553*	**.568***	.563*	.552*	.493	.779*	**.781***	.771*	.761*	.675	.517*	.530*	**.531***	.523*	.479	**.789***	.782*	.774*	.745*	.672
TAS-B	.555*	**.569***	.562*	.551*	.476	.818*	**.826***	.815*	.804*	.718	.503*	.516*	**.521***	.515*	.475	.783*	**.797***	.786*	.765*	.684
	DL HD										RB '04									
ANCE	.027	.152	**.196**	.195	.181	.062	.328	**.384**	.365	.328	.018	.094	.147	**.151**	.141	.062	.276	.368	**.376**	.362
Contriever	.360*	**.370***	.359*	.343*	.245	.590*	**.601***	.574*	.542*	.381	.289*	.312*	**.319***	.317*	.245	.621*	**.650***	.647*	.639*	.499
TAS-B	.357*	**.364***	.353*	.340*	.238	.607*	**.608***	.594*	.568*	.384	.267*	.281*	**.282***	.275*	.212	.594*	.606*	**.609***	.586*	.453



(a) ANCE **(b) Contriever** **(c) TAS-B**

Figure 2: Distribution of performance on DL '19 for u^{act} when using different relevant documents. The dashed line represents the original performance (i.e., 100% dimensions retained). Contriever and TAS-B improves always, ANCE only when at least 60% dimensions are retained.The dot corresponds to the performance reported in Table 3.

most important to induce the optimal document ranking. At the same time, we define a novel class of models, the *Dimension IMportance Estimators (DIMEs)*. We propose an oracle DIME which allows us to show that, by appropriately selecting optimal dimensions, we improve the original retrieval performance up to 120% (from 0.140 to 0.308). While the oracle DIME effectively highlights that the MC hypothesis has ground in reality, it relies on the availability of relevant documents and cannot be used in practice. Therefore, we propose a set of DIMEs that exploits different heuristics, such as the magnitude of the dimensions of the query representation, pseudo-relevant feedback documents, and pseudo-relevant documents generated by a LLM. Even in this case, the improvement is impressive, allowing to gain +11.5% in the best scenario, moving from 0.675 to 0.752 of nDCG@10. Finally, we propose an active-feedback DIME that, by using a single relevant document provided as active feedback, is capable of largely improving the retrieval performance of dense IR models. The improvement in this scenario is as big as +52.8% (moving from 0.238 to 0.364 of AP) and +58.6% (moving from 0.384 to 0.608 of nDCG@10). Furthermore, a major advantage of DIME models is that they can be applied in any existing dense IR pipeline – either for ranking or re-ranking. It is sufficient to use a DIME model to identify the per-query subset of dimensions to retain to obtain a significant performance improvement.

Among future developments, we plan to tackle the automatic selection of the optimal number of dimensions to be retained. Additionally, we plan to explore DIME based on other signals, such as previous utterances in the conversational search scenario or query reformulations. Finally, we plan to develop DIMEs based on linear combinations of the dimensions.

ACKNOWLEDGMENTS

This work is supported, in part, by the Spoke "FutureHPC & BigData" of the ICSC – Centro Nazionale di Ricerca in High-Performance Computing, Big Data and Quantum Computing, the Spoke "Human-centered AI" of the M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research", and the EFRA project funded by the European Commission under the grant agreement n. 101093026, the FoReLab project (Departments of Excellence), the NEREO and CAMEO PRIN projects funded by the Italian Ministry of Education and Research Grant no. 2022AEFHAZ and 2022ZLL7MW. However, the views and opinions expressed are those of the authors only and do not necessarily reflect those of the EU or European Commission-EU. Neither the EU nor the granting authority can be held responsible for them.

REFERENCES

[1] Giambattista Amati. 2003. *Probability models for information retrieval based on divergence from randomness.* Ph. D. Dissertation. University of Glasgow, UK. http://theses.gla.ac.uk/1570/

[2] Gianni Amati and C. J. van Rijsbergen. 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.* 20, 4 (2002), 357–389. https://doi.org/10.1145/582415.582416

[3] Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. 2013. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 8 (2013), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50

[4] Trevor J. Bihl, Kenneth W. Bauer Jr., and Michael A. Temple. 2016. Feature Selection for RF Fingerprinting With Multiple Discriminant Analysis and Using ZigBee Device Emissions. *IEEE Trans. Inf. Forensics Secur.* 11, 8 (2016), 1862–1874. https://doi.org/10.1109/TIFS.2016.2561902

[5] Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. 2021. Isotropy in the Contextual Embedding Space: Clusters and Manifolds. In *International Conference on Learning Representations.* https://openreview.net/forum?id=xYGNO86OWDH

[6] Emily Cheng, Corentin Kervadec, and Marco Baroni. 2023. Bridging Information-Theoretic and Geometric Compression in Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 12397–12420. https://doi.org/10.18653/v1/2023.emnlp-main.762

[7] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2021. Overview of the TREC 2020 deep learning track. *CoRR* abs/2102.07662 (2021). arXiv:2102.07662 https://arxiv.org/abs/2102.07662

[8] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2020. Overview of the TREC 2019 deep learning track. *CoRR* abs/2003.07820 (2020). arXiv:2003.07820 https://arxiv.org/abs/2003.07820

[9] Maurizio Ferrari Dacrema, Fabio Moroni, Riccardo Nembrini, Nicola Ferro, Guglielmo Faggioli, and Paolo Cremonesi. 2022. Towards Feature Selection for Ranking and Classification Exploiting Quantum Annealers. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai (Eds.). ACM, 2814–2824. https://doi.org/10.1145/3477495.3531755

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. 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).* ACL, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[11] R Dhanya, Irene Rose Paul, Sai Sindhu Akula, Madhumathi Sivakumar, and Jyothisha J Nair. 2020. F-test feature selection in Stacking ensemble model for breast cancer prediction. *Procedia Computer Science* 171 (2020), 1561–1570. https://doi.org/10.1016/j.procs.2020.04.167 Third International Conference on Computing and Network Communications (CoCoNet'19).

[12] Fernando Diaz, Bhaskar Mitra, and Nick Craswell. 2016. Query Expansion with Locally-Trained Word Embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers.* The Association for Computer Linguistics. https://doi.org/10.18653/V1/P16-1035

[13] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 2288–2292. https://doi.org/10.1145/3404835.3463098

[14] Andrea Gigli, Claudio Lucchese, Franco Maria Nardini, and Raffaele Perego. 2016. Fast Feature Selection for Learning to Rank. In *Proceedings of the 2016 ACM on International Conference on the Theory of Information Retrieval, ICTIR 2016, Newark, DE, USA, September 12- 6, 2016*, Ben Carterette, Hui Fang, Mounia Lalmas, and Jian-Yun Nie (Eds.). ACM, 167–170. https://doi.org/10.1145/2970398.2970433

[15] Evan Hernandez and Jacob Andreas. 2021. The Low-Dimensional Linear Geometry of Contextualized Word Representations. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, Arianna Bisazza and Omri Abend (Eds.). Association for Computational Linguistics, Online, 82–93. https://doi.org/10.18653/v1/2021.conll-1.7

[16] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 113–122. https://doi.org/10.1145/3404835.3462891

[17] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Towards Unsupervised Dense Information Retrieval with Contrastive Learning. *CoRR* abs/2112.09118 (2021). arXiv:2112.09118 https://arxiv.org/abs/2112.09118

[18] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-Scale Similarity Search with GPUs. *IEEE Trans. Big Data* 7, 3 (2021), 535–547. https://doi.org/10.1109/TBDATA.2019.2921572

[19] Alan Jovic, Karla Brkic, and Nikola Bogunovic. 2015. A review of feature selection methods with applications. In *38th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2015, Opatija, Croatia, May 25-29, 2015*, Petar Biljanovic, Zeljko Butkovic, Karolj Skala, Branko Mikac, Marina Cicin-Sain, Vlado Sruk, Slobodan Ribaric, Stjepan Gros, Boris Vrdoljak, Mladen Mauher, and Andrej Sokolic (Eds.). IEEE, 1200–1205. https://doi.org/10.1109/MIPRO.2015.7160458

[20] Omar Khattab, Christopher Potts, and Matei Zaharia. 2021. Relevance-guided Supervision for OpenQA with ColBERT. *Trans. Assoc. Comput. Linguistics* 9 (2021), 929–944. https://doi.org/10.1162/TACL_A_00405

[21] Saar Kuzi, Anna Shtok, and Oren Kurland. 2016. Query Expansion Using Word Embeddings. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, Snehasis Mukhopadhyay, ChengXiang Zhai, Elisa Bertino, Fabio Crestani, Javed Mostafa, Jie Tang, Luo Si, Xiaofang Zhou, Yi Chang, Yunyao Li, and Parikshit Sondhi (Eds.). ACM, 1929–1932. https://doi.org/10.1145/2983323.2983876

[22] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu. 2018. Feature Selection: A Data Perspective. *ACM Comput. Surv.* 50, 6 (2018), 94:1–94:45. https://doi.org/10.1145/3136625

[23] Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, Dense, and Attentional Representations for Text Retrieval. *Trans. Assoc. Comput. Linguistics* 9 (2021), 329–345. https://doi.org/10.1162/tacl_a_00369

[24] Iain Mackie, Jeffrey Dalton, and Andrew Yates. 2021. How Deep is your Learning: the DL-HARD Annotated Deep Learning Dataset. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 2335–2341. https://doi.org/10.1145/3404835.3463262

[25] Jonathan Mamou, Hang Le, Miguel A Del Rio, Cory Stephenson, Hanlin Tang, Yoon Kim, and SueYeon Chung. 2020. Emergence of Separable Manifolds in Deep Language Representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20).* JMLR.org, Article 623, 11 pages.

[26] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016 (CEUR Workshop Proceedings, Vol. 1773)*, Tarek Richard Besold, Antoine Bordes, Artur S. d'Avila Garcez, and Greg Wayne (Eds.). CEUR-WS.org. https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf

[27] Rodrigo Frassetto Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *CoRR* abs/1901.04085 (2019). arXiv:1901.04085 http://arxiv.org/abs/1901.04085

[28] OpenAI. 2023. ChatGPT [Large language model]; Accessed on December 2023.

[29] Hanchuan Peng, Fuhui Long, and Chris H. Q. Ding. 2005. Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 8 (2005), 1226–1238. https://doi.org/10.1109/TPAMI.2005.159

[30] Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. 2021. The Intrinsic Dimension of Images and Its Impact on Learning. In *International Conference on Learning Representations.* https://openreview.net/forum?id=XJk19XzGq2J

[31] Alberto Purpura, Karolina Buchner, Gianmaria Silvello, and Gian Antonio Susto. 2021. Neural Feature Selection for Learning to Rank. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 12657)*, Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani (Eds.). Springer, 342–349. https://doi.org/10.1007/978-3-030-72240-1_34

[32] Ashwini Rahangdale and Shital A. Raut. 2019. Deep Neural Network Regularization for Feature Selection in Learning-to-Rank. *IEEE Access* 7 (2019), 53988–54006. https://doi.org/10.1109/ACCESS.2019.2902640

[33] Joseph John Rocchio Jr. 1971. Relevance feedback in information retrieval. *The SMART retrieval system: experiments in automatic document processing* (1971).

[34] Irene Rodríguez-Luján, Ramón Huerta, Charles Elkan, and Carlos Santa Cruz. 2010. Quadratic Programming Feature Selection. *J. Mach. Learn. Res.* 11 (2010), 1491–1516. https://doi.org/10.5555/1756006.1859900

[35] Dwaipayan Roy, Debasis Ganguly, Sumit Bhatia, Srikanta Bedathur, and Mandar Mitra. 2018. Using Word Embeddings for Information Retrieval: How Collection and Term Normalization Choices Affect Performance. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, Alfredo Cuzzocrea, James Allan, Norman W. Paton, Divesh Srivastava, Rakesh Agrawal, Andrei Z. Broder, Mohammed J. Zaki, K. Selçuk Candan, Alexandros Labrinidis, Assaf Schuster, and Haixun Wang (Eds.). ACM, 1835–1838. https://doi.org/10.1145/3269206.3269277

[36] Dwaipayan Roy, Debjyoti Paul, Mandar Mitra, and Utpal Garain. 2016. Using Word Embeddings for Automatic Query Expansion. *CoRR* abs/1606.07608 (2016). arXiv:1606.07608 http://arxiv.org/abs/1606.07608

[37] Andrew Rutherford. 2011. *ANOVA and ANCOVA: a GLM approach.* John Wiley & Sons.

[38] Noelia Sánchez-Maroño, María Caamaño-Fernández, Enrique F. Castillo, and Amparo Alonso-Betanzos. 2006. Functional Networks and Analysis of Variance for Feature Selection. In *Intelligent Data Engineering and Automated Learning - IDEAL 2006, 7th International Conference, Burgos, Spain, September 20-23, 2006, Proceedings (Lecture Notes in Computer Science, Vol. 4224)*, Emilio Corchado, Hujun Yin, Vicente J. Botti, and Colin Fyfe (Eds.). Springer, 1031–1038. https://doi.org/10.1007/11875581_123

[39] Kari Torkkola. 2003. Feature Extraction by Non-Parametric Mutual Information Maximization. *J. Mach. Learn. Res.* 3 (2003), 1415–1438. http://jmlr.org/papers/v3/torkkola03a.html

[40] John W. Tukey. 1949. Comparing Individual Means in the Analysis of Variance. *Biometrics* 5, 2 (1949), 99–114. http://www.jstor.org/stable/3001913

[41] C. J. van Rijsbergen. 1979. *Information Retrieval.* Butterworth.

[42] Ellen Voorhees. 2005. Overview of the TREC 2004 Robust Retrieval Track. https://doi.org/10.6028/NIST.SP.500-261

[43] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=zeFrfgyZln

[44] Hamed Zamani and W. Bruce Croft. 2016. Embedding-based Query Language Models. In *Proceedings of the 2016 ACM on International Conference on the Theory of Information Retrieval, ICTIR 2016, Newark, DE, USA, September 12- 6, 2016*, Ben Carterette, Hui Fang, Mounia Lalmas, and Jian-Yun Nie (Eds.). ACM, 147–156. https://doi.org/10.1145/2970398.2970405

[45] Zilin Zeng, Hongjun Zhang, Rui Zhang, and Chengxiang Yin. 2015. A novel feature selection method considering feature interaction. *Pattern Recognit.* 48, 8 (2015), 2656–2666. https://doi.org/10.1016/j.patcog.2015.02.025

[46] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. Optimizing Dense Retrieval Model Training with Hard Negatives. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (<conf-loc>, <city>Virtual Event</city>, <country>Canada</country>, </conf-loc>) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 1503–1512. https://doi.org/10.1145/3404835.3462880