Turning on a DIME: Estimating Dimension Importance for Dense Information Retrieval*

Extended Abstract

Guglielmo Faggioli[1], Nicola Ferro[1], Raffaele Perego[2] and Nicola Tonellotto[3,2]

[1]University of Padua, Italy
[2]National Research Council, Italy
[3]University of Pisa, Italy

Abstract

Dense Information Retrieval approaches are considered state-of-the-art and are based on projecting the queries and documents in a latent space, where each dimension encodes a latent characteristic of the text. In this paper, we enunciate the Manifold Clustering (MC) Hypothesis: projecting queries and documents onto a subspace of the original representation space can improve retrieval effectiveness. Based on the MC hypothesis, we define the Dimension IMportance Estimators (DIME). DIMEs operate on the query representation to estimate the expected importance of each dimension. Such DIMEs can be used to truncate the representation only to the most important dimensions. We describe two DIMEs, one based on the response generated by a Large Language Model (LLM), and one that relies on the user's active feedback. Our experiments show that the LLM-based DIME 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. Even more impressively, the DIME based on the active feedback allows us to outperform the baseline by up to +0.224 nDCG@10 points (+58.6%, moving from 0.384 to 0.608).

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 [2]. These new IR systems leverage neural networks to acquire a comprehensive understanding of documents and queries [3]. Among them, the dense IR systems rely on learning semantic representations for queries and documents, called contextualised word embeddings. These representations 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. 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. 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

IIR 2024: 14th Italian Workshop on Information Retrieval, September 5–6, 2024, Udine, Italy

*This is an extended abstract of [1].

document embeddings nearest to the query one in the latent space and subsequently ranking them according to the specified similarity measure.

We conjecture that it is possible to find a subspace of the original latent space that best represents the query and the associated relevant documents. Thus, 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. 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.

In this work, we describe two methods to estimate which dimensions to retain and which ones to discard, and we call them *Dimension IMportance Estimators (DIMEs)*. 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.

1.1. The Dimension Importance Estimation Framework

According to the MC hypothesis, by reducing the number of dimensions considered when computing the similarity between dense query and documents' representations, we can improve the retrieval result. Importantly, this does not mean that we train a representation model with fewer dimensions, but we take a complete representation and remove some of the dimensions from it at query time. The task is now to determine which dimensions to be removed. To this end, we employ some heuristics, which we refer to as *Dimension IMportance Estimators (DIMEs)*. A DIME is a function u that takes in input the representation $\mathbf{q} \in \mathbb{R}^d$ of a query q and – possibly – some additional information and outputs a real number for each dimension i of $\mathbf{q}$ describing its importance. 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. This operationalization allows for DIMEs 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.

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

2. Experimental Results

In our experimental analysis[1], we examine three dense retrieval models: ANCE [4], Contriever [5], and TAS-B [6]. In terms of datasets, we consider four experimental collections: TREC Deep Learning '19 (DL '19) [7], TREC Deep Learning '20 (DL '20) [8], Deep Learning Hard (DL HD) [9], and TREC Robust '04 (RB '04) [10]. To instantiate the DIME based on LLMs, we used GPT4 [11].

Table 1
Performance of the proposed DIMEs. Significant improvements over the baseline using all the dimensions (Retained = 1) are indicated with *.

		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^{LLM}	.032	.260	.351	.370	.361	.081	.569	.651	.663	.643	.084	.284	.374	.397	.392	.171	.537	.629	.655	.644
	u^{REL}	.034	.271	.363	**.381**		.075	.565	**.672**	.668		.059	.279	.378	**.393**		.134	.571	.645	**.668**	
Contriever	u^{LLM}	.516	.528	.534*	.527	.493	.720	.742*	.752*	.750*	.675	.503*	.512*	.511*	.504*	.479	.719*	.722*	.725*	.710*	.672
	u^{REL}	.553*	**.568***	.563*	.552*		.779*	**.781***	.771*	.761*		.517*	.530*	**.531***	.523*		**.789***	.782*	.774*	.745*	
TAS-B	u^{LLM}	.512*	.529*	.527*	.521*	.476	.747	.749	.760*	.755*	.718	.483	.498	.501*	.500*	.475	.708	.706	.710	.712	.684
	u^{REL}	.555*	**.569***	.562*	.551*		.818*	**.826***	.815*	.804*		.503*	.516*	**.521***	.515*		.783*	**.797***	.786*	.765*	
		DL HD										RB '04									
ANCE	u^{LLM}	.012	.129	.175	.186	.181	.042	.284	.339	.348	.328	.020	.092	.134	.146	.141	.078	.280	.354	.371	.362
	u^{REL}	.027	.152	**.196**	.195		.062	.328	**.384**	.365		.018	.094	.147	**.151**		.062	.276	.368	**.376**	
Contriever	u^{LLM}	.259	.267	.270*	.270*	.245	.392	.409	.414*	.412*	.381	.257*	.267*	.269*	.265*	.245	.527*	.539*	.539*	.530*	.499
	u^{REL}	.360*	**.370***	.359*	.343*		.590*	**.601***	.574*	.542*		.289*	.312*	**.319***	.317*		.621*	**.650***	.647*	.639*	
TAS-B	u^{LLM}	.243	.254	.258	.250	.238	.385	.397	.401	.397	.384	.217	.233*	.232*	.231*	.212	.462	.487*	.488*	.485*	.453
	u^{REL}	.357*	**.364***	.353*	.340*		.607*	**.608***	.594*	.568*		.267*	.281*	**.282***	.275*		.594*	.606*	**.609***	.586*	

[1]source code available at: https://github.com/guglielmof/DIME-SIGIR-2024

Table 1 shows the performance achieved if we retain a varying fraction of the representation dimensions based on the two DIMEs described before.

Concerning the DIME based on an LLM (u^{LLM}), we notice that, on ANCE, the improvement ranges from +0.005 for AP on DL '20, to +0.020 for nDCG@10 on DL '19– while the improvement is present, it is not 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%) (*Average Precision (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. 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.

We now consider the scenario in which the user provides us some feedback, using the DIME u^{rel}. To simulate such feedback, for each query, we randomly pick a document with maximum relevance among those annotated for the query. 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).

3. 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 address the task of finding such a space, we define the problem of Dimension Importance Estimation and a novel class of models, the DIMEs. Given a dense IR model and a query, a DIME identifies the most important dimensions to induce the optimal document ranking. We propose a DIME that exploits a pseudo-relevant document generated by a LLM which allows us to gain +11.5% in the best scenario, moving from 0.675 to 0.752 of nDCG@10. We also propose an active-feedback DIME that, by using a single relevant document is capable of largely improving the retrieval performance of dense IR models. The improvement 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).

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.

References

[1] G. Faggioli, N. Ferro, R. Perego, N. Tonellotto, Dimension importance estimation for dense information retrieval, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval July 14-18, 2024 (Washington D.C., USA), 2024.

[2] J. Devlin, M. Chang, K. Lee, K. Toutanova, 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, 2019, pp. 4171–4186. URL: https://doi.org/10.18653/v1/n19-1423. doi:10.18653/v1/n19-1423.

[3] Y. Luan, J. Eisenstein, K. Toutanova, M. Collins, Sparse, dense, and attentional representations for text retrieval, Trans. Assoc. Comput. Linguistics 9 (2021) 329–345. URL: https://doi.org/10.1162/tacl_a_00369. doi:10.1162/tacl_a_00369.

[4] L. Xiong, C. Xiong, Y. Li, K. Tang, J. Liu, P. N. Bennett, J. Ahmed, A. Overwijk, 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, 2021. URL: https://openreview.net/forum?id=zeFrfgyZln.

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

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

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

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

[9] I. Mackie, J. Dalton, A. Yates, How deep is your learning: the DL-HARD annotated deep learning dataset, in: F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, T. Sakai (Eds.), SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021, ACM, 2021, pp. 2335–2341. URL: https://doi.org/10.1145/3404835.3463262. doi:10.1145/3404835.3463262.

[10] E. Voorhees, Overview of the trec 2004 robust retrieval track, 2005. doi:https://doi.org/10.6028/NIST.SP.500-261.

[11] OpenAI, Chatgpt [large language model]; accessed on december 2023, 2023.