From Neural Re-Ranking to Neural Ranking: Learning a Sparse Representation for Inverted Indexing

Hamed Zamani[1] Mostafa Dehghani[2] W. Bruce Croft[1] Erik Learned-Miller[1] Jaap Kamps[2]
[1]University of Massachusetts Amherst [2]University of Amsterdam

1 Extended Abstract*

Retrieving unstructured documents in response to a natural language query is the core task in information retrieval (IR). Due to the importance of this task, the IR community has put a significant emphasis on designing efficient and effective retrieval models since the early years. The recent and successful development of deep neural networks for various tasks has also impacted IR applications. In particular, neural ranking models (NRMs) have recently shown significant improvements in a wide range of IR applications, such as ad-hoc retrieval, question answering, context-aware retrieval, mobile search, and product search. Most of the existing neural ranking models have a specific property in common: they are employed for *re-ranking* a small set of potentially relevant documents for a given query, provided by an efficient first stage ranker. In other words, since most neural ranking models rely on semantic matching that can be achieved using distributed dense representations, computing the retrieval score for all the documents in a large-scale collection is generally infeasible. Queries are short and terms have a highly skewed Zipfian distribution making each term relatively selective, resulting in a simple join over very few relatively short posting lists. In contrast, dense representations have an almost uniform distribution, with every term (to some degree) matching essentially all documents —similar to extreme stopwords that we cannot filter out.

Our approach addresses this head-on: by enforcing and rewarding sparsity in the representation learning, we create a latent representation that aims to capture meaningful semantic relations while still parsimoniously matching documents. This is illustrated in Figure 1, showing that the Zipfian distribution of the term space is matching far fewer documents than the dense representation returning collection-length posting lists for every term, dramatically increasing index size and query processing time. However, the latent sparse representation proposed in this paper mimics the posting list length distribution of the term based model, even matching fewer documents than term based models. That is, unlike existing neural ranking models, we propose to learn *high-dimensional sparse* representations for query and documents in order to allow for an inverted index based *standalone neural ranker* (SNRM). Our model does not require a first stage ranker and can retrieve documents from a large-scale collection as efficient as conventional term matching models.

Our main goal is learning representations for documents and queries that result in better matching compared to the original term vectors and exact matching models, while we still inherit the efficiency rooted in the sparsity of those representations. So there are two objectives, *introducing sparsity* and *capturing latent semantic meanings*. We first maps ngrams of queries and documents to a low-dimensional dense representation to compress the information, and then transform it to a high-dimensional representation pursuing the sparsity as a desired characteristic for these representations. By aggregating over sparse ngram representations, we obtain a sparse representation for a text with an arbitrary length, whose sparsity is a function of the input length: this implies higher sparsity for queries in comparison with documents, achieving an efficient retrieval model. We achieve a sparsity ratio in the learned representations that is comparable to the sparsity ratio in original term vectors.

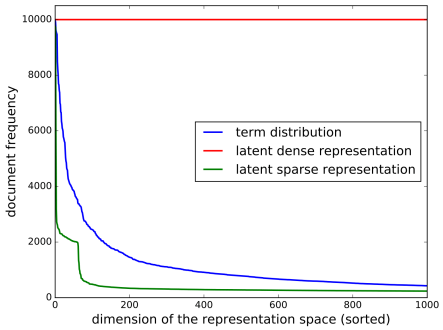

Figure 1: Document frequency in the term space (blue), the latent dense space (red), and the latent sparse space (green).

Once our latent sparse representation is trained offline, we initiate an inverted index construction phase that looks at each dimension of the learned representation as a "latent term" and builds an inverted index from each latent term to each document of the collection. At query time, we transform a given query to the learned latent high-dimensional space, and obtain its sparse representation. Given the small number of non-zero elements of the query representation and the constructed inverted index, we are able to retrieve documents from the entire collection efficiently. In addition, we can perform traditional pseudo-relevance feedback in the learned semantic space. We conduct extensive experiments with SNRM on TREC Robust and ClueWeb collections demonstrating the effectiveness of the proposed model. In summary, we show that SNRM gains in efficiency without loss of effectiveness: it not only outperforms the existing term matching baselines, but also performs similarly to the recent re-ranking based neural models with dense representations. Our model can also take advantage of pseudo-relevance feedback for further improvements. More generally, our results demonstrate the importance of sparsity in neural IR models and show that dense representations can be pruned effectively, giving new insights about essential semantic features and their distributions.

References

[1] Hamed Zamani, Mostafa Dehghani, W. Bruce Croft, Erik Learned-Miller, and Jaap Kamps. 2018. From Neural Re-Ranking to Neural Ranking: Learning a Sparse Representation for Inverted Indexing. In *CIKM'18: Proceedings of the 2018 ACM on Conference on Information and Knowledge Management*. https://doi.org/10.1145/3269206.3271800

*This is an extended abstract of Zamani et al. [1].

DIR '18, November 23, 2018, Leiden, the Netherlands