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], and Jaap Kamps[2]

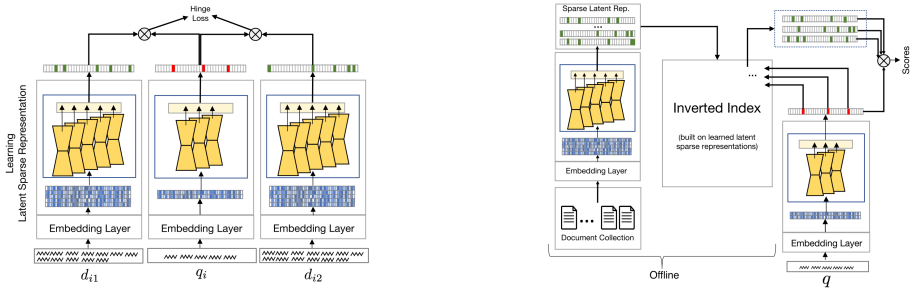
[1] University of Massachusetts Amherst
[2] University of Amsterdam

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. 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 unfeasible. 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. 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*. 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. Our model aims at maximizing the *sparsity ratio*, which is defined as the fraction of zero elements in each latent vector. Defining $0^0 = 0$, maximizing the sparsity ratio is equivalent to minimizing the L_0 norm. However, the L_0 norm is non-differentiable, which makes it impossible to train our model with the backpropagation algorithm. Therefore,

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



(a) Training time (b) Inference time

Fig. 1: General schema of the proposed SNRM model.

a tractable surrogate loss function should be considered. An alternative would be minimizing L_1 norm.

Figure 1 shows the approach. Based on the two objectives, given a document or a query, the proposed model first maps each ngram to a low-dimensional dense representation to compress the information and learn the low dimensional manifold of the data, and then learns a function to transform it to a high-dimensional representation pursuing the sparsity as a desired characteristic for these representations. By aggregating the 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, which is desired to have an efficient retrieval model. We achieve a sparsity ratio in the learned representations that is comparable to the sparsity ratio in original term vectors of documents and queries, while we are also able to better capture the semantic relevance of queries and documents using simple and efficient vector space matching functions on the learned representations.

Once our latent sparse representation is trained, 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. This is an offline process and the constructed inverted index allows for efficient retrieval. 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 a large collection efficiently. We also study pseudo-relevance feedback in the learned semantic space. We conduct extensive experiments with SNRM on newswire and large-scale web collections and demonstrate the effectiveness of the proposed model. In summary, we show that SNRM not only performs as efficiently as term matching models, e.g., query likelihood, but also performs as effectively as re-ranking based neural ranking models. Our model can take advantage of pseudo-relevance feedback and significantly outperforms competitive baselines.

References

[1] H. Zamani, M. Dehghani, W. B. Croft, E. Learned-Miller, and J. Kamps. 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*, 2018.