ReNeuIR: Reaching Efficiency in Neural Information Retrieval

Sebastian Bruch
Pinecone
New York, United States
sbruch@acm.org

Claudio Lucchese
Ca' Foscary University of Venice
Venice, Italy
claudio.lucchese@unive.it

Franco Maria Nardini
ISTI-CNR
Pisa, Italy
francomaria.nardini@isti.cnr.it

ABSTRACT

Perhaps the applied nature of information retrieval research goes some way to explain the community's rich history of evaluating machine learning models holistically, understanding that efficacy matters but so does the computational cost incurred to achieve it. This is evidenced, for example, by more than a decade of research on efficient training and inference of large decision forest models in learning-to-rank. As the community adopts even more complex, neural network-based models in a wide range of applications, questions on efficiency have once again become relevant. We propose this workshop as a forum for a critical discussion of efficiency in the era of neural information retrieval, to encourage debate on the current state and future directions of research in this space, and to promote more sustainable research by identifying best practices in the development and evaluation of neural models for information retrieval.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; **Retrieval models and ranking**.

KEYWORDS

efficiency, neural IR, sustainable IR, retrieval, ranking, algorithms and data structures

ACM Reference Format:
Sebastian Bruch, Claudio Lucchese, and Franco Maria Nardini. 2022. ReNeuIR: Reaching Efficiency in Neural Information Retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3477495.3531704

1 MOTIVATION AND THEME

Over a decade ago, machine learning transformed how we approach the ranking problem where a number of "documents" are to be ordered with respect to a "query." Many applications of ranking such as *ad hoc* search benefited substantially from a paradigm shift from early statistical methods and hand-crafted rules to what would later be called *learning to rank* (LTR) [27]. This leap was perhaps best exemplified by LambdaMART [10] in the Yahoo! Learning-to-Rank Challenge [12].

The success of LambdaMART and subsequent decision forest-based descendents [9, 15, 17, 33] in improving the quality of rankings came at the expense of the efficiency of training and inference. The training of such models is expensive because we must often learn ensembles of hundreds to thousands of deep decision trees sequentially with gradient boosting [16], with each node in every tree requiring a search in the feature space [7]. To become accurate, these large models need to be trained on vast amounts of data, often represented as complex, costly features. Inference, too, is computationally intensive because computing a score for a single query-document pair requires the traversal of paths, from roots to leaves, of every decision tree in the model.

The enormity of the new computational costs was not lost on the information retrieval (IR) community; in fact, it spawned a line of research to systematically investigate questions of efficiency and explore the trade-offs between efficiency and effectiveness in ranking models, leading to several innovations. The community widely adopted multi-stage rankers, separating light-weight ranking on large sets of documents from costly re-ranking of top candidates to speed up inference at the expense of quality [1, 3, 13, 14, 26, 34, 45]. From probabilistic data structures [2, 4], to cost-aware training and *post hoc* pruning of decision forests [5, 15, 29, 32], to early-exit strategies and fast inference algorithms [6, 11, 30, 31], the information retrieval community thoroughly considered the practicality and scalability of complex ranking algorithms. In addition to volumes of publications, the output of this research effort included standardized algorithms and reusable software packages [23, 30]. Perhaps more crucially, the community developed an understanding that quality is not the be-all and end-all of IR research, and that model complexity must be managed (through more efficient training and inference) and justified (e.g., by contextualizing quality gains in terms of the amount of computation resources required).

A decade on, deep neural networks, and in particular, Transformer-based [44] models advanced the state-of-the-art in ranking dramatically [24, 38–40]. Learnt representations of queries and documents by deep networks, too, offer a range of opportunities including the development of a new generation of retrieval methods [22, 48], document expansion techniques [41], and others. These recent developments mark the beginning of a new era known as Neural Information Retrieval (NIR).

NIR methods are a leap forward, reaching new highs in quality. Whatever the reason behind their success may be, they achieve a greater effectiveness than the previous wave of machine learning models like decision forests on many IR tasks, but with orders of magnitude more learnable parameters and much greater amounts of data. The new scale drastically increases the computational and economic costs of model training and inference. GPT-3[8], for example, required 285,000 CPU cores and 10,000 GPUs to train, with

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531704

an estimated economic cost of $4.6M [1]. Once trained, the use of such large models similarly requires a nontrivial amount of tensor multiplications and other complex operations.

Accuracy by way of ever-increasing complexity once again presents a new, but nonetheless familiar challenge that necessitates the exploration of the Pareto frontier of the two competing objectives: efficiency and effectiveness—echoes of the past decade of research albeit in a different context.

It is thus unsurprising that there is renewed interest in this space, not just in the IR community but also in related branches such as natural language processing. Interestingly, many of the proposals put forward to date to tame efficiency are reincarnations of past ideas such as stage-wise ranking with BERT-based models [35, 40], early-exit strategies in Transformers [43, 46, 47], and neural connection pruning [19, 25, 28, 36]. Other novel but general ideas such as knowledge distillation [18, 21, 42] have also proved effective in reducing the size of deep models. Yet other innovative ideas developed particularly for ranking include efforts to reinvent Transformers from the ground-up [20, 37].

While researchers in various communities concurrently investigate efficiency-related questions posed by neural network-based methods, we believe that the information retrieval community would benefit from an organized effort that is focused on NIR, as some of the works cited above show [18, 20, 37]. To facilitate that, we propose this workshop as a forum for the discussion of efficient and effective models in NIR such as ranking and dense retrieval. In particular, we wish to promote the following notions and encourage the IR community to raise and debate questions on these themes:

- **Justification**: We believe it is important to justify the ever-growing model complexity through empirical analysis.
- **Training and inference efficiency**: We encourage the development of models that require less data or computational resources for training and fine-tuning, and that offer similarly fast inference. We also ask if there are meaningful simplifications of the existing training processes or model architectures that lead to comparable quality.
- **Evaluation and reporting**: We draw attention to the lessons learnt from past IR studies and encourage a multi-faceted evaluation of NIR models from quality to efficiency, and the design of reusable benchmarks and standardized metrics.

Finally, we believe that the ACM SIGIR conference is an appropriate venue for our proposed workshop. That is because, this gathering of information retrieval researchers—who increasingly use neural network-based models in their work—would help identify specific questions in this space and shape future directions. We hope our forum would foster collaboration across interested groups.

2 TOPICS

With the objective of promoting the themes discussed in the preceding section and enabling a critical analysis and debate of each point, we solicit contributions on the following topics, including but not limited to:

- Novel NIR models that reach competitive quality but are designed to provide fast training or fast inference;
- Efficient NIR models for decentralized IR tasks such as conversational search;
- Strategies to speed up training or inference of existing NIR models;
- Sample-efficient training of NIR models;
- Efficiency-driven distillation, pruning, quantization, retraining, and transfer learning;
- Empirical investigation and justification of the complexity of existing NIR models through an analysis of quality, interpretability, or robustness; and
- Evaluation protocols for efficiency in NIR.

3 ORGANIZATION

Sebastian Bruch completed his Ph.D. in Computer Science at the University of Maryland, College Park in 2013. His research since has centered around the application of machine learning to information retrieval with a particular focus on efficiency. He has published in and served on the program committees and senior program committees of premier IR and data mining conferences like SIGIR, WSDM, SIGKDD, and the Web Conference. He currently works at Pinecone in the United States as a staff research scientist.

Claudio Lucchese is professor with the Università Ca' Foscari di Venezia - Dartiment of Environmental Sciences, Informatics and Statistics (DAIS). His main research activities are in the areas of Information Retrieval, Explainable AI, Data Mining. He has published more than 100 papers on these topics in peer reviewed international journals, conferences and other venues. He won the Best Paper Award at the ACM SIGIR Conference on Research and Development in Information Retrieval 2015. He participated to and coordinated activities in European and Italian national projects. Since 2018 he is Delegate of the Head of the Department for Research activities. He is a member of the Data Mining and Information Retrieval Lab.

Franco Maria Nardini is a senior researcher with ISTI-CNR in Pisa (Italy). He received the Ph.D. in Information Engineering from the University of Pisa in 2011. His research interests are focused on Web Information Retrieval (IR), Machine Learning (ML), and Data Mining (DM). He authored more than 70 papers in peer-reviewed international journal, conferences and other venues. In the past, he has been Tutorial Co-Chair of ACM WSDM 2021, Demo Papers Co-Chair of ECIR 2021, Program Committee Chair of the Italian Information Retrieval Workshop (IIR) in 2016 and General Chair of the International Workshop on Tourism Facilities (co-located with IEEE/WIC/ACM Web Intelligence) in 2012. He also participated to the organization of ACM SIGIR 2016 held in Pisa, Italy. Moreover, he participated to and coordinated activities in European and Italian national projects. He is co-recipient of the ACM SIGIR 2015 Best Paper Award and of the ECIR 2014 Best Demo Paper Award. He is member of the program committee of several top-level conferences in IR, ML and DM, like ACM SIGIR, ECIR, ACM SIGKDD, ACM CIKM, ACM WSDM, IJCAI, ECML-PKDD.

[1] https://lambdalabs.com/blog/demystifying-gpt-3/

4 RELATED WORKSHOPS

This workshop has not been held in the past and its co-location with SIGIR would be its first occurrence. However, to offer a sense of potential interest and participation, we would like to use this space to review existing workshops on similar themes in neighboring communities and elaborate how our proposal is different.

One noteworthy gathering is the EMC^2 workshop [2] held in recent years in conjunction with CVPR and NeurIPS by the broader machine learning community. The objective of this forum is to explore "energy efficient techniques and architectures for cognitive computing and machine learning," with a particular focus on "systems running at the edge" such as embedded systems. While we share similar values and some of our topics overlap, the scope of EMC^2 is different and, in many ways, orthogonal to ours.

The same is true of the Efficient Deep Learning in Computer Vision workshop [3] held recently at CVPR. With a focus on computer vision, the workshop serves as a venue to discuss efficiency issues, particularly those faced in on-device machine learning.

Perhaps the workshop most closely-aligned with ours is SustaiNLP [4] held jointly with EMNLP and ACL in the last two years. This workshop on Simple and Efficient Natural Language Processing promotes "more sustainable NLP research and practices" by encouraging two lines of research: (a) the development of more efficient NLP models—both in terms of training and inference efficiency; and (b) empirical justification of model complexity.

It is true that the Venn diagram of NLP and IR publications overlap considerably—this is particularly true of Transformer-based models. It may, therefore, seem redundant and unnecessary to hold a similar workshop at SIGIR. But it is also true that NIR faces unique challenges which the information retrieval community is better placed to address. This includes areas where the two fields do not intersect (e.g., dense retrieval) as well as where they do (e.g., using Transformers to rank long documents with respect to short queries). For this reason, we believe the IR community would benefit from a similar forum dedicated to NIR.

REFERENCES

[1] Nima Asadi. 2013. *Multi-Stage Search Architectures for Streaming Documents.* University of Maryland.

[2] Nima Asadi and Jimmy Lin. 2012. Fast Candidate Generation for Two-Phase Document Ranking: Postings List Intersection with Bloom Filters. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management* (Maui, Hawaii, USA). 2419–2422.

[3] Nima Asadi and Jimmy Lin. 2013. Effectiveness/Efficiency Tradeoffs for Candidate Generation in Multi-Stage Retrieval Architectures. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Dublin, Ireland). 997–1000.

[4] Nima Asadi and Jimmy Lin. 2013. Fast Candidate Generation for Real-Time Tweet Search with Bloom Filter Chains. *ACM Trans. Inf. Syst.* 31, 3, Article 13 (aug 2013), 36 pages.

[5] Nima Asadi and Jimmy Lin. 2013. Training efficient tree-based models for document ranking. In *European Conference on Information Retrieval.* Springer, 146–157.

[6] Nima Asadi, Jimmy Lin, and Arjen P. de Vries. 2014. Runtime Optimizations for Tree-Based Machine Learning Models. *IEEE Transactions on Knowledge and Data Engineering* 26, 9 (2014), 2281–2292.

[7] Leo Breiman, Jerome Friedman, Charles J. Stone, and R.A. Olshen. 1984. *Classification and Regression Trees.* Chapman and Hall/CRC.

[8] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. (2020). arXiv:2005.14165 [cs.CL]

[9] Sebastian Bruch. 2021. An Alternative Cross Entropy Loss for Learning-to-Rank. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia). 118–126.

[10] Christopher J.C. Burges. 2010. *From RankNet to LambdaRank to LambdaMART: An Overview.* Technical Report MSR-TR-2010-82. Microsoft Research.

[11] B. Barla Cambazoglu, Hugo Zaragoza, Olivier Chapelle, Jiang Chen, Ciya Liao, Zhaohui Zheng, and Jon Degenhardt. 2010. Early Exit Optimizations for Additive Machine Learned Ranking Systems. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining* (New York, New York, USA). 411–420.

[12] Olivier Chapelle and Yi Chang. 2011. Yahoo! Learning to Rank Challenge Overview. 1–24.

[13] J Shane Culpepper, Charles LA Clarke, and Jimmy Lin. 2016. Dynamic cutoff prediction in multi-stage retrieval systems. In *Proceedings of the 21st Australasian Document Computing Symposium.* ACM, 17–24.

[14] Van Dang, Michael Bendersky, and W Bruce Croft. 2013. Two-Stage learning to rank for information retrieval. In *Advances in Information Retrieval.* Springer, 423–434.

[15] Domenico Dato, Claudio Lucchese, Franco Maria Nardini, Salvatore Orlando, Raffaele Peřego, Nicola Tonellotto, and Rossano Venturini. 2016. Fast Ranking with Additive Ensembles of Oblivious and Non-Oblivious Regression Trees. *ACM Trans. Inf. Syst.* 35, 2, Article 15 (Dec. 2016), 31 pages. https://doi.org/10.1145/2987380

[16] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics* 29, 5 (2001), 1189–1232.

[17] Yasser Ganjisaffar, Rich Caruana, and Cristina Videira Lopes. 2011. Bagging gradient-boosted trees for high precision, low variance ranking models. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in I nformation Retrieval.* ACM, 85–94.

[18] Luyu Gao, Zhuyun Dai, and Jamie Callan. 2020. Understanding BERT Rankers Under Distillation. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval* (Virtual Event, Norway). 149–152.

[19] Mitchell Gordon, Kevin Duh, and Nicholas Andrews. 2020. Compressing BERT: Studying the Effects of Weight Pruning on Transfer Learning. In *Proceedings of the 5th Workshop on Representation Learning for NLP.* 143–155.

[20] Sebastian Hofstätter, Hamed Zamani, Bhaskar Mitra, Nick Craswell, and Allan Hanbury. 2020. Local Self-Attention over Long Text for Efficient Document Retrieval. In *Proc. of SIGIR.*

[21] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for Natural Language Understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020.*

[22] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

[23] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems 30.* 3146–3154.

[24] Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2020. Pretrained Transformers for Text Ranking: BERT and Beyond. *CoRR* abs/2010.06467 (2020). arXiv:2010.06467 https://arxiv.org/abs/2010.06467

[25] Zi Lin, Jeremiah Liu, Zi Yang, Nan Hua, and Dan Roth. 2020. Pruning Redundant Mappings in Transformer Models via Spectral-Normalized Identity Prior. In *Findings of the Association for Computational Linguistics: EMNLP 2020.*

[26] Shichen Liu, Fei Xiao, Wenwu Ou, and Luo Si. 2017. Cascade Ranking for Operational E-commerce Search. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 1557–1565.

[27] Tie-Yan Liu. 2009. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval* 3, 3 (2009), 225–331.

[28] Zejian Liu, Fanrong Li, Gang Li, and Jian Cheng. 2021. EBERT: Efficient BERT Inference with Dynamic Structured Pruning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021.* Association for Computational Linguistics, Online, 4814–4823. https://doi.org/10.18653/v1/2021.findings-acl.425

[29] Claudio Lucchese, Franco Maria Nardini, Salvatore Orlando, Raffaele Perego, Fabrizio Šilvestri, and Salvatore Trani. 2016. Post-Learning Optimization of Tree Ensembles for Efficient Ranking. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy) *(SIGIR '16).* ACM, New York, NY, USA, 949–952. https://doi.org/10.1145/2911451.2914763

[30] Claudio Lucchese, Franco Maria Nardini, Salvatore Orlando, Raffaele Perego, Nicola T̃onellotto, and Rossano Venturini. 2015. QuickScorer: A Fast Algorithm

[2] https://www.emc2-ai.org/

[3] https://workshop-edlcv.github.io/

[4] https://sites.google.com/view/sustainlp2021/

to Rank Documents with Additive Ensembles of Regression Trees. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. ACM, 73–82. https://doi.org/10.1145/2766462.2767733

[31] Claudio Lucchese, Franco Maria Nardini, Salvatore Orlando, Raffaele Perego, Nicola Tonellotto, and Rossano Venturini. 2016. Exploiting CPU SIMD Extensions to Speed-up Document Scoring with Tree Ensembles. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy) *(SIGIR '16)*. ACM, New York, NY, USA, 833–836. https://doi.org/10.1145/2911451.2914758

[32] Claudio Lucchese, Franco Maria Nardini, Salvatore Orlando, Raffaele Perego, and Salvatore Trani. 2017. X-DART: Blending Dropout and Pruning for Efficient Learning to Rank. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. ACM, New York, NY, USA, 1077–1080. https://doi.org/10.1145/3077136.3080725

[33] Claudio Lucchese, Franco Maria Nardini, Raffaele Perego, Salvatore Orlando, and Salvatore Trani. 2018. Selective Gradient Boosting for Effective Learning to Rank. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA). 155–164.

[34] Joel Mackenzie, J Shane Culpepper, Roi Blanco, Matt Crane, Charles LA Clarke, and Jimmy Lin. 2018. Query Driven Algorithm Selection in Early Stage Retrieval. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 396–404.

[35] Yoshitomo Matsubara, Thuy Vu, and Alessandro Moschitti. 2020. *Reranking for Efficient Transformer-Based Answer Selection.* 1577–1580.

[36] J. S. McCarley, Rishav Chakravarti, and Avirup Sil. 2021. Structured Pruning of a BERT-based Question Answering Model. arXiv:1910.06360 [cs.CL]

[37] Bhaskar Mitra, Sebastian Hofstätter, Hamed Zamani, and Nick Craswell. 2021. *Improving Transformer-Kernel Ranking Model Using Conformer and Query Term Independence.* 1697–1702.

[38] Rodrigo Nogueira and Kyunghyun Cho. 2020. Passage Re-ranking with BERT. arXiv:1901.04085 [cs.IR]

[39] Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document Ranking with a Pretrained Sequence-to-Sequence Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 708–718.

[40] Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019. Multi-Stage Document Ranking with BERT. arXiv:1910.14424 [cs.IR]

[41] Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document Expansion by Query Prediction. *arXiv preprint arXiv:1904.08375* (2019).

[42] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108 [cs.CL]

[43] Luca Soldaini and Alessandro Moschitti. 2020. The Cascade Transformer: an Application for Efficient Answer Sentence Selection. In *ACL*.

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[45] Lidan Wang, Jimmy Lin, and Donald Metzler. 2011. A cascade ranking model for efficient ranked retrieval. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in I nformation Retrieval*. ACM, 105–114.

[46] Ji Xin, Raphael Tang, Jaejun Lee, Yaoliang Yu, and Jimmy Lin. 2020. DeeBERT: Dynamic Early Exiting for Accelerating BERT Inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

[47] Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. 2021. BERxiT: Early Exiting for BERT with Better Fine-Tuning and Extension to Regression. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 91–104.

[48] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In *International Conference on Learning Representations*.