The Neural Hype, Justified! A Recantation

Jimmy Lin
David R. Cheriton School of Computer Science, University of Waterloo

1 Introduction

One year ago, in the SIGIR Forum issue of December 2018, I ranted about the "neural hype" [9]. One year later, I write again to publicly recant my heretical beliefs. What a difference a year makes! In accelerated "deep learning" time, a year seems like an eternity—so much exciting progress has been made in the previous months!

My opinion piece from a year ago has gotten some amount of attention in the community, and I think it is important to clarify my current position, based on new developments and empirical evidence that has emerged. To be clear, my previous essay made two main points, the first about the effectiveness of neural models for document retrieval:

> [...] It is unclear to me, at least for "classic" *ad hoc* retrieval problems without vast quantities of training data from behavior logs, whether neural techniques are actually more effective in absolute terms.

The second point concerned the state of empirical rigor in our field:

> [...] I am disappointed that it is not difficult to find neural ranking papers that demonstrate winning by showing statistically significant improvements over weak or inadequately-tuned baselines.

Before reflecting on the first point in depth, let me offer some quick commentary about the second: I am heartened that my message seems to have resonated with the broader community. A research study with a similar message, but focused on recommender systems, was recognized with the best paper award at RecSys 2019 [4]. The original "call to arms" by Armstrong et al. [2] seems to have received renewed vigor and attention, but whether this leads to concrete action and cultural change, it's still too early to tell.

2 Clarity

With respect to my first point, I believe there is now clarity: deep transformer models, heavily pretrained via language modeling tasks, have significantly and substantially improved the effectiveness of document retrieval, *even in the absence of vast amounts of training data*. Although BERT [6] exemplifies such models (and remains today the most popular instantiation), it is important to recognize previous work such as ELMo [15] and GPT [17] that paved the way for BERT,

and to acknowledge a panoply of related models such as XLNet [24], RoBERTa [10], T5 [18], and many more, that have followed.

For rhetorical convenience, I'll use the term "high-resource regime" to refer to web search engine companies and other large organizations with access to large amounts of behavioral log data or have the substantial financial resources necessary to gather large amounts of human editorial judgments. To contrast, I'll use the term "low-resource regime" to refer to the level of data access available to most academic research groups—relevance judgments from TREC and TREC-like evaluations, perhaps supplemented by locally-developed resources. To be clear, for the present discussion, I consider using the MS MARCO dataset [3] to be operating in the high-resource regime, although it is available to researchers today: The creation of that dataset is well beyond the resources available to NIST, other TREC-like campaign organizers, and academic research groups. We are quite fortunate that Microsoft generously shared the data with the research community, and are indebted to the track organizers for shepherding the data release.

Whereas before I presented empirical evidence, later expanded upon in Yang et al. [22], that (pre-BERT) neural models, with limited training data, struggle to even beat well-tuned "bag of words" query expansion techniques, I now believe that pretrained transformer models are unequivocally more effective than those baselines, as well as pre-BERT neural models, in the low-resource regime. Over the past year, a number of studies [23, 5, 11, 16, 14, 25] have independently established this empirical result. While some of these studies are unrefereed preprints, a sufficient number of these papers have gone through the gauntlet of peer review that the improvements observed can be considered robust.[1]

What's the catch, you might ask? Well, BERT inference is really slow compared to other neural models (and needless to say, much slower than term-based techniques). This was self-evident from the very beginning, and in the context of search, long query latencies for BERT-based models are well documented [13, 8]. Addressing this weakness is the focus of much activity, but even a cursory overview lies beyond the scope of this piece.

3 Reflection

What did I get completely wrong? It was the assumption—thinking back, a perfectly reasonable one—that an effective document (re-)ranking model needed to be fed vast amounts of relevance judgments, either derived from behavioral log data or human editorial judgments at the scale beyond what any academic institution can afford.

What I failed to foresee was a number of rapid developments, beginning near the end of 2018 and continuing through to the present.[2] The most important, of course, was the advent of models that were, essentially, self-supervised—negating the premise of my arguments at the very outset. The brilliance of the language modeling task is that the text itself serves as the prediction target, and thus the only necessary ingredients were large (unannotated) corpora and lots of compute horsepower! The first is easy to acquire, and the second is readily available at large companies. Furthermore, with BERT, Devlin et al. [6] demonstrated that starting with a pretrained model, a modest amount of annotated data (within the reach of academic research groups) is sufficient

[1] I am purposely leaving out papers that primarily use the MS MARCO dataset for the reason discussed above.

[2] Yes, I was being proven wrong at the very moment I was penning my piece last year!

to fine-tune the model for NLP tasks ranging from sentence classification to sequence labeling. These models were quite expensive to pretrain from scratch, but Google decided to make them publicly available, kicking off the seemingly endless parade of research studies whacking everything in sight with BERT. In fact, passage ranking was among the first nails pounded with the BERT hammer [12] (albeit in the high-resource regime). Multiple attempts at good 'ol document retrieval soon followed. The rest, as they say, is history.

I don't think that every progression in the above narrative was necessarily pre-ordained, but the community is very fortunate that things turned out this way. For example, Google could have elected *not* to share the source code or the pretrained models; there are plenty of previous innovations that Google has shared only in papers, but left it up to the community to build its own open-source implementation. Quite often, in these cases, the community has had to figure out from scratch many details and "tricks" that were missing from the papers. Google could have decided to share the source code but not the pretrained models, forcing everyone who wanted to use BERT to pretrain their own models from scratch. In this case, we'd hope that some other well-resourced group would pretrain models and share with everyone else. I'm glad that we don't live in one of these alternate realities, and Google should be given tremendous credit for their openness and contributions to the community.

Finally, it wasn't obvious that BERT, specifically designed for NLP tasks, would "work" for document retrieval. The history of IR is littered with ideas from NLP that intuitively "should work", but never panned out, at least with the implementations of the time. Just to give two examples from the 1990s when I was a student: word sense disambiguation (WSD) and linguistic indexing. Surely, knowing that an occurrence of "bank" refers to a financial institution rather than the side of a river *should* help improve search? Nope. A number of papers tackled this problem but couldn't demonstrate the benefits of WSD, at least robustly [21, 19]. Surely, queries that exploit linguistic relations (as opposed to bags of words) *should* improve retrieval effectiveness? Once again, the results were, at best, inconclusive [7, 20, 1]. I only provide these two examples to illustrate that it isn't obvious an NLP breakthrough would benefit IR. I'm certainly glad that it does though: What a disappointment it would have been otherwise...

4 Exciting Times

As researchers in information retrieval, and more broadly, the human language technologies, we live in exciting times. In the annals of research, written many years from now, we'll look back fondly on these few years as an incredibly innovative and creative period full of meaningful advances. Just look at all the exciting developments that have occurred since my previous piece! Future researchers, looking back, will be able to identify distinct "eras" in ranking technology: traditional term-based techniques (e.g., query expansion), learning to rank driven by feature engineering, pre-BERT neural models, and post-BERT neural models.

By no means does the story end here, though! There remain numerous unresolved issues and open research problems, many of which were discussed at TREC 2019. At the TREC workshop, I attended some of the most memorable plenary sessions in recent memory—full of everything from clever tricks to insightful proposals, capped off with the unveiling of exciting developments for next year. I returned from Gaithersburg feeling invigorated in a way that I haven't felt about TREC in a long time (and I've been participating in TREC for nearly two decades). Ian Soboroff

would probably chide me if I said more[3] (and I agree with his sentiments), so I'll end with a pitch for TREC 2020: Yes, everyone now has and uses BERT for document retrieval, but there are many ways to exploit it, and even more exciting models to examine. There are plenty more interesting avenues for exploration and ideas to pursue. Come join us on this exciting journey!

Acknowledgments

I'd like to thank Ashutosh Adhikari, Martin Gauch, Kuang Lu, Rodrigo Nogueira, and Andrew Yates for comments on earlier drafts of this piece. Feedback, however, does not imply agreement or endorsement.

References

[1] A. Arampatzis, T. Tsoris, C. H. A. Koster, and T. P. van der Weide. Phrase-based information retrieval. *Information Processing and Management*, 34(6):693–707, December 1998.

[2] T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel. Improvements that don't add up: Ad-hoc retrieval results since 1998. In *Proceedings of the 18th International Conference on Information and Knowledge Management (CIKM 2009)*, pages 601–610, Hong Kong, China, 2009.

[3] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, and T. Wang. MS MARCO: A human generated MAchine Reading COmprehension dataset. *arXiv:1611.09268v3*, 2018.

[4] M. F. Dacrema, P. Cremonesi, and D. Jannach. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems (RecSys '19)*, pages 101–109, Copenhagen, Denmark, 2019.

[5] Z. Dai and J. Callan. Deeper text understanding for IR with contextual neural language modeling. In *Proceedings of the 42nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019)*, pages 985–988, Paris, France, 2019.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019.

[7] J. L. Fagan. Experiments in automatic phrase indexing for document retrieval: A comparison of syntactic and non-syntactic methods. Technical Report TR87-868, Cornell University, Department of Computer Science, September 1987.

[3] `https://twitter.com/ian_soboroff/status/1194816360465010688`

[8] S. Hofstätter and A. Hanbury. Let's measure run time! Extending the IR replicability infrastructure to include performance aspects. In *Proceedings of the Open-Source IR Replicability Challenge (OSIRRC 2019): CEUR Workshop Proceedings Vol-2409*, pages 12–16, Paris, France, 2019.

[9] J. Lin. The neural hype and comparisons against weak baselines. *SIGIR Forum*, 52(2):40–51, 2018.

[10] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692*, 2019.

[11] S. MacAvaney, A. Yates, A. Cohan, and N. Goharian. CEDR: Contextualized embeddings for document ranking. In *Proceedings of the 42nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019)*, pages 1101–1104, Paris, France, 2019.

[12] R. Nogueira and K. Cho. Passage re-ranking with BERT. *arXiv:1901.04085*, 2019.

[13] R. Nogueira, W. Yang, J. Lin, and K. Cho. Document expansion by query prediction. *arXiv:1904.08375*, 2019.

[14] H. Padigela, H. Zamani, and W. B. Croft. Investigating the successes and failures of BERT for passage re-ranking. *arXiv:1905.01758*, 2019.

[15] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018.

[16] Y. Qiao, C. Xiong, Z. Liu, and Z. Liu. Understanding the behaviors of BERT in ranking. *arXiv:1904.07531*, 2019.

[17] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training, 2018.

[18] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv:1910.10683*, 2019.

[19] M. Sanderson. Word-sense disambiguation and information retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1994)*, pages 142–151, Dublin, Ireland, 1994.

[20] A. F. Smeaton, R. O'Donnell, and F. Kelledy. Indexing structures derived from syntax in TREC-3: System description. In *Proceedings of the Third Text REtrieval Conference (TREC-3)*, Gaithersburg, Maryland, 1994.

[21] E. M. Voorhees. Using WordNet to disambiguate word senses for text retrieval. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1993)*, pages 171–180, Pittsburgh, Pennsylvania, 1993.

[22] W. Yang, K. Lu, P. Yang, and J. Lin. Critically examining the "neural hype": weak baselines and the additivity of effectiveness gains from neural ranking models. In *Proceedings of the 42nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019)*, pages 1129–1132, Paris, France, 2019.

[23] W. Yang, H. Zhang, and J. Lin. Simple applications of BERT for ad hoc document retrieval. In *arXiv:1903.10972*, 2019.

[24] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le. XLNet: Generalized autoregressive pretraining for language understanding. *arXiv:1906.08237*, 2019.

[25] Z. A. Yilmaz, W. Yang, H. Zhang, and J. Lin. Cross-domain modeling of sentence-level evidence for document retrieval. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3481–3487, Hong Kong, China, 2019.