LLMs and the Future of Search

Anjan Goswami

Search engines fulfill our fundamental desire for knowledge by indexing the web's extensive content and delivering accurate information almost instantly, effortlessly connecting our questions to the right answers. Behind their user-friendly interfaces are sophisticated AI systems. These systems integrate complex engineering components such as web crawling, parsing, information extraction, indexing, retrieval, ranking, and query understanding. They also require innovative hardware solutions to handle the storage and processing of enormous amounts of indexed data and to ensure reliable operations for the billions of queries they receive each second.

One of the complex problems in machine learning algorithm design has been improving search result quality. Modern search engines use extensive behavioral data and semantic analysis of documents to process large volumes of web content. They assist users in crafting effective queries by providing advanced spell correction and efficient auto-completion engines, both of which are built upon vast user data and are grounded in language model technologies.

Furthermore, today's search engines boast advanced query understanding modules. These can identify named entities in user queries, refine them for increased relevance, and often enhance them with metadata drawn from historical data, given the high frequency of repeated queries.

To delve a little deeper, search engines typically operate with multiple layers of ranking. The initial retrieval layer processes the user's query, now optimized by query understanding and service modules, and searches the inverted index to pull a sufficiently large set of potentially relevant documents and then provides a ranked list of all such documents to the next layer of ranking. While the retrieval layer once relied on the classic BM25f, a simple function of term frequency and inverted document frequency normalized by document lengths, it's now more common to use sophisticated BERT-based embeddings along with simpler low latency machine learning models.

The subsequent layer, learning to rank, is where the core of machine-learned ranking unfolds. While simple one or two-layer neural networks, support vector machines, or gradient-boosted trees

have been long-standing popular algorithms for ranking, search engine companies have increasingly started to integrate signals from deep learning models based on transformer architectures. In essence, the technology underpinning modern search engines forms the basis of contemporary Large Language Models (LLMs).

The advent of Large Language Models (LLMs) marks a revolutionary stride in the domain of information retrieval. Think of these models as a single intricate recursive mathematical equation, proficient in navigating the web's breadth and responding with an expert's precision, captured in its millions or billions of weights and connections. They transcend traditional search engines, offering exhaustive, articulate answers that enhance user insight and interaction. The prevailing hypothesis for LLMs' human-like proficiency is their ability to assimilate statistical regularities, logical reasoning, and adept next-word prediction within the data. Supplementary, task-specific training further hones their question-answering prowess. Some scholars, however, posit that the exact mechanics of LLM learning and potential evolution are not fully grasped. It seems that these models can internalize grammatical and logical semantics from a multitude of texts, improving significantly when fed with clear, less ambiguous content, such as computer code, mathematical proofs, or logically structured essays. This mirrors the human progression from basic communicative needs to the articulate expression of complex notions, influenced by the semantic environment. These topics warrant a dedicated discussion; herein, our focus is the striking resemblance between the mathematics of LLMs and search engine systems — to the extent that LLMs might be seen as a mathematization of search engines, transforming multifaceted engineering systems into a singular, highly efficient mathematical function.

At the heart of LLMs is the transformer architecture, a component that contemporary search engine technologies use several of its components. The transformer's core function, 'Attention', can be delineated as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Here, 'Q' symbolizes the query, 'K' represents the keys analogous to a search engine's indexing database, and 'V' is the value. The interplay of 'Q' and 'K' initiates the retrieval process, comparable to how search engines commence potential document matching. This interaction is refined by the softmax function, reflecting the ranking stage of search results. The defining quality of LLMs — their human-like interactive capability — becomes apparent during 'instruction tuning'. This is predominantly facilitated by proximal policy optimization (PPO), a reinforcement learning technique that utilizes a reward function created by training a traditional machine learning model with human-rated LLM responses to maximize the relevance of the ranked answers. This component known as reinforcement learning from human feedback (RLHF), mirrors the training algorithm for refining search engine's learning to rank models. Thus, the foundational principles of LLMs are intricately bound with the mathematical underpinnings of search engine operations, indicating their prospective role in revolutionizing information retrieval.

Envision a transformer block as an individual search engine that retrieves information in a vectorized manner. A sequence of these blocks operates like a cascade of search engines, each layer refining the search, culminating in not just possible answers, but the most relevant response to a query.

Reflecting on LLMs' language processing prowess, some models interpret contexts up to 32,000 tokens in length, a capability that far exceeds the constraints of traditional search engines, which process far fewer words per query. While this distinction is noteworthy, it's a subtle point in the context of LLMs' broader capabilities. LLMs signify not just a gradual advancement but an

exponential leap in computational might and sophistication over the AI frameworks of existing search engines.

LLMs are lauded for their ability to generate responses that remarkably mimic human interaction, a feat they accomplish by delving into extensive text corpora. In contrast to traditional search engines, which evolve their algorithms by analyzing user clicks, LLMs bypass this necessity. They are refined through reinforcement learning from human feedback (RLHF), a method that allows them to rapidly adjust and provide more precise answers. This advancement indicates that LLMs can achieve excellence without relying on copious amounts of user data—a pertinent factor, especially when considering the potential for bias this data can introduce, a topic well-tread in search engine optimization circles. It is important to note, however, that LLMs are not immune to biases present in their training material, and the development of LLMs that minimize such biases remains an area of active research and discussion.

Geoffrey Hinton, one of the most prominent researchers in the field of deep learning, has highlighted the potential of LLMs like GPT-4. Hinton's own tests, including a riddle requiring logical and temporal reasoning, demonstrated GPT-4's comprehension and planning capabilities, Hinton talked about this in his recent 60-minute interview on TV. This suggests that the capacities of LLMs extend far beyond simple next-word predictions. Hinton sees these models as precursors to substantial benefits in sectors like healthcare, where AI has already shown promise in medical imaging and drug design. Yet, he warns of potential risks, such as job displacement, bias amplification, and the use of autonomous weapons.

Hinton calls for a prudent approach to AI development, including rigorous experimentation, regulation, and international agreements to limit military applications of AI. His call to action is a reminder that we are at a pivotal juncture, a moment that calls for careful consideration as we integrate these powerful systems into society.

The evolution of Large Language Models (LLMs) is ushering in a new era, one in which the very fabric of human experience is poised for redefinition. As LLMs grow more sophisticated, their ability to generate structured and coherent responses is not merely an incremental improvement but a transformative leap forward. This leap extends far beyond the current capabilities of technology, signaling profound changes in how we handle complex information requests and undertake a diverse range of tasks. The potential applications are vast and impactful—reimagining education through tailored learning paths, revolutionizing healthcare with predictive diagnostics, sparking innovation in the creative arts through AI collaboration, and breaking down linguistic barriers to unify global communication.

As we weave additional sensory inputs into the LLM tapestry, these AI systems will not only supplement human efforts but profoundly enhance them. This enhancement heralds a significant augmentation of human abilities, promising to overhaul traditional methodologies in problem-solving, creativity, and decision-making. In the hands of users, LLMs will become more than just tools; they will act as ever-present partners in navigating the complexities of everyday life, transforming the landscape of human-computer interaction.

In this emerging landscape, the role of conventional search engines is poised to diminish. The once-dominant search engine, while not becoming obsolete, is likely to become just one element within a broader constellation of LLM capabilities. These AI systems will provide comprehensive assistance across numerous domains, fundamentally reshaping how we search, process, and engage with information. As we stand on the cusp of this transformative era, the integration of LLMs into society calls for careful consideration. It is imperative that we harness their capabilities to amplify the human experience, all the while maintaining a vigilant stance on the ethical and societal implications of such powerful technologies.

References

[1] C. D. Manning, P. Raghavan, and H. Schütze, "Introduction to information retrieval," An Introduction To Information Retrieval, vol. 151, no. 177, p. 5.

[2] T.-Y. Liu, "Learning to rank for information retrieval," Foundations and Trends in Information Retrieval, vol. 3, no. 3, pp. 225–331.

[3] C. J. Burges, "From RankNet to LambdaRank to LambdaMART: An Overview," Microsoft Research, 2010.

[4] T. Joachims, A. Swaminathan, and T. Schnabel, "Unbiased Learning-to-Rank with Biased Feedback," CoRR, vol. abs/1608.04468, 2016.

[5] L. Wang and T. Joachims, "User Fairness, Item Fairness, and Diversity for Rankings in Two-Sided Markets," ICTIR '21: Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval, pp. 23–41, July 2021.

[6] S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," Computer Networks, vol. 30, pp. 107–117, 1998.

[7] A. Vaswani et al., "Attention is All you Need," Advances in Neural Information Processing Systems, vol. 30, 2017.

[8] A. Radford et al., "Language models are unsupervised multitask learners," OpenAI blog, 2019.

[9] J. Hoffmann et al., "Training compute-optimal large language models," arXiv preprint arXiv:2203.15556, 2022.

[10] J. Scholten et al., "Deep reinforcement learning with feedback-based exploration," 2019 IEEE 58th Conference on Decision and Control (CDC), pp. 803–808, IEEE, 2019.

[11] J. W. Rae et al., "Scaling language models: Methods, analysis & insights from training gopher," arXiv preprint arXiv:2112.11446, 2021.

[12] Y. Goldberg, "A primer on neural network models for natural language processing," Journal of Artificial Intelligence Research, vol. 57, pp. 345–420, 2016.

[13] M. Trabelsi, Z. Chen, B. D. Davison, et al., "Neural ranking models for document retrieval," Information Retrieval Journal, vol. 24, pp. 400–444, 2021. https://doi.org/10.1007/s10791-021-09398-0

[14] P. Nayak, "Understanding searches better than ever before," Google Blog, 2019. https://blog.google/products/search/search-language-understanding-bert/