بسمه تعالی

۴۰۰-الف

دانشگاه علم و صنعت ایران

دانشکده مهندسی کامپیوتر

پردیس دانشگاهی

خلاصه پیشنهاد سمینار کارشناسی ارشد

عنوان سمینار: بررسی روش‌های تولید تقویت‌شده از طریق بازیابی برای مدل های زبانی بزرگ

۱- شرح مساله (با ارجاع به مراجع)

مدل‌های زبانی بزرگ[1] با پارامترهای قابل‌توجهی که دارند، به عنوان فناوری امیدبخش برای پیشبرد پردازش زبان طبیعی[2] مطرح شده‌اند. این مدل‌ها با عملکردی برجسته در فهم و تولید زبان، نقش مهمی در تحول دنیای دیجیتال ایفا می‌کنند. با این حال، این مدل‌ها در استفاده‌های عملی با چالش‌هایی مواجه هستند، که مهم‌ترین آنها توهم[3] علمی است. این مشکل زمانی بروز می‌کند که محتوای تولیدی توسط این مدل‌ها با داده‌های موجود یا دانش قبلی همخوانی نداشته باشد، که می‌تواند به اطلاعات نادرست یا گمراه‌کننده منجر شود[۱]. تکنیک تولید تقویت‌شده از طریق بازیابی[4] با بهبود مدل‌های زبانی بزرگ از طریق بازیابی تکه سندهای مرتبط از پایگاه‌های دانش خارجی با استفاده از محاسبه شباهت معنایی، به این چالش‌ها می‌پردازد. این فرآیند به طور موثر تولید محتوای نادرست را به حداقل می رساند. ادغام تکنیک تولید تقویت‌شده از طریق بازیابی مدل های زبانی بزرگ منجر به پذیرش گسترده آن به عنوان یک فناوری کلیدی برای پیشرفت چت بات ها و برنامه های کاربردی مبتنی بر مناسبت مدل‌های زبانی بزرگ در دنیای واقعی شده است[۲].

۲- مباحث تحت پوشش سمینار(با ارجاع به مراجع)

۱. شرح کلی تکنیک تولید تقویت‌شده از طریق بازیابی:

تکنیک تولید تقویت‌شده از طریق بازیابی یا به اختصار رگ از طریق بازیابی به عنوان پلی بین مدل زبانی بزرگ و منابع دانش خارجی عمل می کند. پاسخ‌های به‌روز و جامع‌تری را حتی در مورد موضوعاتی فراتر از داده‌های از پیش آموزش دیده شده در مدل‌های زبانی بزرگ ارائه می‌گردد. این رویکرد به رفع محدودیت‌های اتکای مدل زبانی به داده‌های از پیش آموزش دیده کمک می‌کند و اجازه می‌دهد تا پاسخ‌های آگاهانه‌تر و مرتبط‌تری به پرسش‌های کاربر ارائه شود[۳]. برای پیاده سازی تولید تقویت‌شده از طریق بازیابی سه نوع دسته بندی مطرح می‌شود که شرح آن به صورت زیر است:

۱.۱. رگ ساده[5] :

اولین روشی است که مدت کوتاهی پس از پذیرش گسترده چت جی پی تی شهرت یافت. رگ ساده از یک فرآیند سنتی پیروی می کند که شامل نمایه سازی[6]، بازیابی و تولید است که و به عنوان چارچوب «بازیابی-خواندن[7]»[۴]نیز مشخص می شود.

۱.۲. رگ پیشرفته[8]:

رگ پیشرفته با استفاده از استراتژی‌های پیش بازیابی و پس بازیابی[9]، کیفیت بازیابی را ارتقا می‌دهد، تکنیک‌های شاخص‌گذاری خود را با استفاده از پنجره لغزان، تقسیم‌بندی دقیق و ادغام فرا داده بهبود می‌بخشد. علاوه بر این، از روش‌های بهینه‌سازی برای سرعت‌بخشی فرآیند بازیابی استفاده می‌کند و محدودیت‌های رگ ساده را برطرف می‌کند[۴].

۱.۳. رگ ماژولار[10]:

---

[1] Large language Model (LLM)
[2] Natural Language Processing (NLP)
[3] Hallucination
[4] Retrieval-Augmented Generation (RAG)
[5] Naive
[6] Indexing
[7] Retrieve-Read
[8] Advanced
[9] Pre-retrieval and Post-retrieval
[10] Modular

معماری رگ ماژولار، فراتر و پیشرفته‌تر از دو معماری قبلی است و سازگاری و تطبیق‌پذیری بیشتری را ارائه می دهد. این شامل استراتژی های متنوعی برای بهبود اجزای مانند افزودن یک ماژول جستجو برای جستجوهای مشابه و پالایش بازیابی‌کننده[11] از طریق تنظیم دقیق است.

۲. کاربردها:

کاربرد اصلی انواع رگ پرسش و پاسخ[12] است که شامل پرسش و پاسخ سنتی تک‌گامی[13] یا چندگامی[14] ، پرسش و پاسخ چند انتخابی و همچنین سناریو های طولانی و مناسب برای رگ است. علاوه بر پرسش پاسخ، رگ به طور مداوم به چندین وظیفه پایین‌دستی مانند استخراج اطلاعات[15]، تولید گفتگو، جستجوی کد و غیره گسترش می‌یابد[۵].

۳. روش‌های بهینه سازی در بازیابی:

در زمینه رگ، بازیابی کارآمد اسناد مربوطه از منبع داده بسیار مهم است. چندین موضوع کلیدی، مانند منبع بازیابی، دانه‌بندی بازیابی[16][۷,۶]، پردازش پیش از بازیابی، و انتخاب مدل تعبیه مربوط وجود دارد.

۴. تولید محتوا[17]:

در ارتباط با بازیابی اطلاعات برای مدل زبان بزرگ، مهم است که اصلاحات را از دو دیدگاه تهیه محتوا و تنظیم دقیق مدل‌های زبانی در نظر گرفت. تهیه محتوا شامل مرتب‌سازی مجدد بلوک‌های سند برای برجسته‌کردن نتایج مهم و فشرده‌سازی محتوا برای کاهش نویز می‌شود، در حالی که تنظیم دقیق مدل‌های زبانی امکان انجام تنظیمات براساس سناریوها و ویژگی‌های داده خاص را فراهم می‌کند[۸]. همچنین، این امکان را فراهم می‌کند که با استفاده از یادگیری تقویتی، همسان‌سازی با ترجیحات انسان یا بازیابی صورت گیرد. علاوه بر این، تنظیم دقیق می‌تواند با ترجیحات بازیابی هماهنگ شود و از تقویت مدل‌های قدرتمندتر در صورت محدودیت دسترسی به مدل‌های مخصوص یا پارامترهای بزرگتر استفاده شود[۹]

۵. فرآیند تقویتی[18] در رگ:

در تکنیک رگ، روش استاندارد اغلب شامل یک مرحله (یکبار) بازیابی و سپس تولید می‌شود که می‌تواند منجر به ناکارآمدی شود و گاهی اوقات برای مسائل پیچیده نیازمند استدلال چند مرحله‌ای هستند. این رویکرد دامنه محدودی از اطلاعات را فراهم می‌کند زیرا تنها یک بار محتوای بازیابی شده را قبل از تولید پاسخ در نظر می‌گیرد[۱۰].

۶. ارزیابی در رگ:

توسعه سریع و انتشار روزافزون رگ در حوزه پردازش زبان طبیعی باعث ارتقاء ارزیابی مدل‌های مبتنی بر رگ در جامعه مدل‌های زبانی بزرگ شده است. هدف اصلی این ارزیابی، درک و بهینه‌سازی عملکرد مدل‌های رگ در سناریوهای کاربردی متنوع است که اجزای حیاتی این فرآیند ارزیابی، شامل هدف ارزیابی، جنبه‌های ارزیابی، معیارهای ارزیابی و ابزارها هستند که امکان ارزیابی جامع و بهینه‌سازی مدل‌های رگ را فراهم می‌کنند[۱۱].

۷. چالش‌ها و جهت‌های توسعه آینده فناوری در رگ:

با وجود پیشرفت قابل توجه در فناوری رگ، چندین چالش باقی‌مانده است که هرکدام نیازمند تحقیقات عمیق هستند. دراین بخش چالش‌های فعلی و جهت‌های تحقیقات آینده حوزه رگ مانند، مقایسه رگ با متن‌های طولانی، استحکام رگ[۷]، رویکردهای ترکیبی[۹] قوانین مقیاس‌پذیری رگ[۱۲]، رگ آماده تولید و رگ چند وجهی معرفی می‌گردد.

۳- اهمیت موضوع

رگ با ترکیب دانش از پایگاه‌های داده خارجی با مدل‌های زبانی بزرگ، بهبود دقت و اعتبار تولید محتوا، به‌ویژه برای کارهای دانش‌محور، به یک راه‌حل مناسب برای تولید متن قابل اعتماد تبدیل شده است. توانایی آن در تفسیر و پردازش انواع داده‌ها، مانند تصاویر، ویدیوها و

[11] Retriever
[12] Question Answering
[13] Single-Hop
[14] Multi-Hop
[15] Information Extraction
[16] Retrieval Granularity
[17] Content Generation
[18] Augmentation

کد، کاربرد آن را در حوزه‌های چندوجهی گسترش داده است مانند مدل های چند وجهی که قادر به بازیابی و تولید متن و تصاویر هستند. این به این معنی است که هم با تصاویر و هم با متن کار می کنند و نتایج عملی قابل توجه، آن را برای استقرار هوش مصنوعی برجسته کرده است. این پیشرفت توجه بخش‌های دانشگاهی و صنعتی را به دلیل پتانسیل رگ برای به‌روزرسانی مداوم دانش و ادغام اطلاعات مربوط به دامنه به خود جلب کرده است. از برنامه های آینده رگ می توان به خلاصه سازی اسناد، پرسش و پاسخ و ادامه دهنده متن اشاره کرد[۲].

۴- منابع

1. N. Kandpal, H. Deng, A. Roberts, E. Wallace, and C. Raffel, "Large language models struggle to learn long-tail knowledge," in International Conference on Machine Learning. PMLR, 2023, pp. 15 696–15 707.
2. Lyu, Yuanjie, Zhiyu Li, Simin Niu, Feiyu Xiong, Bo Tang, Wenjin Wang, Hao Wu, et al. "CRUD-RAG: A Comprehensive Chinese Benchmark for Retrieval-Augmented Generation of Large Language Models." arXiv, February 18, 2024.
3. Gao, Yunfan, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. "Retrieval-Augmented Generation for Large Language Models: A Survey." arXiv, March 27, 2024.
4. X. Ma, Y. Gong, P. He, H. Zhao, and N. Duan, "Query rewriting for retrieval-augmented large language models," arXiv preprint arXiv:2305.14283, 2023.
5. F. Shi, X. Chen, K. Misra, N. Scales, D. Dohan, E. H. Chi, N. Sch¨arli, and D. Zhou, "Large language models can be easily distracted by irrelevant context," in International Conference on Machine Learning.PMLR, 2023, pp. 31 210–31 227.
6. W. Yu, H. Zhang, X. Pan, K. Ma, H. Wang, and D. Yu, "Chain-of-note: Enhancing robustness in retrieval-augmented language models," arXiv preprint arXiv:2311.09210, 2023.
7. X. Du and H. Ji, "Retrieval-augmented generative question answering for event argument extraction," arXiv preprint arXiv:2211.07067, 2022
8. X. V. Lin, X. Chen, M. Chen, W. Shi, M. Lomeli, R. James, P. Rodriguez, J. Kahn, G. Szilvasy, M. Lewis et al., "Ra-dit: Retrieval augmented dual instruction tuning," arXiv preprint arXiv:2310.01352, 2023.
9. O. Yoran, T. Wolfson, O. Ram, and J. Berant, "Making retrievalaugmented language models robust to irrelevant context," arXiv preprint arXiv:2310.01558, 2023.
10. Y. Hoshi, D. Miyashita, Y. Ng, K. Tatsuno, Y. Morioka, O. Torii, and J. Deguchi, "Ralle: A framework for developing and evaluating retrieval-augmented large language models," arXiv preprint arXiv:2308.10633, 2023.
11. T. Zhang, S. G. Patil, N. Jain, S. Shen, M. Zaharia, I. Stoica, and J. E.Gonzalez, "Raft: Adapting language model to domain specific rag," arXiv preprint arXiv:2403.10131, 2024.
12. U. Alon, F. Xu, J. He, S. Sengupta, D. Roth, and G. Neubig, "Neurosymbolic language modeling with automaton-augmented retrieval," in International Conference on Machine Learning. PMLR 2022, pp. 468–485.

۵- نتیجه ارزیابی در گروه:

قبول ☐ رد ☐ تصحیح ☐ ارسال برای داوری ☐ تاریخ --/--/---- امضاء مدیر گروه: