УДК 004.89

Б. Ю. Варер
В. Б. Мокін
Б. І. Мокін

ОГЛЯД ТЕХНІК ПІДВІЩЕННЯ ЕФЕКТИВНОСТІ ВИКОРИСТАННЯ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ ДЛЯ РОЗВ'ЯЗАННЯ ПРИКЛАДНИХ ЗАДАЧ

Вінницький національний технічний університет

Анотація

Розглянуто використання великих мовних моделей (LLMs) спільно з різними новітніми методами покращення їх ефективності. Дослідження акцентує увагу на нових підходах та стратегіях, які включають використання реляційної бази даних разом з LLMs, кооперацію моделей та використання векторних баз даних для побудови Case-Based Reasoning. Дослідження підкреслює важливість розробки нових алгоритмів, архітектур та підходів, що сприятимуть подальшому розвитку використання великих мовних моделей.

Ключові слова: машинне навчання, великі мовні моделі, бази даних, кооперація моделей, Case-Based Reasoning.

Abstract

The use of large language models (LLMs) in conjunction with various state-of-the-art methods to enhance their effectiveness is examined. The research focuses on novel approaches and strategies, including the utilization of relational databases with LLMs, model cooperation, and the integration of vector databases for Case-Based Reasoning. The study underscores the importance of developing new algorithms, architectures, and approaches to further advance the utilization of large language models.

Keywords: machine learning, large language models, databases, model cooperation, Case-Based Reasoning.

Вступ

У сучасному інформаційному суспільстві значимість та потужність штучного інтелекту зростають в істотній мірі. Машинне навчання та великі мовні моделі, як вагомі складові сучасного штучного інтелекту, проявляють величезний потенціал для трансформації різноманітних сфер життєдіяльності, зокрема освіти, науки, бізнесу, медицини, інформаційних технологій тощо [1].

Втім, незважаючи на значні досягнення у цьому напрямку, використання великих мовних моделей вимагає вирішення проблем, пов'язаних з обмеженими можливостями контекстуального розуміння, особливо в складних задачах, які вимагають глибшого аналізу [2]. У такому контексті з'являється необхідність проведення досліджень нових підходів і методологій, що допоможуть подолати ці обмеження. Таким чином, важливим є розроблення нових алгоритмів, архітектур та підходів, які дозволять покращити контекстуальне розуміння та здатність генерації тексту великими мовними моделями.

Враховуючи ці виклики, можна підкреслити два основних шляхи до покращення ефективності великих мовних моделей. По-перше, можна вносити зміни безпосередньо до моделі. Це може включати модифікацію архітектури моделі, використання спеціалізованих технік навчання, або тонку настройку (fine-tuning), що передбачає додаткове навчання моделі на специфічному для задачі наборі даних. Однак, цей підхід може бути складним для рядового користувача, оскільки він часто вимагає значних обчислювальних ресурсів для перенавчання моделей та глибокого розуміння машинного навчання.

По-друге, можна покращити ефективність за допомогою різних методологій використання великих мовних моделей, поєднуючи їх з іншими інструментами та іншими моделями. Цей підхід має як теоретичну, так і практичну цінність, оскільки він відкриває нові шляхи для ефективного використання великих мовних моделей в різних додатках.

Огляд методологій підвищення ефективності великих мовних моделей

У сфері використання великих мовних моделей для вирішення завдань все більше з'являються дослідження, спрямовані на підвищення їх ефективності та функціональності. Ці дослідження охоплюють як існуючі підходи, так і новітні методи, що допомагають удосконалити роботу великих мовних моделей у різних контекстах і завданнях. Огляд існуючих підходів дозволяє нам зрозуміти, як великі мовні моделі використовуються на сьогоднішній день та які результати досягнуті. Але на фоні стрімкого розвитку цієї галузі, новітні дослідження пропонують нові підходи та методи, які можуть значно покращити ефективність та точність використання великих мовних моделей.

Метою даного дослідження є аналіз технік та методологій, які дозволяють покращити контекстуальне розуміння та можливості великих мовних моделей щодо вирішення задач. Цей огляд спрямований на те, щоб надати уявлення про поточний стан досліджень у сфері використання великих мовних моделей та вказати на можливості та виклики, які відкриваються перед науковою спільнотою.

Одним із варіантів прогресу у цій області є розробка та втілення концепцій співпраці великих мовних моделей. Ця ідея має потенціал до створення систем, що спроможні розробляти рішення для задач більшої складності, ніж ті, що доступні одиничним моделям, посилюючи розуміння контексту та збільшуючи загальну ефективність. Це відкриває нові горизонти для оптимізації вирішення складних задач та може допомогти досягнути глибшого, більш інтегрованого розуміння. У дослідженні [3] було продемонстровано, що кооперація трьох моделей GPT-3.5, які виконували ролі "аналітика", "програміста" та "тестувальника", приводило до вищої ефективності в розв'язанні задач з програмування в порівнянні з використанням однієї моделі GPT-3.5 і навіть перевищило ефективність використання моделі наступного покоління GPT-4. Це підкреслює величезний потенціал співпраці великих мовних моделей у вирішенні важких завдань. Розподіл різних ролей між моделями, як було зазначено в дослідженні, може істотно покращити загальний контекст та глибину розуміння завдання, враховуючи різні перспективи та деталі. Цей результат акцентує на необхідності подальших досліджень цього напрямку, незалежно від наявності більш сучасних моделей, таких як GPT-4. Подальші дослідження цього підходу можуть відкрити нові можливості для оптимізації вирішення складних системних задач за допомогою співпрацюючих мовних моделей.

Великі моделі мови (LLM), не дивлячись на свою потужність, зіткнулися з обмеженнями щодо зберігання та використання контексту в багатоходових взаємодіях. Зокрема, це обмеження кількості токенів на вхід, яке можуть обробляти ці моделі. У відповідь на ці проблеми, дослідники запропонували використання баз даних як символічної пам'яті для LLM у своєму фреймворку ChatDB [4]. У цьому фреймворку пам'ять може зберігати історичну інформацію в структурованій формі та сприяти точним операціям з даними за допомогою SQL-інструкцій, що генеруються LLM. ChatDB також вводить підхід "ланцюг пам'яті" (chain-of-memory), який спрощує складні проблеми, розкладаючи їх на послідовність проміжних операцій з пам'яттю. Цей підхід поліпшує здатність LLM виконувати точні маніпуляції з базами даних. Таким чином, фреймворк ChatDB представляє собою інноваційний підхід до покращення ефективності LLM, використовуючи символічну пам'ять для підтримки складного багатоступеневого розуміння. Цей підхід дозволяє моделям більш точно та ефективно виконувати комплексні маніпуляції з даними, включаючи роботу з багатьма таблицями в базі даних. Такий підхід до обробки даних виявився особливо корисним для реальних застосувань, які включають в себе складні та точні взаємодії з історичними даними, наприклад, ведення записів і аналіз даних в управлінні.

Таким чином, ChatDB використовує бази даних як символічну пам'ять, що дозволяє великим мовним моделям краще виконувати свої завдання. Використання зовнішньої символічної пам'яті в базах даних може стати важливим напрямком у вирішенні складних проблем та викликів, пов'язаних з історичними даними і довготривалими процесами обробки даних.

Застосування векторних баз даних спільно з великими мовними моделями представляє значний потенціал для методу Case-Based Reasoning. Метод Case-Based Reasoning (CBR) є підходом до розуміння і розв'язання проблем на основі аналізу аналогічних ситуацій, що вже були вирішені у минулому. В основі CBR лежить ідея використання досвіду з раніше вирішених ситуацій для вироблення адаптованих рішень для нових проблем. У контексті використання векторних баз даних разом з великими мовними моделями, метод CBR отримує новий потенціал. Векторні бази даних забезпечують ефективне зберігання та доступ до інформації, включаючи дані про раніше вирішені ситуації. За допомогою великих мовних моделей, таких як GPT-4, CBR може використовувати цей досвід для аналізу та порівняння аналогічних ситуацій, а також для вироблення адаптованих рішень

для нових проблем. Цей підхід починається з трансформації прецедентів з векторної бази даних за допомогою LLM, перетворюючи кожен випадок в семантичний векторний об'єкт, що зберігається в векторній базі даних (VDB).

При надходженні нового запиту або задачі, LLM генерує відповідне векторне представлення. VDB використовуються для пошуку векторів, що найкращим чином відповідають вектору запиту, служачи ідентифікаторами схожих прецедентів з бази даних. Вибрані випадки потім використовуються як контекст для LLM, розширюючи запит контекстуальними даними, які відображають спосіб вирішення подібних проблем в минулому.

На основі цього контексту, LLM здійснює прогнозування або генерує висновок. Таким чином, цей підхід ефективно поєднує гнучкість і масштабність LLMs з точністю і ефективністю пошуку схожих випадків, яку надають VDBs.

Ця методологія представляє значні можливості для практичних застосувань, зокрема в рекомендаційних системах, системах підтримки прийняття рішень, обробці природної мови та інтелектуальному аналізі даних, водночас підвищуючи ефективність використання великих мовних моделей в контексті Case-Based Reasoning.

Схожа методологія використовувалась у дослідженні використання Case-Based Reasoning з LLM для класифікації логічних помилок [5]. У цьому дослідженні автори використовували великі мовні моделі для перетворення текстових випадків логічних помилок у багатовимірні вектори. Використовуючи косинусну схожість між цими векторами, вони здійснювали пошук схожих випадків у базі даних. Однак, автори цього дослідження не використовували спеціалізовану інфраструктуру векторної бази даних, але натомість використовували свої власні системи для зберігання і пошуку векторів. Тим не менш, засновуючись на цих принципах, методика, що включає використання векторної бази даних, може поліпшити ефективність і швидкість процесу пошуку схожих випадків.

Висновки

У підсумку, використання великих мовних моделей, таких як GPT-4, в комбінації з різними методологіями та інструментами, має значний потенціал для покращення ефективності та точності роботи цих моделей. Дослідження в цій області відкривають нові можливості для розширення контекстуального розуміння, глибшого аналізу та ефективної генерації тексту.

Застосування методів, які поєднують великі мовні моделі з різними інструментами, такими як векторні бази даних чи методологія Case-Based Reasoning, використання кооперації великих мовних моделей або інтеграція LLM з реляційною базою даних, дозволяє вирішувати складні задачі з більшою ефективністю та точністю. Продовження досліджень у цьому напрямку буде сприяти розвитку нових методологій, архітектур та підходів, які допоможуть покращити функціональність великих мовних моделей. Подальші дослідження вимагатимуть зосередженості на вирішенні проблем, пов'язаних з ефективним поєднанням різних методів та забезпечення етичного використання цих моделей.

Загалом, зростаюча роль великих мовних моделей у сучасному інформаційному суспільстві вимагає постійного дослідження та розвитку нових методологій для покращення їх ефективності та точності. Здатність використовувати різні підходи та інструменти в комбінації з великими мовними моделями відкриває шляхи для нових інноваційних застосувань та розв'язання складних проблем у різних галузях життєдіяльності.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Liu, Yiheng, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, et al. "Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models." arXiv, May 10, 2023. Accessed May 10, 2023. http://arxiv.org/abs/2304.01852.
2. Ray, Partha Pratim. "ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope." Internet of Things and Cyber-Physical Systems 3 (2023): 121–54. doi:10.1016/j.iotcps.2023.04.003.
3. Dong, Yihong, Xue Jiang, Zhi Jin, and Ge Li. "Self-Collaboration Code Generation via ChatGPT." arXiv, May 24, 2023. Accessed May 24, 2023. http://arxiv.org/abs/2304.07590.
4. Hu, Chenxu, Jie Fu, Chenzhuang Du, Simian Luo, Junbo Zhao, and Hang Zhao. "ChatDB: Augmenting LLMs with Databases as Their Symbolic Memory." arXiv, 2023. Accessed June 20, 2023. http://arxiv.org/abs/2306.03901.
5. Sourati, Zhivar, Filip Ilievski, Hông-Ân Sandlin, and Alain Mermoud. "Case-Based Reasoning with Language Models for Classification of Logical Fallacies." arXiv, 2023. Accessed June 20, 2023. http://arxiv.org/abs/2301.11879.

Варер Борис Юхимович **–** аспірант кафедри системного аналізу та інформаційних технологій, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: androbor17@gmail.com

Мокін Віталій Борисович **–** д-р. техн. наук, проф., завідувач кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, Вінниця, e-mail: vbmokin@vntu.edu.ua

Мокін Борис Іванович – академік НАПН України, д-р техн. наук, професор кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, Вінниця, e-mail: borys.mokin@gmail.com

Varer Borys Y. **–** postgraduate student of the System Analysis and Information Technologies, Faculty of Intelligent Information Technologies and Automation, Vinnytsia National Technical University, Vinnytsia, e-mail; androbor17@gmail.com

Mokin Vitalii B. **–** Dr. tech. Sciences, Prof., Head of the Department of System Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: vbmokin@vntu.edu.ua

Mokin Borys I. — Academician of NAPS of Ukraine, Dr. tech. Sc. (Eng.), Professor of the Chair of System Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia e-mail: borys.mokin@gmail.com