continues to play a pivotal role in enhancing data management and decision-making processes, shaping the future of information systems.

References

[1] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

[2] Chen, J., Song, L., Liu, L., & Qin, Y. (2019). Big data deep learning: challenges and perspectives. IEEE Access, 7, 47528-47555.

[3] Jurafsky, D., & Martin, J. H. (2020). Speech and Language Processing. Pearson.

Денис Ган
аспірант кафедри КН, НЛТУ України, Львів, Україна,
han.denys@nltu.lviv.ua

Використання векторних баз даних для додавання нової інформації до запитів до великої мовної моделі (LLM)

Анотація – Зараз все більшої популярності з кожним днем набувають великі мовні моделі. Вони здатні вирішувати велику низку проблем набагато краще в порівнянні з існуючими підходами машинного навчання та штучного інтелекту, зокрема розпізнавання мови, токенізація, відповіді на запитання, формування стислого переказу тексту і багато інших. Також вони часто досягають неочікувано хороших результатів навіть у сферах на які велика мовна модель не є розрахованою – наприклад математичні проблеми [3].

Майже всі великі мовні моделі що є у відкритому доступі мають недоліки у вигляді того, що вони не мають пам’ять про попередні запити, не мають доступу до інформації про приватні дані користувача та не мають доступу до нових подій і знань що трапилися після тренування моделі.

Використовуючи векторні бази даних можна уникнути проблеми описаної вище зберігаючи потрібні дані у БД та передаючи їх в велику мовну модель під час запиту.

Ключові слова – велика мовна модель, LLM, векторна база даних, llama_index, Faiss.

* науковий керівник

Люди зазвичай сприймають текст як зв’язаний набір слів, однак для комп’ютерів текст – це лише послідовність символів. Перші спроби дозволити машинам розуміти текст були побудовані на основі рекурентних нейронних мереж. Ця модель обробляє по одному слову або символу та надає вихід, коли весь вхідний текст буде використано. Цей підхід працює відносно добре, за винятком того, що модель іноді “забуває” те, що сталося на початку послідовності, коли досягнуто кінця.

У 2017 році Vaswani та інші опублікували статтю «Attention is all you need» [1], що описувала створення моделі трансформатора. В його основі лежить механізм уваги. На відміну від рекурентних нейронних мереж, механізм уваги дозволяє бачити все речення (або навіть абзац) відразу, а не одне слово за раз. Це дозволяє моделі трансформатора краще розуміти контекст слова. Багато найсучасніших моделей обробки мови засновані на трансформаторній архітектурі.

Векторне вбудовування (*Embedding - англ.*) – це тип представлення даних, який несе семантичну інформацію, що допомагає системам штучного інтелекту краще розуміти дані, а також мати можливість підтримувати довготривалу пам’ять. У всьому новому що можна вивчити важливими елементами є розуміння теми та її запам’ятовування.

Векторне вбудовування за допомогою традиційних скалярних баз даних є проблемою, оскільки LLM не може впоратися з масштабом і складністю даних. У зв’язку з усією складністю, яка пов’язана з векторним вбудовуванням є необхідність в спеціалізованій базі даних, що дозволить спростити векторне вбудовування для великих мовних моделей.

Векторна база даних використовує різні алгоритми, які допомагають у пошуку приблизного найближчого сусіда (ANN). Це робиться за допомогою хешування, пошуку на основі графів або квантування, які збираються в конвеєр для отримання сусідів запитуваного вектора.

Результати базуються на тому, наскільки вони близькі або наближені до запиту, тому основними елементами, які враховуються, є точність і швидкість.

* науковий керівник

Одним з прикладів векторної БД, що може використовуватися для задач вбудовування, є Faiss що реалізує алгоритм “інвертованого мульти-індексу” [2]. Її можна підключати до великих мовних моделей як напряму так і через різні бібліотеки, що дозволяють абстрагувати використання різних LLM та БД, такі як llama_index.

Висновок. Загалом, використовуючи вищеописаний підхід можна зберегти будь-які дані у векторній БД і згодом використовувати їх при запитах до LLM, що вирішує один з основних недоліків які вони мають, а саме відсутність доступу до нових подій і даних після створення.

Список використаних літературних джерел

[1] Vaswani, Ashish, et al. Attention Is All You Need. 1, arXiv, 12 June 2017. arXiv.org, https://doi.org/10.48550/arXiv.1706.03762.

[2] Babenko, Artem, and Victor Lempitsky. ‘The Inverted Multi-Index’. 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 3069–76. IEEE Xplore, https://doi.org/10.1109/CVPR.2012.6248038.

[3] Improving mathematical reasoning with process supervision [Електронний ресурс]. – Режим доступу: https://openai.com/research/improving-mathematical-reasoning-with-process-supervision.

Максим Іванов, Ігор Пірко*

магістрант кафедри КН, НЛТУ України, Львів, Україна.

* к.ф.-м.н., доцент кафедри КН, НЛТУ України, Львів, Україна.

Розроблення інформаційно-аналітичної системи прогнозу цін на комп’ютерну техніку

Абстракт – В роботі представлено систему прогнозування цін на ноутбуки за допомогою методів машинного навчання. Для аналізу використано методи прогнозування машинного навчання Support Vector Machine, KNN, Decision Tree, Random Forest. Розроблено інформаційно-аналітичну систему, в якій ціна є залежною змінною, що формується на основі таких факторів як модель ноутбука та його конфігурація.

* науковий керівник