досягнуто ліміту дозволених помилок. В адаптивному ж тестуванні це може бути при досягненні певної межі «рівня знань» в обидва боки. В цьому полягає одна з ключових переваг адаптивного тестування, а саме – точність оцінювання знань користувача.

Вдосконалення системи тестування доцільно здійснювати за рахунок проектування та навчання моделі машинного навчання, яка має виконувати роль механізму оцінки користувача. Крім того, складність самих завдань має змінюватись в залежності від відповідей користувачів.

Список використаних джерел

1. Howard W. Computerized adaptive testing: A Primer (2nd Edition). Mahwah, NJ: ELawrence Erlbaum Associates, 2000, 361p.
2. Олійник М. М. Тест як інструмент кількісної діагностики рівня знань в сучасних технологіях навчання. Донецьк: Донецький національний університет, 2001. 83 с.

ЛУЦЕНКО Р. С.,
РОМАНЮК О. В.,
Вінницький національний технічний університет

ПЕРСПЕКТИВИ ЗАСТОСУВАННЯ ГІБРИДНОГО АДАПТИВНОГО СКОРОЧЕННЯ РАНГУ ДЛЯ ОПТИМІЗАЦІЇ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ НА МОБІЛЬНИХ ПРИСТРОЯХ

Анотація: У роботі розглянуто метод гібридного адаптивного скорочення рангу (HARR) як новітній підхід до оптимізації великих мовних моделей (LLM) для мобільних пристроїв. Метод базується на поєднанні технік Low-Rank Adaptation (LoRA) та прунингу, що дозволяє ефективно адаптувати моделі під обмежені ресурси мобільних платформ, таких як обсяг пам'яті та потужність процесора. Впровадження HARR забезпечує зменшення обсягу моделей без втрати точності, сприяє підвищенню енергоефективності та швидкодії додатків на основі великих мовних моделей.

Ключові слова: мовні моделі, скорочення рангу, мобільні платформи, LoRA, прунинг, оптимізація.

Потреба в адаптації великих мовних моделей для мобільних платформ значно зросла останнім часом, що обумовлено широким впровадженням технологій штучного інтелекту в мобільні додатки. Однак мобільні пристрої, такі як смартфони, мають обмежені апаратні ресурси, що створює низку проблем для реалізації повномасштабних моделей. Скорочення розміру моделей є необхідною умовою для забезпечення ефективного виконання та економії енергії на таких пристроях. Метод HARR пропонує інноваційний підхід до зниження рангу матриць параметрів моделей, що дозволяє скоротити обчислювальне навантаження, зберігаючи при цьому високу точність.

Відсутність оптимізованих під мобільні платформи моделей призводить до необхідності розробки нових методів, здатних адаптувати LLM для пристроїв з обмеженими ресурсами. На сьогодні техніки Low-Rank Adaptation та прунингу застосовуються окремо, однак їх інтеграція може забезпечити більш гнучке та ефективне рішення, яке одночасно враховуватиме потреби зменшення розміру та забезпечення точності моделі.

Техніки LoRA та прунингу широко досліджуються у контексті оптимізації глибоких нейронних мереж. LoRA, як метод зниження рангу, дозволяє зменшити кількість параметрів шляхом використання низькорівневих матриць для основних компонентів моделі, що дає змогу значно зекономити пам'ять [1]. Прунинг же зосереджений на видаленні незначущих нейронів та зв'язків, що додатково скорочує модель та знижує її обчислювальну складність [2]. Проте, питання сумісного використання LoRA та прунингу в умовах обмежених ресурсів мобільних платформ досі залишається недостатньо дослідженим. Запропонований метод HARR заповнює цю прогалину, забезпечуючи ефективне поєднання цих технік та врахування апаратних обмежень.

Метод HARR складається з кількох основних етапів. Перший етап передбачає застосування техніки LoRA для створення низькорівневих матриць, що забезпечує початкове скорочення рангу. Далі застосовується адаптивний прунинг, що додатково видаляє елементи моделі, які мають низький вплив на загальну точність. Оскільки мобільні платформи мають індивідуальні обмеження щодо обсягу пам'яті та доступної потужності процесора, HARR включає етап збору інформації про апаратні характеристики пристрою. Ця інформація використовується для адаптивного налаштування

рівня прунингу та зниження рангу в залежності від обчислювальних можливостей конкретного пристрою [3]. Це забезпечує збалансований компроміс між точністю моделі та швидкодією.

Завдяки гнучкості HARR, модель може виконуватися на більшій кількості пристроїв, включаючи старіші моделі смартфонів, де обмеження обчислювальних ресурсів є особливо відчутними. Метод дозволяє значно зекономити пам'ять та забезпечує стабільність роботи, що робить його перспективним для мобільних додатків з високим попитом на обробку природної мови, таких як чат-боти [4], інтелектуальні помічники та системи перекладу.

Запропонований метод HARR демонструє ефективність у зменшенні розміру мовних моделей для мобільних пристроїв без значної втрати точності, що забезпечує більш стабільну роботу додатків на платформах з обмеженими ресурсами. Його застосування відкриває нові можливості для розробки високоякісних інтелектуальних мобільних додатків, що підтримують обробку природної мови. Подальші дослідження можуть бути спрямовані на удосконалення HARR для його адаптації під нові, більш складні моделі, що підвищить його ефективність у майбутніх поколіннях мобільних пристроїв.

Список використаних джерел

1. Романюк О. В., Луценко Р. С. «Вплив оптимізації великих мовних моделей для iOS на розвиток освітніх, медичних і розважальних додатків», в матеріалах конференції Комп'ютерні ігри і мультимедіа як інноваційний підхід до комунікації - 2024, Одеса, 2024. – с. 294-296.
2. Han, Song, Huizi Mao, and William J. Dally. "Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding." in Proc. Int. Conf. on Learning Representations (ICLR), 2016, pp. 1-14.
3. Sze, Vivienne, Yu-Hsin Chen, Tien-Ju Yang, and Joel S. Emer. "Efficient Processing of Deep Neural Networks: A Tutorial and Survey" in Foundations and Trends in Electronic Design Automation, vol. 12, no. 3, Morgan & Claypool Publishers, 2020, pp. 239-252.
4. Tang, Yanming, Fu-Dong Zhang. "Low-Power Deep Learning and Applications to Mobile and Embedded Devices.", Springer, 2021, pp. 78-95.

МАЙДАНЮК В. П.
Вінницький національний технічний університет

ВИКОРИСТАННЯ СЕРВІСУ MATLAB ONLINE В НАВЧАЛЬНОМУ ПРОЦЕСІ

Анотація: Показано, що використання онлайн сервісів може бути хорошою альтернативою дорогим пакетам прикладних програм, придбання яких часто через нестачу коштів неможливе для навчальних закладів. Зокрема, це стосується і MATLAB. Наведено приклад використання MATLAB online для вивчення цифрових фільтрів та візуалізації результатів фільтрації.

Ключові слова: онлайн сервіс, MATLAB, цифровий фільтр.

Abstract: It is shown that the use of online services can be a good alternative to expensive packages of application programs, the purchase of which is often impossible for educational institutions due to lack of funds. In particular, this applies to MATLAB. An example of using MATLAB online to study digital filters and visualize filtering results is given.

Keywords: online service, MATLAB, digital filter.

Вступ. Використання безкоштовних онлайн-сервісів в навчальному процесі є хорошою альтернативою дорогим пакетам прикладних програм, придбання яких часто через нестачу коштів неможливе для навчальних закладів. Зокрема це стосується і MATLAB.

MathWorks Inc. надає доступ до MATLAB online (20 годин на місяць, доступ до 10 наборів інструментів), який дозволяє працювати в браузері (без встановлення), реєстрація та вхід за посиланням: https://www.mathworks.com/products/matlab-online/matlab-online-versions.html [1].

Перевагами MATLAB-онлайн такі [2]:

- використання MATLAB без завантаження чи встановлення;
- співпраця з іншими учасниками шляхом обміну інформацією та публікації в Інтернет;
- зберігання файлів, керування ними та отримання доступу до них у будь-якому місці.

MATLAB Online надає доступ до MATLAB з будь-якого стандартного веб-браузера з доступом в Інтернет - необхідно просто увійти до системи. Він ідеально підходить для викладання, навчання та зручного, легкого доступу.