УДК 004.8:004.912

Дзюба В.В., ст. викладач
Український державний університет науки і технологій

РОЗРОБКА ТА НАВЧАННЯ КОМПАКТНИХ МОВНИХ МОДЕЛЕЙ ДЛЯ ЗАСТОСУВАННЯ В ОСВІТНІХ СИСТЕМАХ ТА КІБЕРБЕЗПЕЦІ

У прикладних сценаріях освіти й кібербезпеки, де важливими є автономність, конфіденційність даних і можливість роботи на локальному обладнанні, перспективним є напрям створення компактних мовних моделей, придатних для запуску на споживчих GPU та локальних серверних вузлах. Додатковою перевагою такого підходу є можливість адаптації моделі до предметної області закладу освіти або до спеціалізованих наборів даних із кібербезпеки без постійної залежності від зовнішніх хмарних сервісів.

Сучасні мовні моделі великого масштабу забезпечують високу якість розв'язання широкого кола завдань обробки природної мови, проте їх розробка та навчання стримується значними вимогами до обсягу відеопам'яті графічного адаптера та наявності стабільного електроживлення обчислювальної системи на якій проводиться навчання.

Разом із тим дослідження архітектур компактних моделей має власну методологічну складність: кожна зміна конфігурації моделі, токенізатора, довжини контексту чи гіперпараметрів потребує окремого тренувального запуску, а отже – окремого обчислювального ресурсу.

Для моделі порядку 350 млн параметрів повноцінне навчання вже не є «швидким експериментом», особливо в умовах нестабільного електроживлення та обмеженого локального обладнання. У цьому контексті критично важливим стає не лише підбір архітектури, а й планування конкретних експериментів, різні методи та підходи до скорочення часу обчислень.

Роботи з обчислювально-оптимального масштабування доводять, що якість моделі визначається не лише кількістю параметрів, а й співвідношенням між розміром моделі та обсягом тренувальних токенів [1]. На практиці у низці робіт показано, що оцінка обчислювально-оптимального режиму узгоджується з практичним правилом близько 20 токенів на параметр у межах підходу Chinchilla [2].

З огляду на це у роботі пропонується дворівневий підхід до розробки моделі. На першому етапі виконується пошук архітектури та перевірка навчального контуру на дуже малих моделях у діапазоні 5-10

млн параметрів. Такий режим дає змогу швидко виявляти помилки в реалізації, оцінювати вплив токенізатора, параметрів оптимізації та структури блоку моделі без багатотижневих тренувальних циклів. Доцільність такого підходу підтверджується, зокрема, результатами TinyStories, де показано, що навіть моделі з числом параметрів нижче 10 млн за наявності добре підібраного корпусу можуть генерувати зв’язний текст [3].

На другому етапі, після фіксації базової архітектури, токенізатора та режиму навчання, доцільно переносити повне тренування цільової моделі в хмарне середовище. Такий підхід дозволяє використовувати локальну інфраструктуру для дешевого й швидкого архітектурного пошуку, а дорожчі хмарні ресурси – лише для фінального навчання моделі на вже перевіреному стеку.

В умовах нестабільного електроживлення обов’язковою складовою такого процесу є механізм частого збереження стану моделі, оптимізатора та станів генераторів випадкових чисел, що забезпечує детерміноване відновлення після переривань. Такий поділ мінімізує ризик втрати довгих запусків і робить дослідження більш надійним.

Розробка компактної мовної моделі для локального використання в освіті та кібербезпеці має спиратися не лише на зменшення кількості параметрів, а на системний підхід: контроль якості корпусу, проєктування токенізатора, використання еталонної моделі для верифікації, багатоетапний архітектурний пошук на малих конфігураціях і перенесення повного навчання у хмару після стабілізації всіх компонентів. Саме така методика дозволяє зробити дослідження компактних мовних моделей технічно реалістичним, енергоефективним і придатним до практичного впровадження в умовах обмежених ресурсів.

Список використаних джерел

1. Hoffmann J., Borgeaud S., Mensch A. та ін. Training Compute-Optimal Large Language Models. 2022. arXiv:2203.15556. DOI: 10.48550/arXiv.2203.15556.

2. Besiroglu T., Erdil E., Barnett M., You J. Chinchilla Scaling: A Replication Attempt. 2024. arXiv:2404.10102

3. Eldan R., Li Y. TinyStories: How Small Can Language Models Be and Still Speak Coherent English? 2023. arXiv:2305.07759. DOI: 10.48550/arXiv.2305.07759