ПРИКЛАДНА МАТЕМАТИКА, КОМП'ЮТЕРНІ НАУКИ

УДК 004.421 https://doi.org/10.31713/vt1202511

Ляшко Д. А., аспірант, Турбал Ю. В., д.т.н., професор, Климюк Ю. Є., к.т.н., доцент (Національний університет водного господарства та природокористування, м. Рівне, d.a.lyashko@nuwm.edu.ua ; y.v.turbal@nuwm.edu.ua ; yu.ye.klymiuk@nuwm.edu.ua)

ПАРАМЕТРИЧНО ЕФЕКТИВНІ МЕТОДИ АДАПТАЦІЇ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ

Епоха великих мовних моделей (ВММ) ознаменувала собою значний прогрес у галузі обробки природної мови. Донавчання стало невід'ємною складовою адаптації цих моделей до різноманітних спеціалізованих застосувань. У статті розглядаються сучасні методи донавчання великих мовних моделей (ВММ), що відіграють ключову роль у їх адаптації до спеціалізованих застосувань. Аналізуються параметрично ефективні підходи (PEFT), які дозволяють значно знизити обчислювальні витрати, зберігаючи при цьому продуктивність, порівнянну з повнопараметричним донавчанням. Особливу увагу приділено адаптерним методам, техніці LoRA (Low-Rank Adaptation) та префіксному налаштуванню, які дозволяють скоротити обсяг необхідних обчислювальних ресурсів.

Окремий розділ присвячено стратегіям оптимізації процесу донавчання, зокрема використанню навчання зі змішаною точністю, градієнтної акумуляції та ефективних оптимізаторів (AdamW, Adafactor). Також розглядається баланс між спеціалізацією моделей та їх узагальнювальною здатністю після донавчання, що є критично важливим для забезпечення високої продуктивності на різних завданнях.

У статті окреслено перспективні напрямки досліджень, як-от інтеграція PEFT із методами компресії моделей, адаптація до мультимодальних задач та автоматизація вибору оптимальних стратегій донавчання. Узагальнюючи, робота акцентує увагу на необхідності подальшого розвитку інноваційних методів для підвищення ефективності, надійності та універсальності ВММ.

Ключові слова: **мовна модель; адаптація; донавчання; оптимізація; метод.**

1. Вступ

Великі мовні моделі (LLM), з їхніми багатомільярдними параметричними архітектурами, продемонстрували революційний потенціал у вирішенні задач, пов'язаних з розумінням та генерацією природної мови. Їхня здатність до навчання на величезних масивах неструктурованих текстових даних дозволяє отримувати моделі з широким спектром загальних мовних знань. Однак для ефективного застосування LLM у конкретних, вузькоспеціалізованих доменах, необхідний процес донавчання (fine-tuning). Донавчання виступає як критично важливий етап трансферного навчання, дозволяючи перенести загальні знання, отримані моделлю на етапі попереднього навчання, у площину специфічних вимог цільового завдання. Цей процес є принциповим для практичного розгортання LLM, забезпечуючи їхню адаптацію до різноманітних реальних сценаріїв.

Масштаби сучасних ВММ створюють значні обчислювальні та економічні виклики. Традиційне повнопараметричне донавчання, що передбачає оновлення всіх параметрів моделі, стає дедалі більш ресурсовитратним, часто перевищуючи можливості наявних обчислювальних інфраструктур, особливо для дослідницьких груп з обмеженим доступом до високопродуктивних обчислень. Ця проблема стимулювала інтенсивний розвиток *параметрично ефективних методів донавчання (PEFT)*. PEFT-методи представляють собою інноваційний клас технік, спрямованих на досягнення продуктивності, порівнянної з повнопараметричним донавчанням, але при цьому оновлюючи лише незначну частку параметрів моделі. Таким чином, PEFT-методи пропонують стратегічне рішення для зниження обчислювальних витрат, зменшення вимог до пам'яті та прискорення процесу донавчання, відкриваючи можливості для більш широкого та демократичного використання потужності великих мовних моделей.

2. Систематизація та критичний аналіз методів донавчання

Сучасний науковий ландшафт досліджень у сфері донавчання ВММ характеризується різноманітністю підходів, які можна систематизувати за кількома ключовими категоріями, відображаючи різні стратегії оптимізації та аспекти ефективності.

Огляд та систематизація методів PEFT є предметом активних наукових досліджень, як підтверджує вичерпна аналітична робота [1]. У цій статті представлено систематизований огляд понад 40 ключових наукових публікацій, охоплюючи період з 2019 по 2023 рік, класифікуючи існуючі PEFT-методи за таксономією та надаючи

порівняльний аналіз їхньої емпіричної ефективності. Автори дійшли до висновку, що використання методів PEFT є чи не єдиним способом донавчання LLM в умовах нестачі обчислювальних ресурсів. До найбільш розповсюджених та науково обґрунтованих PEFT-методів належать:

Адаптерні методи (Adapters): Ця техніка [2] передбачає інтеграцію невеликих, додаткових нейронних мереж – адаптерів – у архітектуру існуючої мовної моделі. Адаптери, як правило, вставляються після існуючих шарів уваги або feed-forward шарів у трансформерних моделях. Під час донавчання, параметри оригінальної LLM залишаються замороженими, і лише параметри нововведених адаптерів підлягають оновленню.

Принцип роботи: Адаптери діють як модулі спеціалізації, що навчаються адаптувати вихідні представлення від замороженої ВММ до специфічних вимог цільового завдання. Архітектура адаптерів може варіюватися, включаючи прості feed-forward мережі або більш складні структури. Адаптерні методи демонструють ефективність завдяки принципу модульності. Вони дозволяють зберігати загальні мовні знання, закодовані в основній LLM, одночасно забезпечуючи гнучкість у адаптації до конкретних задач через навчання спеціалізованих адаптерних модулів

Низькорангове адаптування (LoRA – Low-Rank Adaptation): Метод LoRA [3] ґрунтується на гіпотезі про низькорангову природу змін ваг у попередньо навчених ВММ при адаптації до нових завдань. LoRA заморожує оригінальні вагові матриці ВММ, натомість донавчаючи низькорангові матриці розкладання, які представляють апроксимацію змін у вагових матрицях.

Принцип роботи: Для кожної ваговій матриці в оригінальній моделі, LoRA представляє зміни як низькорангове розкладання: W = BA. Під час донавчання, лише матриці A та B підлягають оновленню, тоді як W залишається незмінною. LoRA експлуатує спостереження, що зміни, необхідні для адаптації ВММ, часто лежать у низькоранговому підпросторі. Це дозволяє ефективно захоплювати необхідні адаптації, використовуючи набагато менше донавчуваних параметрів, зберігаючи при цьому більшість параметрів моделі незмінними. LoRA продемонстрував високу емпіричну ефективність у різноманітних завданнях, ставши одним з найбільш популярних PEFT-методів.

Префіксне налаштування (Prefix-Tuning): Підхід Prefix-Tuning [4] вводить концепцію додавання невеликих, донавчуваних векторів –

префіксів – до вхідних даних моделі на кожному шарі трансформера. Донавчанню підлягають лише параметри цих префіксів, тоді як основні параметри ВММ залишаються замороженими.

Принцип роботи: Префікси впливають на активації моделі на кожному шарі, спрямовуючи її поведінку без прямої зміни параметрів моделі. Довжина та структура префіксів можуть бути налаштовані для досягнення оптимальної продуктивності. Prefix-tuning ефективний завдяки здатності префіксів [4] модифікувати внутрішні представлення моделі, керуючи її увагою та генерацією тексту у бажаному напрямі, при цьому зберігаючи цілісність попередньо навчених знань.

У своєму дослідженні [12], автори доходять до висновку, що тонке налаштування з використанням PEFT може покращити продуктивність моделей на мовах з обмеженими ресурсами, хоча іноді це може призвести до зниження продуктивності на мовах з високими ресурсами. Таким чином, методи PEFT можуть допомогти адаптувати мовну модель до малоресурсної мови, така як українська.

3. Оптимізація процесу донавчання: стратегії зменшення обчислювальних витрат

Зниження обчислювальної складності та підвищення ефективності процесу донавчання є ключовим напрямком досліджень, детально проаналізованим у статті [5]. Автори систематично досліджують різні фактори, що впливають на ефективність донавчання, включаючи обсяг обчислювальних ресурсів, час виконання, розмір моделі та довжину контексту. Особлива увага приділяється методам, спрямованим на мінімізацію використання пам'яті та прискорення швидкості донавчання. Ключові стратегії оптимізації включають:

Навчання зі змішаною точністю (Mixed Precision Training): Використання форматів чисел меншої точності, зокрема напівточності (FP16), для представлення ваг та активацій моделі, а також для виконання обчислень [6].

FP16 вимагає вдвічі менше пам'яті, ніж стандартна одинарна точність (FP32), і може використовуватись для прискорення обчислень на сучасних графічних процесорах, що підтримують операції з напівточною. Для забезпечення числової стабільності навчання, особливо при роботі з малими градієнтами, застосовується техніка **log scaling**, що масштабує функцію втрат перед обчисленням градієнтів, та обернене масштабування після. Навчання зі змішаною точністю дозволяє суттєво зменшити обсяг пам'яті, необхідної для

зберігання моделі та проміжних результатів, а також прискорити матричні операції, що є домінуючими в нейромережевих обчисленнях, забезпечуючи при цьому збереження або незначне зниження продуктивності моделі [6].

Градієнтна акумуляція (Gradient Accumulation): Техніка, що дозволяє ефективно донавчати моделі з великими розмірами пакетів даних, навіть при обмеженому обсязі пам'яті графічного процесора. Великі пакети даних часто сприяють стабільнішому навчанню та кращій збіжності.

Градієнтна акумуляція полягає у розбитті великого пакету даних на серію менших підпакетів. Для кожного підпакету обчислюється градієнт функції втрат, але оновлення параметрів моделі відкладається. Градієнти, обчислені для кожного підпакету, акумулюються (сумуються або усереднюються). Після обробки всіх підпакетів, накопичений градієнт використовується для оновлення параметрів моделі. Цей метод емулює навчання з великим пакетом даних, використовуючи пам'ять, необхідну лише для обробки підпакету. Це дозволяє використовувати переваги навчання з великими пакетами (наприклад, краща оцінка градієнта, стабільність навчання) в умовах обмежених ресурсів пам'яті.

Ефективні оптимізатори: Використання оптимізаційних алгоритмів, розроблених спеціально для навчання великих нейронних мереж, таких як AdamW чи Adafactor [7].

AdamW модифікує стандартний оптимізатор Adam, відокремлюючи вартість регуляризації ваг (weight decay) від адаптивного темпу навчання, що сприяє кращій узагальнювальній здатності. Adafactor зменшує витрати пам'яті оптимізатора шляхом факторизації матриць адаптивного темпу навчання.

Ефективні оптимізатори розроблені для вирішення специфічних проблем, що виникають при навчанні великих моделей, таких як нестабільність навчання, повільна збіжність та високі вимоги до пам'яті оптимізатора. Використання цих оптимізаторів може призвести до прискорення донавчання, кращої збіжності та покращення загальної продуктивності.

Варто зауважити, що донавчання LLM, призначених для обробки документів великого обсягу або діалогів з довгим контекстом, природно вимагає обробки довгих послідовностей вхідних даних, що призводить до зростання обсягу пам'яті, необхідної для зберігання активацій та градієнтів. Застосування навчання зі змішаною точністю у поєднанні з градієнтною

акумуляцією дозволяє ефективно донавчити такі моделі навіть на графічних процесорах з відносно обмеженою пам'яттю [8], значно прискорюючи ітераційний цикл розробки та зменшуючи загальні обчислювальні витрати.

4. Узагальнювальна здатність LLM після донавчання: баланс між спеціалізацією та універсальністю

Критичним аспектом донавчання є розуміння його впливу на здатність LLM до узагальнення, тобто на здатність моделі демонструвати високу продуктивність не лише на даних донавчання, але й на нових, раніше небачених даних, що походять з того ж або схожих доменів. Автори статті [9] у своїй роботі емпірично досліджують цей важливий аспект, аналізуючи продуктивність моделей після донавчання на широкому спектрі мовних завдань та доменів. Аналізуючи результати експериментів авторів, можна виділити ключові висновки досліджень у цій області:

- ***Дилема спеціалізації-узагальнення:*** Донавчання, хоча і необхідне для досягнення високої продуктивності в цільовому домені, воно потенційно може призвести до перенавчання та зниження узагальнювальної здатності моделі в інших доменах або на загальних мовних задачах. Це явище, відоме як катастрофічне забування, виникає, коли модель, фокусуючись на новому завданні, втрачає раніше набуті знання.
- ***Вплив даних та стратегій донавчання:*** Якість, обсяг та репрезентативність даних донавчання, а також вибір стратегії донавчання (наприклад, інтенсивність регуляризації, швидкість навчання, вибір PEFT-методу), відіграють вирішальну роль у балансуванні продуктивності в цільовому домені та збереженні узагальнювальної здатності. Недостатній обсяг або низька якість даних донавчання може призвести до перенавчання та обмеженої узагальнювальної здатності.
- ***Необхідність стратегій збереження узагальнення:*** Актуальним напрямком досліджень є розробка нових методологій донавчання, спрямованих на збереження узагальнювальної здатності. Це може включати використання регуляризаційних технік, методів мультизадачного навчання, або стратегій, що експліцитно заохочують модель до збереження загальних мовних знань під час донавчання.

5. Перспективні напрямки та невирішені проблеми

Багато дослідників акцентують увагу на важливості розробки нових архітектур нейронних мереж [10], що є більш придатними для

PEFT-методів, на інтеграції PEFT з методами компресії моделей для досягнення ще більшої ресурсної ефективності, та на розширенні застосування PEFT-підходів до мультимодальних задач, де моделі повинні обробляти та інтегрувати інформацію з різних модальностей (текст, зображення, аудіо, відео).

Ключові наукові виклики та перспективні напрямки досліджень:

- **Інноваційні PEFT-методи:** Подальший пошук та розробка нових, більш ефективних та універсальних PEFT-методологій, що мінімізують кількість донавчуваних параметрів, одночасно забезпечуючи високу продуктивність, надійність та узагальнювальну здатність моделей. Особливу увагу варто приділити модифікації сімейства методів LoRA (DoRA, QLoRA, MoRA) Дослідження можуть бути спрямовані на розробку адаптивних PEFT-методів, що динамічно налаштовують стратегію донавчання залежно від специфіки завдання та доступних ресурсів.
- **Синергія PEFT та компресії моделей:** Вивчення можливостей інтеграції PEFT-методів з техніками компресії моделей, такими як квантизація, прунінг (weight pruning) та дистиляція знань [11], для створення надзвичайно компактних та енергоефективних LLM. Комбінування цих підходів може відкрити шлях до розгортання потужних моделей на периферійних пристроях з обмеженими обчислювальними ресурсами, сприяючи їхньому поширенню у мобільних та вбудованих системах.
- **PEFT для мультимодальних моделей:** Розширення теоретичних засад та емпіричних досліджень PEFT-методів для задач донавчання мультимодальних моделей, здатних обробляти та інтегрувати гетерогенні дані. Це включає розробку PEFT-стратегій, що ефективно адаптують мультимодальні ВММ до нових комбінацій модальностей та специфічних мультимодальних завдань, таких як візуальне запитування, створення описів зображень, або відеорозуміння.
- **Автоматизація та адаптивність у виборі PEFT-методів та гіперпараметрів:** Створення автоматизованих методологій та алгоритмів для інтелектуального вибору оптимальних PEFT-методів, їх гіперпараметрів та стратегій донавчання для конкретних завдань, типів моделей, наборів даних та обмежень на обчислювальні ресурси. Це включає розробку методів машинного навчання, здатних аналізувати характеристики

завдання та моделі, і на основі цього аналізу пропонувати найбільш ефективні конфігурації донавчання.

6. Висновки

Донавчання великих мовних моделей є фундаментальним процесом, що визначає їхню практичну цінність та можливості застосування у різноманітних галузях. Сучасні наукові дослідження активно просувають розробку параметрично ефективних методів донавчання, адже вони не потребують повного перенавчання моделі. Велика увага приділяється ресурсній оптимізації, де знайшли застосування як і класичні mixed precision методи так і оптимізатори створені виключно для PEFT методів. Сучасні дослідження поглиблюють розуміння узагальнювальної здатності моделей після адаптації. Майбутні наукові зусилля повинні бути спрямовані на створення ще більш ефективних, універсальних, надійних та ресурсоекономних методологій донавчання. Ключовим вектором прогресу є розробка інноваційних підходів, що інтегрують різні техніки, враховують специфіку завдань, доменів та обмеження на обчислювальні ресурси.

1. V. Lialin, V. Deshpande. Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning, 2023. **2.** Houlsby, N. et al, Parameter-efficient transfer learning for NLP. *arXiv preprint arXiv:1902.00751*. 2019. **3.** Hu E. J., Shen Y., Wallis P., Allen-Zhu Z. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*. 2021. **4.** Li, X., P. P. Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00132*. 2021. **5.** A. Singh et al, A Study of Optimizations for Fine-tuning Large Language Models. *arXiv preprint arXiv:2406.02290.* 2024. **6.** Micikevicius P., Narang S., Alben J., Diamos G., Elsen E., Garcia J. Mixed precision training. *arXiv preprint arXiv:1710.03740*. 2017. **7.** Loshchilov I., Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1812.03233*. 2019. **8.** J. Kim, Y. Lee. A Gradient Accumulation Method for Dense Retriever under Memory Constraint. *arXiv preprint arXiv:2406.12356.* 2024. **9.** H. Yang et al. Unveiling the Generalization Power of Fine-Tuned Large Language Models. *arXiv preprint arXiv:2403.09162*. 2024. **10.** Wang et al. Towards Better Parameter-Efficient Fine-Tuning for Large Language Models: A Position Paper. *arXiv preprint arXiv:2311.13126*. 2023. **11.** G. Hinton et al. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*. 2015. **12.** D. Aggarval, A. Sathe, I. Watts, S. Sitaram. MAPLE: Multilingual Evaluation of Parameter Efficient Finetuning of Large Language Models. *arXiv preprint arXiv:2401.07598.* 2024.

REFERENCES:

1. V. Lialin, V. Deshpande. Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning, 2023. **2.** Houlsby, N. et al, Parameter-efficient transfer learning for NLP. *arXiv preprint arXiv:1902.00751*. 2019. **3.** Hu E. J., Shen Y., Wallis P., Allen-Zhu Z. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*. 2021. **4.** Li, X., P. P. Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00132*. 2021. **5.** A. Singh et al, A Study of Optimizations for Fine-tuning Large Language Models. *arXiv preprint arXiv:2406.02290.* 2024. **6.** Micikevicius P., Narang S., Alben J., Diamos G., Elsen E., Garcia J. Mixed precision training. *arXiv preprint arXiv:1710.03740*. 2017. **7.** Loshchilov I., Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1812.03233*. 2019. **8.** J. Kim, Y. Lee. A Gradient Accumulation Method for Dense Retriever under Memory Constraint. *arXiv preprint arXiv:2406.12356.* 2024. **9.** H. Yang et al. Unveiling the Generalization Power of Fine-Tuned Large Language Models. *arXiv preprint arXiv:2403.09162*. 2024. **10.** Wang et al. Towards Better Parameter-Efficient Fine-Tuning for Large Language Models: A Position Paper. *arXiv preprint arXiv:2311.13126*. 2023. **11.** G. Hinton et al. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*. 2015. **12.** D. Aggarval, A. Sathe, I. Watts, S. Sitaram. MAPLE: Multilingual Evaluation of Parameter Efficient Finetuning of Large Language Models. *arXiv preprint arXiv:2401.07598.* 2024.

Liashko D. A., Post-graduate Student, Turbal Yu. V., Doctor of Engineering, Professor, Klumiuk Yu. Ye., Candidate of Engineering (Ph.D.), Associate Professor (National University of Water and Environmental Engineering, Rivne, d.a.lyashko@nuwm.edu.ua ; y.v.turbal@nuwm.edu.ua ; yu.ye.klymiuk@nuwm.edu.ua)

PARAMETRICALLY EFFICIENT METHODS FOR ADAPTATION OF LARGE LANGUAGE MODELS

The era of large language models (LLMs) has marked significant progress in the field of natural language processing. Fine-tuning has become an essential component in adapting these models to various specialized applications. This paper explores modern fine-tuning methods for LLMs, which play a crucial role in their adaptation to specific domains. The study analyzes parameter-efficient fine-tuning (PEFT) approaches that significantly reduce computational costs while maintaining performance comparable to full-parameter fine-tuning. Particular attention is given to adapter-based methods, Low-Rank

Adaptation (LoRA), and prefix-tuning, which help minimize the required computational resources.

A dedicated section focuses on strategies for optimizing the fine-tuning process, including mixed precision training, gradient accumulation, and efficient optimizers (AdamW, Adafactor). The paper also examines the balance between model specialization and generalization after fine-tuning, which is critical for ensuring high performance across different tasks.

The study outlines promising research directions such as integrating PEFT with model compression methods, adapting LLMs for multimodal tasks, and automating the selection of optimal fine-tuning strategies. In summary, the paper highlights the need for further development of innovative approaches to enhance the efficiency, reliability, and universality of LLMs.

***Keywords*: language model; adaptation; retraining; optimization; method.**