УДК 004.89:004.032.26

ПОРІВНЯЛЬНИЙ АНАЛІЗ СУЧАСНИХ МЕТОДІВ ЕФЕКТИВНОГО ЗА ПАРАМЕТРАМИ ДОНАВЧАННЯ LLM

Мічурін І.Є.
e-mail: ihor.michurin@nure.ua
Науковий керівник – к.т.н., доц. Рябова Н.В.
Харківський національний університет радіоелектроніки, каф. ШІ
м. Харків, Україна

This work presents a comparative analysis of modern Parameter-Efficient Fine-Tuning (PEFT) methods for Large Language Models (LLMs). Existing approaches are systematized into four categories: adapter-based methods, low-rank adaptation (LoRA) and its variants, prompt-based tuning, and hybrid approaches. For each category, the mathematical foundations, computational complexity, and practical trade-offs between parameter count, training efficiency and downstream task performance are analyzed. A comparative analysis on the GLUE benchmark is presented. Key open challenges including rank collapse, static compute allocation, and limited multilingual evaluation are identified. Recommendations for selecting PEFT methods depending on computational resources and task characteristics are formulated.

Дана робота є частиною досліджень, які здійснюються в рамках міжнародного проєкту Dual Degree Programme MSc-by-Research in Computer Science, та присвячена порівняльному аналізу сучасних методів ефективного за параметрами донавчання великих мовних моделей (LLM).

Актуальність дослідження обумовлена, перш за все, появою архітектури Transformer та побудованих на ній LLM, таких як BERT, GPT та LLaMA, що докорінно змінили підходи до NLP. Ці моделі містять до сотень мільярдів параметрів і демонструють високу якість на широкому спектрі задач. Проте повне донавчання (Full Fine-Tuning) стає обчислювально неприпустимим через великі обсяги GPU-пам'яті та навантаження на сховище для зберігання окремих копій моделі [1]. Зокрема, донавчання GPT-3 (175B параметрів) потребує сотні гігабайт відеопам'яті, що є недоступним для більшості дослідницьких груп.

У відповідь на ці виклики сформувався напрям Parameter-Efficient Fine-Tuning (PEFT), що передбачає оновлення менше 1% параметрів при замороженому backbone. Стрімкий розвиток цього напряму призвів до появи десятків різних методів, що створює потребу в систематичному порівняльному аналізі їх переваг, обмежень та оптимальних сценаріїв застосування. Саме цій задачі присвячена дана робота.

Адаптерні методи. Houlsby et al. [2] запропонували bottleneck-модулі між шарами Transformer: $f(h) = W_{up} \cdot \varphi(W_{down} \cdot h)$, додаючи лише $2dr$ параметрів. Pfeiffer оптимізували до одного адаптера на шар, Compacter – гіперкомплексне множення для стиснення ваг.

Низькорангова адаптація (LoRA). Hu et al. [3] моделюють оновлення ваг як $\Delta W = BA$ без затримки при інференсі. Розвитком стали: AdaLoRA – динамічний розподіл рангу; QLoRA – 4-бітна квантизація для моделей 65B на одній GPU [4]; DoRA та VeRA – подальше стиснення параметрів.

Методи на основі промптів. Prefix Tuning додає віртуальні токени до входу шарів Transformer. Prompt Tuning працює лише з вхідним шаром. IA^3 масштабує активації навчальними векторами у few-shot режимі.

Гібридні підходи. He et al. [5] показали, що адаптери, LoRA та Prefix Tuning є частковими випадками єдиного framework, що відкрило шлях до комбінованих методів: гейтинг з регуляризацією спектру, MoE-адаптери та інтеграція PEFT зі стисненням моделей.

Порівняльний аналіз на GLUE показує, що LoRA є найбільш збалансованим методом при оновленні 1–2% параметрів. Адаптери мають переваги на задачах з обмеженими даними. Prefix Tuning ефективний при бюджеті менше 0,1%. QLoRA дозволяє донавчати моделі 70B на споживчих GPU.

В роботі також досліджуються такі проблеми як колапс рангу у низькорангових методах, статичний розподіл обчислень між токенами та обмежена валідація на мультимовних задачах. Перспективні напрямки продовження досліджень включають адаптивні методи з динамічним керуванням інформацією та інтеграцію PEFT з квантизацією для edge-пристроїв.

Список використаних джерел:

1. Ding N., Qin Y., Yang G. et al. Parameter-Efficient Fine-Tuning of Large-Scale Pre-Trained Language Models. Nature Machine Intelligence. 2023. Vol. 5. P. 220–235. DOI: 10.1038/s42256-023-00626-4.

2. Houlsby N., Giurgiu A., Jastrzebski S. et al. Parameter-Efficient Transfer Learning for NLP // Proceedings of the 36th International Conference on Machine Learning (ICML). 2019. P. 2790–2799.

3. Hu E. J., Shen Y., Wallis P. et al. LoRA: Low-Rank Adaptation of Large Language Models // Proceedings of the 10th International Conference on Learning Representations (ICLR). 2022.

4. Dettmers T., Pagnoni A., Holtzman A., Zettlemoyer L. QLoRA: Efficient Finetuning of Quantized LLMs // Advances in Neural Information Processing Systems. 2023. Vol. 36.

5. He J., Zhou C., Ma X., Berg-Kirkpatrick T., Neubig G. Towards a Unified View of Parameter-Efficient Transfer Learning // Proceedings of the 10th International Conference on Learning Representations (ICLR). 2022.