Бондар В.В., аспірант,
Бабенко В.Г., д.т.н., професор,
Черкаський державний технологічний університет

СУЧАСНІ ПІДХОДИ ПОКРАЩЕННЯ ЕФЕКТИВНОСТІ МЕХАНІЗМУ ЕМБЕДІНГІВ ТРАНСФОРМЕРІВ

Моделі трансформерів стали основою сучасних систем штучного інтелекту. Важливою частиною цієї архітектури є механізми ембедінгів, які перетворюють вхідні дані у високорозмірні вектори, придатні для self-attention операцій. Але ці шари ембедінгів часто займають значну частину параметрів моделі (35-40%), що зумовлює зниження ефективності в середовищах з обмеженими ресурсами. Tensor-train метод вирішує ці проблеми шляхом розкладання матриць ембедінгів, зменшуючи час обчислень, наприклад розкладаючи тензори розміром 512×768 на тензори розміром 4×16×12×6, зберігаючи при цьому 98,1% початкової перплексії в моделях BERT [1]. Ця факторизація використовує надлишковість параметрів, що особливо ефективно в задачах обробки природної мови [1, 2]. Підходи квантизації пропонують інший шлях до ефективності ембедінгів. SQ-Transformer запроваджує структурно квантовані ембедінги, які групують слова за синтаксичними ролями (наприклад, дієслова/іменники) використовуючи векторну квантизацію. У семантичному аналізі COGS, цей метод досягає 86,7% точності з 4-бітними ембедінгами проти 82,1% для стандартної 8-бітної квантизації [3]. Ключовою інновацією є Systematic Attention Layer, який застосовує однакові шаблони уваги до синтаксично еквівалентних структур, що архітектурно зменшує розмір ембедінгів у 4 рази, одночасно покращуючи композиційну генералізацію [2, 3]. Динамічне призначення бітової ширини базується на підходах квантизації, застосовуючи коригування точності на основі важливості, що включає обчислення значимості токенів через норми градієнтів, виділення 8 бітів для високозначимих токенів, таких як дієслова, і лише 4 бітів для менш значимих токенів, таких як артиклі, обраховуючи компенсації помилок через залишкові з'єднання від високоточних ембедінгів. Попередній аналіз на наборі даних GLUE показує, що цей метод зберігає 97,3% точності BERT-подібних моделей, зменшуючи при цьому середню бітову ширину до 5,2 бітів — на 35% краще стиснення порівняно зі статичною 8-бітною квантизацією [2, 3]. Міждоменне перенесення ембедінгів в роботі [5] з вирівнювання геномів та білків включає навчання універсального енкодера на мультимодальних даних (текст, послідовності білків, мовлення), отримання доменно-специфічних ембедінгів через низькорангові адаптери (LoRA), та спільне використання основних параметрів ембедінгів для всіх модальностей. Цей метод дозволяє єдиному 768-вимірному простору ембедінгів обслуговувати декілька задач, використовуючи на 40% меншу кількістю параметрів [5, 6]. Розділені ембедінги пропонують покращення в систематичній генералізації шляхом розділення ембедінгів на інваріантні компоненти (синтаксичні ролі) та варіативні компоненти (семантичний зміст), що підвищує систематичність у тестах SCAN на 22% [3]. Сучасні методи покращення механізму ембедінгів балансують ефективність та виразність через квантизацію, декомпозицію та доменну адаптацію. Майбутні напрямки включають динамічні мережі ембедінгів, двостадійний обрахунок ембедінгів, квантизацію простору ембедінгів та інше розширять можливості застосування трансформерів у середовищах з обмеженими ресурсами, зберігаючи архітектурні переваги [1, 4, 3].

Список використаних джерел:

1. Y. Wang et al., "Characterization of MPC-based Private Inference for Transformer-based Models," 2022 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), Singapore, 2022, pp. 187-197, doi: 10.1109/ISPASS55109.2022.00025.

2. Liu, Shih-yang, et al. "Llm-fp4: 4-bit floating-point quantized transformers." arXiv preprint arXiv:2310.16836 (2023). https://doi.org/10.18653/v1/2023.emnlp-main.39

3. Jiang, Yichen, Xiang Zhou, and Mohit Bansal. "Inducing systematicity in transformers by attending to structurally quantized embeddings." arXiv preprint arXiv:2402.06492 (2024). https://doi.org/10.48550/arXiv.2402.06492

4. Liang, Z., Wang, P., Zhang, R., Xu, N., Zhang, S., Xing, L., Bai, H., & Zhou, Z. (2024). MERGE: Fast Private Text Generation. Proceedings of the AAAI Conference on Artificial Intelligence, 38(18), 19884-19892. https://doi.org/10.1609/aaai.v38i18.29964

5. Nogin, Yevgeni, et al. "OM2Seq: Learning retrieval embeddings for optical genome mapping." Bioinformatics Advances, Volume 4, Issue 1, 2024, vbae079, https://doi.org/10.1093/bioadv/vbae079

6. Wu, Yueh-Kao, Ching-Yu Chiu, and Yi-Hsuan Yang. "JukeDrummer: Conditional beat-aware audio-domain drum accompaniment generation via Transformer VQ-VAE." arXiv preprint arXiv:2210.06007 (2022). https://doi.org/10.48550/arXiv.2210.06007