МЕЖДУНАРОДНЫЙ ЖУРНАЛ
ТЕОРЕТИЧЕСКИХ И ПРИКЛАДНЫХ
ВОПРОСОВ ЦИФРОВЫХ ТЕХНОЛОГИЙ

P-ISSN: 2181-3086 E-ISSN: 2181-3094

Самаркандский филиал Ташкентского университета информационных технологий имени Мухаммада ал-Хоразмий

WEB: https://ijdt.uz/index.php/ijdt

ЭФФЕКТИВНОСТЬ LORA В ЗАДАЧАХ ОБОБЩЕНИЯ ТЕКСТОВ: СРАВНЕНИЕ МОДЕЛЕЙ T5 И UZT5

Фатима Адилова [1], Рифкат Давронов [1], Самариддин Кушмуратов [1]
[1] Институт Математики им В.И. Романовского АН Республики Узбекистан, Ташкент, Узбекистан
fatadilova@matinst.uz, rifqat.davronov@mathinst.uz, bekmezonali@gmail.com

Цитирование: *Адилова Ф.Т., Давронов Р.Р., Кушмуратов С.И. (2024).* Эффективность Эффективность LoRA в за-дачах обобщения текстов: сравнение моделей T5 и uzT5. Международный Журнал Теоретических и Прикладных Вопросов Цифровых Технологий, 7(3), –С. 112-116. https://doi.org/10.62132/ijdt.v7i3.204

Дата поступления: 31.06.2024
Дата принятия: 12.07.2024
Дата печати: 30.09. 2024

DOI: https://doi.org/10.62132/ijdt.v7i3.204

УДК 577.29

ЭФФЕКТИВНОСТЬ LORA В ЗАДАЧАХ ОБОБЩЕНИЯ ТЕКСТОВ: СРАВНЕНИЕ МОДЕЛЕЙ T5 И UZT5

Адилова Ф.Т.[1], Давронов Р.Р.[1], Кушмуратов С.И.[1]

[1] Институт Математики имени В.И. Романовского АН Республики Узбекистан, Ташкент, Узбекистан

fatadilova@matinst.uz, rifqat.davronov@mathinst.uz, bekmezonali@gmail.com

Аннотация. *В данной работе представлен анализ применения метода низкоранговой адаптации (Low-Rank Adaptation, LoRA) для задачи одноязычной генерации текстов на узбекском языке. Мы использовали модели T5-base, T5-Large и uzT5, чтобы определить, какая из них показывает наилучшие результаты при использовании LoRA, а также сравнили их производительность с традиционной тонкой настройкой. В качестве набора данных использовали текст из 5000 новостей с платформы Kun.uz, из которых 4000 были использованы для обучения, а 1000 — для тестирования. Производительность моделей оценивалась с помощью метрик BLEU, ROUGE-1, ROUGE-2, ROUGE-L и ROUGE-LSUM. Результаты показали, что модель uzT5-base с параметрами LoRA равными r=256 и α=512, демонстрирует наивысшие показатели среди всех рассмотренных моделей, обеспечивая наилучшие значения метрик ROUGE и BLUE при умеренном количестве параметров для обучения, что делает её более вычислительно эффективной по сравнению с mT5-large.*

Ключевые слова: *низкоранговые адаптации, T5-base, T5-Large, uzT5, сжатые модели.*

I. ВВЕДЕНИЕ

Появление предварительно обученных больших языковых моделей (LLM), таких как PaLM2 [1], LLaMA2 [2], T5 [3] и семейства GPT от OpenAI, значительно продвинуло состояние обработки естественного языка (NLP). Однако увеличение размера LLM создает значительные проблемы для традиционной тонкой настройки, особенно, если необходимо обрабатывать множество задач или задачи с большим объемом памяти, например, при обработке длинных входных последовательностей.

Методы параметрической эффективной тонкой настройки (ПЭТН) недавно показали свою перспективность для адаптации предварительно обученной модели к различным задачам путем избирательной тонкой настройки небольшого числа дополнительных параметров. Широко применяемые методы ПЭТН включают адаптеры [4], низкоранговую адаптацию [5], настройку префиксов [6] и подсказок [7]. Среди них LoRA стала одной из самых популярных подходов, достигая высоких результатов без увеличения задержки при выводе. Большинство исследований ПЭТН сосредоточены на понимании естественного языка, например, на задачах классификации, как это показано в бенчмарках GLUE [8] и SuperGLUE [9], и монолингвальной генерации, например, генерации текста по таблицам или суммаризации [6].

В данной статье мы представляем результаты исследования возможностей метода LoRA для решения задач одноязычного обобщения на основе узбеко-язычных данных, созданных нашей командой. Мы исследуем применение LoRA к моделям T5-base, T5-Large и узT5, стремясь определить, в какой из них LoRA демонстрирует наилучшие результаты, а также выделить преимущества LoRA по сравнению с традиционной тонкой настройкой (fine-tuning).

В рамках нашего исследования мы оцениваем производительность моделей с использованием известных метрик качества, таких как BLEU, ROUGE-1, ROUGE-2, ROUGE-L и ROUGE-LSUM. Основной целью является определить, какие подходы и модели наиболее эффективно справляются с задачами обобщения текста на узбекском языке, и оценить потенциальные улучшения, которые могут быть достигнуты с помощью LoRA.

II. ОСНОВНАЯ ЧАСТЬ

Эффект применения LoRA состоит в уменьшении количества обучаемых параметров путем обучения пар матриц рангового разложения при замораживании исходных весов модели. Это значительно снижает требования к хранению больших языковых моделей, адаптированных к конкретным задачам, и позволяет эффективно переключаться между задачами

без увеличения задержки при выводе. Недавние исследования показывают, как адаптивно регулировать ранг (r) матриц [10], предлагают обобщения LoRA и связанных подходов ПЭТН в рамках единой структуры [11], и комбинируют LoRA с квантизацией [12]. Большинство этих исследований сосредоточены на задачах классификации и монолингвальной генерации, однако не используют. LoRA или не включают комплексные эксперименты с моделью T5.

Кросс-языковой перенос требует от модели изучения задачи на основе размеченных данных на одном языке (обычно английском), а затем выполнения аналогичной задачи на другом языке, для которого нет или почти нет размеченных данных [14]. Предыдущие исследования, сосредоточенные на методах ПЭТН для кросс-языкового переноса, исследовали подходы на основе адаптеров [13] и комбинируемую разреженную тонкую настройку. В [18] оценивают настройку подсказок в условиях нулевого обучения для кросс-язычной суммаризации, сосредотачиваясь на наборе данных Wikilingua [19], но исследование не включает LoRA и сценарии с большим или с малым количеством доступных данных.

LoRA: пусть $W_0 \in R^{d\times k}$ обозначает весовую матрицу предварительно обученной LLM (где d - размерность входных данных, а k - размерность выходных данных). Ключевая идея LoRA заключается в представлении настроенной WWW с помощью низкорангового разложения $W_0 + \Delta W = W_0 + BA$, где $B \in R^{d\times r}$ и $A \in R^{r\times k}$, и $r \ll min(d,k)$, что делает BA низкоранговой матрицей по сравнению с W_0. Во время обучения W_0 остается замороженной, в то время как B и A содержат обучаемые параметры, которые эффективно являются частью $(2r/d)$ параметров по сравнению с полной тонкой настройкой. Хотя LoRA может быть применена к любому подмножеству весовых матриц, в [5] обновляют только весовые матрицы в модуле самовнимания архитектуры Transformer. В отличие от этого,мы в экспериментах обновляем все четыре матрицы внимания (т.е. query, key, value и out).

LoraHub: это недавно предложенный подход к обучению с небольшим количеством данных без использования градиентов [19], который фокусируется на комбинировании индивидуально обученных модулей LoRA для обобщения на новые задачи. Учитывая, что w состоит из относительно небольшого числа параметров, авторы выбрали без градиентные методы оптимизации вместо градиентного спуска. Доступные модули LoRA m_i синтезируются в модуль $\hat{m} = \sum_{i=1}^{N} w_i m_i$, где w_i- скалярный вес, который может принимающий положительные и отрицательные значения. Процесс оптимизации управляется перекрестными потерями энтропии, целью которого является поиск наилучшего набора {w1, w2, . . . , wN }, который уменьшает потери L в нескольких коротких примерах Q. Кроме того, мы включили регуляризацию L1 для ограничения суммы абсолютных значений w, что помогает предотвратить получение экстремальных значений. Следовательно, конечной целью Lora Hub является минимизация $L + \alpha \sum_{i=1}^{N} |w_i|$, где α является гиперпараметром.

Для наших экспериментов мы использовали набор данных, состоящий из 5000 новостей, взятых с новостной платформы Kun.uz. Эти данные были предназначены для задачи суммаризации: 4000 из них использовались для обучения и 1000 для тестирования.

Эксперименты были сосредоточены на модели T5, представляющей собой LLM с энкодером-декодером использования метода LoRA на рисунке 1. В частности, мы использовали три размера модели T5 (mT5 [3], mT5-large и uzT5 [21-24]).

Все эксперименты проводились на компьютере DGX Station с различной скоростью обучения в диапазоне от 1e-3 до 2e-5. Модель uzT5 была обучена на основе 19 ГБ данных на узбекском языке, мы оцениваем эффективность моделей LoRA, основанных на mT5, mT5-large и uzT5-base [24], в решении задачи суммаризации. Для оценки качества созданных текстов мы применяем различные метрики:

BLEU (Bilingual Evaluation Understudy): Метрика, которая измеряет точность совпадения n-грамм с эталонными текстами и широко используется для оценки задач машинного перевода и суммаризации текста [26].

ROUGE-1: Метрика, оценивающая совпадение униграмм (отдельных слов) между сгенерированным текстом и эталонным резюме [25].

ROUGE-2: Метрика, оценивающая совпадение биграмм (пар последовательных слов) между сгенерированным текстом и эталонным резюме [25].

ROUGE-L: Метрика, основанная на наибольшей общей подпоследовательности (LCS), которая оценивает, насколько хорошо сгенерированное резюме сохраняет длинные последовательности слов из эталонного текста [25].

ROUGE-LSUM: Вариант метрики ROUGE-L, адаптированный специально для задач суммаризации, который учитывает структуру документа и длину сгенерированных резюме [27].

Эти метрики представляют всесторонний анализ качества сгенерированных моделей LoRA, позволяя объективно оценивать их производительность и сравнивать с другими моделями.

III. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

В соответствии с сформулированной выше целью исследования был проведен ряд вычислительных экспериментов, результаты которых показаны ниже. Мы оцениваем производительность LoRA на базе моделей mT5, mT5-large и uzT5-base. В качестве метрик оценки NLP моделей использовали показатели BLUE, ROUGE-1, ROUGE-2, ROUGE-L и ROUGELSUM.

Таблица 1. Результаты модели на mT5-base

Модель и метод	rouge1	rouge2	rougeL	rougeLsum	BLUE	Обучаемые параметры
mt5-base	0.1541	0.0578	0.1139	0.1139	0.0206	582.4M
mt5-base(r=128,alfa=256)	0.1556	0.0593	0.1217	0.1217	0.0223	14.1M
mt5-base(r=256,alfa=256)	0.1566	0.0579	0.1189	0.1189	0.0213	28.3M
mt5-base(r=256,alfa=512)	0.1591	0.0620	0.1243	0.1243	0.0245	28.3M
mt5-base(r=512,alfa=512)	**0.1592**	**0.0632**	**0.1244**	**0.1244**	**0.0250**	56.6M

Из таблицы 1 следует, что с использованием LoRA модель mt5-base при значениях (r=512, alfa=512) достигла наивысшего значения rouge1 (0.1592) среди базовых моделей, показывая сбалансированную производительность по всем метрикам (rouge2, rougeL, rougeLsum, BLUE) и требуя относительно небольшого количества параметров для обучения (56.6M).

Таблица 2. Результаты модели на mT5-large

Модель и метод	rouge1	rouge2	rougeL	rougeLsum	BLUE	Обучаемые параметры
mt5-large	**0.2081**	**0.0934**	**0.1585**	**0.1585**	**0.0464**	1229.5M
mt5-large(r=128,alfa=256)	0.1859	0.0757	0.1371	0.1371	0.0290	37.7M
mt5-large(r=256,alfa=256)	0.1883	0.0785	0.1425	0.1425	0.0339	75.4M
mt5-large(r=256,alfa=512)	0.1035	0.0256	0.0777	0.0777	0.0094	75.4M
mt5-large(r=512,alfa=512)	0.0849	0.0137	0.0638	0.0638	0.0003	150.9M

Из таблицы 2 следует, что результаты с полной точной настройкой были лучше, чем полученные с LoRA. Модель mt5-large достигла наивысшего значения rouge1 (0.2081) среди крупных моделей, демонстрируя превосходную производительность по всем метрикам, но требуя большого количества параметров для обучения (1229.5M), что указывает на высокую вычислительную стоимость.

Таблица 3. Результаты модели на uzT5-base

Модель и метод	rouge1	rouge2	rougeL	rougeLsum	BLUE	Обучаемые параметры
uzt5-base	0.2133	0.0807	0.1338	0.1338	0.0287	247.5M
uzt5-base(r=128,alfa=256)	0.2069	0.0739	0.1303	0.1303	0.0242	14.1M
uzt5-base(r=256,alfa=256)	0.2079	0.0760	0.1316	0.1316	0.0259	28.3M
uzt5-base(r=256,alfa=512)	**0.2152**	**0.0820**	**0.1377**	**0.1377**	0.0314	28.3M
uzt5-base(r=512,alfa=512)	0.2116	0.0801	0.1358	0.1358	**0.0315**	56.6M

Очевидно, что наивысший результат метрики ROUGE был достигнут при значениях r и alfa равных 256 и 512 соответственно. Наилучшие результаты по метрике BLUE были достигнуты при значениях r и alfa равных 512. Таким образом, модель uzT5-base с подключением LoRA достигает наивысшего значения rouge1 (0.2152) среди моделей uzT5, показывая производительность с лучшим показателем BLUE (0.0314) и работая на умеренном количестве параметров для обучения (28.3M), что делает её более эффективной по сравнению с моделью mt5-large.

Таблица 4. Сводная таблица

Модель и метод	rouge1	rouge2	rougeL	rougeLsum	BLUE	Обучаемые параметры
mt5-base(r=512, alfa=512)	0.1592	0.0632	0.1244	0.1244	0.0250	56.6M
mt5-large	0.2081	0.0934	0.1585	0.1585	0.0464	1229.5M
uzt5-base(r=256, alfa=512)	0.2152	0.0820	0.1377	0.1377	0.0314	28.3M

Таким образом, из сводной таблицы 4 следует, что модель uzT5-base с параметрами r=256 и alfa=512 обеспечивает лучшую общую производительность с наивысшими показателями rouge1 и BLUE, при этом являясь более вычислительно эффективной по сравнению с моделью mt5-large.

IV. ЗАКЛЮЧЕНИЕ

В результате проведенного исследования были выявлены преимущества использования метода LoRA для задач генерации текстов на узбекском языке. Модель uzT5-base, обученная с использованием LoRA, продемонстрировала наилучшие результаты по большинству метрик, обеспечивая высокую производительность и эффективность. Эти результаты подтверждают потенциал LoRA в качестве перспективного подхода для адаптации больших языковых моделей к специфическим задачам с минимальными вычислительными затратами. Дальнейшие исследования могут быть направлены на изучение применения LoRA в других областях обработки естественного языка (Казахский, кыргызский, туркменский и таджикский), а также на оптимизацию параметров адаптации для достижения ещё более высоких результатов.

ЛИТЕРАТУРА

[1] *Anil, R., et al.* (2023). PaLM 2: Pre-trained Large Model for Language Understanding. ArXiv. DOI: 10.48550/arXiv.2305.10403

[2] *Touvron, H., et al.* (2023). LLaMA 2: Open and Efficient Foundation Language Models. ArXiv. DOI: 10.48550/arXiv.2302.13971

[3] *Raffel, C., et al.* (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arxiv.org/abs/1910.10683.

[4] *Houlsby, N., et al.* (2019). Parameter-efficient Transfer Learning for NLP. https://arxiv.org/abs/1902.00751.

[5] *Hu, E. J., et al.* (2022). LoRA: Low-Rank Adaptation of Large Language Models. https://arxiv.org/abs/2106.09685.

[6] *Li, X. L., Liang, P.* (2021). Prefix-Tuning: Optimizing Continuous Prompts for Generation. https://arxiv.org/abs/2101.00190.

[7] *Lester, B., et al.* (2021). The Power of Scale for Parameter-Efficient Prompt Tuning. https://arxiv.org/abs/2104.08691.

[8] *Wang, A., et al.* (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. https:arxiv.org/abs/1804.07461.

[9] *Wang, A., et al.* (2019). SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. https://arxiv.org/abs/1905.00537.

[10] *Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao.* 2023b. Adaptive Budget Allocation for Parameter-Efficient Fine-Tuning. In The Eleventh International Conference on Learning Representations.

[11] *Arnav Chavan, Zhuang Liu, Deepak Gupta, Eric Xing, and Zhiqiang Shen.* 2023. One-for-All: Generalized LoRA for Parameter-Efficient Fine-tuning. arXiv preprint arXiv:2306.07967.

[12] *Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer.* 2023. QLoRA: Efficient finetuning of quantized LLMs. In Thirty-seventh Conference on Neural Information Processing Systems.

[13] *Ansell, B., et al.* (2021). Composable Sparse Fine-Tuning for Cross-Lingual Transfer. https://arxiv.org/abs/2110.07560.

[14] *Artetxe, M., et al.* (2020). Translation Artifacts in Cross-lingual Transfer Learning. ArXiv.

[15] *Karthikeyan, K., et al.* (2020). Cross-lingual Transfer Learning for Multilingual Task-Oriented Dialog. arxiv.org/abs/2004.04721.

[16] *Lauscher, A., et al.* (2020). From Zero to Hero: On the Limitations of Zero-Shot Cross-Lingual Transfer. arxiv.org/abs/2005.00633.

[17] *Whitehouse, P., et al.* (2022). Cross-lingual Transfer Learning for Text Classification with Multilingual BERT. https://arxiv.org/abs/2104.08645.

[18] *Vu, X. T., et al.* (2022). Zero-Shot Cross-Lingual Transfer with AdapterFusion. https://arxiv.org/abs/2402.14778.

[19] *Ladhak, F., et al.* (2020). Wikilingua: A Multilingual Abstractive Summarization Dataset. https://arxiv.org/abs/2010.03093.

[20] *Huang, Y., et al.* (2023). LoRAHub: Combining Individually Trained LoRA Modules for Generalization. https://arxiv.org/html/2307.13269v2.

[21] *Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin.* 2023. LoraHub: Efficient Cross-Task Generalization via Dynamic LoRA Composition. https://arxiv.org/abs/2307.13269.

[22] *Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu.* 2022. Black-Box Tuning for Language-Model-as-a-Service. In Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, pages 20841–20855. PMLR.

[23] https://huggingface.co/rifkat/t5-base-uzbek

[24] *Davronov, R., Adilova, F.* UzRoberta: A Pre-Trained Language Model for Uzbek / AIP Conference Proceedings., 2024, 3004(1), 050001

[25] *Lin, Chin-Yew.* 2004. ROUGE: a Package for Automatic Evaluation of Summaries. In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain, July 25 - 26, 2004.

[26] *Kishore Papineni and Salim Roukos and Todd Ward and Wei-jing Zhu* BLEU: a Method for Automatic Evaluation of Machine Translation / Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318.

Поступила в редакцию 31.06.2024

Цитирование: *Адилова Ф.Т., Давронов Р.Р., Кушмуратов С.И.* (2024). Эффективность Эффективность LoRA в задачах обобщения текстов: сравнение моделей T5 и uzT5. *Международный Журнал Теоретических и Прикладных Вопросов Цифровых Технологий*, 7(3), –С. 112-116. https://doi.org/10.62132/ijdt.v7i3.204

LORA EFFECTIVENESS IN TEXT GENERALIZATION TASKS: COMPARISON OF T5 AND UZT5 MODELS

Adilova F.T.[1], *Davronov R.R.*[1], *Kushmuratov S.I.*[1]

[1] V.I. Romanovsky Institute of Mathematics of the Academy of Sciences of the Republic of Uzbekistan Republic of Uzbekistan, Tashkent, Uzbekistan

Abstract. *This paper presents an analysis of the application of the Low-Rank Adaptation method (LoRA) for the task of monolingual text generation in Uzbek. We used the T5-base, T5-Large and uzT5 models to determine which one shows the best results when using LoRA, and also compared their performance with traditional fine tuning. The text from 5,000 news items from the Kun.uz platform was used as a dataset of which 4,000 were used for training and 1,000 for testing. The performance of the models was evaluated using the metrics BLEU, ROUGE-1, ROUGE-2, ROUGE-L and ROUGE-LSUM. Our results showed that the uzT5-base model with LoRA r=256 and α=512 parameters demonstrate the highest performance among all the considered models, providing the best values of the ROUGE and BLEU metrics with a moderate number of training parameters, which makes it more computationally efficient compared to mT5-Large.*

Keywords: *Low-rank adaptation, T5-base, T5-Large, uzT5, model compression***.**