УДК 004.89

ОПТИМИЗАЦИЯ НЕЙРОННОЙ СЕТИ ТРАНСФОРМЕР С СОХРАНЕНИЕМ КОНТЕКСТА ДЛЯ ЗАДАЧ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

Большим М.А. (ИТМО)
Научный руководитель – кандидат технических наук Кугаевских А.В.
(ИТМО)

Введение. В настоящее время наблюдается растущий интерес к использованию нейросетей, базирующихся на архитектуре трансформеров, для анализа текстовых данных. Значимость данных сетей неоспорима: их способность к эффективной обработке обширных объемов текстовой информации, включая решение разнообразных задач естественного языка (NLP), делает их одними из наиболее востребованных инструментов в современном информационном мире.

Однако повышение качества решений подобных задач зачастую достигается за счет увеличения числа параметров нейронных сетей и объема анализируемого контекста. Это, в свою очередь, приводит к значительным затратам ресурсов и снижению эффективности работы таких сетей. В настоящее время базовые модели трансформеров, способные решать сложные задачи в области обработки естественного языка, требуют высокопроизводительных вычислительных ресурсов и специализированных ускорителей, таких как NVidia [1,2].

Основная часть. Для решения заявленной проблемы выделяются следующие этапы работ:

1) Проведение начального анализа существующих улучшений в области архитектуры трансформеров.
2) Исследование возможных комплексных оптимизаций структуры трансформеров.

В результате начального анализа и тестирования выдвигается гипотеза о том, что трансформер, в котором слой внимания строго отбирает входные признаки и обладает долгой ассоциативной памятью, способен эффективно выделять наиболее значимые слова в контексте и эффективно их сохранять на протяжении обработки всего входного массива данных. Это достигается путем комбинации различных улучшений из статей Performer[3], Switch Transformer[4], Compressive Transformer[5], Recurrent Attention Network[6], Sparse Attention Patterns[7].

Предложенное комплексное решение прежде всего направлено на решение следующих проблем:

1) Снижение количества параметров в сети путем уменьшения параметров в слое внимания, отбирая только необходимые и важные элементы входных данных.
2) Увеличение эффективности вычислений. Это достигается благодаря снижению вычислительной сложности механизма внимания с квадратичной до линейной за счет улучшений, предложенных в работе Performer. Кроме того, сжатое пространство параметров также сокращает количество вычислений, что позволяет получать выводы из сети за более короткое время.

Однако данное решение имеет проблему, которая приводит к значительному ухудшению контекста сети и, следовательно, может снизить точность. Для решения этой проблемы предлагается применение методики RAG[8], в которой контекст дополняется сетью DBN[9], предварительно обученной на специфическом домене данных. Использование DBN, обладающей хорошей ассоциативностью, позволит дополнять контекст сети качественными данными.

Выводы. Проведен анализ оптимизаций сетей архитектуры трансформер и разработана комплексная оптимизация данной архитектуры.

Список использованных источников:

1. Yi T., Dehghani M., Dara B., Donald M. et al. Efficient Transformers: A Survey //arXiv preprint arXiv:2009.0673. - 2022.
2. Satwik B., Arkil P. Navin G. et al. On the Computational Power of Transformers and its Implications in Sequence Modeling //arXiv preprint arXiv:2006.09286. - 2020.
3. Choromanski K. et al. Rethinking attention with performers //arXiv preprint arXiv:2009.14794. – 2020.
4. Fedus W., Zoph B., Shazeer N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity //The Journal of Machine Learning Research. – 2022. – Т. 23. – №. 1. – С. 5232-5270.
5. Rae J. W. et al. Compressive transformers for long-range sequence modelling //arXiv preprint arXiv:1911.05507. – 2019.
6. Graves A. Adaptive computation time for recurrent neural networks //arXiv preprint arXiv:1603.08983. – 2016.
7. Liu H. et al. Mitigating gender bias for neural dialogue generation with adversarial learning //arXiv preprint arXiv:2009.13028. – 2020.
8. Lewis P. et al. Retrieval-augmented generation for knowledge-intensive nlp tasks //Advances in Neural Information Processing Systems. – 2020. – Т. 33. – С. 9459-9474.
9. Hinton G. E. eep belief networks //Scholarpedia. – 2009. – Т. 4. – №. 5. – С. 5947.