УДК 004.89

Гайдуков Сергей Викторович, магистрант, Федеральное государственное бюджетное образовательное учреждение высшего образования «МИРЭА - Российский технологический университет»

ТРАНСФОРМЕРНЫЕ АРХИТЕКТУРЫ И ПРЕДОБУЧЕННЫЕ ЯЗЫКОВЫЕ МОДЕЛИ В ОБРАБОТКЕ ЕСТЕСТВЕННОГО ЯЗЫКА

Аннотация

В данной работе представлен обзор трансформерных архитектур и предобученных языковых моделей, получивших широкое распространение в обработке естественного языка. Рассматриваются механизмы внимания, устройство блока трансформера, принципы предобучения моделей типа BERT, а также методы тонкой настройки для прикладных задач. Проведён анализ современных тенденций, включая масштабирование моделей и развитие больших языковых моделей (LLM).

Annotation

This paper provides an overview of transformer architectures and pre-trained language models that have become widely used in natural language processing. It discusses attention mechanisms, the structure of the transformer block, the principles of pre-training models like BERT, and fine-tuning methods for applied tasks. The paper also analyzes current trends, including the scaling of models and the development of large language models (LLM).

Ключевые слова: трансформер, механизм внимания, BERT, языковые модели, предобучение, тонкая настройка, NLP.

Keywords: transformer, attention mechanism, BERT, language models, pre-training, fine-tuning, NLP.

Введение

Обработка естественного языка (Natural Language Processing, NLP) – одно из наиболее динамично развивающихся направлений искусственного интеллекта. На протяжении десятилетий исследователи использовали статистические методы и рекуррентные нейронные сети для решения задач перевода, классификации текстов и извлечения информации. Однако в 2017 году публикация работы «Attention is All You Need» [1] ознаменовала принципиальный сдвиг парадигмы: архитектура трансформера полностью вытеснила рекуррентные подходы в большинстве задач NLP.

Ключевым событием стал выход модели BERT (Bidirectional Encoder Representations from Transformers) в 2018 году [2]. Благодаря двунаправленному пониманию контекста и предобучении на больших корпусах, BERT установил рекорды качества сразу в 11 задачах NLP. Сегодня трансформерные модели составляют основу современных систем машинного перевода, диалоговых агентов и больших языковых моделей (LLM).

Цель работы – собрать и конкретизировать ключевые принципы трансформерных и предобученных языковых моделей, рассмотреть механизмы их работы, а также проанализировать применение данных подходов на практике.

От рекуррентных сетей к трансформерам

До появления трансформеров основным инструментом для работы с последовательными данными были рекуррентные нейронные сети (RNN), а также их усовершенствованные варианты – LSTM и GRU. Их главный недостаток состоит в том, что обработка данных происходит только последовательно: вычисление каждого нового шага возможно только после окончания предыдущего. Это сильно ограничивает возможности параллельных вычислений на GPU [3]. Помимо этого, при работе с большими последовательностями возникает эффект затухающего градиента, из-за которого модель теряет информацию о контексте, находящемся далеко от текущего элемента.

Механизм внимания (attention) был предложен для решения проблемы учета дальних зависимостей. Благодаря ему модель может явно оценивать важность каждого элемента входной последовательности при формировании выходных данных. Изначально этот механизм применялся вместе с RNN в задачах машинного перевода [4], однако в работе [1] было показано, что attention способен полностью заменить рекуррентные структуры, при этом обеспечивая значительно более высокий уровень параллелизма.

Архитектура трансформера

Основным строительным блоком трансформерной модели является механизм самовнимания (self-attention). Для каждого входного токена x_i вычисляются три вектора: запрос (query, Q), ключ (key, K) и значение (value, V) – путём умножения на обучаемые матрицы W^Q, W^K, W^V. Оценка внимания токена i к токену j вычисляется как скалярное произведение запроса q_i и ключа k_j, нормированное на корень из размерности ключей показано в формуле (1):

$$\alpha_i j = softmax\left(q_i \cdot k_j / \sqrt{d_k}\right) \tag{1}$$

Выходное представление токена i является взвешенной суммой векторов значений всех позиций: $a_i = \Sigma_j \alpha_i j \cdot v_j$. Таким образом, каждый токен «видит» весь контекст одновременно, что решает проблему дальних зависимостей.

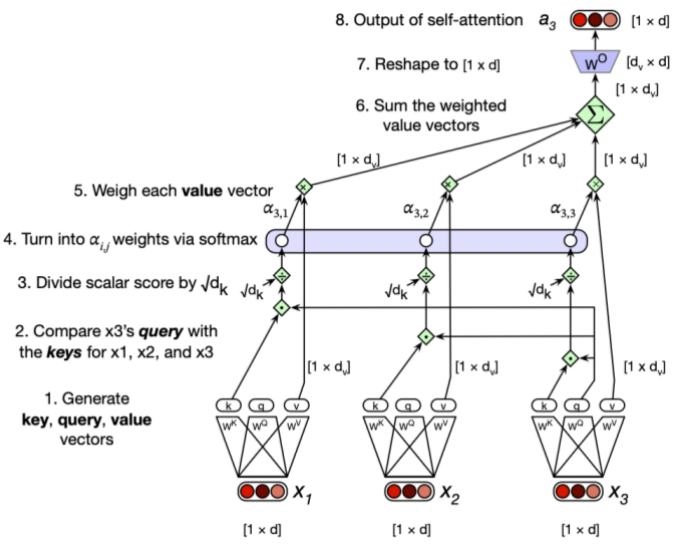

Рисунок 1. Вычисление вектора самовнимания a_3 для третьего токена. Показан процесс формирования весов α через скалярное произведение запроса и ключей

На практике применяется многоголовое внимание (multi-head attention): A независимых голов внимания работают параллельно, каждая со своими матрицами параметров. Такая функция позволяет модели извлечь одновременно разные виды семантических и синтаксических зависимостей. Выходы голов объединяются и проецируются матрицей W^O.

Полный блок трансформерной модели включает в себя слой многоголового внимания, остаточное соединение, нормировку слоя (LayerNorm), позиционно-независимую полносвязную сеть (FFN), снова остаточное соединение и LayerNorm. Трансформер представляет собой композицию из L повторяющихся блоков. Большие языковые модели используют от 12 (GPT-2 small) до 96 и более слоёв (GPT-3 large) [5].

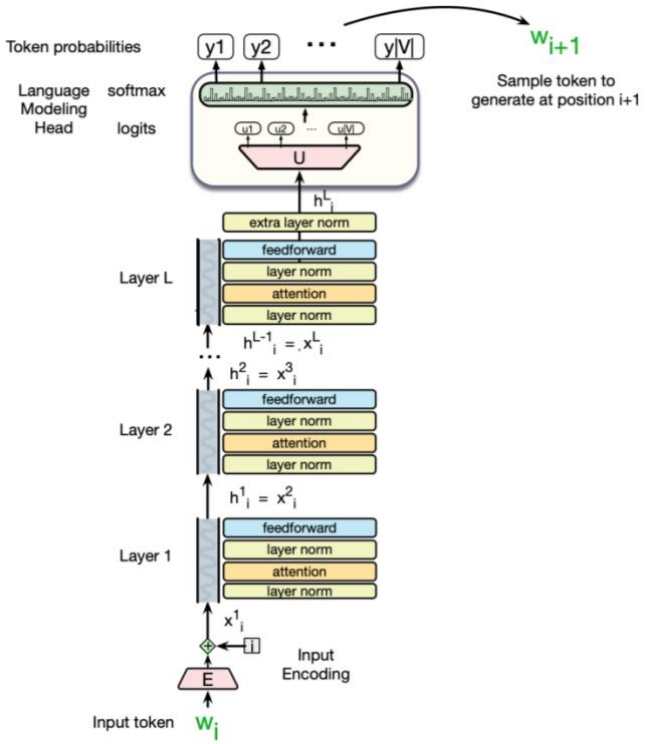

Рисунок 2. Полная архитектура трансформера для декодинга

Предобученные языковые модели: BERT

Модель BERT (Devlin et al., 2019) является двунаправленным энкодером, чье обучение базировалось на задаче маскированного языкового моделирования (MLM), где модель учится восстанавливать скрытые токены в последовательности. Во время предобучения 15% от входных токенов случайным маскируются специальным токеном, так называемой маской, и модель старается восстановить исходные токены по контексту. Авторегрессионные модели работают с текстом только в одном направлении, а BERT смотрит на него сразу с двух сторон. Благодаря этому он лучше понимает смысл каждого слова, учитывая всё, что идет до и после него. BERT обучается на задаче предсказывания предложения (Next Sentence Prediction, NSP). Модель определяет является ли второе предложение продолжением первого. Это всё помогает усваивать модели связи между предложениями. Основная версия BERT содержит 12 слоёв трансформера, 12 голов внимания и 110 млн параметров, а расширенная (BERT-Large) – 24 слоя и 340 млн параметров [2].

После предобучения модель адаптируется к задачам с помощью тонкой настройки (fine-tuning). Для классификации последовательностей к началу добавляется специальный токен CLS, а его выходной вектор h_CLS из последнего слоя трансформера передаётся в голову классификации. Обучение происходит на размеченных данных благодаря кросс-энтропийной функции потерь. Обычно, чтобы доработать уже обученную модель, достаточно внести небольшие изменения, затрагивающие лишь самые верхние слои [6].

BERT продемонстрировал значительное улучшение результатов. Тонкая настройка для анализа тональности (GLUE benchmark) обеспечила прирост точности в несколько процентных пунктов по сравнению с предыдущими показателями. Аналогично для задач распознавания именованных сущностей (NER), ответа на вопросы (SQuAD) и распознавания текстовых импликаций (NLI).

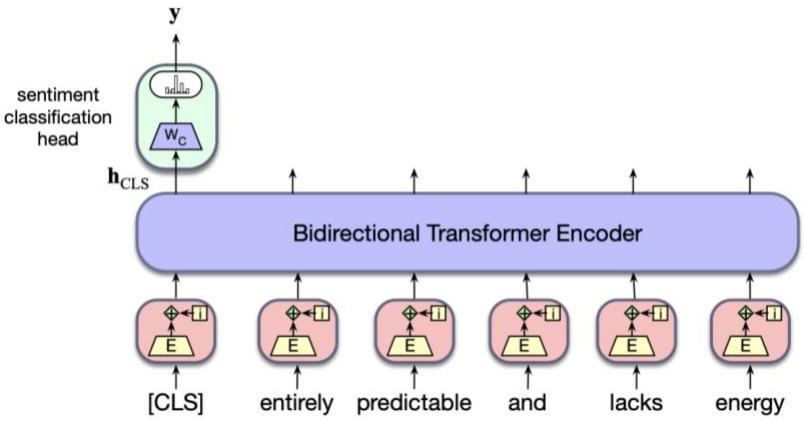

Рисунок 3. Архитектура тонкой настройки BERT для классификации: выходной вектор токена CLS направляется в классификационную головку W_C

Применения трансформеров в задачах NLP

Трансформерные модели на сегодняшний день занимают лидирующие позиции в широком спектре задач NLP. Анализ тональности (sentiment analysis) – задача определения эмоциональной окраски текста – была одной из первых областей, где BERT продемонстрировал превосходство над классическими методами логистической регрессии и CNN [6]. Для двоичной классификации (положительный / отрицательный отзыв) или трёхклассовой задачи (positive / neutral / negative) достаточно добавления одного линейного слоя поверх [CLS]-вектора.

В задаче распознавания именованных сущностей (NER) каждому токену последовательности сопоставляется метка класса (Person, Organization, Location и т.д.). Для этого вместо глобального [CLS]-вектора используются все выходные векторы энкодера; над каждым из них устанавливается классификатор. Аналогичная схема применяется при разметке частей речи (POS-tagging) и синтаксическом анализе.

Машинный перевод потребовал полной энкодер-декодерной архитектуры. Декодер дополнительно содержит слои «перекрёстного внимания» (cross-attention), которые позволяют каждому генерируемому токену обращаться к представлениям исходного предложения. Благодаря использованию этого

принципа, модели T5 и mBART смогли обеспечить качество перевода для высокоресурсных языковых пар, практически неотличимое от человеческого [5].

Современные тенденции: масштабирование и LLM

Исследования законов масштабирования [5] выявили, что чем больше параметров у языковой модели, чем больше данных для ее обучения и чем больше вычислительных мощностей используется, тем лучше она работает. Причем это улучшение происходит по определенному степенному закону. Это открытие начало гонку за размером моделей: от 117 млн параметров (GPT-1) до более чем 400 млрд.

Современные большие языковые модели основаны на декодерной архитектуре трансформера с авторегрессионным обучением. Главными особенностями этих моделей являются обучение с подкреплением на основе обратной связи от людей (RLHF), инструкционная настройка и способности к контекстному обучению. Вместе с расширением масштабов возникают трудности, такие как значительные вычислительные расходы, вероятность ложных срабатываний и сложности с осмыслением работы моделей. Для применений с ограниченными ресурсами используется дистилляция знаний (DistilBERT, TinyBERT), позволяющая уменьшить размер модели в 2–4 раза при минимальной потере качества.

Заключение

В работе рассмотрены основные принципы трансформерных архитектур и предобученных языковых моделей. Механизм самовнимания может позволить эффективно моделировать зависимости в тексте при полной параллелизации вычислений. Концепция предобучения на больших корпусах с последующей настройкой обеспечивает высокое качество на разнообразных задачах NLP при ограниченных размеченных данных.

Модель BERT и её производные заложили основу современного NLP. Дальнейшее масштабирование привело к появлению больших языковых

моделей, демонстрирующих замечательные возможности в широком спектре задач. Открытыми остаются вопросы интерпретируемости, снижения вычислительных затрат и борьбы с галлюцинациями – что делает данную область активной темой для дальнейших исследований.

Список литературы

1. *Vaswani, A., Shazeer, N., Parmar, N. [et al.].* Attention is all you need // Advances in Neural Information Processing Systems. – 2017. – Vol. 30.
2. *Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.* BERT: Pre-training of deep bidirectional transformers for language understanding // Proceedings of NAACL-HLT. – 2019. – P. 4171–4186.
3. *Jurafsky, D., Martin, J. H.* Speech and Language Processing. Draft of January 6, 2026. – Stanford University, 2026. – Chapter 8: Transformers.
4. *Bahdanau, D., Cho, K. H., Bengio, Y.* Neural machine translation by jointly learning to align and translate // ICLR. – 2015.
5. *Kaplan, J., McCandlish, S., Henighan, T. [et al.].* Scaling laws for neural language models // arXiv preprint arXiv:2001.08361. – 2020.
6. *Jurafsky, D., Martin, J. H.* Speech and Language Processing. Draft of January 6, 2026. – Stanford University, 2026. – Chapter 10: Masked Language Models.
7. *Brown, T., Mann, B., Ryder, N. [et al.].* Language models are few-shot learners // Advances in NeurIPS. – 2020. – Vol. 33.
8. *Pang, B., Lee, L.* Opinion mining and sentiment analysis // Foundations and Trends in Information Retrieval. – 2008. – Vol. 2, No. 1-2. – P. 1–135.
9. *Liu, Y., Ott, M., Goyal, N. [et al.].* RoBERTa: A robustly optimized BERT pretraining approach // arXiv preprint arXiv:1907.11692. – 2019.
10. *Sanh, V., Debut, L., Chaumond, J., Wolf, T.* DistilBERT, a distilled version of BERT // arXiv preprint arXiv:1910.01108. – 2019.
11. *Llama Team.* The Llama 3 herd of models // arXiv preprint. – 2024.
12. *Elhage, N., Nanda, N., Olsson, C. [et al.].* A mathematical framework for transformer circuits // Transformer Circuits Thread. – 2021.

Literature

1. *Vaswani, A., Shazeer, N., Parmar, N. [et al.].* Attention is all you need // Advances in Neural Information Processing Systems. – 2017. – Vol. 30.
2. *Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.* BERT: Pre-training of deep bidirectional transformers for language understanding // Proceedings of NAACL-HLT. – 2019. – P. 4171–4186.

3. *Jurafsky, D., Martin, J. H.* Speech and Language Processing. Draft of January 6, 2026. – Stanford University, 2026. – Chapter 8: Transformers.
4. *Bahdanau, D., Cho, K. H., Bengio, Y.* Neural machine translation by jointly learning to align and translate // ICLR. – 2015.
5. *Kaplan, J., McCandlish, S., Henighan, T. [et al.].* Scaling laws for neural language models // arXiv preprint arXiv:2001.08361. – 2020.
6. *Jurafsky, D., Martin, J. H.* Speech and Language Processing. Draft of January 6, 2026. – Stanford University, 2026. – Chapter 10: Masked Language Models.
7. *Brown, T., Mann, B., Ryder, N. [et al.].* Language models are few-shot learners // Advances in NeurIPS. – 2020. – Vol. 33.
8. *Pang, B., Lee, L.* Opinion mining and sentiment analysis // Foundations and Trends in Information Retrieval. – 2008. – Vol. 2, No. 1-2. – P. 1–135.
9. *Liu, Y., Ott, M., Goyal, N. [et al.].* RoBERTa: A robustly optimized BERT pretraining approach // arXiv preprint arXiv:1907.11692. – 2019.
10. *Sanh, V., Debut, L., Chaumond, J., Wolf, T.* DistilBERT, a distilled version of BERT // arXiv preprint arXiv:1910.01108. – 2019.
11. *Llama Team.* The Llama 3 herd of models // arXiv preprint. – 2024.
12. *Elhage, N., Nanda, N., Olsson, C. [et al.].* A mathematical framework for transformer circuits // Transformer Circuits Thread. – 2021.