Фролов Д. О., Петрова А. Н.
D. O. Frolov, A. N. Petrova

СОВРЕМЕННЫЕ ПОДХОДЫ К АРХИТЕКТУРЕ НЕЙРОННЫХ СЕТЕЙ ДЛЯ ПОВЫШЕНИЯ РЕЛЕВАНТНОСТИ ПОИСКА В БОЛЬШИХ ОБЪЁМАХ ДАННЫХ

MODERN APPROACHES TO NEURAL NETWORK ARCHITECTURE TO INCREASE THE RELEVANCE OF SEARCH IN LARGE DATA VOLUME

Фролов Дмитрий Олегович – аспирант Комсомольского-на-Амуре государственного университета (Россия, Комсомольск-на-Амуре). E-mail: optcompanys@mail.ru.
Dmitriy O. Frolov – Postgraduate Student, Komsomolsk-na-Amure State University (Russia, Komsomolsk-on-Amur). E-mail: optcompanys@mail.ru.

Петрова Анна Николаевна – кандидат технических наук, заведующая кафедрой «Проектирование, управление и развитие информационных систем» Комсомольского-на-Амуре государственного университета (Россия, Комсомольск-на-Амуре). E-mail: PetrovaAN2006@yandex.ru.
Anna N. Petrova – PhD in Engineering, Head of Design, Management and Development of Information Systems Department, Komsomolsk-na-Amure State University (Russia, Komsomolsk-on-Amur). E-mail: PetrovaAN2006@yandex.ru.

Аннотация. В статье исследуются современные нейросетевые архитектуры, направленные на повышение точности поисковых результатов при обработке больших объёмов данных. Особое внимание уделяется трансформерам, методам обучения с учителем и без, а также стратегиям оптимизации моделей. Для оценки эффективности этих архитектур был проведён эксперимент, в рамках которого сравнивались различные модели по показателям точности и релевантности, включая BERT, RoBERTa и Longformer. Полученные результаты демонстрируют, что усовершенствованные модели на основе трансформеров обеспечивают значительные улучшения качества поисковых запросов.

Summary. This paper explores modern neural network architectures aimed at improving the accuracy of search results when processing large amounts of data. Particular attention is paid to transformers, supervised and unsupervised learning methods, and model optimization strategies. To evaluate the effectiveness of these architectures, an experiment was conducted in which various models were compared in terms of accuracy and relevance, including BERT, RoBERTa, and Longformer. The results demonstrate that improved transformer-based models provide significant improvements in the quality of search queries.

Ключевые слова: поиск информации, машинное обучение, функция потерь, квантовая информация.

Key words: search for information, machine learning, loss function, quantum information.

УДК 004.41

Введение. Современные поисковые системы, которые обрабатывают большие объёмы данных, сталкиваются с трудностью точного извлечения информации. Задача нахождения релевантных данных становится особенно важной для динамичных баз данных, где необходимо не только обеспечить высокую точность, но и адаптироваться к предпочтениям пользователей и новым запросам. В данной работе рассматривается, как трансформерные архитектуры и методы их оптимизации могут способствовать улучшению релевантности поиска. В статье также представлен эксперимент, направленный на оценку различных нейросетевых моделей и их влияния на эффективность поиска.

Современные подходы к поиску информации. Трансформеры, такие как BERT (Bidirectional Encoder Representations from Transformers) и его усовершенствованная версия RoBERTa (Robustly Optimized BERT Pretraining Approach), являются одними из ключевых прорывов в области обработки естественного языка (NLP) за последние годы. Эти модели значительно

трансформировали традиционные методы анализа и понимания текста, открыв новые возможности для решения разнообразных задач, включая анализ настроений, извлечение информации, машинный перевод и др.

Главной особенностью трансформеров является их способность учитывать контекст запроса на всех уровнях текста. Это особенно важно, поскольку традиционные модели часто не справлялись с улавливанием долгосрочных зависимостей между словами в предложении. Например, если слово в начале текста влияет на значение слова в конце, старые модели не всегда могли отразить такую зависимость. В отличие от них, трансформеры способны учитывать как локальные, так и глобальные зависимости, что обеспечивает более точную интерпретацию смысла текста [1].

Основным элементом, который обеспечивает эффективность трансформеров в решении различных задач, является механизм self-attention (самовнимание). Этот механизм позволяет модели оценивать важность каждого слова в контексте других слов в предложении независимо от их расстояния друг от друга. Он достигается путём вычисления весов внимания для каждой пары слов, что позволяет модели анализировать взаимодействие между всеми словами в предложении при обработке каждого отдельного элемента текста.

BERT использует два основных подхода для обучения: masked language modeling (MLM) и next-sentence prediction (NSP) [10]. В первом случае часть слов в предложении скрывается и модель должна предсказать их, опираясь на контекст. Это позволяет модели учитывать двусторонний контекст, включая как предыдущие, так и последующие слова. Во втором подходе модель обучается предсказывать, будет ли следующее предложение логичным продолжением текущего, что помогает ей учитывать связи между предложениями.

RoBERTa является усовершенствованной версией BERT, которая была оптимизирована с целью повышения эффективности предобучения. Эта модель отличается от BERT более агрессивными подходами к обучению, такими как увеличение объёма данных, расширение длины последовательностей и отказ от задачи NSP, которая показала низкую эффективность в реальных приложениях [2]. Вместо этого RoBERTa сосредотачивается на более глубоком обучении с использованием длинных последовательностей и улучшенных оптимизационных методов, что способствует улучшению производительности модели в различных задачах NLP.

Ключевое преимущество моделей типа BERT и RoBERTa заключается в их способности воспринимать текст как целостную структуру, а не как набор отдельных элементов, что значительно повышает точность и обоснованность выводов. Эти модели стали основой для разработки множества успешных решений, таких как поисковые системы, системы рекомендаций, чат-боты и другие, которые требуют глубокого понимания контекста текста для обеспечения более точных и релевантных откликов. Для задач поиска с заранее размеченными данными часто используется метод обучения с учителем, который включает дообучение моделей, предварительно обученных на больших объёмах данных [8]. В условиях нехватки размеченных данных активно применяется обучение без учителя, включающее методы самообучения, такие как контрастивное обучение и псевдомаркировка. Обучение нейронных сетей можно разделить на два основных подхода: с учителем (supervised learning) и без учителя (unsupervised learning). Эти методы активно применяются для повышения релевантности поиска и оптимизации обработки данных в нейронных сетях, включая трансформерные модели.

Обучение с учителем (Supervised Learning). Обучение с учителем предполагает, что модель обучается на данных, которые уже имеют соответствующие метки (labels). Это означает, что для каждой обучающей выборки известно, какой результат должен быть получен. Модель получает входные данные и их соответствующие ответы, затем обучается минимизировать разницу между предсказанными и реальными метками. Такой подход активно используется для задач ранжирования и оценки релевантности в поисковых системах. Например, если нужно оценить, насколько страница соответствует запросу, модель обучается на большом наборе данных, содержащем различные запросы и их релевантные результаты. На основе этого обучающего материала модель может впоследствии ранжировать новые данные, учитывая релевантность, ориентируясь на примеры из обучающего набора.

Процесс обучения с учителем в трансформерах [7]:

1. Подготовка данных. Составляется набор данных, где для каждого запроса (входного текста) есть несколько меток релевантности (например, документ оценивается по шкале от 1 до 5).

2. Тренировка модели. Модель, например на основе BERT, обрабатывает текст запроса и документа, генерируя представление, которое затем сравнивается с метками.

3. Оптимизация. Ошибки, возникшие из-за несоответствия предсказанных и фактических меток, минимизируются с помощью алгоритмов оптимизации (например, обратного распространения ошибки).

4. Применение. Обученная модель может предсказывать релевантность документов для новых запросов.

Преимущества обучения с учителем:

- высокая точность при наличии качественных данных;
- обучение на конкретных задачах, где метки позволяют модели стать более предсказуемой;
- возможность легко оценить качество модели через метрики, такие как точность и полнота.

Ограничения:

- требуется большое количество размеченных данных, что может быть дорого и трудоёмко;
- модель может быть избыточно специфичной для данных и плохо обобщаться на новые, непривычные случаи.

Обучение без учителя (Unsupervised Learning). Обучение без учителя осуществляется на данных, не содержащих меток или заранее известных ответов. В этом случае модель обучается выявлять скрытые структуры, закономерности и группы данных, опираясь исключительно на их содержание. Модель должна самостоятельно извлекать значимую информацию, например классифицируя схожие тексты или выявляя тематические паттерны. Такой подход широко используется для кластеризации документов и тематического анализа. Это помогает эффективно организовывать данные для упрощения их дальнейшего поиска [3]. Например, модель без учителя может быть обучена на большом наборе данных для создания тематической карты, где она группирует документы по схожести, что впоследствии помогает оптимизировать ранжирование результатов поиска по теме и смысловой близости.

Процесс обучения без учителя в трансформерах [6]:

1. Инициализация данных. Загружается большой объём неразмеченного текста, например набор документов без указания их тем.

2. Выявление закономерностей. Модель, например на основе GPT или BERT, обрабатывает текст, выстраивая внутренние представления слов и фраз, выявляя закономерности (например, темы).

3. Кластеризация. Модель группирует тексты или предложения, выявляя скрытые паттерны или темы.

4. Применение. Результаты могут быть использованы для улучшения рекомендаций или ранжирования документов, чтобы выдавать наиболее подходящие по теме результаты.

Преимущества обучения без учителя:

- отсутствие необходимости в метках позволяет экономить ресурсы на разметке данных;
- модель может выявлять неизвестные ранее категории или темы;
- модель подходит для обработки больших объёмов данных, находя связи и закономерности, которые могут быть неочевидны для человека.

Ограничения:

- полученные результаты могут быть трудно интерпретируемыми, т. к. модель находит закономерности, не всегда соответствующие реальной смысловой структуре данных;
- невозможно оценить точность без меток, что делает анализ качества работы модели субъективным.

Постановка эксперимента. Для исследования методов повышения релевантности поиска с применением современных нейросетевых архитектур был проведён эксперимент, включающий обучение и оценку моделей на реальных поисковых данных. В эксперименте использовались

нейронные сети, основанные на трансформерах, включая стандартные архитектуры, такие как BERT и RoBERTa, а также их модификации, адаптированные для задач поиска. В ходе работы было оценено влияние архитектурных усовершенствований на улучшение релевантности поиска по сравнению с базовой моделью.

Целью эксперимента было определить, какие архитектурные модификации трансформеров способны улучшить релевантность поиска в условиях обработки больших объёмов данных. Мы проводили сравнение производительности моделей, учитывая метрики качества поиска, такие как точность и полнота, на задачах ранжирования и поиска. Для эксперимента был использован датасет MS MARCO – популярный набор данных, включающий запросы и соответствующие релевантные ответы для текстового поиска, что делает его удобным для тестирования моделей ранжирования [4]. В нём содержатся несколько десятков тысяч запросов и ответов, а также оценки релевантности, которые позволили обучать и проверять модели на реальных примерах. Затем была проведена предварительная обработка данных:

- все запросы и тексты были очищены от лишних символов и подготовлены для обработки в виде токенов, которые могли воспринять BERT и его производные. Запросы нормализованы для единообразия;

- был проведён анализ часто встречающихся токенов, который помог улучшить токенизацию.

Для эксперимента были обучены несколько моделей на основе BERT: стандартная версия BERT, RoBERTa и BERT с улучшениями архитектуры, включая более глубокие слои и оптимизированные механизмы внимания. Все модели использовали одинаковый процесс обучения с оптимизацией гиперпараметров для обеспечения стабильности. В процессе эксперимента были применены стандартные метрики для оценки релевантности поиска, такие как MRR (Mean Reciprocal Rank) и NDCG (Normalized Discounted Cumulative Gain), которые были выбраны за их способность точно оценивать релевантность и учитывать аспекты ранжирования [5]. После оценки производительности моделей на основе точности и полноты результаты были систематизированы в табл. 1 для дальнейшего сравнения.

Таблица 1

Оценка производительности моделей

Модель	MRR@10	NDCG@10	Параметры	Обучение, ч
BERT-base	0.26	0.53	110M	3
RoBERTa	0.29	0.57	125M	4
BERT с модификацией	0.32	0.59	135M	4.5

Из табл. 1 видно, что модель BERT с модификацией показала лучшие результаты из-за более тщательной предобученной архитектуры, что помогло ей более эффективно справляться с задачей поиска.

На рис. 1 изображён график сравнения MRR@10 и NDCG@10.

Модифицированная версия BERT, которая включала дополнительные слои внимания и улучшенные механизмы обработки информации, продемонстрировала лучшие результаты. Её значение метрики NDCG@10 составило 0.59, что свидетельствует о значительном улучшении релевантности результатов по сравнению с базовой моделью.

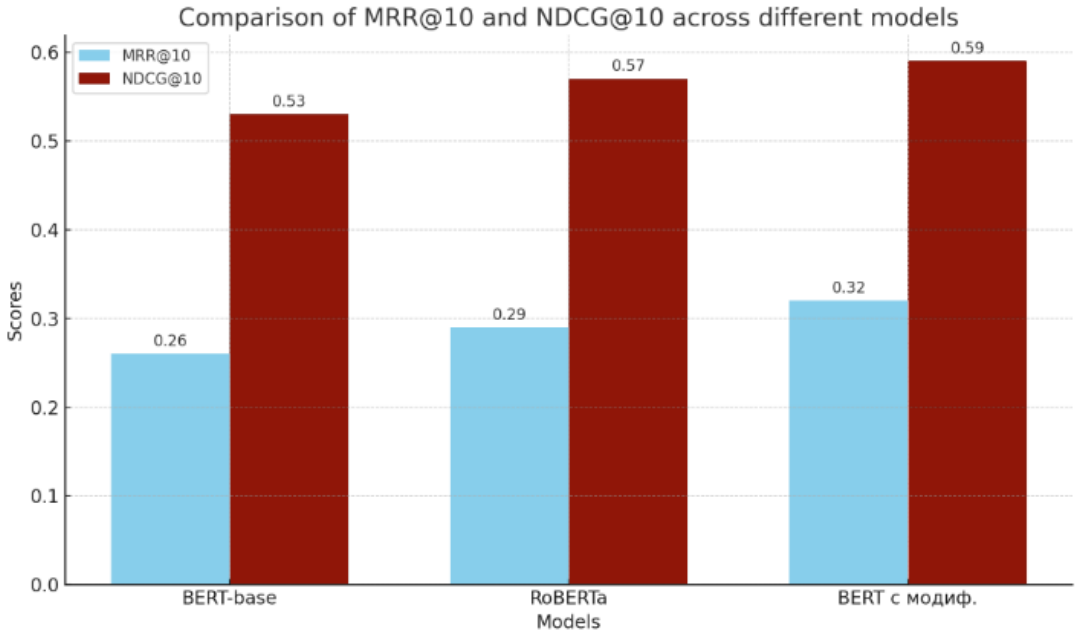

Рис. 1. График сравнения MRR@10 и NDCG@10

Заключение. В ходе исследования были рассмотрены современные подходы к архитектуре нейронных сетей, направленные на улучшение релевантности поиска в больших объёмах данных, с акцентом на трансформеры и их модификации. Проведённый эксперимент подтвердил эффективность улучшений в архитектуре моделей, таких как BERT и RoBERTa, для повышения точности и качества поисковых запросов. Результаты показали, что модели с усовершенствованной архитектурой, включая дополнительные слои внимания и оптимизированные механизмы обработки информации, обеспечивают значительное улучшение релевантности поиска. Экспериментальная оценка моделей, включая стандартный BERT и его модификации, показала, что более глубокие версии и улучшенные механизмы внимания значительно повышают качество поисковых систем, способных обрабатывать сложные запросы.

ЛИТЕРАТУРА

1. Vaswani, A., Shardlow, M., & Parmar, N. (2017). Attention is All You Need. In Advances in Neural Information Processing Systems.
2. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In NAACL-HLT.
3. Raffel, C., Shinn, C., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. In Journal of Machine Learning Research.
4. Liu, Y., Ott, M., Goyal, N., & Zhang, X. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.
5. Lan, Z., Chen, M., Goodman, S., & Gimpel, K. (2020). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In International Conference on Learning Representations.
6. Khashabi, D., et al. (2020). UNIFIEDQA: Crossing Format Boundaries with a Single QA System. In NeurIPS.
7. Radford, A., et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. In International Conference on Machine Learning.
8. Beltagy, I., Lo, K., & Cohan, A. (2020). SciBERT: A Pretrained Language Model for Scientific Text. In Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics.
9. Zaheer, M., Ruochen, H., & Ranganath, R. (2020). Deep Sets. In Advances in Neural Information Processing Systems.
10. Zhang, Y., & Wallace, B. (2019). A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.