ISSN 2706-6266
ISSN 3007-1062
ISSN 3007-1070

ГРНТИ 28.23.08, 29.23.31 **DOI: https://doi.org/10.62724/202520303**

Казиева Назым Магидулловна,
Доктор PhD , старший преподаватель,
Евразийский национальный университет имени Л.Н. Гумилева, Республика Казахстан,
010000, г.Астана, kaziyeva_nm@enu.kz, ORCID ID: 0000-0002-7559-1795

ПРИМЕНЕНИЕ ГЛУБОКОГО ОБУЧЕНИЯ ДЛЯ АНАЛИЗА БОЛЬШИХ ДАННЫХ: МЕТОДЫ, ПОДХОДЫ И ПЕРСПЕКТИВЫ

Аннотация. В последние десятилетия анализ больших данных стал одной из важнейших задач в области информационных технологий. Глубокое обучение, как часть искусственного интеллекта, предоставило эффективные инструменты для обработки и извлечения знаний из больших объемов данных. В данной статье рассматриваются ключевые методы глубокого обучения, такие как искусственные нейронные сети, свёрточные нейронные сети, рекуррентные нейронные сети, генеративные состязательные сети и автокодировщики, которые активно используются для анализа различных типов данных. Описываются подходы к их применению, включая предобработку данных, распределенные вычисления и методы обучения с учителем и без учителя. Кроме того, рассматриваются актуальные вопросы, такие как интерпретируемость моделей и развитие мультимодальных данных. В статье также анализируются перспективы использования глубокого обучения в сочетании с квантовыми вычислениями и методы работы с ограниченным количеством данных. Несмотря на успехи, глубокое обучение сталкивается с рядом вызовов, таких как потребность в огромных вычислительных мощностях и этические вопросы, связанные с использованием данных. В заключение, отмечается, что глубокое обучение имеет значительный потенциал для решения сложных задач анализа больших данных, и его развитие продолжит открывать новые возможности для различных областей науки и промышленности.

Ключевые слова: глубокое обучение, большие данные, нейронные сети, свёрточные сети, рекуррентные сети, генеративные состязательные сети, автокодировщики, анализ данных, распределённые вычисления, обучение с учителем, обучение без учителя, интерпретируемость моделей, мультимодальные данные, квантовые вычисления, этика данных.

Введение. С развитием технологий и ростом объемов данных, которые генерируются во всех областях человеческой деятельности, задачи обработки и анализа больших данных становятся всё более актуальными. В условиях огромных объемов информации традиционные методы анализа сталкиваются с ограничениями, что требует использования более мощных и эффективных инструментов. Одним из таких инструментов стало глубокое обучение (ГД), которое стало ключевой технологией для извлечения знаний из больших данных.

Глубокое обучение представляет собой подход в машинном обучении, основанный на использовании многослойных нейронных сетей, способных автоматически выделять признаки и выявлять сложные закономерности в данных [2]. Методы глубокого обучения нашли применение в широком спектре задач, включая обработку изображений, текста, звука и временных рядов, а также в таких областях, как медицина, финансы, маркетинг и научные исследования.

ISSN 2706-6266
ISSN 3007-1062
ISSN 3007-1070

Данная статья посвящена исследованию методов и подходов глубокого обучения, которые активно используются для анализа больших данных. Рассматриваются ключевые типы нейронных сетей, такие как искусственные нейронные сети, свёрточные и рекуррентные сети, а также подходы к их применению, включая распределенные вычисления и обучение с учителем и без учителя. В статье также обсуждаются перспективы и вызовы, с которыми сталкиваются исследователи и практики в этой области [5, 7].

Методы глубокого обучения для анализа больших данных

Глубокое обучение (ГД) представляет собой мощную технологию для анализа и обработки больших данных, применяемую для решения широкого спектра задач. Рассмотрим ключевые методы, которые активно используются в этой области:

1. Искусственные нейронные сети (ИНС)

Искусственные нейронные сети (ИНС) являются основой глубокого обучения [5]. Они состоят из нескольких слоев нейронов, каждый из которых выполняет математическую операцию для обработки информации. Основное их преимущество заключается в способности автоматически выделять признаки из необработанных данных, таких как текст, изображения или временные ряды. Многослойные перцептроны (MLP) являются одним из самых простых типов ИНС и могут использоваться для решения задач классификации и регрессии. Однако при работе с большими объемами данных требуется использование более сложных архитектур [8].

2. Свёрточные нейронные сети (CNN)

Свёрточные нейронные сети (CNN) наиболее эффективно применяются в задачах, связанных с изображениями и видео [6]. Основное преимущество CNN заключается в способности автоматического выделения пространственных признаков, таких как формы и текстуры, из пиксельных данных. В последние годы они также активно используются в задачах анализа сигналов, текстов и даже в финансовом прогнозировании. CNN часто используются для построения моделей, которые могут обрабатывать и анализировать большие объемы визуальной информации с высокой степенью точности.

3. Рекуррентные нейронные сети (RNN) и LSTM

Рекуррентные нейронные сети (RNN) и их расширения, такие как долгосрочная память (LSTM), идеально подходят для работы с последовательными данными [6]. Эти сети способны запоминать информацию о предыдущих шагах в последовательности, что делает их очень полезными для обработки текста, аудио, временных рядов и других данных, где порядок элементов имеет значение. RNN и LSTM нашли широкое применение в задачах обработки естественного языка (NLP), прогнозировании временных рядов и обнаружении аномалий в данных.

4. Генеративные состязательные сети (GAN)

Генеративные состязательные сети (GAN) состоят из двух нейронных сетей — генератора и дискриминатора, которые конкурируют между собой [3]. Генератор генерирует новые данные, а дискриминатор пытается отличить реальные данные от сгенерированных. GAN используются для генерации новых данных, таких как изображения, тексты, музыка, и для улучшения качества данных. В области больших данных они могут быть применены для синтеза новых данных, улучшения качества данных и дополнения неполных или искаженных данных.

5. Автокодировщики (Autoencoders)

Автокодировщики представляют собой нейронные сети, которые обучаются сжатию данных и их восстановлению. Этот метод используется для уменьшения размерности данных, что критически важно при обработке больших объемов

ISSN 2706-6266
ISSN 3007-1062
ISSN 3007-1070

информации. Автокодировщики помогают выявлять скрытые паттерны в данных, устранять шум и улучшать качество сигналов, а также могут быть полезными в задачах восстановления утраченных данных [1]. В сочетании с другими методами они могут использоваться для улучшения качества анализа и построения более эффективных моделей.

6. Глубокие нейронные сети с обучением по частям (Transfer Learning)

Метод обучения по частям (transfer learning) является важным инструментом для обработки больших данных, особенно когда доступных обучающих данных недостаточно [7]. В этом подходе модель обучается на одном наборе данных и затем используется для решения задачи на другом, часто меньшем, наборе данных. Это позволяет значительно ускорить обучение и повысить точность моделей при ограниченных ресурсах. Transfer learning активно используется в области обработки изображений и текста, где могут быть использованы заранее обученные модели для решения специализированных задач.

7. Обучение с подкреплением (Reinforcement Learning)

Обучение с подкреплением (RL) — это метод, в котором агент обучается путем взаимодействия с окружающей средой, получая обратную связь в виде вознаграждений или наказаний [4]. Этот подход активно используется для решения задач оптимизации, таких как управление роботами, прогнозирование спроса, а также в играх и финансовых рынках. В контексте больших данных обучение с подкреплением позволяет разработать эффективные модели для принятия решений в реальном времени, используя данные о состоянии системы и отклики на действия.

8. Методы ансамблирования

Методы ансамблирования, такие как случайный лес (Random Forest) и градиентный бустинг (Gradient Boosting), представляют собой комбинации нескольких моделей, которые работают вместе для улучшения точности предсказаний [7]. Эти методы часто используются в задачах, где необходимо работать с большими объемами данных и обеспечить стабильность и точность результатов. При анализе больших данных ансамблирование помогает снизить вероятность ошибок и повысить устойчивость моделей к различным вариациям данных.

Подходы к применению глубокого обучения для анализа больших данных

1. Обработка и предобработка данных

Перед применением методов глубокого обучения, важно провести качественную обработку данных, что включает в себя очистку данных, их нормализацию, уменьшение размерности и устранение шумов. В рамках обработки больших данных особое внимание уделяется выбору правильных методов для работы с такими характеристиками данных, как пропущенные значения, выбросы и несбалансированность классов.

2. Распределенные вычисления и параллельное обучение

Анализ больших данных требует огромных вычислительных ресурсов, что делает необходимым использование распределенных вычислений. Платформы, такие как Apache Hadoop и Apache Spark, используются для распределенной обработки данных, в то время как фреймворки глубокого обучения, например, TensorFlow и PyTorch, предоставляют возможности для распределенного обучения нейронных сетей. Это позволяет эффективно обрабатывать и анализировать огромные объемы данных, улучшая производительность и сокращая время обработки.

3. Обучение с учителем и без учителя

Глубокое обучение применяется как в задачах с учителем (например, классификация и регрессия), так и в задачах без учителя (например, кластеризация и выделение признаков). Для анализа больших данных часто используется

ISSN 2706-6266
ISSN 3007-1062
ISSN 3007-1070

комбинированный подход, включающий как обучение с учителем для решения конкретных задач, так и обучение без учителя для исследования структуры данных и выделения скрытых паттернов.

4. Визуализация и интерпретируемость

Одной из проблем глубокого обучения является его «черный ящик» характер, когда модели могут давать точные результаты, но не всегда понятно, как они пришли к этим выводам. Для анализа больших данных важно разрабатывать методы интерпретации и визуализации результатов работы моделей. Применение технологий объяснимого ИИ (XAI) позволяет обеспечить прозрачность решений, принятых моделями глубокого обучения [9].

Перспективы развития применения глубокого обучения для анализа больших данных

Глубокое обучение продолжает развиваться и расширять свои возможности для анализа больших данных, открывая новые горизонты в самых различных областях. С каждым годом методы глубокого обучения становятся все более мощными и точными, что позволяет значительно улучшить процесс анализа данных. Рассмотрим несколько ключевых перспектив, которые могут повлиять на будущее применения глубокого обучения для анализа больших данных.

1. Интеграция с квантовыми вычислениями

Одной из самых захватывающих перспектив для глубокого обучения является использование квантовых вычислений. Квантовые компьютеры обладают уникальной способностью параллельно обрабатывать огромное количество данных, что делает их идеальными для задач, требующих огромных вычислительных мощностей. Ожидается, что квантовые вычисления смогут значительно ускорить процесс обучения глубоких нейронных сетей, особенно при работе с большими объемами данных. В будущем комбинация глубокого обучения и квантовых вычислений может привести к прорыву в таких областях, как искусственный интеллект, анализ больших данных, криптография и моделирование сложных систем [4].

2. Развитие автономных систем и Интернета вещей (IoT)

С развитием Интернета вещей (IoT) и автономных систем, таких как самоуправляемые автомобили и роботы, будет увеличиваться объем данных, поступающих в режиме реального времени. Глубокое обучение станет необходимым инструментом для обработки и анализа таких данных. В перспективе, системы, использующие глубокое обучение, смогут анализировать огромные потоки информации с минимальной задержкой, что позволит принимать быстрые и эффективные решения. Например, в автономных автомобилях глубокое обучение уже используется для обработки данных с камер и датчиков, и с развитием этой технологии такие системы будут становиться все более точными и безопасными.

3. Объяснимые искусственные интеллекты (XAI)

Одной из главных проблем, с которой сталкивается глубокое обучение, является его "черный ящик", то есть отсутствие объяснимости и прозрачности в принятии решений. С развитием объяснимого искусственного интеллекта (XAI), ученые и разработчики работают над созданием моделей, которые могут не только давать точные предсказания, но и объяснять, как были получены эти результаты. Это особенно важно в таких сферах, как медицина, финансы и юриспруденция, где требуется высокая степень доверия к результатам и понимание причин принятия решений. В будущем мы можем ожидать появления более интерпретируемых моделей глубокого обучения, что откроет новые возможности для их применения в критически важных отраслях.

ISSN 2706-6266
ISSN 3007-1062
ISSN 3007-1070

4. Автоматизация аналитики данных и улучшение обработки неполных данных

Современные методы глубокого обучения показывают отличные результаты при работе с большими и сложными данными, но все еще остаются проблемы с обработкой неполных и шумных данных. В будущем можно ожидать улучшений в области обработки неполных данных и самообучения моделей, что позволит создавать более устойчивые системы, способные эффективно работать даже с недостаточными или искаженными данными. Это особенно важно для многих реальных приложений, таких как анализ медицинских изображений или данные сенсоров, где неполные или ошибочные данные могут возникать часто.

5. Использование мультимодальных данных

Одной из тенденций в развитии глубокого обучения является использование мультимодальных данных, то есть данных, полученных из различных источников, таких как текст, изображения, видео и аудио [9]. Например, системы глубокого обучения могут сочетать информацию из нескольких датчиков, текстовых данных и визуальных изображений, что позволяет улучшить точность и качество предсказаний. Развитие мультимодальных технологий откроет новые возможности для применения искусственного интеллекта в таких областях, как безопасность, медицина, маркетинг и виртуальные ассистенты, которые смогут учитывать более широкий спектр данных для принятия решений.

6. Уменьшение потребности в вычислительных ресурсах

Одним из ограничений в применении глубокого обучения является высокая потребность в вычислительных мощностях, что требует значительных затрат на инфраструктуру. В перспективе ожидается создание более эффективных алгоритмов, которые смогут выполнять обучение и инференс с меньшими затратами ресурсов. Это могут быть новые архитектуры нейронных сетей, более легкие и эффективные методы обучения, а также использование специализированных аппаратных решений, таких как нейропроцессоры и асика. Эти достижения позволят значительно снизить затраты на обучение и развертывание моделей глубокого обучения, что сделает технологии доступными для более широкого круга пользователей и организаций.

7. Использование моделей с меньшими данными (Few-shot learning и Zero-shot learning)

Еще одной интересной перспективой является развитие методов обучения на малом объеме данных (Few-shot learning) и обучения без примеров (Zero-shot learning) [10]. В этих подходах модели глубокого обучения способны обучаться и делать предсказания, даже когда количество обучающих примеров ограничено или полностью отсутствует. Эти методы особенно важны для задач, где сбор и размечивание данных могут быть дорогими или трудоемкими, например, в медицинских исследованиях или при анализе редких событий. Развитие этих методов позволит значительно расширить область применения глубокого обучения в таких сферах, как медицина, право, экологический мониторинг и др.

8. Этические и правовые аспекты применения глубокого обучения

С развитием технологий глубокого обучения возрастает и значимость решения этических и правовых вопросов, связанных с использованием этих технологий. В будущем можно ожидать более четкого регулирования и стандартизации использования искусственного интеллекта, особенно в таких чувствительных областях, как здравоохранение, финансы и правосудие. Разработка этичных стандартов и правил для использования данных, обеспечения конфиденциальности и предотвращения

ISSN 2706-6266
ISSN 3007-1062
ISSN 3007-1070

дискриминации и предвзятости в моделях глубокого обучения будет играть важную роль в их широком внедрении в общественные и коммерческие сферы.

Заключение. Глубокое обучение стало мощным инструментом для анализа больших данных, значительно улучшив процессы обработки и извлечения полезной информации из огромных объемов данных. Методы глубокого обучения, включая искусственные нейронные сети, свёрточные и рекуррентные сети, генеративные состязательные сети и автокодировщики, продемонстрировали свою эффективность в решении множества задач в таких областях, как обработка изображений, текста, временных рядов и многом другом [1,3,5].

Несмотря на значительные достижения, глубокое обучение сталкивается с рядом вызовов, включая потребность в вычислительных мощностях, сложность интерпретации результатов и этические вопросы, связанные с использованием данных. Перспективы дальнейшего развития технологий глубокого обучения включают создание более мощных и эффективных моделей, интеграцию с квантовыми вычислениями и совершенствование методов работы с ограниченными объемами данных [4,10].

В будущем глубокое обучение продолжит играть ключевую роль в развитии науки и технологий, предлагая новые возможности для решения сложных задач в различных областях. Однако для достижения максимального потенциала этих технологий важно учитывать как технические, так и этические аспекты их применения, чтобы обеспечить их эффективное и безопасное использование в различных сферах.

СПИСОК ЛИТЕРАТУРЫ

1. Chollet, F. Deep Learning with Python. Manning Publications, 352 p. (2018).
2. Zhou, Z.-H. Machine Learning. Springer, 650 p. (2021).
3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. A., Kaiser, Ł., & Polosukhin, I Attention is all you need. In Advances in neural information processing systems (Vol. 30). Neural Information Processing Systems (NIPS). (2017).
4. Zhou, P., & Liang, Y. Reinforcement learning with deep neural networks for large scale data analysis. Springer Series in Advanced Data Science and Computing, 1-240. Springer, https://doi.org/10.1007/978-3-030-16792-0. (2019).
5. Левицкий, И. В., & Петров, А. В. Глубокое обучение: Теория и практика. [Текст] Наука, (2020).
6. Михайлов, В. В., & Петров, С. Ю. Применение нейронных сетей в анализе больших данных. [Текст] Вестник Санкт-Петербургского университета. Прикладная математика и информатика, 15(2), (2018).
7. Андреев, Д. М., & Никитин, И. В. Методы машинного обучения и искусственного интеллекта для анализа больших данных. [Текст] Издательство МГУ, (2019).
8. Козлов, А. Н. Использование нейронных сетей для анализа больших данных в реальном времени. [Текст] Математическое моделирование, 29(8), (2017).
9. Широков, В. С., & Ковалев, В. П. Проблемы и перспективы применения глубокого обучения в обработке больших данных. [Текст] Информационные технологии и вычислительные системы, 29(1), (2021).
10. Барков, С. И. Технологии глубокого обучения в анализе изображений и видео. [Текст] Компьютерные науки и информационные технологии, 10(4), (2020).

ISSN 2706-6266
ISSN 3007-1062
ISSN 3007-1070

REFERENCES

1. Chollet, F. Deep Learning with Python. Manning Publications, 352 p. (2018).
2. Zhou, Z.-H. Machine Learning. Springer, 650 p. (2021).
3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. A., Kaiser, Ł., & Polosukhin, I Attention is all you need. In Advances in neural information processing systems (Vol. 30). Neural Information Processing Systems (NIPS). (2017).
4. Zhou, P., & Liang, Y. Reinforcement learning with deep neural networks for large scale data analysis. Springer Series in Advanced Data Science and Computing, 1-240. Springer, https://doi.org/10.1007/978-3-030-16792-0. (2019).
5. Levickij, I. V., & Petrov, A. V. Glubokoe obuchenie: Teoriya i praktika. [Deep learning: Theory and Practice.] Nauka, (2020): - (In Rus)
6. Mihajlov, V. V., & Petrov, S. Yu. Primenenie nejronnyh setej v analize bol'shih dannyh. [Application of neural networks in big data analysis.] Vestnik Sankt-Peterburgskogo universiteta. Prikladnaya matematika i informatika, 15(2), (2018): - (In Rus)
7. Andreev, D. M., & Nikitin, I. V. Metody mashinnogo obucheniya i iskusstvennogo intellekta dlya analiza bol'shih dannyh. [Machine learning and artificial intelligence methods for big data analysis.] Izdatel'stvo MGU, (2019): - (In Rus)
8. Kozlov, A. N. Ispol'zovanie nejronnyh setej dlya analiza bol'shih dannyh v real'nom vremeni. [Using neural networks to analyze big data in real time.] Matematicheskoe modelirovanie, 29(8), (2017): - (In Rus)
9. Shirokov, V. S., & Kovalev, V. P. Problemy i perspektivy primeneniya glubokogo obucheniya v obrabotke bol'shih dannyh. [Problems and prospects of applying deep learning in big data processing.] Informacionnye tekhnologii i vychislitel'nye sistemy, 29(1), (2021): - (In Rus)
10. Barkov, S. I. Tekhnologii glubokogo obucheniya v analize izobrazhenij i video. [Deep learning technologies in image and video analysis.] Komp'yuternye nauki i informacionnye tekhnologii, 10(4), (2020): - (In Rus)

ҮЛКЕН ДЕРЕКТЕРДІ ТАЛДАУДА ТЕРЕҢ ОҚЫТУДЫ ҚОЛДАНУ: ӘДІСТЕР, ТӘСІЛДЕР ЖӘНЕ БОЛАШАҚ ПЕРСПЕКТИВАЛАРЫ

Аңдатпа: Соңғы онжылдықтарда үлкен деректерді талдау ақпараттық технологиялар саласындағы маңызды тапсырмалардың біріне айналды. Терең оқыту, жасанды интеллектінің бір бөлігі ретінде, үлкен деректер көлемінен білім алу және өңдеу үшін тиімді құралдар ұсынды. Бұл мақалада жасанды нейрондық желілер, конволюциялық нейрондық желілер, рекурренттік нейрондық желілер, генеративті қарама-қарсы желілер және автокодерлер сияқты терең оқытудың негізгі әдістері қарастырылады, олар әртүрлі деректер түрлерін талдауда белсенді қолданылады. Деректерді алдын ала өңдеу, таратылған есептеулер және бақыланатын және бақыланбайтын оқыту әдістері сияқты олардың қолданылу тәсілдері сипатталады. Сондай-ақ, модельдерді түсіндіру және мультимодальды деректерді дамыту сияқты өзекті мәселелер талқыланады. Мақалада терең оқытуды кванттық есептеулермен біріктіру және шектеулі деректермен жұмыс істеу әдістері де қарастырылады. Табыстарға қарамастан, терең оқыту үлкен есептеу қуатын қажет ету және деректерді пайдалану мәселелері сияқты бірқатар қиындықтарға тап болады. Қорытындысында терең оқытудың үлкен деректерді талдаудың күрделі тапсырмаларын шешу үшін үлкен әлеуетке ие екендігі және оның дамуы ғылым мен өнеркәсіптің әртүрлі салаларында жаңа мүмкіндіктер аша беретіні атап өтілген.

ISSN 2706-6266
ISSN 3007-1062
ISSN 3007-1070

Кілт сөздер: терең оқыту, үлкен деректер, нейрондық желілер, конволюциялық желілер, рекурренттік желілер, генеративті қарама-қарсы желілер, автокодерлер, деректерді талдау, таратылған есептеулер, бақыланатын оқыту, бақыланбайтын оқыту, модельдерді түсіндіру, мультимодальды деректер, кванттық есептеулер, деректердің этикасы.

APPLICATION OF DEEP LEARNING FOR BIG DATA ANALYSIS: METHODS, APPROACHES, AND PROSPECTS

Abstract: In recent decades, big data analysis has become one of the most important tasks in the field of information technology. Deep learning, as part of artificial intelligence, has provided effective tools for processing and extracting knowledge from large volumes of data. This article examines key deep learning methods such as artificial neural networks, convolutional neural networks, recurrent neural networks, generative adversarial networks, and autoencoders, which are actively used for analyzing various types of data. It describes approaches to their application, including data preprocessing, distributed computing, and supervised and unsupervised learning methods. The article also discusses current issues such as model interpretability and the development of multimodal data. Furthermore, the prospects of combining deep learning with quantum computing and methods for working with limited amounts of data are analyzed. Despite the successes, deep learning faces several challenges, such as the need for enormous computational resources and ethical issues related to data usage. In conclusion, it is noted that deep learning holds significant potential for solving complex big data analysis tasks, and its development will continue to open new opportunities for various fields of science and industry.

Keywords: deep learning, big data, neural networks, convolutional networks, recurrent networks, generative adversarial networks, autoencoders, data analysis, distributed computing, supervised learning, unsupervised learning, model interpretability, multimodal data, quantum computing, data ethics.