

СРАВНЕНИЕ ПОДХОДОВ К ОБУЧЕНИЮ МОДЕЛЕЙ НА НЕСТРУКТУРИРОВАННЫХ ДАННЫХ

Е.С. Ильин¹

¹ФГБОУ ВО «Воронежский государственный лесотехнический университет
имени Г.Ф. Морозова»

Аннотация. В статье рассмотрены и сравнены основные подходы к обучению моделей машинного и глубокого обучения на неструктурированных данных: контролируемое, неконтролируемое, самообучение и перенос обучения. Анализ проводится с точки зрения требований к объёму и качеству разметки, поддержки мультимодальных данных, вычислительной сложности и практической применимости в информационных системах. Показано, что самообучение и перенос обучения с использованием крупных предварительно обученных моделей позволяют эффективно использовать большие массивы неразмеченных данных и снижать затраты на разметку при сохранении высокого качества решений.

Ключевые слова: неструктурированные данные; машинное обучение; глубокое обучение; контролируемое обучение; неконтролируемое обучение; самообучение; self-supervised learning; перенос обучения; крупные языковые модели; мультимодальные модели; информационные системы.

COMPARISON OF APPROACHES TO MODEL TRAINING ON UNSTRUCTURED DATA

E.S. Ilyin¹

¹Voronezh State University of Forestry and Technologies named after G.F. Morozov

Abstract. The paper examines and compares the main approaches to training machine learning and deep learning models on unstructured data, including supervised, unsupervised, self-supervised learning and transfer learning. The analysis focuses on the required amount and quality of labeled data, support for multimodal inputs, computational cost, and practical applicability in information systems. The study shows that self-supervised learning and transfer learning with large pretrained models make it possible to effectively exploit large volumes of unlabeled data and reduce annotation costs while maintaining high solution quality.

Keywords: unstructured data; machine learning; deep learning; supervised learning; unsupervised learning; self-supervised learning; transfer learning; large language models; multimodal models; information systems

Введение

Сегодняшние информационные системы постоянно производят и копируют огромные объёмы неструктурированных данных: тексты, изображения, аудио- и видеозаписи, журналы событий, медицинские записи, сообщения в социальных сетях и многое другое. Большая часть корпоративной и веб-информации относится именно к этому типу, поэтому неструктурированные данные становятся ключевым объектом для аналитики и поддержки принятия решений. В отличие от структурированных наборов, такие данные лишены заранее заданной схемы, сильно различаются по форме и содержанию и характеризуются высокой размерностью, что значительно осложняет их использование в классических системах хранения и анализа.

На фоне этих вызовов за период 2020–2025 гг. заметно продвинулось развитие подходов к обучению моделей машинного и глубокого обучения, ориентированных на работу с неструктурированной информацией. В центре внимания находятся классическое контролируемое обучение (supervised learning), неконтролируемое обучение (unsupervised learning), различные варианты самообучения (self-supervised learning), а также перенос обучения и дообучение крупных предварительно обученных моделей (transfer learning, fine-tuning), включая языковые и мультимодальные архитектуры. Каждый из этих подходов предлагает свой баланс между требуемым объёмом разметки, вычислительными затратами, качеством обобщения и способностью работать с разнородными, в том числе мультимодальными, данными.

Цель данной работы состоит в том, чтобы сопоставить основные подходы к обучению моделей на неструктурированных данных, показать их сильные и слабые стороны и продемонстрировать, зачем разработчику или исследователю важно осознанно выбирать стратегию обучения. Теоретическая значимость работы заключается в том, что предлагается упорядоченный взгляд на современные стратегии обучения и критерии их сравнения. Практическая значимость связана с возможностью опираться на полученные выводы при проектировании реальных информационных систем и аналитических решений.

Обзор литературы

В современной литературе неструктурированные данные рассматриваются как один из главных факторов, усложняющих внедрение аналитики больших данных и систем искусственного интеллекта в прикладных областях. Для работы с текстовой информацией центральное место заняли трансформеры и крупные языковые модели (LLM), которые показали высокую эффективность в задачах классификации, извлечения сущностей, суммаризации, построения диалоговых систем и извлечения фактов из свободного текста. Принципиальное преимущество таких моделей в том, что они позволяют опираться на масштабное предварительное обучение и лишь затем дообучаться на специализированных, часто весьма компактных, размеченных выборках.

При работе с изображениями и мультимедиа основные успехи связаны с глубокими сверточными и гибридными архитектурами, а также с методами самообучения и контрастивного обучения, которые используют большие коллекции неразмеченных данных. Обзорные исследования показывают, что self-supervised и contrastive learning способны формировать устойчивые и переносимые представления, нередко превосходящие по качеству классические supervised-решения, особенно в условиях дефицита маркированных примеров. Параллельно активно развивается направление мультимодального обучения, где в единой модели совместно обрабатываются текст, изображение, аудио и табличные данные; это особенно важно для интеграции структурированных и неструктурированных источников в единую аналитическую среду.

Отдельный пласт исследований посвящён применению крупных моделей для извлечения структурированной информации из неструктурированных медицинских записей, юридических документов, финансовых отчётов и других сложных текстов.

Публикации в области цифровой медицины и корпоративной аналитики демонстрируют, что дообученные LLM и мультимодальные модели способны заметно сократить объём ручной работы, обеспечивая при этом высокое качество извлечения ключевых параметров и признаков. При этом авторы подчёркивают, что выбор конкретного подхода (контролируемое обучение, неконтролируемое обучение, самообучение или перенос обучения) должен опираться на доступность разметки, требования к интерпретируемости и ограничения по ресурсам.

Объект и предмет исследования

В рамках данной работы в качестве объекта рассматриваются подходы к обучению моделей машинного и глубокого обучения, ориентированных на обработку неструктурированных данных: текстов, изображений и их мультимодальных комбинаций. Предметом исследования являются сравнительные характеристики этих подходов с точки зрения:

- объёма и качества необходимой разметки;
- поддержки чисто неструктурированных и мультимодальных данных;
- вычислительной стоимости обучения и внедрения;
- способности к обобщению и устойчивости к смещению данных;
- практической применимости в реальных информационных системах.

Классификация подходов

Для упорядочения анализа выделим четыре ключевых класса подходов:

1. **Контролируемое обучение (supervised learning).** Модель обучается по размеченной выборке и нацелена на решение конкретной задачи, например классификации текстов, изображений, анализа настроений или медицинской диагностики. Такой подход традиционно даёт высокую точность, но сильно зависит от объёма и качества разметки.

2. **Неконтролируемое обучение (unsupervised learning).** Методы кластеризации, тематического моделирования и обучения представлений без меток применяются для первичного структурирования больших корпусов документов, поиска скрытых закономерностей и снижения размерности. Они полезны на этапе разведочного анализа, но сами по себе редко решают прикладные задачи «от начала до конца».

3. **Самообучение (self-supervised, в том числе контрастивное).** Модель учится на неразмеченных данных, решая вспомогательные, искусственно сформулированные задачи (восстановление маскированных фрагментов, предсказание контекста, сопоставление разных представлений одного объекта и т.п.). Такие методы позволяют задействовать большие объёмы сырых данных и получать представления, пригодные для последующего дообучения.

4. **Перенос обучения и fine-tuning.** В этом случае используется уже предварительно обученная крупная модель (например, языковая или мультимодальная). Её параметры адаптируются под конкретную задачу и домен с помощью относительно небольшого набора размеченных примеров. Это направление

особенно актуально в связи с широким распространением LLM и крупных мультимодальных архитектур.

Критерии сравнения и представление результатов

Сравнение рассматриваемых подходов выполняется по нескольким осям:

- объём и тип требуемой разметки (полная, частичная, отсутствие разметки);
- поддержка мультимодальных данных;
- типичная вычислительная сложность;
- основные области применения в информационных системах;
- зрелость инструментальной экосистемы (доступность библиотек, готовых моделей и т.п.).

Сравнительный анализ подходов

В сводной таблице ниже в сжатом виде показаны ключевые характеристики каждого подхода.

Контролируемое обучение остаётся естественным выбором, когда есть достаточно большая и качественно размеченная выборка. В таких условиях оно даёт хорошие результаты и позволяет строить точные модели под конкретную задачу. Однако в случае неструктурированных данных ключевой проблемой становится стоимость разметки: требуется участие квалифицированных специалистов, причём объём данных может исчисляться миллионами экземпляров. Это резко снижает привлекательность чисто supervised-подходов в ряде доменов.

Неконтролируемые методы удобны для первичного ориентирования в больших массивах неструктурированной информации. С их помощью можно выделить группы схожих документов, темы, латентные факторы, что помогает понять структуру данных и сформировать гипотезы. Но сами по себе такие методы редко дают готовое решение прикладной задачи, особенно если требуется точное предсказание или строгая интерпретируемость. Обычно unsupervised-результаты используют как основу для последующего контролируемого или самообучения.

Самообучение и контрастивные подходы за последние годы стали одним из центральных направлений в обработке изображений, текста и мультимодальных данных. Это объясняется тем, что в реальных проектах гораздо проще собрать большой массив сырых данных, чем обеспечить для него качественную разметку. Self-supervised-подходы позволяют использовать такие массивы для обучения универсальных представлений, которые затем легко адаптировать к

конкретной задаче с помощью сравнительно небольшого размеченного поднабора. На практике это даёт выигрыш как в качестве, так и в устойчивости моделей к изменениям распределения данных, но требует серьёзных вычислительных ресурсов и продуманной методики обучения.

Таблица 1 – Сравнение подходов к обучению моделей на неструктурированных данных

Подход	Требования к разметке	Работа с мультимодальностью	Вычислительная сложность	Типичные области применения
Контролируемое обучение	Нужен значительный объём размеченных данных, качество меток критично	Возможна при наличии размеченных мультимодальных корпусов	Средняя/высокая, зависит от архитектуры	Классификация текстов и изображений, диагностика, анализ настроений
Неконтролируемое обучение	Разметка не требуется, интерпретация кластеров выполняется постфактум	Обычно применяется к отдельным модальностям	Средняя/высокая при больших коллекциях	Тематическое моделирование, кластеризация документов, разведочный анализ
Самообучение (self-supervised)	Разметка не нужна, используются искусственно заданные предтекстовые задачи	Эффективно для текста и изображений, в том числе мультимодальных	Высокая, требуются серьёзные вычислительные ресурсы	Предобучение представлений, база для последующего supervised-дообучения
Перенос обучения, fine-tuning	Используются предварительно обученные модели, достаточно ограниченной разметки	Широко применяется в языковых и мультимодальных моделях	Высокая при предобучении, умеренная при дообучении	Извлечение данных из документов, анализ медицинских записей, корпоративная аналитика

Перенос обучения и fine-tuning крупных предварительно обученных моделей, в том числе языковых и мультимодальных, во многом стали ответом на потребности прикладных проектов. Широкая доступность готовых LLM и открытых моделей компьютерного зрения позволяет разработчикам сосредоточиться на адаптации уже существующих моделей к своим задачам, а не на обучении с нуля. В медицине, праве и корпоративной аналитике показано, что дообучение

таких моделей на относительно небольшой выборке доменных данных нередко позволяет достичь или превзойти качество специализированных решений при существенно меньших затратах на разметку и инфраструктуру. В то же время возрастает потребность в контроле качества, интерпретируемости и соблюдении требований безопасности и конфиденциальности.

Теоретическая и практическая значимость

С теоретической точки зрения сравнение рассмотренных подходов помогает более чётко понимать их место и роль в экосистеме методов анализа неструктурированных данных. Предложенный набор критериев – от требований к разметке до поддержки мультимодальности и вычислительной сложности – может служить основой для оценки новых методов и гибридных схем, которые всё активнее появляются в литературе. Кроме того, становится проще формулировать исследования, направленные на сочетание сильных сторон разных парадигм, например объединение self-supervised-предобучения, unsupervised-кластеризации и лёгкого supervised-дообучения.

Практическая значимость работы выражается в том, что её выводы дают разработчикам и исследователям понятные ориентиры при выборе стратегии обучения для конкретного проекта. Понимая, какие ресурсы доступны (разметка, вычислительная инфраструктура, экспертиза), какие данные предстоит обрабатывать (одна или несколько модальностей) и какие требования предъявляются к качеству и интерпретируемости, можно более осознанно выбирать между контролируемым обучением, неконтролируемыми методами, самообучением и переносом обучения. В ряде случаев оптимальным оказывается не один подход, а их сочетание, что тоже отражено в проведённом сравнении.

Вывод

Проведённый анализ показывает, что при работе с неструктурированными данными невозможно выделить единственный «лучший» подход к обучению моделей. Контролируемое обучение обеспечивает высокую точность, но плохо масштабируется в условиях дефицита разметки и быстро меняющихся доменов. Неконтролируемые методы ценны для разведочного анализа и структурирования данных, однако сами по себе чаще выступают как вспомогательный инструмент, а не как завершённое решение.

Самообучение и контрастивные подходы позволяют использовать большие объёмы сырых данных и получать универсальные представления, что делает

их особенно привлекательными в современных задачах анализа текстов, изображений и мультимодальных источников. Перенос обучения и fine-tuning крупных предварительно обученных моделей, в свою очередь, дают практикам мощный инструмент, позволяющий быстро и относительно недорого внедрять интеллектуальные компоненты в информационные системы, не начиная каждый раз «с нуля».

Теоретическая значимость работы заключается в систематизации ролей основных парадигм обучения в контексте неструктурированных данных и в предложении удобных критериев их сравнения. Практическая значимость связана с тем, что сформулированные выводы и акценты можно использовать как методические рекомендации при выборе подхода для конкретного проекта – от анализа медицинских записей до мониторинга социальных медиа и обработки корпоративных документов. В качестве перспективных направлений дальнейших исследований можно выделить разработку более эффективных и экономичных процедур самообучения, глубокую интеграцию структурированных и неструктурированных источников в единую мультимодальную архитектуру, а также методы повышения интерпретируемости и надёжности моделей в критически важных областях.

Список литературы

1. Роннебергер О. U-Net: Свёрточные сети для биомедицинской сегментации изображений // Медицинская компьютерная томография и компьютерное вмешательство (MICCAI). 2015. URL: <https://lmb.informatik.uni-freiburg.de/people/ronneber/u-net/> (дата обращения: 13.02.2025).
2. Фюлоп А., Хорват А. Сквозное обучение глубоких нейронных сетей в частотной области // Mathematics. 2022. Т. 10. С. 2132. DOI: 10.3390/math10132132.
3. Лу Х., Хуан Я., Чжан Ч. и др. Достижения в глубоком обучении для микроскопии сверхвысокого разрешения (приглашённый доклад) // Laser & Optoelectronics Progress. 2024. Т. 61, № 16. Ст. 1611002. DOI: 10.3788/LOP6124161002.
4. Машинное обучение для физических наук: воркшоп на NeurIPS 2022 [ML for Physical Sciences Workshop at NeurIPS 2022]. URL: https://ml4physicalsciences.github.io/2022/files/NeurIPS_ML4PS_2022_92.pdf (дата обращения: 13.02.2025).

5. Рудской А.И., Волков К.Н., Соколов Ю.А., Кондратьев С.Ю. Цифровые производственные системы: технологии, моделирование, оптимизация. СПб.: СПбПУ, 2020. 828 с.
6. Маркидис С. Старое и новое: Могут ли физически информированные методы глубокого обучения заменить традиционные линейные решатели? // Mathematics. 2022. Т. 10. С. 603. DOI: 10.3390/math9100603.
7. NVIDIA Solutions. URL: <https://www.macnica.co.jp/en/business/semiconductor/manufacturers/nvidia/products/134046/> (дата обращения: 13.02.2025).
8. Сизая А. Моделирование теплопередачи и напряжений при лазерной сварке с использованием пакета FENICS // Английский язык в сфере профессиональных коммуникаций: Сб. докл. VII Всерос. молодежной науч. конф. Казань, 2021. С. 249-251.
9. Appercase News. URL: <https://www.appercase.ru/news/27220/> (дата обращения: 13.02.2025).
10. Assured Systems. URL: <https://www.assured-systems.com/nvidia-jetson-agx-xavier-for-new-era-of-ai-in-robotics/> (дата обращения: 13.02.2025).
11. Sigma Labs News. URL: <https://3dprintingindustry.com/news/sigma-labs-printrite3d-platform-made-compatible-with-polymer-3d-printing-201503/> (дата обращения: 13.02.2025).
12. Николенко, С. Д. Моделирование возникновения внутренних напряжений в сложной структуре материала / С. Д. Николенко, С. П. Козодаев, С. А. Сазонова // Моделирование систем и процессов. – 2024. – Т. 17, № 2. – С. 50-61. – DOI 10.12737/2219-0767-2024-17-2-50-61.
13. Сазонова, С. А. Моделирование процесса диагностики утечек на основе двухальтернативной гипотезы с учетом помех от стохастичности потребления в гидравлической системе / С. А. Сазонова, И. В. Щербакова, Г. И. Сметанкина // Моделирование систем и процессов. – 2024. – Т. 17, № 1. – С. 111-120.
14. Регрессионный анализ оценки загрузки платного отделения стоматологической поликлиники / Т. П. Новикова, А. А. Бодин, А. И. Заревич // Научно-технический вестник Поволжья. – 2023. – № 3. – С. 77–80.
15. Математическая модель индуктивного делителя напряжения с электронной компенсацией / А. И. Заревич, С. В. Муравьев // Известия Томского политехнического университета. – 2010. – Т. 317, № 4. – С. 129–133.
16. Approximation of the absorption spectrum of indium phosphide in the context of simulation of the process of sensation / Ph. V. Makarenko, V. K. Zolnikov, A.

I. Zarevich [et al.] // Russian Microelectronics. – 2024. – Vol. 53, № 4. – P. 329–338. – DOI 10.1134/S1063739724040041.

17. Формирование ИКТ-компетентности студентов специальности «Информатика» / А. В. Полуэктов // Информационно-коммуникационные технологии в педагогическом образовании. – 2009. – № 3 (3). – С. 8–11. – EDN SAXGJX.

References

1. Ronneberger O. U-Net: Convolutional Networks for Biomedical Image Segmentation // Medical Image Computing and Computer-Assisted Intervention (MICCAI). 2015. URL: <https://lmb.informatik.uni-freiburg.de/people/ronneber/u-net/> (accessed: 13.02.2025).

2. Fülöp A., Horváth A. End-to-End Training of Deep Neural Networks in the Fourier Domain // Mathematics. 2022. Vol. 10. P. 2132. DOI: 10.3390/math10132132.

3. Lu X., Huang Y., Zhang Z. et al. Advances in Deep Learning for Super-Resolution Microscopy (Invited) // Laser & Optoelectronics Progress. 2024. Vol. 61, No. 16. Art. 1611002. DOI: 10.3788/LOP6124161002.

4. Machine Learning for Physical Sciences: Workshop at NeurIPS 2022 [ML for Physical Sciences Workshop at NeurIPS 2022]. URL: https://ml4physicalsciences.github.io/2022/files/NeurIPS_ML4PS_2022_92.pdf (accessed: 13.02.2025).

5. Rudskoy A.I., Volkov K.N., Sokolov Y.A., Kondratyev S.Y. Digital Manufacturing Systems: Technologies, Modeling, Optimization. St. Petersburg: SPbPU, 2020. 828 p.

6. Markidis S. The Old and the New: Can Physics-Informed Deep Learning Replace Traditional Linear Solvers? // Mathematics. 2022. Vol. 10. P. 603. DOI: 10.3390/math9100603.

7. NVIDIA Solutions. URL: <https://www.macnica.co.jp/en/business/semiconductor/manufacturers/nvidia/products/134046/> (accessed: 13.02.2025).

8. Sizaya A. Simulation of Heat Transfer and Stresses in Laser Welding Using the FENICS Package // English in Professional Communication: Proc. VII All-Russian Youth Sci. Conf. Kazan, 2021. P. 249-251.

9. Appercase News. URL: <https://www.appercase.ru/news/27220/> (accessed: 13.02.2025).

10. Assured Systems. URL: <https://www.assured-systems.com/nvidia-jetson-agx-xavier-for-new-era-of-ai-in-robotics/> (accessed: 13.02.2025).

11. Sigma Labs News. URL: <https://3dprintingindustry.com/news/sigma-labs-printrite3d-platform-made-compatible-with-polymer-3d-printing-201503/> (accessed: 13.02.2025).
12. Nikolenko, S.D., Kozodaev, S.P., Sazonova, S.A. Modeling of Internal Stresses Formation in Complex Material Structures // Modeling of Systems and Processes. 2024. Vol. 17, No. 2. Pp. 50-61. DOI: 10.12737/2219-0767-2024-17-2-50-61.
13. Sazonova, S.A., Shcherbakova, I.V., Smetankina, G.I. Modeling of Leakage Diagnostics Process Based on Binary Hypothesis with Consideration of Stochastic Consumption Interference in Hydraulic Systems // Modeling of Systems and Processes. 2024. Vol. 17, No. 1. Pp. 111-120. DOI: 10.12737/2219-0767-2024-17-1-111-120.
14. Regression analysis of the assessment of the workload of the paid department of the stomatological polyclinic / T. P. Novikova, A. A. Bodin, A. I. Zarevich // Scientific and Technical Bulletin of the Volga region. – 2023. – No. 3. – pp. 77-80.
15. Mathematical model of an inductive voltage divider with electronic compensation / A. I. Zarevich, S. V. Muravyev // Proceedings of Tomsk Polytechnic University. - 2010. – Vol. 317, No. 4. – pp. 129-133.
16. Approximation of the absorption spectrum of indium phosphide in the context of simulation of the process of sensation / Ph. V. Makarenko, V. K. Zolnikov, A. I. Zarevich [et al.] // Russian Microelectronics. – 2024. – Vol. 53, No. 4. – P. 329-338. – DOI 10.1134/S1063739724040041.
17. Formation of ICT competence of students of the specialty "Informatics" / A.V. Poluektov // Information and communication technologies in teacher education. – 2009. – № 3 (3). – Pp. 8-11. – EDN SAXGJX.