УДК: 004.89

Худайберидева Г. Б., магистр, ассистент кафедры
«Информатика и информационные технологии»
Московский Политехнический Университет,
Россия, г. Москва

Кожухов Д. А., магистр, ассистент кафедры
«Информатика и информационные технологии»
Московский Политехнический Университет,
Россия, г. Москва

Пименкова А. А., студент-бакалавр кафедры
«Информатика и информационные технологии»
Московский Политехнический Университет,
Россия, г. Москва

ЭКСТРЕМАЛЬНАЯ СПЕЦИАЛИЗАЦИЯ КРУПНЫХ ЯЗЫКОВЫХ МОДЕЛЕЙ НА ОСНОВЕ ОНТОЛОГИЧЕСКОЙ РЕЛЕВАНТНОСТИ ДЛЯ ПРОМЫШЛЕННЫХ ЗАДАЧ

***Аннотация**: Предлагается методология экстремальной компрессии крупных языковых моделей (LLM) посредством целевого удаления функциональных возможностей, не релевантных конкретной узкопрофильной промышленной задаче. В отличие от традиционных подходов к сжатию, ориентированных на сохранение общих способностей модели, данный подход фокусируется на идентификации и последующем устранении параметров и внутренних представлений, ответственных за обработку знаний, выходящих за пределы необходимой предметной области. Метод предполагает анализ семантической важности данных относительно целевой онтологии задачи (например, диагностика неисправностей станка на основе логов), применение структурированного прунинга и селективного замораживания модулей сети. Результатом является значительное уменьшение вычислительных требований и размера модели при сохранении требуемой специализированной функциональности.*

Данный подход обеспечивает практическую возможность внедрения LLM в ресурсоограниченные промышленные среды, требующие высокой эффективности и предсказуемости.

Ключевые слова: *большие языковые модели, компрессия моделей, экстремальная специализация, промышленное применение, онтологическая релевантность, структурированный прунинг, замораживание параметров, диагностика оборудования, анализ логов, эффективность вычислений, ресурсоограниченные среды.*

Khudaiberideva G. B.
master and department assistant at the department of
"Computer Science and Information Technology"
Moscow Polytechnic University
Moscow, Russia

Kozhukhov D. A.
master and department assistant at the department of
"Computer Science and Information Technology"
Moscow Polytechnic University
Moscow, Russia

Pimenkova A. A.
bachelor's student at the department of
"Computer Science and Information Technology"
Moscow Polytechnic University
Moscow, Russia

EXTREME SPECIALIZATION OF LARGE LANGUAGE MODELS BASED ON ONTOLOGICAL RELEVANCE FOR INDUSTRIAL TASKS

Annotation: *A methodology for extreme compression of large language models (LLM) is proposed by purposefully removing functionality that is not relevant to a specific narrow-*

profile industrial task. Unlike traditional compression approaches that focus on preserving the overall capabilities of the model, this approach focuses on identifying and subsequently eliminating the parameters and internal representations responsible for processing knowledge beyond the required domain. The method involves analyzing the semantic importance of data relative to the target ontology of the task (for example, fault diagnosis of a machine based on logs), the use of structured pruning and selective freezing of network modules. The result is a significant reduction in computational requirements and model size while maintaining the required specialized functionality. This approach provides a practical opportunity to implement LLM in resource-limited industrial environments that require high efficiency and predictability.

***Keywords**: large language models, model compression, extreme specialization, industrial application, ontological relevance, structured pruning, parameter freezing, equipment diagnostics, log analysis, computational efficiency, resource-limited environments.*

Введение

Широкое распространение крупных языковых моделей (LLM) [1] выявило существенный разрыв между их потенциальными возможностями и практическими требованиями промышленного внедрения. Основным препятствием выступают чрезвычайно высокие вычислительные и энергетические затраты, необходимые для функционирования данных моделей, особенно в условиях реального времени или на периферийных устройствах (edge computing) [10]. Традиционные методы компрессии LLM, такие как прунинг весов, квантование, дистилляция [2, 7, 8], направлены на общее уменьшение размера и сложности модели с сохранением как можно более широкого спектра её исходных способностей. Однако для многих специализированных промышленных сценариев, таких как автоматизированный анализ текстовых логов оборудования для диагностики сбоев, прогнозирование отказов на основе технических описаний или семантический поиск в базах знаний инженерной поддержки, обширные общие знания модели являются избыточными [12]. Требуется лишь узконаправленная функциональность,

строго соответствующая онтологии конкретной предметной области и решаемой задачи. Возникает гипотеза о возможности достижения существенно более высокой степени компрессии за счет целенаправленного удаления тех компонентов модели, которые отвечают за обработку информации, не имеющей отношения к целевой промышленной задаче [16]. Данный подход, обозначаемый как экстремальная специализация, предполагает переход от сохранения общих способностей к сохранению исключительно критически необходимых для конкретного use case.

Проблематика Общей Компрессии vs. Требования Промышленности

Стандартные подходы к сжатию LLM сталкиваются с фундаментальным ограничением при адаптации к узкоспециализированным промышленным задачам. Методы глобального прунинга стремятся удалить наименее значимые веса по всей сети, основываясь на эвристиках величины веса или его влияния на общую функцию потерь [2, 3]. Квантование снижает битность представления параметров, влияя на все слои модели равномерно или адаптивно [8]. Дистилляция передает знания большой модели (учителя) в меньшую (ученика), стремясь аппроксимировать общее поведение учителя [7]. Общей чертой этих методов является цель минимизировать деградацию производительности модели на широком наборе общих задач (например, GLUE, SuperGLUE) [9]. Однако в контексте промышленного применения, где модель должна решать одну четко определенную задачу в строго ограниченной предметной области, поддержание широких общих способностей становится неоправданной роскошью [10, 12]. Значительная часть вычислительных ресурсов тратится на обработку и поддержание внутренних представлений, абсолютно нерелевантных для, например, классификации кодов ошибок станка ЧПУ по их текстовым описаниям в логах. Эта "онтологическая избыточность" [16] представляет собой основной резерв для достижения радикального сокращения размера и

сложности модели применительно к специализированному сценарию использования.

Концепция Экстремальной Специализации на основе Онтологической Релевантности

Предлагаемый подход фундаментально отличается от традиционной компрессии [2, 7, 8, 16]. Его ядром является принцип функционального прунинга, направленного не на малозначимые веса в глобальном смысле, а на целенаправленное удаление возможностей модели, связанных с обработкой данных, лежащих вне целевой онтологии. Под "онтологической релевантностью" понимается соответствие знаний, фактов, концепций и языковых конструкций той узкой предметной области, которая необходима для решения конкретной промышленной задачи. Например, для системы диагностики по логам критичны знания о кодах ошибок, технических терминах, специфических последовательностях событий, номенклатуре компонентов оборудования и их взаимосвязях. Знания о литературе, истории, общей науке или даже о смежных, но не используемых в данной задаче инженерных дисциплинах являются нерелевантными [12]. Цель экстремальной специализации — идентифицировать параметры и структурные элементы LLM (нейроны, группы нейронов, слои внимания, целые слои) [4, 5, 13], ответственные за кодирование и манипулирование нерелевантными знаниями, и насильственно устранить или деактивировать их, оставив только минимально необходимый для целевой задачи функционал [16].

Методы Идентификации Нерелевантных Функциональных Возможностей

Ключевым этапом реализации экстремальной специализации является разработка надежных методов для идентификации частей модели, ответственных за нерелевантные знания. Один перспективный путь — анализ влияния на целевую задачу [3, 6, 9]. Используя специализированный датасет, строго соответствующий промышленной

задаче (например, аннотированные логи ошибок станков), можно применять методики, подобные вычислению градиентов по функции потерь задачи относительно активаций нейронов или выходов слоев. Нейроны или слои, демонстрирующие стабильно низкое абсолютное значение градиента или низкую вариативность активаций при обработке релевантных входных данных, могут рассматриваться как потенциальные кандидаты на удаление, так как их вклад в решение целевой задачи минимален [6, 9]. Другой подход основан на семантическом зондировании [13]. Создаются специализированные пробные наборы (probes), содержащие примеры, явно принадлежащие к релевантной онтологии (технические описания, коды ошибок) и к нерелевантным областям (общие новости, художественные тексты, описания из других отраслей). Анализ паттернов активации модели при обработке этих проб позволяет выявить специфические компоненты сети, избирательно реагирующие на нерелевантные входные данные [4, 5]. Третий метод предполагает анализ внутренних представлений [13]. Используя методы снижения размерности (такие как t-SNE, UMAP) или кластеризацию, можно визуализировать и проанализировать, как различные типы входных данных (релевантные и нерелевантные) проецируются во внутренние пространства активаций различных слоев или голов внимания модели. Области пространства представлений, преимущественно занятые проекциями нерелевантных данных, указывают на модули, ответственные за их обработку [5, 13]. Комбинация этих методов повышает надежность идентификации [16].

Техники Удаления и Деактивации Нерелевантных Компонентов

После идентификации компонентов, ассоциированных с нерелевантными функциональными возможностями, применяются методы их устранения. Наиболее радикальным является структурированный прунинг на уровне нейронов, групп нейронов (channels) или целых слоев [2, 3, 16]. В отличие от неструктурированного прунинга, удаляющего отдельные веса, структурированный подход удаляет целые структурные

единицы, что приводит к более значительному уменьшению размера модели и упрощению её архитектуры, а также обеспечивает лучшую аппаратную эффективность при инференсе [10, 16]. Решение об удалении принимается на основе метрик важности, полученных на этапе идентификации (например, средняя величина градиента, дисперсия активаций, вклад в кластеризацию нерелевантных данных) [3, 6, 9]. Более консервативной альтернативой является селективное замораживание (freezing). Параметры идентифицированных как нерелевантные модулей (например, определенных слоев трансформера или голов внимания) фиксируются, их веса не обновляются в процессе возможного последующего дообучения (fine-tuning) на целевом промышленном датасете [9, 16]. Это исключает вычислительные затраты на их обновление и может упростить архитектуру для вывода, хотя и не уменьшает физический размер модели [10]. Замораживание предпочтительнее, если существует гипотетическая, но маловероятная в рамках конкретной задачи, необходимость в сохранении удаляемых знаний. Применение данных техник должно сопровождаться валидацией на целевом датасете для контроля за сохранением требуемой функциональности [9, 16].

Ожидаемые Преимущества и Практическая Значимость

Основным ожидаемым преимуществом экстремальной специализации является достижение существенно более высоких степеней сжатия по сравнению с традиционными методами [2, 7, 8], применительно к узкой промышленной задаче [16]. Удаление значительных массивов параметров, ответственных за нерелевантные знания, напрямую ведет к уменьшению объема памяти, необходимого для хранения модели [10, 16]. Упрощение архитектуры сети (удаление целых слоев или блоков) сокращает количество операций, требуемых для вывода, что критически важно для развертывания в системах реального времени или на периферийных устройствах с ограниченными вычислительными ресурсами и энергопотреблением [10]. Уменьшение сложности модели также

потенциально снижает требования к пропускной способности памяти и задержкам, что повышает скорость отклика системы [10]. Помимо вычислительной эффективности, экстремальная специализация может способствовать повышению предсказуемости и надежности модели в рамках её узкой задачи [11]. Устранение компонентов, ответственных за обработку нерелевантной информации, теоретически снижает риск генерации нежелательных или неконтролируемых выходных данных (hallucinations), связанных с непредусмотренным использованием знаний из нецелевых областей [11, 12]. Фокусировка модели исключительно на релевантной онтологии упрощает её интерпретацию и валидацию для конкретного промышленного контекста [13, 16].

Проблемы и Ограничения Подхода

Несмотря на потенциальные преимущества, подход экстремальной специализации сопряжен с рядом существенных проблем и ограничений. Первичной проблемой является разработка точных и надежных методов идентификации нерелевантных компонентов [13, 16]. Современные LLM представляют собой сложные высокоинтерконнектированные системы, где знания распределены по сети, а не локализованы строго в отдельных модулях [4, 5, 13]. Существует риск ошибочного удаления компонентов, косвенно важных для целевой задачи, или неполного удаления нерелевантных, что приведет к неоптимальной компрессии или деградации качества [6, 9, 11]. Важным ограничением является узкая применимость результата [16]. Модель, подвергнутая экстремальной специализации для одной конкретной задачи (например, диагностика ошибок станка А), будет непригодна или крайне неэффективна для решения даже близкородственных задач (например, диагностика станка Б другой модели или прогнозирование износа того же станка А) [9]. Потеря общих способностей делает модель ригидной и неспособной к адаптации без полного пересмотра процесса специализации [12]. Процесс самой специализации (идентификация + удаление) требует вычислительных

ресурсов и наличия качественного, репрезентативного датасета для целевой задачи, что может быть затратно [16]. Существуют также фундаментальные вопросы, связанные с определением границ "релевантности" и потенциальным влиянием удаления, казалось бы, нерелевантных знаний на общую когерентность и связность генерируемых моделью выходных данных в рамках целевой онтологии [11, 12, 13].

Заключение

Экстремальная специализация крупных языковых моделей на основе онтологической релевантности представляет собой перспективный подход к радикальному сжатию LLM для узкоспециализированных промышленных применений. Смещение фокуса с сохранения общих способностей на хирургическое удаление функциональных возможностей, не критичных для конкретной задачи, открывает путь к достижению значительно более высоких степеней компрессии по сравнению с традиционными методами. Ключевыми элементами методологии являются разработка точных методов идентификации параметров и структурных компонентов модели, ответственных за обработку нерелевантных данных, и применение техник структурированного прунинга или селективного замораживания для их устранения. Ожидаемым результатом является модель с резко уменьшенными вычислительными требованиями и размером, сохраняющая при этом необходимую функциональность для целевого промышленного сценария, что критически важно для внедрения в ресурсоограниченные среды. Однако успешная реализация данного подхода требует преодоления существенных вызовов, связанных со сложностью точной идентификации распределенных знаний в LLM, риском потери косвенно важной функциональности и фундаментальной потерей адаптивности модели.

СПИСОК ЛИТЕРАТУРЫ:

1. Brown T., Mann B., Ryder N. et al. Language Models are Few-Shot Learners // Advances in Neural Information Processing Systems. 2020.

Vol. 33. P. 1877–1901. URL: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf DOI: 10.48550/arXiv.2005.14165

2. Han S., Mao H., Dally W.J. Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding // International Conference on Learning Representations (ICLR). 2016. URL: https://arxiv.org/abs/1510.00149
3. Frankle J., Carbin M. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks // International Conference on Learning Representations (ICLR). 2019. URL: https://arxiv.org/abs/1803.03635
4. Michel P., Levy O., Neubig G. Are Sixteen Heads Really Better than One? // Advances in Neural Information Processing Systems. 2019. Vol. 32. URL: https://proceedings.neurips.cc/paper/2019/file/2c601ad9d2ff9bc8b282670cdd54f69f-Paper.pdf
5. Voita E., Talbot D., Moiseev F. et al. Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019. P. 5797–5808. URL: https://aclanthology.org/P19-1580/. DOI: 10.18653/v1/P19-1580
6. Prasanna S., Rogers A., Rumshisky A. When BERT Plays the Lottery, All Tickets Are Winning // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020. P. 3208–3229. URL: https://aclanthology.org/2020.emnlp-main.259/ DOI: 10.18653/v1/2020.emnlp-main.259
7. Sanh V., Debut L., Chaumond J., Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter : препринт. 2019. arXiv:1910.01108 [cs.CL]. URL: https://arxiv.org/abs/1910.01108
8. Zafrir O., Boudoukh G., Izsak P., Wasserblat M. Q8BERT: Quantized 8Bit BERT // 2019 Fifth Workshop on Energy Efficient Machine

Learning and Cognitive Computing - NeurIPS Edition (EMC2-NIPS). 2019. P. 36–39. DOI: 10.1109/EMC2-NIPS53020.2019.00012

9. Gordon M.A., Duh K., Andrews N. Compressing BERT: Studying the Effects of Weight Pruning on Transfer Learning // Proceedings of the 5th Workshop on Representation Learning for NLP. 2020. P. 173–180. URL: https://aclanthology.org/2020.repl4nlp-1.20/ DOI: 10.18653/v1/2020.repl4nlp-1.20
10. Gupta U., Akin B. Architectural Strategies for Energy Efficiency // Hardware Accelerator Systems for Artificial Intelligence and Machine Learning / ed. by A. Reuther et al. (Advances in Computers ; vol. 122). Elsevier, 2021. P. 31–75. DOI: 10.1016/bs.adcom.2021.01.001
11. Hooker S., Moorosi N., Clark G. et al. Characterising Bias in Compressed Models : препринт. 2020. arXiv:2010.03058 [stat.ML]. URL: https://arxiv.org/abs/2010.03058
12. Bender E.M., Gebru T., McMillan-Major A., Shmitchell S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? // FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021. P. 610–623. DOI: 10.1145/3442188.3445922
13. Rogers A., Kovaleva O., Rumshisky A. A Primer in BERTology: What We Know About How BERT Works // Transactions of the Association for Computational Linguistics. 2021. Vol. 8. P. 842–866. DOI: 10.1162/tacl_a_00349
14. Liu Y., Ott M., Goyal N. et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach : препринт. 2019. arXiv:1907.11692 [cs.CL]. URL: https://arxiv.org/abs/1907.11692
15. Fedus W., Zoph B., Shazeer N. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity // Journal of Machine Learning Research (JMLR). 2022. Vol. 23(120). P. 1–39. URL: https://jmlr.org/papers/v23/21-0998.html

16. Ganesh P., Chen Y., Lou X. et al. Compressing Large-Scale Transformer-Based Models: A Case Study on BERT // Transactions of the Association for Computational Linguistics. 2021. Vol. 9. P. 1061–1080. DOI: 10.1162/tacl_a_00413
17. Xu C., McAuley J. A Survey on Model Compression for Natural Language Processing : препринт. 2023. arXiv:2302.07105 [cs.CL]. URL: https://arxiv.org/abs/2302.07105