УДК 004.8

Скворцов Роберт Родионович,
и.о. директора по научно-техническому развитию, АО «Алмаз»
Skvortsov Robert Rodionovich,
Acting Director for Scientific and Technical Development, Almaz JSC

ИССЛЕДОВАНИЕ МЕХАНИЗМОВ ПРЕДОТВРАЩЕНИЯ КАТАСТРОФИЧЕСКОГО ЗАБЫВАНИЯ В НЕЙРОННЫХ ЯЗЫКОВЫХ МОДЕЛЯХ НА ОСНОВЕ КОНТРОЛИРУЕМОГО СОХРАНЕНИЯ ПРИОБРЕТЁННЫХ ЗНАНИЙ
RESEARCH INTO MECHANISMS FOR PREVENTING CATASTROPHIC FORGETTING IN NEURAL LANGUAGE MODELS THROUGH CONTROLLED PRESERVATION OF ACQUIRED KNOWLEDGE

Аннотация. Рассмотрена проблема катастрофического забывания при последовательном освоении новых задач компактными нейронными языковыми моделями. Цель исследования – разработать и описать метод контролируемого сохранения приобретённых компетенций, предусматривающий фиксацию защищённого состояния модели, изолированное получение экспериментальных изменений и их допуск только после проверки ранее освоенных задач. Представлена формальная постановка задачи, предложены критерии сохранности и пластичности, а также протокол раздельного оценивания неизменённой базовой конфигурации и конфигурации с активными дополнительными компонентами. На предварительном этапе апробирована процедура регрессионной проверки на внутреннем наборе из 616 контрольных заданий. Установлены границы интерпретации этих наблюдений: воспроизводимость защищённого состояния не эквивалентна доказательству отсутствия забывания после интеграции новых компонентов. Для проверки эффективности метода предложен сравнительный эксперимент с независимой тестовой выборкой и контрольными способами последовательного обучения. Работа носит методико-экспериментальный характер; показатели сравнительного превосходства до завершения такого эксперимента не заявляются.

Abstract. This paper addresses catastrophic forgetting in compact neural language models during sequential acquisition of new tasks. We present a controlled knowledge-preservation methodology combining a protected reference state, isolated experimental changes, and a validation gate based on retained capabilities. We formalize retention and plasticity criteria and distinguish between evaluating a frozen baseline and evaluating a model with newly activated components. A preliminary regression-control procedure employed an internal set of 616 checks; this observation establishes the feasibility of monitoring a protected state but does not prove forgetting-free continual learning. A comparative protocol with independent holdout tests and established baselines is outlined. Claims of superiority are deliberately deferred until controlled experiments are completed.

Ключевые слова: Нейронные языковые модели, катастрофическое забывание непрерывное обучение, сохранение знаний, регрессионное тестирование, изоляция параметров, контроль изменений, устойчивость компетенций.

Keywords: Neural language models, catastrophic forgetting, continual learning, knowledge retention, regression testing, parameter isolation, controlled updates.

Введение

Последовательное обучение нейронных языковых моделей предполагает расширение спектра решаемых задач без полного повторного обучения на всех ранее доступных данных. Однако адаптация параметров к новой предметной области способна ухудшать выполнение

задач, освоенных ранее. Такое ухудшение описывается как катастрофическое забывание (catastrophic forgetting) и остаётся одним из центральных ограничений непрерывного обучения [1, 4]. В обработке естественного языка проблема осложняется контекстной зависимостью смысла, вариативностью синтаксиса и необходимостью обобщать знания на формулировки, отсутствовавшие в обучающей выборке [5].

Особое значение проблема приобретает для компактных языковых моделей: ограниченный объём параметров и вычислительных ресурсов затрудняет одновременное поддержание ранее освоенных компетенций и освоение новых. Существующие методы используют регуляризацию, воспроизведение ранее встречавшихся примеров, ограничения градиентов либо выделение отдельных обучаемых параметров [1–4, 6–9]. Тем не менее защита параметров и фактическая сохранность поведения модели не тождественны: неизменность части весов не гарантирует сохранения ответов системы после подключения новых вычислительных компонентов.

Цель работы – сформулировать и методически обосновать процедуру контролируемого сохранения приобретённых компетенций при последовательном расширении возможностей языковой модели. Объект исследования – нейронные языковые модели в режиме последовательного обучения; предмет – организационные и вычислительные условия допуска новых изменений при заданном ограничении на ухудшение ранее освоенных задач. Проверяемая гипотеза состоит в том, что сочетание защищённого базового состояния, локализованного обучения и независимой поведенческой проверки может снизить риск деградации ранее приобретённых компетенций по сравнению с неконтролируемым дополнительным обучением.

1. Подходы к предотвращению катастрофического забывания

В классической постановке последовательного обучения модель получает задачи T_1, T_2, …, T_n поочерёдно. Оптимизация на задаче T_t изменяет параметры в направлении, выгодном для новой обучающей выборки, но потенциально конфликтующем с областями параметрического пространства, необходимыми для предыдущих задач. Метод Elastic Weight Consolidation ограничивает отклонение параметров, наиболее значимых для ранее решённых задач [1]. Подход Learning without Forgetting использует ограничения на поведение сети, направленные на сохранение ранее сформированных возможностей [2].

Другое направление связано с хранением примеров предыдущих задач и контролем направления обновлений. Gradient Episodic Memory (GEM) ограничивает обновления с учётом качества на сохранённых примерах; подход A-GEM предлагает вычислительно упрощённый вариант [3, 9]. Методы изоляции параметров позволяют уменьшить прямую интерференцию между задачами: в частности, PackNet фиксирует существенную часть параметров для ранее освоенных задач, высвобождая доступную ёмкость для последующих [6].

Работы по непрерывному обучению в NLP подчёркивают, что оценивать только результат на новой задаче недостаточно [5, 8]. Для предметной адаптации языковых моделей важны как качество на новых данных, так и сохранение общей языковой компетентности [7, 8]. По этой причине в настоящей работе фокус смещён с единственного способа оптимизации параметров на проверяемое правило допуска изменений, объединяющее меры защиты и независимый контроль итогового поведения модели.

Предлагаемый подход не претендует на первенство в использовании заморозки параметров, модульного обучения или регрессионного тестирования по отдельности. Предмет предлагаемого исследования – их совместное применение как формализованной процедуры отбора новых состояний модели по двум независимым критериям: приобретение новой компетенции и сохранность ранее освоенных.

2. Концепция контролируемого сохранения приобретённых знаний

2.1. Защищённое состояние и экспериментальный контур

За отправную точку принимается эталонное состояние M_0, для которого заранее зафиксированы результаты на совокупности ранее освоенных задач. Под «сохранением знаний» в работе понимается не физическое хранение сведений в неизменных параметрах, а сохранение наблюдаемой способности системы корректно выполнять заранее определённый класс заданий. Такая операциональная трактовка позволяет связывать утверждения о сохранности знаний с измеряемыми показателями.

Предлагается разделять эталонное состояние и экспериментальные изменения. В защищённом контуре хранятся идентификатор версии, контрольные данные, правила оценивания и исходные результаты. В экспериментальном контуре формируется кандидат на расширение возможностей за счёт обучаемого дополнительного компонента или другого локализованного изменения. До завершения проверки кандидат не заменяет эталонное состояние.

Принцип изоляции уменьшает риск непосредственной модификации защищённых параметров, но не исключает косвенной интерференции на уровне выходов. Поэтому необходимы два различных режима проверки: проверка воспроизводимости M_0 и проверка поведения полной экспериментальной конфигурации M_t при активном новом компоненте. Смешение этих режимов способно привести к завышенной оценке эффективности предотвращения забывания.

2.2. Пятиэтапная процедура допуска изменений

Контролируемое сохранение приобретённых компетенций предполагает последовательное прохождение пяти этапов. На первом этапе создаются неизменяемая базовая версия модели и описание защищённых навыков. На втором формируются контрольные задания, включая случаи с близкой лексикой, отрицанием, изменением порядка компонентов предложения и изменением формулировок. На третьем новые компетенции осваиваются вне защищённого состояния. На четвёртом кандидат оценивается по независимым критериям сохранности и приобретения навыков. На пятом принимается решение о принятии, отклонении либо доработке кандидата; при отклонении сохраняется возможность возврата к предыдущему состоянию.

Существенным условием является запрет на многократную настройку кандидата по окончательной тестовой выборке. Для выбора и отладки допускается отдельная валидационная часть; независимая тестовая часть используется для финальной оценки, а не для текущей оптимизации. В противном случае набор защищённых компетенций может превратиться в ещё одну скрытую обучающую выборку, а наблюдаемая сохранность – в следствие переобучения на конкретных проверках.

Таблица 1

Функции элементов метода контролируемого сохранения знаний

Элемент	Функция	Контролируемый риск
Защищённое состояние	Фиксирует исходное качество и допускает возврат	Необратимая деградация эталона
Изолированное обучение	Локализует экспериментальные обновления	Непредусмотренное изменение базовых параметров
Контроль старых навыков	Измеряет качество при активных изменениях	Скрытое ухудшение прежних компетенций
Контроль новых навыков	Подтверждает полезность обновления	Формальное сохранение при отсутствии обучения
Критерий допуска	Связывает оба типа проверок в одно решение	Принятие частично успешного кандидата

3. Формальная постановка и критерии оценки

Пусть $M_0 = f(x; \theta_0)$ – исходная модель с фиксированными параметрами θ_0, а $M_t = f(x; \theta_0, \varphi_t)$ – экспериментальная конфигурация после t-го этапа расширения, где φ_t обозначает совокупность изменяемых параметров или подключаемых компонентов. Обозначение не задаёт конкретной архитектуры: дополнительный компонент рассматривается как абстрактная часть вычислительной системы. Пусть Dold и Dnew – независимые оценочные совокупности для ранее освоенных и новых задач соответственно.

Для заданной метрики Q, в которой большее значение соответствует лучшему результату, определим изменение качества ранее освоенных компетенций и величину забывания:

$$\Delta old\ (t) = Q\ (M_t, Dold) - Q\ (M_0, Dold), \quad (1)$$

$$F\ (t) = \max\{0; Q\ (M_0, Dold) - Q\ (M_t, Dold)\}. \quad (2)$$

При $\Delta old\ (t) < 0$ наблюдается снижение средней результативности на выбранной совокупности задач. Для языковых моделей агрегированная оценка может скрывать существенные ошибки на отдельных типах высказываний. Поэтому дополнительно рассматривается набор подгрупп Dold,k (например, отрицание, перестановка ролей, пассивные конструкции, вложенные высказывания) и рассчитывается наихудшее изменение качества по подгруппам:

$$\Delta min\ (t) = min_k\ \{Q\ (M_t, Dold,k) - Q\ (M_0, Dold,k)\}. \quad (3)$$

Наличие нового навыка предлагается фиксировать отдельным условием – достижением минимально приемлемого результата на Dnew, который не должен выводиться из результатов на Dold. При заданных порогах $\varepsilon \geq 0$, $\varepsilon_k \geq 0$ и τ правило допуска кандидата может быть записано так:

$G\ (t) = 1$, если $F\ (t) \leq \varepsilon$, $\Delta_k\ (t) \geq -\varepsilon_k$ для всех k и $Q\ (M_t, Dnew) \geq \tau$; иначе $G\ (t) = 0$.

Пороговые значения должны задаваться до анализа окончательных тестовых результатов. В исследовании допускается устанавливать более жёсткие пороги для критичных компетенций и отдельно оценивать уровень неопределённости. Если ответы недетерминированы, одинаковые задания необходимо выполнять при стандартизированных параметрах генерации и/или с повторениями. Для парных бинарных исходов уместно использовать анализ таблиц сопряжённости, в том числе критерий Мак-Немара; доверительные интервалы можно оценивать бутстрэпом с учётом группировки сходных заданий.

Важное ограничение: сохранение значения Q при сравнении M_0 с его неизменённой копией характеризует корректность процедуры фиксации состояния, но не демонстрирует эффективное предотвращение забывания. Содержательное испытание возникает только при сопоставлении M_0 и M_t с включёнными экспериментальными изменениями.

4. Материалы и методика исследования

4.1. Состав контрольных заданий

Оценочная совокупность должна включать задачи, на которых модель обладала приемлемым качеством до начала очередного этапа обучения. Для анализа сохранности языковых компетенций предлагаются классы заданий на установление отношений «субъект – действие – объект», определение области действия отрицания, интерпретацию активных и пассивных конструкций, разрешение простых контекстных зависимостей и устойчивость к эквивалентным переформулировкам. При формировании выборки необходимо исключать прямые и близкие дубликаты примеров нового обучения.

Один из предварительных этапов внутреннего исследования использовал контрольную совокупность объёмом 616 проверок. Она служила инструментом регрессионного контроля ранее освоенных задач. Это число характеризует объём проверок, но само по себе не является числом независимых наблюдений: наличие тематически или лексически связанных примеров необходимо отдельно учитывать при статистической интерпретации. Описание выборки в

общедоступной публикации может быть дано через типы заданий, правила отбора и обезличенные примеры без раскрытия внутреннего набора данных.

4.2. Сравниваемые режимы

Для проверки гипотезы требуется сравнение в идентичных условиях как минимум трёх режимов: (А) обычного дополнительного обучения с изменением доступных параметров; (Б) признанного базового метода сохранения компетенций; (В) контролируемого обучения с изоляцией изменений и правилом допуска. Отдельным контролем служит неизменённая исходная модель M_0. Для всех режимов следует использовать одинаковые обучающие последовательности, исходные веса, ограничения на вычислительные ресурсы и независимые оценочные наборы. Сравнение следует повторять при нескольких начальных состояниях генератора случайных чисел и различных порядках предъявления новых задач.

Таблица 2

План сравнительного эксперимента (показатели заполняются после испытаний)

Режим	**Качество старых задач**	**Качество новых задач**	**Сравнение с эталоном**
M_0: эталон	Исходная точка	Исходная точка	Контроль воспроизводимости
А: обычное дообучение	Измеряется	Измеряется	Оценка деградации без защиты
Б: базовый CL-метод	Измеряется	Измеряется	Сопоставление с известным подходом
В: предлагаемый контроль	Измеряется при активном изменении	Измеряется	Оценка эффекта критерия допуска

4.3. Порядок проверки и воспроизводимость

Протокол исследования предполагает предварительную регистрацию порогов допуска, сохранение конфигураций и контрольных сумм артефактов, формирование отдельной выборки для настройки порогов и независимой финальной выборки, а также журналирование результатов по каждому классу задач. После каждого этапа обучения оценивается вектор качества старых компетенций; при прохождении защитного критерия дополнительно проверяется усвоение новой задачи и измеряется стоимость обновления по времени и дополнительной памяти.

В целях научной воспроизводимости публикация должна содержать сведения о классах задач, правилах разбиения данных, используемых метриках, числе повторений и процедурах статистического анализа. Для защиты интеллектуальной собственности допускается не раскрывать инженерную реализацию дополнительного компонента, однако такой выбор ограничивает внешнюю воспроизводимость и должен быть явно обозначен. Частичная компенсация ограничения возможна за счёт тестирования описанной процедуры на общедоступной модели и открытых контрольных данных.

5. Предварительные наблюдения и границы их интерпретации

В ходе исследовательской разработки была организована процедура контроля сохранности ранее освоенных компетенций, включающая фиксацию защищённого состояния, раздельное экспериментирование с дополнительными изменениями и повторное выполнение контрольных заданий. На предварительном этапе использовался внутренний набор из 616 проверок. В рамках выполненных проверок отмечалась воспроизводимость защищённого состояния и сохранность исходной конфигурации при изоляции экспериментальных изменений.

Полученное наблюдение подтверждает практическую реализуемость регрессионного контроля и процедуры возврата к защищённому состоянию. Вместе с тем оно не позволяет утверждать, что подключение новых компонентов гарантированно не ухудшает качество ранее освоенных задач. Для такого вывода необходима оценка активной экспериментальной конфигурации на независимых данных, сопоставление с обычным дообучением и известными методами непрерывного обучения, а также проверка устойчивости результата по различным видам лингвистических задач.

По состоянию на подготовку настоящей рукописи контролируемый сравнительный эксперимент с единым набором открытых базовых моделей, независимыми тестовыми выборками и статистическим анализом не завершён. Поэтому численные оценки эффективности метода относительно альтернатив, доверительные интервалы и выводы о статистической значимости в статье не приводятся. Следующий этап исследования должен обеспечить именно эту доказательную часть.

6. Обсуждение

Предложенный подход позволяет развести три утверждения, которые часто ошибочно объединяются: (1) параметры защищённой модели не изменились; (2) её поведение воспроизводимо; (3) новая конфигурация с дополнительными компонентами сохранила прежние компетенции. Первые два утверждения можно подтвердить контролем версии и регрессионным тестированием эталона. Третье требует независимой поведенческой оценки новой конфигурации; именно оно относится к предотвращению катастрофического забывания в строгом экспериментальном смысле.

Преимуществом контролируемого допуска представляется возможность формализовать требования к сохранности по отдельным классам задач, а не ориентироваться исключительно на средний показатель качества. Кроме того, процедура допуска допускает отказ от неудачного изменения без потери работоспособной версии. Это особенно важно для небольших моделей, где изменение ограниченной части параметров может приводить к локальным, но критичным языковым ошибкам.

Однако метод не устраняет фундаментальный конфликт устойчивости и пластичности автоматически. При чрезмерно строгом пороге допуска новая компетенция может не быть принята, даже если она полезна; при слабом пороге возможно накопление малых потерь по мере роста числа задач. Кроме того, рост числа дополнительных компонентов увеличивает вычислительные затраты, а фиксированный набор проверок не охватывает всех возможных проявлений деградации. Поэтому дальнейшее развитие метода предполагает адаптивное пополнение контрольной совокупности, исследование совокупного эффекта нескольких последовательных обновлений и оценку задержки вычислений при активации новых компонентов.

С точки зрения научной новизны необходима осторожность. Изоляция параметров, сохранение эталона, регрессионное тестирование и ограничения обновлений изучались ранее [1–6, 9]. Потенциальный самостоятельный результат данной работы заключается в формализации совместного решения о допуске изменений по двум независимым группам метрик и в проверке этого решения для компактных языковых моделей на лингвистически структурированных контрольных задачах. Степень новизны и практической эффективности должна быть подтверждена сравнением с сильными базовыми методами.

Заключение

Предложена методика контролируемого сохранения приобретённых знаний при последовательном обучении компактных нейронных языковых моделей. Метод объединяет защищённое исходное состояние, локализацию экспериментальных изменений, раздельную оценку старых и новых компетенций и формализованное правило допуска кандидата.

Определены показатели снижения качества ранее освоенных задач, предложен учёт наихудшего результата по лингвистическим подгруппам и описаны условия независимой экспериментальной проверки.

Предварительный этап позволил организовать контроль сохранности исходной конфигурации с использованием внутреннего набора из 616 проверок. Данный результат подтверждает реализуемость процедуры регрессионного контроля, но не служит доказательством полного устранения катастрофического забывания. Для проверки гипотезы о снижении деградации компетенций при последовательном обучении необходимо завершить сравнительные испытания с активными новыми компонентами, независимой тестовой выборкой, несколькими повторениями и статистической оценкой различий.

Практическая ценность предлагаемой методики состоит в возможности использовать сохранность ранее освоенных компетенций как обязательный критерий принятия решений о расширении возможностей компактных языковых систем.

Список литературы:

1. Kirkpatrick J., Pascanu R., Rabinowitz N. et al. Overcoming catastrophic forgetting in neural networks // Proceedings of the National Academy of Sciences. 2017. Vol. 114, no. 13. P. 3521–3526. DOI: 10.1073/pnas.1611835114.

2. Li Z., Hoiem D. Learning without Forgetting // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2018. Vol. 40, no. 12. P. 2935–2947. DOI: 10.1109/TPAMI.2017.2773081.

3. Lopez-Paz D., Ranzato M. Gradient Episodic Memory for Continual Learning // Advances in Neural Information Processing Systems. 2017. Vol. 30. P. 6467–6476.

4. De Lange M., Aljundi R., Masana M. et al. A Continual Learning Survey: Defying Forgetting in Classification Tasks // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2022. Vol. 44, no. 7. P. 3366–3385. DOI: 10.1109/TPAMI.2021.3057446.

5. Biesialska M., Biesialska K., Costa-jussà M. R. Continual Lifelong Learning in Natural Language Processing: A Survey // Proceedings of the 28th International Conference on Computational Linguistics. 2020. P. 6523–6541. DOI: 10.18653/v1/2020.coling-main.574.

6. Mallya A., Lazebnik S. PackNet: Adding Multiple Tasks to a Single Network by Iterative Pruning // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018. P. 7765–7773. DOI: 10.1109/CVPR.2018.00810.

7. Gururangan S., Marasović A., Swayamdipta S. et al. Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. P. 8342–8360. DOI: 10.18653/v1/2020.acl-main.740.

8. Ke Z., Shao Y., Lin H. et al. Continual Pre-Training of Language Models // Proceedings of the International Conference on Learning Representations. 2023.

9. Chaudhry A., Ranzato M., Rohrbach M., Elhoseiny M. Efficient Lifelong Learning with A-GEM // Proceedings of the International Conference on Learning Representations. 2019.