УДК 004.032.26

МЕТОД ОБУЧЕНИЯ ГРУППЫ ЭКСПЕРТОВ НА ОСНОВЕ АВТОМАТИЧЕСКОГО РАСШИРЕНИЯ АРХИТЕКТУРЫ

А. К. Клименко, аспирант МГТУ им. Н.Э. Баумана, Москва, Россия;
orcid.org/0009-0009-2412-0641, e-mail: klimenkoak@student.bmstu.ru
К. А. Майков, д.т.н., профессор кафедры ИУ7 МГТУ им. Н.Э. Баумана, Москва, Россия;
orcid.org/0000-0003-1864-2397, e-mail: maikov@bmstu.ru
В. В. Тишкина, к.т.н., доцент кафедры ВПМ РГРТУ, Рязань, Россия;
orcid.org/0000-0002-6320-3513, e-mail: LeraTishkina@mail.ru

Архитектуры группы (смеси) экспертов (англ. Mixture-of-Experts, MoE) позволяют масштабировать языковые модели без пропорционального роста вычислительных затрат, активируя лишь подмножество параметров для каждого токена. Однако, известные классические подходы требуют априорного выбора числа экспертов, что может приводить к субоптимальной емкости модели и замедлению сходимости. В данной работе представлен метод обучения нейросетей на основе архитектуры смеси экспертов, который автоматически расширяет множество экспертов на стадии обучения. Предложенный механизм добавляет новых экспертов при выходе на плато метрики качества, используя стратегию «теплого старта» для ускорения адаптации. Эксперименты на задачах GLUE демонстрируют ускорение сходимости на 5 – 8 % по сравнению с аналогичными стратегиями обучения при сопоставимом конечном размере обученных моделей. Метод обеспечивает теоретически обоснованную возможность повышения качественных характеристик решения и снижения ресурсоемкости.

Ключевые слова: *группа экспертов, смесь экспертов, MoE, адаптивное обучение, динамическое расширение архитектуры.*

DOI: 10.21667/1995-4565-2026-95-143-147

Введение

Архитектуры типа «смесь экспертов» (Mixture-of-Experts, MoE) стали важным инструментом масштабирования больших языковых моделей [1]. В отличие от плотных сетей, где все параметры активны для каждого входного токена, MoE-слои активируют специализированные подсети (эксперты) через механизм маршрутизации, что позволяет значительно увеличить общее число параметров при умеренных вычислительных затратах [2].

Однако известные MoE-архитектуры обладают существенным ограничением: количество экспертов фиксируется до начала обучения [3]. Определение оптимального числа экспертов требует либо трудоемкой оптимизации гиперпараметров, либо их консервативного выбора, что может приводить к недозагруженности экспертов или недостаточной выразительной способности модели. В последние годы появились подходы к динамическому изменению архитектур нейросетей, включая прогрессивное наращивание [4] и методы нейроэволюции, однако они либо требуют заранее заданного плана расширения, либо обладают высокой вычислительной сложностью.

В данной работе представлен метод адаптивного расширения смеси экспертов, который решает следующие задачи:

– автоматическое определение моментов для добавления новых экспертов на основе анализа сходимости метрик обучения;

– минимизация влияния на стабильность обучения при расширении архитектуры;

– обеспечение совместимости с распределенным обучением и существующими оптимизациями архитектуры MoE.

Основной вклад работы включает:

– эффективный эвристический механизм расширения архитектуры MoE;
– стратегию инициализации и «теплого старта» новых экспертов;
– экспериментальное подтверждение эффективности на многозадачных бенчмарках.

Теоретическая часть

Пусть $D = \{(x_i, y_i)\}_{i=1}^{N}$ — обучающая выборка. MoE-слой с K экспертами $\{E_k\}_{k=1}^{K}$ и функцией маршрутизации $g : \mathbb{R} \to \mathbb{R}^K$ минимизирует целевую функцию

$$\mathcal{L}(\theta) = E_{x,y} \sim D\left[\ell\left(f_\theta(x), y\right)\right] + \alpha\mathcal{L}_{balance} + \beta\mathcal{L}_{z-loss}, \quad (1)$$

где ℓ – функция потерь на обучающей выборке, $\mathcal{L}_{balance} = \sum_{k=1}^{K} p_k log(p_k)$ – балансировочный член, способствующий равномерной загрузке экспертов, $\mathcal{L}_{z-loss} = log(\sum_{k=1}^{K} exp(g(x)_k))$ – регуляризатор стабильности маршрутизации, α,β – весовые коэффициенты для соответствующих функций потерь.

Предлагаемый метод динамически увеличивает количество экспертов K в ответ на замедление сходимости. Пусть $\mathcal{L}_{val}(t)$ – значение функции потерь на валидационной выборке после эпохи t. Добавление нового эксперта E_{K+1} осуществляется при условии фиксации выхода на плато качества модели

$$\min_{i=1\ldots\tau} \mathcal{L}_{val}(t-i) - \mathcal{L}_{val}(t) < \varepsilon, \quad (2)$$

где τ — размер окна анализа, ε – порог значимого улучшения.

Для минимизации дестабилизации обучения применяется следующая стратегия инициализации:

– параметры нового эксперта инициализируются случайным отклонением $\theta_{K+1} = \theta_k + \mathcal{N}(0, \sigma^2)$, где θ_k — параметры случайно выбранного существующего эксперта;

– скорость обучения для нового эксперта увеличивается в γ раз на период «теплого старта» T_{warm};

– соответствующий вектор в слое маршрутизации инициализируется малыми случайными значениями.

Пусть $\mathcal{H}_K$ – пространство гипотез MoE с K экспертами, тогда последовательное добавление экспертов образует вложенную структуру $\mathcal{H}_K \subseteq \mathcal{H}_{K+1}$, что гарантирует неснижение точности модели

$$\min_{\theta\in\Theta_{K+1}} \mathcal{L}(\theta) \le \min_{\theta\in\Theta_K} \mathcal{L}(\theta), \quad (3)$$

доказательство чего следует из включения параметрических множеств и выпуклости функций потерь [4].

Добавление экспертов в моменты плато точности способствует преодолению локальных минимумов и ускорению сходимости, что подтверждается приведенными результатами проведенных экспериментальных исследований.

Экспериментальные исследования

Экспериментальные исследования проводились для четырех архитектурных конфигураций: двух моделей статической смеси экспертов с 4 и 8 экспертами, а также двух моделей динамической смеси экспертов с расширением общего числа экспертов от 2 до 4 и 8. В исследуемых архитектурах использовалось 2 активируемых эксперта на входной токен. Выбранное количество экспертов обусловлено невысокой семантической сложностью тренировочных примеров, вычислительной трудоемкостью процесса обучения и стремлением до-

биться улучшений качества моделей за счет предложенного метода обучения, а не за счет количества параметров обучаемых моделей.

Экспериментальные исследования проводились на датасетах для оценки понимания письменной речи GLUE CoLA и SST-2 [5]. Анализируемые тексты предобрабатывались токенизатором и моделью для формирования 384-мерных эмбеддингов e5-small [6]. Обучение и валидация производились пакетами по 64 примера в течение 30 эпох для исследуемых моделей с оптимизатором AdamW с коэффициентом скорости обучения $1 \cdot 10^{-4}$ и «линейным разогревом» в 10 % от общего числа шагов.

Параметры расширения архитектуры: $\tau = 3$, $\varepsilon = 5 \cdot 10^{-4}$. Новому эксперту назначался 10-кратно увеличенный коэффициент скорости обучения до момента добавления следующего эксперта. Значения весовых коэффициентов функции потерь: $\alpha = 5 \cdot 10^{-2}$, $\beta = 1 \cdot 10^{-2}$.

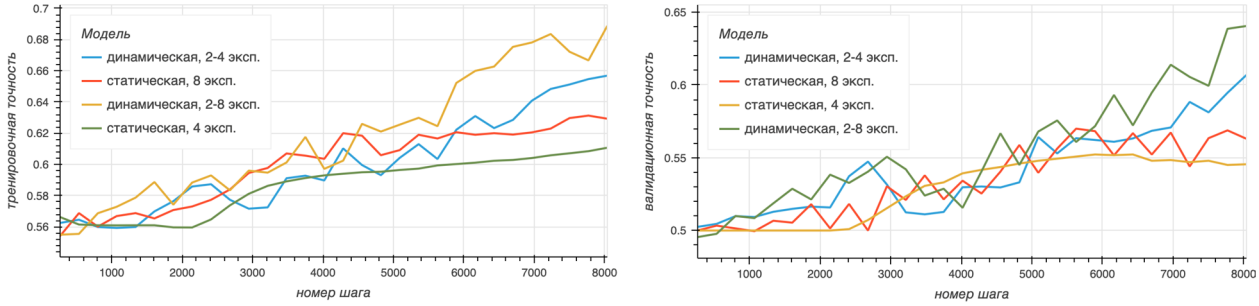

Рисунок 1 – Тренировочная и валидационная точности моделей со статическим и динамическим механизмом выделения экспертов
Figure 1 – Train and validation accuracy of models with static and dynamic experts allocation strategies

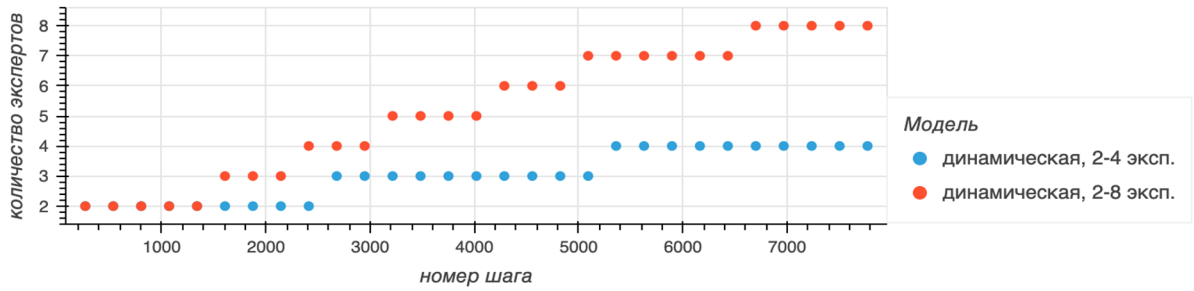

Рисунок 2 – Количество экспертов моделей с динамическим механизмом выделения экспертов
Figure 2 – Experts count for models with dynamic experts allocation strategy

Время проведения экспериментов на чипе Apple Silicon M4 Max 24 ГБ с использованием аппаратного ускорения (Metal) составило 4 часа.

Из приведенных на рисунке 1 зависимостей видно, что модели с фиксированным количеством экспертов демонстрируют стабилизацию точности. Дальнейшее обучение не приводит к значительному увеличению точности, не превышая 3 %. На рисунке 2 группы кривых, соответствующие динамическим моделям, показывают существенный рост точности (5 % и более) при автоматическом добавлении экспертов. Отсутствие заметного снижения валидационной точности на рисунке 1 показывает, что новые эксперты встраиваются без искажения ранее сформированной структуры представлений. Увеличение конечного числа экспертов с 4 до 8 в равных начальных условиях обучения повышает точность классификации обученной модели на 4 % без вырождения состояния модели и переобучения.

Разработанный метод адаптивного обучения обеспечивает повышение точности на 5 – 8 % относительно аналогичных моделей с фиксированным числом экспертов и отсутствие переобучения, что определяет практическую значимость разработанного решения для задач классификации лингвистических структур.

Таблица 1 – Сравнение точности моделей на датасетах CoLA и ST-2
Table 1 – Model accuracy comparison for CoLA and ST-2 datasets

Архитектура модели			Точность на датасетах, %	
Тип	Кол-во экспертов		CoLA	ST-2
Статическая	4	$4{,}7 \cdot 10^6$	58,4	71,0
Статическая	8	$9{,}5 \cdot 10^6$	62,7	73,1
Динамическая	2 – 4	$2{,}4 \cdot 10^6 – 4{,}7 \cdot 10^6$	63,9	73,0
Динамическая	2 – 4	$2{,}4 \cdot 10^6 – 9{,}5 \cdot 10^6$	65,1	74,2

Заключение

В работе описан метод адаптивного расширения экспертной структуры на основе отслеживания точности модели в процессе обучения. Проведенные экспериментальные исследования показали, что предложенный подход обеспечивает повышение точности классификации по сравнению со статическими аналогами, а также предотвращает риск переобучения. Практически значимыми преимуществами предложенного решения являются совместимость с распределенным обучением и теоретически обоснованная возможность повышения точности модели в процессе обучения.

Направления дальнейших исследований включают решение следующих задач:

– инициализацию новых экспертов с помощью методов низкоранговой адаптации, таких как LoRA/DoRA/QLoRA;

– реализацию механизма отсечения редко используемых экспертов;

– исследование применимости метода для экспертно-параллельных сред обучения;

– адаптацию метода для многомодальных и кросс-модальных архитектур моделей.

Библиографический список

1. **Fedus W., Zoph B., Shazeer N.** Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity // Journal of Machine Learning Research. 2022. Vol. 23. No. 120. Pp. 1-39.

2. **Roller S. et al.** Hash Layers for Large Sparse Models // Advances in Neural Information Processing Systems. 2021. Vol. 34. Pp. 17555-17566.

3. **Artetxe M. et al.** Efficient Large Scale Language Modeling with Mixtures of Experts // Journal of Machine Learning Research. 2022. Vol. 23. No. 120. Pp. 1-39.

4. **Mu S., Lin S.** A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications // arXiv preprint arXiv:2503.07137. 2025.

5. **Wang A. et al.** GLUE: A MultiTask Benchmark and Analysis Platform for Natural Language Understanding // Proceedings of the International Conference on Learning Representations (ICLR). 2019.

6. **Wang L. et al.** Text embeddings by weakly-supervised contrastive pre-training // arXiv preprint arXiv:2212.03533. 2022.

UDC 004.032.26

AUTOMATIC ARCHITECTURE-EXPANSION TRAINING FOR MIXTURE-OF-EXPERTS MODELS

A. K. Klimenko, post-graduate student, BMSTU, Moscow, Russia;
orcid.org/0009-0009-2412-0641, e-mail: klimenkoak@student.bmstu.ru
K. A. Maikov, Dr. in technical sciences, full professor, ICS7 BMSTU, Moscow, Russia;
orcid.org/0000-0000-0000-000X, e-mail: maikov@bmstu.ru
V. V. Tishkina, PhD (in technical sciences.), associate professor, RSREU, Ryazan, Russia;
orcid.org/0000-0002-6320-3513, e-mail: LeraTishkina@mail.ru

Mixture-of-Experts (MoE) architectures enable language-model scaling without a proportional increase in computational cost by activating only a subset of parameters per token. Classical approaches, however,

fix the number of experts a priori, often yielding sub-optimal capacity and slower convergence. We propose a training method that automatically grows the expert pool during optimization. A new expert is inserted when the validation metric plateaus; a newcomer is initialized via small random perturbation of an existing expert and is warmed-up with increased learning rate. GLUE benchmarks show 5 – 8 % faster convergence versus static-MoE baselines of comparable final size, while theoretical nesting of hypothesis spaces guarantees non-increasing loss. The method provides a theoretically justified opportunity to improve quality characteristics of the solution and to reduce resource intensity.

Keywords: *mixture of experts, MoE, adaptive training, dynamic architecture expansion.*

DOI: 10.21667/1995-4565-2026-95-143-147

References

1. **Fedus W., Zoph B., Shazeer N.** Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *Journal of Machine Learning Research*. 2022, vol. 23, no. 120, pp. 1-39.

2. **Roller S. et al.** Hash Layers for Large Sparse Models. *Advances in Neural Information Processing Systems.* 2021, vol. 34, pp. 17555-17566.

3. **Artetxe M. et al.** Efficient Large Scale Language Modeling with Mixtures of Experts. *Journal of Machine Learning Research*. 2022, vol. 23, no. 120, pp. 1-39.

4. **Mu S., Lin S.** A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications. *arXiv preprint arXiv*:2503.07137. 2025.

5. **Wang A. et al.** GLUE: A MultiTask Benchmark and Analysis Platform for Natural Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*. 2019.

6. **Wang L. et al.** Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv*:2212.03533. 2022.