Записки научных
семинаров ПОМИ
Том 552, 2026 г.

А. А. Лаэтин, Б. Пулфер, С. К. Смирнов

ИЕРАРХИЧЕСКАЯ ИНИЦИАЛИЗАЦИЯ ВРЕМЕННЫХ МАСШТАБОВ В MAMBA-2 ДЛЯ МНОГОМАСШТАБНОГО МОДЕЛИРОВАНИЯ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

§1. Введение

Архитектура Transformer остается одним из основных инструментов языкового моделирования, однако ее вычислительная сложность по длине последовательности растет квадратично из-за механизма self-attention. Это ограничение особенно заметно в задачах, где требуется устойчиво работать с длинным контекстом. Одной из наиболее перспективных альтернатив стали модели пространства состояний (state space models, SSM), которые обеспечивают линейное или почти линейное масштабирование и при этом способны моделировать дальние зависимости [3, 4].

Среди современных SSM-архитектур особый интерес представляет Mamba и ее развитие Mamba-2 [1,2]. Эти модели совмещают эффективную рекуррентную обработку последовательностей с механизмом избирательной обработки элементов последовательности, позволяющим адаптировать динамику слоя к текущему входу. Важную роль в этой динамике играет параметр временного шага Δt, который определяется через обучаемое смещение Δ_{bias}. Выбор начальных значений этого параметра смещает начальное распределение временных шагов и тем самым влияет на предпочтительный режим обработки информации до начала обучения.

В Mamba-2 шаг дискретизации Δt зависит от входа: он вычисляется на основе линейной проекции входа и смещения Δ_{bias}. Поэтому Δ_{bias} не задает временной масштаб напрямую, но влияет на начальное распределение значений Δt. Интерес к тому, как параметризация и инициализация влияют на поведение SSM, отражен и в специальных работах по инициализации таких моделей [6, 7]. Стандартная инициализация этого смещения не учитывает положение слоя в сети. Между

Ключевые слова: модели пространства состояний, Mamba, инициализация параметров, языковое моделирование, многомасштабная обработка.

тем для иерархических моделей естественно предположить, что нижние слои должны быть более чувствительны к локальным зависимостям, а верхние - к более протяженному контексту. Отсюда возникает гипотеза, что послойная инициализация Δ_{bias} может служить простым способом задать начальную многомасштабную структуру в Mamba-2 без изменения архитектуры.

В работе рассматривается послойная инициализация Δ_{bias}, при которой начальные значения Δt уменьшаются с глубиной: нижним слоям соответствуют большие значения, верхним - меньшие. Эта стратегия сравнивается с базовой равномерной инициализацией и с обратным, возрастающим по слоям вариантом. Сравнение проводится на 45 независимых запусках модели Mamba-2 размером около 40M параметров. Эксперименты показывают, что убывающая инициализация ускоряет сходимость при предварительном обучении и дает более устойчивое качество на внешних наборах данных.

Вклад работы состоит в следующем:

- предложена простая послойная стратегия инициализации Δ_{bias}, основанная на предположении о различии функций нижних и верхних слоев;
- проведено сравнение трех стратегий инициализации при фиксированных архитектуре, данных и процедуре обучения;
- проанализировано влияние выбранной инициализации на сходимость и качество обобщения Mamba-2.

§2. Архитектура Mamba-2 и предлагаемая инициализация

Модели пространства состояний вновь стали активно использоваться после работ HIPPO и S4, где была показана их эффективность при моделировании длинных зависимостей в последовательностях [3, 4]. В Mamba селективный механизм делает параметры динамики зависящими от входа, что повышает выразительность модели по сравнению с линейными SSM [1]. В Mamba-2 временной шаг вычисляется как

$$\Delta t = \text{softplus}(\text{Linear}(x) + \Delta_{\text{bias}}). \tag{1}$$

Параметр Δ_{bias} тем самым задает начальное смещение распределения временных шагов слоя.

Значение Δt влияет на характерный масштаб обработки информации. При фиксированной динамике большие значения Δt соответствуют более быстрому обновлению состояния и лучше подходят для локальных зависимостей. Меньшие значения в типичной интерпретации способствуют более медленному изменению состояния и, следовательно, лучшему накоплению контекста. Поэтому одинаковая стратегия инициализации Δ_{bias} для всех слоев не учитывает возможное функциональное различие между нижними и верхними уровнями сети.

Предлагаемый подход опирается на идею многомасштабной обработки последовательностей [5]. Он также согласуется с работами, где показано, что параметризация и начальная инициализация SSM существенно влияют на обучение модели и на ее способность учитывать дальние зависимости [6,7]. В отличие от подходов, которые добавляют специальные архитектурные механизмы для работы с разными масштабами, мы не изменяем структуру модели. Вместо этого мы задаем начальное послойное распределение временных шагов через инициализацию Δ_{bias}.

Мы сравниваем три стратегии инициализации. Пусть L - число слоев модели, $i \in \{0, \ldots, L-1\}$ - номер слоя, а допустимый диапазон временных шагов задается значениями $\Delta t_{\min} = 0.001$ и $\Delta t_{\max} = 0.1$.

Убывающая стратегия задается формулой

$$\Delta t_{\text{layer}_i} = \Delta t_{\max} - \frac{i}{L-1}(\Delta t_{\max} - \Delta t_{\min}), \tag{2}$$

то есть нижние слои получают большие начальные значения Δt, а верхние - меньшие. Возрастающая стратегия задается симметрично:

$$\Delta t_{\text{layer}_i} = \Delta t_{\min} + \frac{i}{L-1}(\Delta t_{\max} - \Delta t_{\min}). \tag{3}$$

Равномерная стратегия используется как базовая и соответствует стандартной инициализации: значения Δt для всех слоев и голов выбираются из распределения $U[0.001, 0.1]$.

Для каждой стратегии сначала задавалось целевое начальное распределение Δt_{layer} по слоям. Затем соответствующие значения Δ_{bias} выбирались с помощью обратного преобразования softplus:

$$\Delta_{\text{bias},i} = \log\left(\exp(\Delta t_{\text{layer},i}) - 1\right). \tag{4}$$

Тем самым Δ_{bias} задавал начальное смещение распределения временных шагов для соответствующего слоя.

Рабочая гипотеза состоит в том, что убывающая стратегия лучше согласуется с иерархической организацией глубокой модели: нижние слои выделяют локальные признаки, а верхние накапливают более широкий контекст. Возрастающая стратегия задает противоположный порядок масштабов и позволяет проверить, важен ли сам порядок распределения Δt по слоям.

§3. Описание экспериментов

В экспериментах использовалась архитектура Mamba-2 [2] с 12 скрытыми слоями, размерностью скрытого состояния 256 и 24 головами для параметра временного шага. Общее число параметров составляло около 40M. Выбранный масштаб модели позволил провести серию из 45 независимых запусков при фиксированном вычислительном бюджете.

Для предварительного обучения применялся корпус OpenWebText [8] объемом 1B токенов. Все модели обучались с нуля на последовательностях длины 512. Использовались одинаковые гиперпараметры оптимизации для всех сравниваемых стратегий: learning rate $5 \cdot 10^{-4}$ с косинусным расписанием, warmup ratio 0.01, $\beta_1 = 0.9$, $\beta_2 = 0.99$, weight decay 0.01. Эффективный размер батча составлял 986 последовательностей, то есть примерно 0.5 M токенов на шаг. Каждая модель обучалась 2000 шагов.

Для каждой из трех стратегий инициализации - убывающей, равномерной и возрастающей - было выполнено по 15 независимых запусков. Основными метриками на этапе обучения служили Eval Loss и Eval Token Accuracy. Чтобы проверить переносимость эффекта на данных другого типа, дополнительно проводилась оценка perplexity на наборах TinyStories [11], SQuAD [10] и Wiki-EN [9].

Статистическое сравнение стратегий выполнялось с помощью одностороннего t-критерия Стьюдента, поскольку основная проверяемая гипотеза была направленной: убывающая стратегия должна превосходить равномерную, а возрастающая служила обратной контрольной схемой. При интерпретации результатов учитывались средние значения и стандартные ошибки по независимым запускам. Нас интересовал прежде всего вопрос, превосходит ли убывающая инициализация базовую равномерную стратегию и насколько устойчиво возрастающая стратегия оказывается хуже. Такой дизайн позволяет отделить

	Eval Loss		**Eval Token Accuracy**	
Стратегия	Mean	SE	Mean	SE
Равномерная	5.0737	0.0123	0.2091	0.0012
Убывающая	5.0287	0.0169	0.2131	0.0013
Возрастающая	5.1223	0.0257	0.2048	0.0022

Таблица 1. Основные метрики на валидации.

эффект собственно послойной организации временных масштабов от эффекта случайной инициализации.

§4. Результаты и обсуждение

4.1. Основные результаты предварительного обучения. Таблица 1 показывает итоговые метрики на валидационном наборе данных после 2000 шагов обучения, усредненные по 15 независимым запускам для каждой стратегии. Убывающая стратегия дает наилучшее среднее значение Eval Loss и наибольшую Eval Token Accuracy, тогда как возрастающая стратегия стабильно оказывается худшей. Для сравнения убывающей и равномерной инициализации различия статистически значимы: для Eval Loss p-value составляет 0.0198, а для Eval Token Accuracy - 0.0142. Возрастающая стратегия, напротив, ухудшает Eval Loss по сравнению с базовой стратегией (p-value 0.0426).

Перед попарными сравнениями была выполнена предварительная проверка применимости параметрических методов. Для выборок итоговых метрик нормальность распределения не отвергалась: характерные значения p-value в тесте Д'Агостино–Пирсона составляли порядка 0.5–0.7. Это не выявило явных нарушений предпосылки нормальности, хотя при размере выборки 15 запусков на стратегию такая проверка имеет ограниченную мощность.

На рисунках 1 и 2 показана динамика обучения. Убывающая инициализация не только дает лучший финальный результат, но и демонстрирует более быструю сходимость, особенно в первой половине обучения. Возрастающая стратегия отстает от двух других практически на всем интервале.

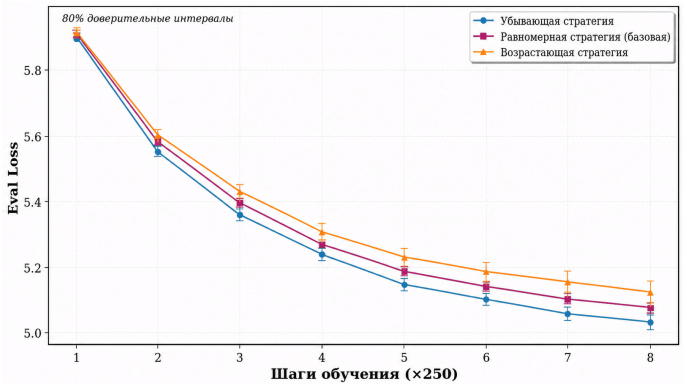

Рис. 1. Динамика среднего значения Eval Loss.

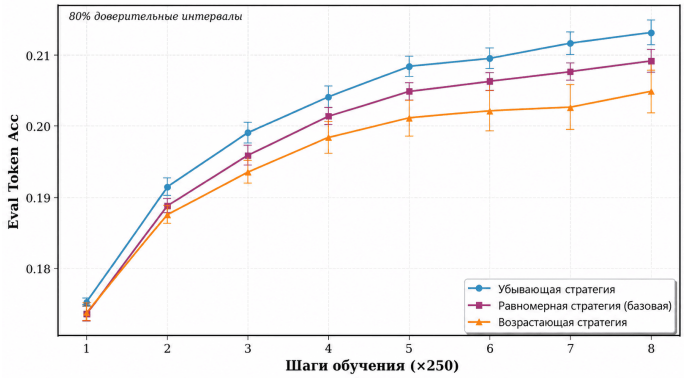

Рис. 2. Динамика среднего значения Eval Token Accuracy.

С точки зрения попарных сравнений результаты можно суммировать следующим образом. Убывающая стратегия статистически значимо превосходит равномерную по обеим основным метрикам. Равномерная стратегия, в свою очередь, статистически значимо превосходит возрастающую по Eval Loss, тогда как по Eval Token Accuracy различие находится на границе значимости (p-value 0.0509). В целом результаты указывают на преимущество убывающей стратегии и на

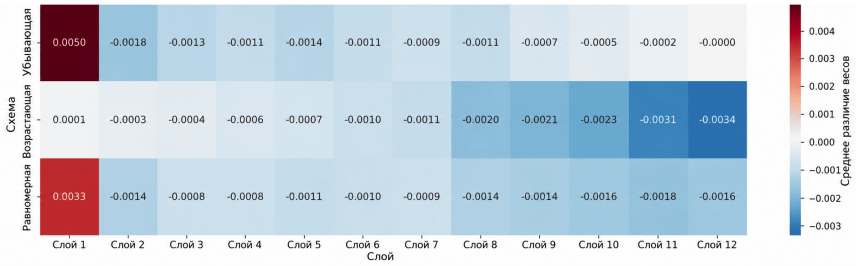

Рис. 3. Разность средних значений Δ_{bias} по слоям до и после обучения для различных стратегий инициализации.

худшее качество возрастающей инициализации по сравнению с двумя другими стратегиями.

4.2. Дополнительные проверки. Чтобы не ограничиваться только финальными метриками на валидационном наборе данных после 2000 шагов, мы рассмотрели дополнительные проверки. Во-первых, был проанализирован сдвиг значений Δ_{bias} по слоям в ходе обучения. Во-вторых, модели были оценены на внешних наборах данных. Эти проверки важны, потому что позволяют понять, сохраняется ли эффект инициализации после оптимизации параметров и переносится ли он за пределы обучающего корпуса.

4.2.1. *Изменение Δ_{bias} в процессе обучения.* На рисунке 3 показана разность средних значений Δ_{bias} до и после обучения. Как для убывающей, так и для равномерной стратегий наиболее заметные изменения наблюдаются на первом слое. При этом для двух лучших стратегий - убывающей и равномерной - среднее значение Δ_{bias} на первом слое имеет тенденцию к росту. Для возрастающей стратегии, напротив, наименьшие изменения сосредоточены в верхних слоях. Это наблюдение согласуется с предположением, что начальное распределение временных масштабов влияет на последующую динамику обучения, а не исчезает сразу после нескольких шагов оптимизации.

Кроме того, характер этой динамики подсказывает естественное направление для будущей работы: помимо линейно убывающей стратегии имеет смысл проверять и более гибкие, например степенные или экспоненциальные профили распределения Δt по слоям.

	TinyStories		**SQuAD**		**Wiki-EN**	
Стратегия	Mean	SE	Mean	SE	Mean	SE
Равномерная	92.2309	1.3437	219.7879	3.2814	279.9258	4.0847
Убывающая	93.2548	2.0674	209.3949	4.8139	267.7820	6.2275
Возрастающая	98.2137	2.3015	229.2305	6.5919	305.1938	10.6265

Таблица 2. Perplexity на внешних наборах данных после предварительного обучения.

4.2.2. *Оценка на внешних наборах данных.* Дополнительная оценка на внешних корпусах позволяет проверить, сохраняется ли наблюдаемая тенденция вне распределения данных предварительного обучения. В таблице 2 приведены средние значения perplexity и стандартные ошибки для трех стратегий. На TinyStories различия между убывающей и равномерной стратегиями малы. На SQuAD и Wiki-EN убывающая стратегия демонстрирует более низкую perplexity, что указывает на лучшую переносимость наблюдаемой тенденции на внешние корпуса.

На наборе TinyStories нельзя сделать вывод о статистически значимом преимуществе убывающей стратегии над равномерной (p-value 0.3406). В то же время возрастающая стратегия приводит к статистически значимому ухудшению perplexity по сравнению с равномерной (p-value 0.0164) и оказывается хуже убывающей на границе значимости (p-value 0.06).

На наборе SQuAD убывающая стратегия дает статистически значимое улучшение perplexity по сравнению как с равномерной инициализацией (p-value 0.0426), так и с возрастающей стратегией (p-value 0.0109).

Для Wiki-EN убывающая стратегия статистически значимо лучше возрастающей (p-value 0.003) и показывает улучшение по сравнению с равномерной стратегией (p-value 0.057), которое не достигает строгого порога значимости, но указывает на устойчивую тенденцию. Равномерная стратегия, в свою очередь, статистически значимо лучше возрастающей (p-value 0.0174).

Таким образом, внешняя оценка показывает, что эффект убывающей инициализации не является полностью однородным по корпусам:

	Eval Loss		Eval Token Accuracy	
Стратегия	Mean	SE	Mean	SE
Равномерная	4.9274	0.0221	0.2171	0.0019
Убывающая	4.8507	0.0234	0.2247	0.0018

Таблица 3. Результаты дополнительной проверки на большем количестве данных.

в одних случаях различия с равномерной стратегией малы, в других наблюдается более выраженное снижение perplexity. В среднем по внешним оценкам, где преимущество убывающей стратегии проявляется сильнее, снижение perplexity по сравнению с равномерной инициализацией составляет около 3–5%.

4.2.3. *Проверка на большем количестве данных.* Отдельно была проведена дополнительная серия экспериментов, в которой сравнивались только две стратегии - равномерная и убывающая - на большем количестве обучающих данных. Эта серия рассматривается как дополнительная проверка устойчивости эффекта и не заменяет основное сравнение трех стратегий, приведенное выше.

Согласно результатам этой серии, убывающая стратегия вновь оказалась лучше базовой равномерной. На валидационной выборке средний Eval Loss снизился с 4.9274 до 4.8507, что соответствует улучшению примерно на 1.56%. Одновременно Eval Token Accuracy выросла с 0.2171 до 0.2247, то есть примерно на 3.51% относительно базовой стратегии.

На рисунке 4 видно, что преимущество убывающей стратегии не сводится к финальной точке: кривая потерь лежит ниже практически на всем интервале обучения. Это важно, потому что согласуется с устойчивостью эффекта при более длительном или более насыщенном обучении и показывает, что предложенная инициализация остается полезной не только в компактной исследовательской постановке, но и в более ресурсоемком режиме.

Внешняя оценка для этой серии также согласуется с основной тенденцией. Для TinyStories perplexity уменьшается с 77.27 до 74.89, для SQuAD - с 213.18 до 201.87, а для Wiki-EN - с 244.99 до 228.19. Иными словами, при переходе к большему количеству данных преимущество

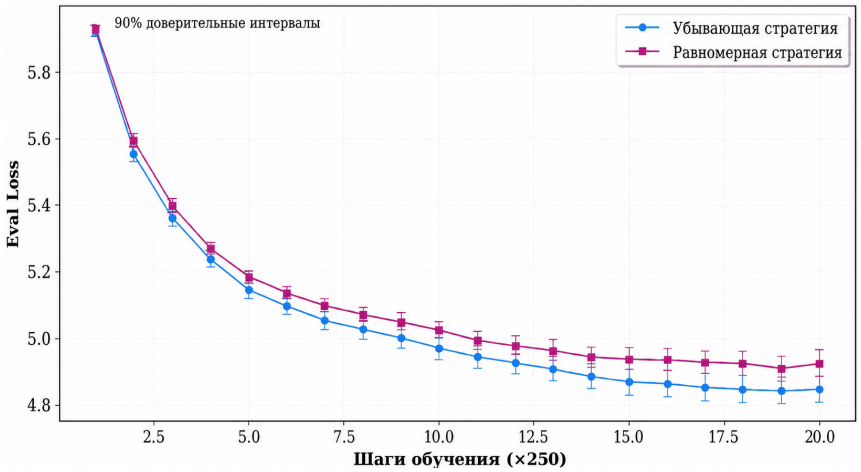

Рис. 4. Динамика Eval Loss в дополнительной серии экспериментов.

убывающей инициализации сохраняется и, по наблюдаемым метрикам, становится более выраженным.

4.3. Обсуждение. Полученные результаты согласуются с исходной гипотезой о целесообразности распределения временных масштабов по глубине модели. Если нижние слои с самого начала ориентированы на более локальную динамику, а верхние - на более глобальную, это может создавать более благоприятные начальные условия для оптимизации. Возрастающая инициализация задает противоположную структуру и, судя по результатам экспериментов, менее согласуется с естественной специализацией слоев. Важно, что наблюдаемый эффект достигается без увеличения числа параметров и без изменения архитектуры Mamba-2: изменяется только способ инициализации одного класса параметров.

§5. Заключение

В работе исследовано, как послойная инициализация параметра Δ_{bias} влияет на обучение Mamba-2. Основной результат состоит в том,

что убывающая по слоям стратегия, в которой нижние уровни получают большие значения Δt, а верхние - меньшие, в проведенных экспериментах оказывается предпочтительнее как равномерной инициализации, так и симметричной возрастающей стратегии.

Эксперименты на 45 запусках модели размером около 40M параметров свидетельствуют, что предложенная стратегия улучшает сходимость на этапе предварительного обучения и в ряде внешних оценок дает более низкую perplexity. При этом эффект достигается без изменений архитектуры и без увеличения числа параметров. С практической точки зрения это делает послойную инициализацию Δ_{bias} простым и дешевым способом улучшения моделей семейства Mamba в рассмотренном масштабе, заслуживающим проверки на более крупных моделях.

Ограничения исследования связаны с относительно небольшим размером модели, ограниченным объемом обучения и рассмотрением только трех простых стратегий инициализации. Следующий естественный шаг состоит в проверке метода на более крупных моделях и в изучении нелинейных стратегий распределения временных масштабов по глубине сети.

Список литературы

1. A. Gu and T. Dao, *Mamba: Linear-Time Sequence Modeling with Selective State Spaces*, ArXiv preprint arXiv:2312.00752 (2024).
2. T. Dao and A. Gu, *Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality*, ArXiv preprint arXiv:2405.21060 (2024).
3. A. Gu, I. Johnson, A. Timalsina, A. Rudra, and C. Ré, *How to Train Your HiPPO: State Space Models with Generalized Orthogonal Basis Projections*, ArXiv preprint arXiv:2206.12037 (2022).
4. A. Gu, K. Goel, and C. Ré, *Efficiently Modeling Long Sequences with Structured State Spaces*, in: International Conference on Learning Representations (ICLR), 2022.
5. S. Subramanian, R. Collobert, M. A. Ranzato, and Y.-L. Boureau, *Multi-Scale Transformer Language Models*, ArXiv preprint arXiv:2005.00581 (2020).
6. A. Gu, A. Gupta, K. Goel, and C. Ré, *On the Parameterization and Initialization of Diagonal State Space Models*, ArXiv preprint arXiv:2206.11893 (2022).
7. A. Trockman, H. Harutyunyan, J. Z. Kolter, S. Kumar, and S. Bhojanapalli, *Mimetic Initialization Helps State Space Models Learn to Recall*, ArXiv preprint arXiv:2410.11135 (2024).
8. OpenWebText dataset, `https://huggingface.co/datasets/Skylion007/openwebtext`.

9. Wikitext-en-de dataset, https://huggingface.co/datasets/LeoLM/wikitext-en-de.
10. SQuAD dataset, https://huggingface.co/datasets/rajpurkar/squad.
11. TinyStories dataset, https://huggingface.co/datasets/roneneldan/TinyStories.

Laetin A., Pulfer B., Smirnov S. Hierarchical initialization of temporal scales in Mamba-2 for Multi-Scale sequence modeling.

This work studies the effect of layer-wise initialization of the Δ_{bias} parameter in the Mamba-2 architecture on language modeling quality. Three initialization strategies are considered: decreasing, increasing, and uniform. The hypothesis under test is that the decreasing strategy, in which lower layers are initialized with larger Δt and upper layers with smaller ones, promotes the formation of multi-scale information processing without changing the model architecture. The experiments are carried out on a Mamba-2 model of about 40M parameters in a pre-training regime on the OpenWebText corpus, with subsequent evaluation on the TinyStories, SQuAD, and Wiki-EN datasets. In the conducted experiments, the decreasing initialization improved convergence compared with the uniform strategy and, in a number of external evaluations, yielded lower perplexity, whereas the increasing strategy degraded quality in most comparisons. The proposed approach does not increase the number of parameters and can be regarded as a simple way to improve the training efficiency of Mamba-family models.

Поступило 4 февраля 2026 г.

СПбГУ, Санкт-Петербург, Россия

Женевский университет, Женева, Швейцария

Женевский университет,
Женева, Швейцария;
СПбГУ, Санкт-Петербург, Россия

E-mail: a.laetin@spbu.ru

E-mail: Brian.Pulfer@unige.ch

E-mail: sksmirnov@gmail.com