ОРИГИНАЛЬНАЯ СТАТЬЯ

DOI: 10.26794/3033-7097-2025-1-4-35-42
УДК 004.85(045)

CC BY 4.0

Динамическая модель внимания в трансформерах

В.Б. Гисин
Финансовый университет при Правительстве Российской Федерации, Москва, Российская Федерация

АННОТАЦИЯ

Механизм внимания является основой трансформеров, ключевого компонента современных искусственных нейронных сетей, используемых при работе с данными различной природы. **В статье изучается** динамическая модель механизма внимания. В рамках этой модели внимание описывается как движение взаимодействующих токенов. Показано, что при подходящих предположениях внимание непрерывно по Липшицу. В частности, непрерывность по Липшицу обеспечивает нормирование токенов. Это служит основанием для исследования решений систем дифференциальных уравнений, описывающих динамику трансформеров. **Целью исследования** является изучение особенностей поведения токенов, составляющих промт, при неограниченном увеличении числа слоев трансформера. В одномерном случае приведено качественное описание траекторий токенов и динамики матрицы внимания. Показано, что если токен в некоторый момент времени выходит за границу достаточно узкого коридора (ширины порядка логарифма размера промта), то этот токен в дальнейшем стремится к бесконечности (положительной или отрицательной в зависимости от того, через какую границу произошел выход). Методология исследования базируется на непрерывной параметризации матрицы внимания. Распространенное представление динамики трансформеров разностными уравнениями заменено представлением с помощью систем обыкновенных дифференциальных уравнений. Описанию и изучению трансформеров посвящено огромное число публикаций, но большинство из них не содержат точных математических описаний архитектуры. **В этой статье сделана** попытка дать математически точное и при этом достаточно простое описание динамики трансформеров. Динамика токенов в одномерном случае, безусловно, значительно проще, чем динамика многомерных токенов. Тем не менее она дает представление о поведении трансформеров и в более общей ситуации создает каркас из точных формулировок.
Ключевые слова: искусственный интеллект; нейронная сеть; трансформер; механизм внимания; траектория; взаимодействие токенов

Для цитирования: **Гисин В.Б. Динамическая модель внимания в трансформерах.** ***Цифровые решения и технологии искусственного интеллекта.*** **2025;1(4):35-42. DOI: 10.26794/3033-7097-2025-1-4-35-42**

ORIGINAL PAPER

Dynamic Model of Attention in Transformers

V.B. Gisin
Financial University under the Governement of the Russian Federation, Moscow, Russian Federation

ABSTRACT

The attention mechanism is a key component of modern artificial neural networks designed to process data of various nature. **The article examines** the dynamic of attention using a continuous model. In this model, attention is described as the movement of interacting tokens. It is shown that, under suitable assumptions, attention is Lipschitz continuous. In particular, Lipschitz continuity may be ensured by token normalization. The dynamics of transformers is modelled by a system of differential equations. Lipschitz continuity guarantees that there exists a solution to this system. **The purpose of the study** is to investigate the behavior of tokens that make up promt under an unlimited increasing in the number of transformer layers. For one-dimensional tokens, a qualitative description of the trajectories of tokens and the dynamics of the attention matrix is given. It is shown that if a token goes beyond a fairly narrow corridor at some point (the width is on the order of the logarithm of the promt size), this token tends to infinity (positive or negative, depending on which border the exit occurred). The research methodology is based on continuous parameterization of the attention matrix. The common representation of transformer dynamics by difference equations has been replaced by a representation using systems of ordinary differential equations. A huge number of publications are devoted to the description and study of transformers, but most of them do not contain accurate mathematical descriptions of architecture. **This article attempts** to give a mathematically meaningful and at the same time fairly simple description of attention. The description dynamics of 1-d tokens is certainly much simpler than the dynamics of multidimensional tokens. Nevertheless, this description gives an idea of the behavior of transformers in a more general situation creates a framework for future investigation.
Keywords: artificial intelligence; neural network; transformer; mechanism of attention; trajectory; interaction of tokens

For citation: **Gisin V.B. Dynamic model of attention in transformers.** ***Digital Solutions and Artificial Intelligence Technologies.*** **2025;1(4):35-42. DOI: 10.26794/3033-7097-2025-1-4-35-42**

© Гисин В.Б., 2025

ВВЕДЕНИЕ

За последние несколько лет произошел невероятный прогресс в обработке естественного языка и в искусственном интеллекте. Модели последних поколений состоят из глубоких нейронных сетей. Их сложная архитектура основана на трансформерах с механизмами внимания [1]. Обучение на огромных по объему и разнородных по составу наборах данных позволило этим моделям достичь беспрецедентных результатов [2, 3].

Несмотря на их определяющую роль в успехе моделей искусственного интеллекта, трансформеры остаются изученными лишь частично. Технологии, связанные с искусственным интеллектом в общем и трансформерами в частности бурно развиваются [4–7]. Архитектура трансформеров совершенствуется и адаптируется к новым задачам [8, 9]. С учетом этого представляется полезным взгляд на трансформеры с точки зрения математики. Такой взгляд позволяет зафиксировать математические идеи, общие для трансформеров разной архитектуры, и наметить рамки, в которых эмпирическое изучение трансформеров может получить математическое обоснование. Одной из (пока отдаленных) целей этого подхода является создание объяснимых моделей искусственного интеллекта [10].

Целью настоящей статьи является создание математически точного и при этом достаточно простого описания динамики трансформеров.

ОСНОВА РАБОТЫ ТРАНСФОРМЕРОВ

Определяющим для трансформеров является механизм внимания [1, 3]. Он позволяет трансформерам обрабатывать не один входной вектор (токен) $x(0)$, а набор токенов

$$(x_1(0),\ldots,x_N(0))\in(R^D)^N,$$

где D — размерность токена; N — число токенов, обрабатываемых трансформером.

Такие наборы называют промтами. При работе с естественным языком каждый токен соответствует слову, а вся последовательность — предложению или другому фрагменту текста. Таким образом промт представляет собой слова вместе с их контекстом.

Слои трансформера $t = 1, 2, \ldots, T$ последовательно преобразуют промт в выходной набор токенов

$$\left(x_1(T),\ldots x_N(T)\right)\in\left(R^D\right)^N.$$

Представим набор токенов $\left(x_1(t),\ldots x_N(t)\right)\in\left(R^D\right)^N$ матрицей $X(t)\in R^{N\times D}$, в которой строка $X_i(t)$ соответствует токену $x_i(t)\in R^D$, так что $X_i(t)=x_i^T(t)$. Тогда работу трансформера можно представить как последовательность отображений пространства $R^{N\times D}$ в себя. Основой трансформера служит механизм самонаблюдения (self-attention), который корректирует координаты токена в зависимости от контекста: относительно усиливаются те признаки, которые в большей степени связаны с контекстом.

Преобразование матрицы X, выполняемое слоем трансформера, имеет вид

$$\Phi(X)=\left[\Phi^1(X),\ldots,\Phi^H(X)\right]W^O, \qquad (1)$$

где H — делитель D; $\Phi^h(X)\in R^{N\times D/H}$ при всех $h = 1, \ldots, H$, и $W^O\in R^{D\times D}$.

Преобразования $X\mapsto\Phi^h(X)$ выполняются наблюдателями (в англоязычной литературе используется термин «head»):

$$\Phi^h(X)=P^hXV^h, \qquad (2)$$

где $P^h\in R^{N\times N}$ — матрица «сходства» токенов (вообще говоря, не симметричная), а $V^h\in R^{N\times D/H}$ — матрица значений, получаемых при обучении.

В простейшем случае, когда W^O — единичная матрица, и при одном наблюдателе соотношение (1) может выглядеть следующим образом:

$$\Phi(X)=Softmax\left[XX^T\right]X \qquad (3)$$

(опущены указания на слой t).

В этом случае

$$P_{ij}=\frac{\exp x_i,x_j}{\sum_{j'=1}^{N}\exp x_i,x_{j'}}. \qquad (4)$$

Более гибкое самонаблюдение получается, если $H>1$ и

$$P^h=Softmax\left[\frac{XQ^h\left(XK^h\right)^T}{\sqrt{D/H}}\right],$$

где Q^h (запросы) и K^h (ключи) — матрицы размерности $D \times D/H$, формируемые при обучении.

Чтобы учесть взаимное расположение токенов в последовательности, используется механизм позиционирования. К вектору x_i добавляется вектор r_i, содержащий информацию о месте i токена x_i в промте. Один из получивших распространение подходов использует векторы r_i, такие что

$$r_{i,m}=\sin\left(\frac{i}{L^{m/D}}\right)$$

для четных значений m, и

$$r_{i,m}=\cos\left(\frac{i}{L^{(m-1)/D}}\right)$$

для нечетных $m = 1, \ldots, D$, где L — достаточно большое число.

При таком подходе модель обучается ориентироваться по относительному расположению токенов [11]. Позиционирование токенов позволяет рассматривать промт как неупорядоченный набор векторов. Поскольку информация о позиции токена включается в токен, она учитывается при обучении.

Как уже было отмечено, отображение, задаваемое трансформером, является функцией не отдельного входного сигнала, а (упорядоченного) набора D-мерных токенов. Эти токены эволюционируют во времени, взаимодействуя друг с другом в соответствии с механизмом самонаблюдения [12].

НЕПРЕРЫВНАЯ МОДЕЛЬ МЕХАНИЗМА ВНИМАНИЯ

Следуя работе [13], можно рассматривать токены как частицы, а динамику трансформера — как систему взаимодействующих частиц, которая описывается уравнениями вида

$$\dot{x}_i(t)=\sum_{j=1}^{N}\frac{\exp\beta\langle Q(t)x_i(t),K(t)x_j(t)\rangle}{Z_i(t)}V(t)x_j(t), \quad (5)$$

где

$$Z_i(t)=\sum_{l=1}^{N}\exp\beta\langle Q(t)x_i(t),K(t)x_l(t)\rangle \quad (6)$$

нормирующий множитель.

Следуя распространенной практике, будем считать, что $\beta=1/\sqrt{D/H}$.

Уравнение (5) позволяет рассматривать самонаблюдение как нелинейный механизм взаимодействия в системе токенов. Коэффициенты в уравнении (5) соответствуют относительному вниманию, которое токен i уделяет токену j. В частности, токен обращает внимание на своих «соседей», а «соседство» определяется матрицами Q и K.

При таком подходе «классические» блоки механизма самонаблюдения

$$x_i\leftarrow x_i+\sum_{j=1}^{N}\frac{\exp\beta\langle Qx_i,Kx_j\rangle}{Z_i(t)}Vx_j,$$

аппроксимируют решение уравнения (5).

Для трансформера с H наблюдателями уравнение (5) приобретает следующий вид:

$$\dot{x}_i(t)=\sum_{h=1}^{H}\sum_{j=1}^{N}\frac{\exp\beta\langle Q^h(t)x_i(t),K^h(t)x_j(t)\rangle}{Z_i^h(t)}V^h(t)x_j(t). \quad (7)$$

УСЛОВИЯ ЛИПШИЦА

Решение уравнения (5) или (7) может рассматриваться как своеобразная компонента объяснения функционирования трансформера. Для существования решения уравнения (7) достаточно, чтобы функция в правой части уравнения удовлетворяла условию Липшица. В общем случае это уже не так.

Чтобы не усложнять обозначения, рассмотрим в качестве примера случай, когда $D = 1$, $N = 2$ и $H = 1$. Соответственно, $Q, K, V \in R$.

Пусть $x_1, x_2 \in R$ — токены. Тогда $X=(x_1,x_2)^T$, а $\dot{x}_i=\frac{V}{Z_i(t)}\left(\exp(QKx_ix_1)x_1+\exp(QKx_ix_2)x_2\right)$.

Далее P — матрица второго порядка, такая, что

$$P_{11}=\frac{e^{\alpha x_1^2}}{e^{\alpha x_1^2}+e^{\alpha x_1x_2}},\quad P_{12}=\frac{e^{\alpha x_1x_2}}{e^{\alpha x_1^2}+e^{\alpha x_1x_2}},$$

$$P_{21}=\frac{e^{\alpha x_1x_2}}{e^{\alpha x_1x_2}+e^{\alpha x_2^2}},\quad P_{22}=\frac{e^{\alpha x_2^2}}{e^{\alpha x_1x_2}+e^{\alpha x_2^2}},$$

где $\alpha=QK$. Положим,

$$f_1(X)=P_{11}x_1+P_{12}x_2,$$

$$f_2(X)=P_{21}x_1+P_{22}x_2.$$

Тогда уравнение (7) приобретает следующий вид:

$$\begin{pmatrix}\dot{x}_1\\ \dot{x}_2\end{pmatrix}=\begin{pmatrix}f_1(X)\\ f_2(X)\end{pmatrix}V.$$

Если функция

$$X\mapsto f(X)=\begin{pmatrix}f_1(X)\\ f_2(X)\end{pmatrix}$$

из R^2 в R^2 удовлетворяет в некоторой области условию Липшица, то частные производные $\frac{\partial f_i}{\partial x_j}$ должны быть ограничены.

В качестве примера найдем $\frac{\partial f_1}{\partial x_1}$ (можно заметить, что $f_1: R^D\to R^D$, так что якобиан $\frac{\partial f_1}{\partial x_1}$ представляет собой квадратную матрицу порядка D).

Имеем:

$$\frac{\partial f_1}{\partial x_1}=$$

$$=P_{11}+\alpha P_{11}x_1\left[x_1-(P_{11}x_1+P_{12}x_2)\right]+$$

$$+\alpha\left[(P_{11}x_1^2+P_{12}x_2^2)-(P_{11}x_1+P_{12}x_2)^2\right].$$

Последнее слагаемое представляет собой дисперсию единственного признака, который имеют оба токена и принимающий значение x_1 с вероятностью P_{11} и значение x_2 с вероятностью P_{12}.

В общем случае несложные, но довольно громоздкие вычисления показывают, что

$$\frac{\partial f_1}{\partial x_1} = P_{11}\cdot I_D + P_{11}\left[x_1 - \sum_{k=1}^{N} P_{1k}x_k\right]x_1^T QK^T + \\ + \left[X^T CX\right]KQ^T,$$

где I_D — единичная матрица порядка D, а матрица C имеет следующий вид:

$$C_{jj} = P_{1j} - P_{1j}^2,\ C_{ij} = -P_{1i}P_{1j} \text{ при } I \neq j$$

(чтобы не усложнять обозначений, опущено указание на номер наблюдателя h).

Следовательно,

$$\left[X^T CX\right]_{kl} = \left(x_{1k},\ldots,x_{Nk}\right)C\left(x_{1l},\ldots x_{Nl}\right)^T = \\ = \sum_{j=1}^{N}\left(P_{1j} - P_{1j}^2\right)x_{jk}x_{jl} - \sum_{i,j=1,i\neq j}^{N} P_{1i}P_{1j}\,x_{ik}x_{jl} = \\ = \sum_{j=1}^{N} P_{1j}x_{jk}x_{jl} - \sum_{i,j=1}^{N} P_{1i}P_{1j}\,x_{ik}x_{jl} = \\ = \sum_{j=1}^{N} P_{1j}x_{jk}x_{jl} - \left(\sum_{j=1}^{N} P_{1j}x_{jk}\right)\cdot\left(\sum_{j=1}^{N} P_{1j}x_{jl}\right).$$

Таким образом, $\left[X^T CX\right]_{kl}$ — ковариация признаков k и l, которые принимают с вероятностью P_{1j} значения, соответственно, x_{jk} и x_{jl}, x_{jk}, $j = 1, \ldots, N$.

Чтобы не усложнять обозначений, вычисления приведены для f_1 и x_1.

Вычисление $\frac{\partial f_i}{\partial x_j}$ для других индексов осуществляется так же.

Если ковариация какой-либо пары признаков или дисперсия какого-то одного признака может быть неограниченно большой, то условие Липшица может не выполняться. Например, если $x_1 = 0$, то, как легко заметить, $P_{1j} = 1/N$ для всех $j = 1, \ldots, N$, и ковариация признаков оказывается просто выборочной дисперсией этих признаков. В то же время умножение на матрицу V может «погасить» признаки с большой дисперсией. С учетом этого заключение теоремы 3.1 в работе [14] оказывается верным лишь при выполнении указанных выше условий.

В приложениях часто применяется нормализация токенов [15]. В этом случае можно считать, что векторы $x_i(t)$ находятся на единичной сфере, так что $X(t) \in \left(S^{D-1}\right)^N$. Можно показать, что при этих условиях отображение $X \mapsto f(X)$ удовлетворяет условию Липшица.

Это можно проиллюстрировать для случая $D = 2$, $N = 2$ при единичных матрицах Q, K и V.

В самом деле, для $X = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix} \in S^2$

имеем $XX^T = \begin{pmatrix} 1 & \cos\varphi \\ \cos\varphi & 1 \end{pmatrix}$,

где φ — угол между векторами x_1 и x_2. Соответственно,

$$P = \frac{1}{e + e^{\cos\varphi}}\begin{pmatrix} e & e^{\cos\varphi} \\ e^{\cos\varphi} & e \end{pmatrix},$$

и

$$PX = \frac{1}{e + e^{\cos\varphi}}\begin{pmatrix} x_{11}e + x_{21}e^{\cos\varphi} & x_{12}e + x_{22}e^{\cos\varphi} \\ x_{21}e + x_{11}e^{\cos\varphi} & x_{22}e + x_{12}e^{\cos\varphi} \end{pmatrix}.$$

Матрица Y получается делением элементов первой строки матрицы PX на длину первого вектора, второй строки — на длину второго вектора. В частности, имеем:

$$y_{11} = \frac{e^2 + e^{1+\cos\varphi}}{e^2 + 2e^{1+\cos\varphi}\cos\varphi + e^{2\cos\varphi}}.$$

Далее

$$\frac{\partial y_{11}}{\partial x_{11}} = \\ = \frac{e^{1+\cos\varphi}\sin\varphi\left(e^2 + 4e^{1+\cos\varphi} + 2e^2\cos\varphi + e^{2\cos\varphi}\right)}{\left(e^2 + 2e^{1+\cos\varphi}\cos\varphi + e^{2\cos\varphi}\right)^2}x_{12}. \quad (8)$$

Так как наименьшее значение знаменателя дроби (8) равно $e^{-2}+e^2-2$, а компонента x_{12} вектора на единичной сфере не превосходит по абсолютной величине единицу, производная $\frac{\partial y_{11}}{\partial x_{11}}$ ограничена.

Точно так же оказываются ограниченными остальные компоненты якобиана.

При нормализации промту $X(t) \in \left(S^{D-1}\right)^N$ можно сопоставить вероятностную меру на S^{D-1}, сосредоточенную в токенах x_i:

$$\mu(t,\cdot) = \frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}(\cdot),$$

где $\delta_{x_i}(x) = 1$ при $x = x_i$ и $\delta_{x_i}(x) = 0$ при $x \neq x_i$ Тогда динамика трансформера может быть описана в терминах мер на S^{D-1} [15].

ТРАЕКТОРИИ ТОКЕНОВ

Динамика $X(t)$ может быть достаточно сложной. Далее рассмотрим упрощенный вариант, когда размерность токена равна единице, так что промт представляет собой набор чисел

$X = (x_1, \ldots, x_N) \in R^N$. Дополнительно будем предполагать, что $QK = 1$ и $V = 1$. Эти предположения не принципиальны, но упрощают выкладки и позволяют более наглядно описать динамику трансформера. Под динамикой трансформера имеется в виду не только (и не столько) траектория токенов $x_i(t)$, но и поведение матрицы

$$P(t) = \left(P_{ij}(t)\right) \in R^{N \times N}.$$

Динамика трансформера $X(t)$ с одномерными токенами описывается системой обыкновенных дифференциальных уравнений:

$$\dot{x}_i = \sum_{j=1}^{N} P_{ij} x_j; \qquad P_{ij} = \frac{e^{x_i x_j}}{\sum_{k=1}^{N} e^{x_i x_k}}. \qquad (9)$$

Покажем сначала, что траектории токенов не сближаются. В самом деле,

$$\frac{d}{dt}\left(x_i(t) - x_j(t)\right)^2 = 2\left(x_i(t) - x_j(t)\right)\left(\dot{x}_i(t) - \dot{x}_j(t)\right).$$

Заметим, что $\dot{x}_i(t) = f'\left(x_i(t)\right)$,

где $$f(x) = \ln\left(\sum_{j=1}^{N} e^{x \cdot x_j}\right).$$

Тогда

$$\frac{d}{dt}\left(x_i(t) - x_j(t)\right)^2 = 2\left(x_i(t) - x_j(t)\right)\left(f'(x_i) - f'(x_j)\right).$$

Функция $f(x)$ выпукла, поэтому $x_i(t) - x_j(t)$ и $f'(x_i) - f'(x_j)$ имеют один и тот же знак, и, значит,

$$\frac{d}{dt}\left(x_i(t) - x_j(t)\right)^2 \geq 0.$$

Следовательно, величина $\left|x_i(t) - x_j(t)\right|$ не убывает при любых $i, j = 1, \ldots, N$.

Если $x_i(0) = x_l(0)$, система уравнений (9) не изменится при перестановке токенов с номерами i и l, а траектории $x_i(t)$ и $x_l(t)$ будут одинаковыми. Поэтому, не ограничивая общности, можно считать, что все токены разные. Позиционирование позволяет включить информацию о взаимном расположении токенов в «содержимое» токена. С учетом этого можно считать, что токены занумерованы в порядке возрастания их значений в начальный момент времени:

$$x_1(0) < x_2(0) < \ldots < x_N(0).$$

Оценим $x_N(t)$. Покажем, что при достаточно больших положительных значениях $x_N(t)$ производная $x_N(t)$ растет не менее быстро, чем $x_N(t)$, при $x_N(t) \to \infty$.

Имеем

$$\dot{x}_N(t) = \sum_{j=1}^{N-1} \frac{e^{x_N(t)x_j(t)}}{\sum_{k=1}^{N} e^{x_N(t)x_k(t)}} x_j(t) + \frac{e^{x_N^2(t)}}{\sum_{k=1}^{N} e^{x_N(t)x_k(t)}} x_N(t). \qquad (10)$$

Второе слагаемое в правой части уравнения (10) можно переписать в виде $\frac{1}{Z_N} x_N(t)$, где

$$Z_N(t) = \sum_{k=1}^{N} e^{-x_N(t)\left(x_N(t) - x_k(t)\right)}.$$

Так как $x_N(0) - x_k(0) > 0$ при $k < N$ и величина $\left(x_N(0) - x_k(0)\right)^2$ не убывает, $x_N(t) - x_k(t) > 0$ для всех t. Рассмотрим случай, когда $x_N(t) \geq 0$. Тогда

$$e^{-x_N(t)\left(x_N(t) - x_k(t)\right)} \leq 1,$$

и, значит,

$$Z_N(t) = \sum_{k=1}^{N-1} e^{-x_N(t)\left(x_N(t) - x_k(t)\right)} + 1 \leq (N-1) + 1 = N.$$

Следовательно,

$$\frac{e^{x_N^2(t)}}{\sum_{k=1}^{N} e^{x_N(t)x_k(t)}} x_N(t) = \frac{1}{Z_N} x_N(t) \geq \frac{1}{N} x_N(t).$$

Оценим первое слагаемое в правой части уравнения (10). Отбрасывая неотрицательные слагаемые, получаем:

$$\sum_{j=1}^{N-1} \frac{e^{x_N(t)x_j(t)}}{\sum_{k=1}^{N} e^{x_N(t)x_k(t)}} x_j(t) \geq \sum_{x_j(t) < 0} \frac{e^{x_N(t)x_j(t)}}{e^{x_N^2(t)} Z_N(t)} x_j(t).$$

Так как $Z_N(t) > 1$, при $x_j(t) < 0$ имеем:

$$\frac{e^{x_N(t)x_j(t)}}{e^{x_N^2(t)} Z_N(t)} x_j(t) =$$

$$= -\frac{e^{-x_N(t)\left|x_j(t)\right|}}{e^{x_N^2(t)} Z_N(t)} \left|x_j(t)\right| \geq -\frac{x_N(t) \cdot \left|x_j(t)\right|}{x_N(t) \cdot e^{x_N(t) \cdot \left|x_j(t)\right|}}.$$

Но $\frac{a}{e^a} < 1$ для любого $a > 0$. Поэтому

$$-\frac{x_N(t) \cdot \left|x_j(t)\right|}{x_N(t) \cdot e^{x_N(t)\left|x_j(t)\right|}} \geq -\frac{1}{x_N(t)}.$$

Окончательно получаем:

$$\sum_{j=1}^{N-1}\frac{e^{x_N(t)x_j(t)}}{\sum_{k=1}^{N}e^{x_N(t)x_k(t)}}x_j(t)\geq-\frac{N}{x_N(t)}$$

и

$$\dot{x}_N(t)\geq\frac{x_N(t)}{N}-\frac{N}{x_N(t)}. \qquad (11)$$

Таким образом, если $x_N(t_0) > N$ в какой-то момент времени t_0, то в дальнейшем $x_N(t)$ устремится к $+\infty$ экспоненциально быстро.

Точно так же, если $x_1(t_0) < -N$ в какой-то момент времени t_0, то в дальнейшем $x_1(t)$ экспоненциально быстро устремится к $-\infty$. Это утверждение получается просто переменой знаков токенов:

$$-x_N(0)<-x_{N-1}(0)<\ldots<-x_1(0).$$

При этом уравнения (10) не меняются.

Замечание. Оценку (11) несложно существенно уточнить. В самом деле, если $x_i(t) > 0$ для некоторого $i = 1,\ldots,N$, то $x_N(t) > 0$. Для j такого, что $x_i(t) < 0$, получаем:

$$\frac{e^{x_i(t)x_j(t)}}{\sum_{k=1}^{N}e^{x_i(t)x_k(t)}}x_j(t)=$$

$$=-\frac{e^{-x_i(t)|x_j(t)|}}{e^{x_i(t)x_N(t)}\sum_{k=1}^{N}e^{-x_i(t)(x_N(t)-x_k(t))}}|x_j(t)|\geq$$

$$\geq-\frac{|x_j(t)|\cdot e^{-x_i(t)|x_j(t)|}}{e^{x_i(t)x_N(t)}}.$$

Далее

$$-\frac{|x_j(t)|\cdot e^{-x_i(t)|x_j(t)|}}{e^{x_i(t)|x_j(t)|}}=$$

$$=-\frac{e^{-x_i(t)x_N(t)}}{x_i(t)}\cdot\frac{x_i(t)|x_j(t)|}{e^{x_i(t)|x_j(t)|}}\geq$$

$$\geq-\frac{e^{-x_i(t)x_N(t)}}{x_i(t)}.$$

Таким образом,

$$\dot{x}_i(t)=\sum_{j=1}P_{ij}(t)x_j(t)\geq$$

$$\geq\frac{1}{N}x_N(t)-\frac{N-1}{e^{x_i(t)x_N(t)}}\cdot\frac{1}{x_i(t)}. \qquad (12)$$

В частности,

$$\dot{x}_N(t)\geq\frac{1}{N}x_N(t)-\frac{N-1}{e^{x_N^2(t)}}\cdot\frac{1}{x_N(t)}.$$

Это означает, что $x_N(t)$ экспоненциально быстро стремится к бесконечности.

Пусть v_0 — решение уравнения

$$\frac{v}{N}-\frac{N-1}{e^{v^2}}\cdot\frac{1}{v}=0. \qquad (13)$$

Заметим, что v_0 растет с увеличением N не быстрее, чем $\sqrt{2\ \ln N}$.

Неравенство (12) показывает, что аналогичное рассуждение применимо к любому токену. Таким образом, для токенов возможны три типа траекторий: траектория может находиться внутри некоторого коридора; если траектория выходит из коридора через верхнюю границу, она устремляется к $+\infty$; если траектория выходит из коридора через нижнюю границу, она устремляется к $-\infty$.

Коридор, о котором идет речь, находится внутри полосы $[-v_0, v_0]\times R_{\geq 0}$ плоскости (x, t), где v_0 — положительное решение уравнения (13).

Рассмотрим теперь поведение элементов матрицы $P(t)$. Предположим, что $i < N$ и $x_i(t)\to+\infty$ при $t\to\infty$. В этом случае также и $x_N(t)\to+\infty$. Кроме того, если $c > 0$ таково, что $c < x_N(0) - x_{N-1}(0)$, то $x_N(t) - x_j(t) > c$ для любого $j < N$.

Следовательно,

$$P_{ij}(t)=\frac{e^{x_i(t)x_j(t)}}{\sum_{k=1}^{N}e^{x_i(t)x_k(t)}}=\frac{e^{-x_i(t)(x_N(t)-x_j(t))}}{\sum_{k=1}^{N}e^{-x_i(t)(x_N(t)-x_k(t))}}\leq e^{-cx_i(t)}$$

при $j < N$. Таким образом, $P_{ij}(t)\to 0$ при $t\to\infty$. Соответственно,

$$P_{iN}=1-\sum_{k=1}^{N-1}P_{ij}\to 1.$$

Аналогичным образом, если $i > 1$ и $x_i(t)\to-\infty$ при $t\to\infty$, то $P_{ij}(t)\to 0$ для $j > 1$ и $P_{i1}(t)\to 1$.

Согласно работе [16], трансформеры могут рассматриваться как универсальные аппроксиматоры дифференциальных уравнений. С этой точки зрения переход к пределу при $t\to\infty$ соответствует неограниченному увеличению числа слоев. Таким образом, увеличение числа слоев трансформера может привести к вырождению матрицы внимания, когда все внимание будет сосредоточиваться на паре признаков с экстремальными значениями в промте.

В статье рассмотрена простейшая одномерная ситуация. Анализ проблемы во всей ее общности — задача дальнейших исследований.

ВЫВОДЫ

Несмотря на центральную роль трансформеров в успехе моделей искусственного интеллекта, теоретические основы, управляющие работой трансформеров, остаются изученными лишь частично. В статье предпринята попытка представить теоретическое осмысление механизмов внимания с использованием математического аппарата «непрерывной» математики. Информационный поток между слоями трансформера естественно описывать с помощью разностных уравнений. Использование в статье дифференциальных уравнений в определенной степени упрощает анализ. В то же время этот анализ остается корректным с учетом аппроксимирующих свойств нейронных сетей. Моделирование механизма внимания с помощью дифференциальных уравнений позволяет говорить об этом механизме в терминах развитой математической теории. Например, естественным образом возникает вопрос об асимптотическом поведении и т.п. Поиск ответов на подобные вопросы может оказаться сложным, но ясно и точно поставленные общие вопросы намечают пути решения проблем.

Когда речь идет об искусственном интеллекте, баланс общности и привязки к конкретной архитектуре в значительной степени должен обусловливаться объяснимостью. С этой точки зрения представляются оправданными сделанные в статье упрощающие предположения относительно размерности и т.п. Используя несложный математический аппарат, они позволяют понять некоторые особенности механизма внимания и динамики трансформеров, которые могут быть использованы при изучении современных моделей искусственного интеллекта во всей их сложности. Дальнейшее исследование пока необъяснимой эффективности трансформеров может быть связано с поиском некоторых общих принципов их работы, сформулированных в рамках развитых математических теорий.

REFERENCES

1. Vaswani A., Shazeer, N., Parmar N., Uszkoreit J., Jones L., Gomez A.N. Kaiser Ł., Polosukhin I. Attention is all you need. In: Guyon I., Von Luxburg U., S. Bengio, et al, eds. *Neural Information Processing Systems*. 2017;30:5998–6008. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
2. Rambelli G., Chersoni E., Testa D., Blache, P., Lenci A. Neural generative models and the parallel architecture of language: A critical review and outlook. *Topics in Cognitive Science*. 2024;17(4):948–961. DOI: 10.1111/tops.12733
3. Turner R.E. An introduction to transformers. *ArXiv preprint*. 2023; arXiv:2304.10557. DOI: 10.48550/arxiv.2304.10557
4. Amatriain X., Sankar A., Bing J., Bodigutla P.K., Hazen T.J., Kazi M. Transformer models: an introduction and catalog. *ArXiv preprint*. 2023; arXiv:2302.07730. DOI: 10.48550/arXiv.2302.07730
5. He S., Sun G., Shen Z., Li A. What matters in transformers? Not all attention is needed. *ArXiv preprint*. 2024; arXiv:2406.15786. DOI: 10.48550/arXiv.2406.15786
6. Passi N., Raj M., Shelke N.A. A review on transformer models: applications, taxonomies, open issues and challenges. 4th Asian Conference on Innovation in Technology (ASIANCON). IEEE, 2024;1–6. DOI: 10.1109/ASIANCON 62057.2024.10838047
7. Joshi S. Evaluation of Large Language Models: Review of Metrics, Applications, and Methodologies. *Preprint*. 2025; DOI: 10.20944/preprints202504.0369.v1
8. Sajun A.R., Zualkernan I., Sankalpa D. A historical survey of advances in transformer architectures. *Applied Sciences*. 2024;14(10):4316. DOI: 10.3390/app14104316
9. Canchila S., Meneses-Eraso C., Casanoves-Boix J., Cortés-Pellicer P., Castelló-Sirvent F. Natural Language Processing: An Overview of Models, Transformers and Applied Practices. *Computer Science and Information Systems.* 2024;21(3):1097–1145. DOI: 10.2298/CSIS 230217031C
10. Ali A., Schnake T., Eberle O., Montavon G., Müller K.R., Wolf L. XAI for transformers: Better explanations through conservative propagation. International Conference on Machine Learning. Proceedings of Machine Learning Research (PMLR). 2022;435–451. DOI: 10.48550/arXiv.2202.07304
11. Dufter P., Schmitt M., Schütze H. Position information in transformers: An overview. *Computational Linguistics*. 2022;48(3):733–763. DOI: 10.1162/coli_a_00445
12. Geshkovski B. Letrouit C., Polyanskiy Y., Rigollet P. The emergence of clusters in self-attention dynamics. *Advances in Neural Information Processing Systems*. 2023;36:57026–57037. DOI: 10.48550/arXiv.2305.05465
13. Sander M.E., Ablin P., Blondel M., & Peyré G. Sinkformers: Transformers with doubly stochastic attention. International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, 2022:3515–3530. DOI: 10.48550/arXiv.2110.11773

14. Kim H., Papamakarios G., Mnih A. The Lipschitz constant of self-attention. International Conference on Machine Learning. Proceedings of Machine Learning Research. 2021;5562–5571. DOI: 10.48550/arXiv.2006.04710
15. Geshkovski B., Letrouit C., Polyanskiy Y., Rigollet P. A mathematical perspective on transformers. *Bulletin of the American Mathematical Society*. 2025;62(3):427–479. DOI: 10.1090/bull/1863
16. Lu Y., Li Z., He D., et al. Understanding and improving transformer from a multi-particle dynamic system point of view. *ArXiv preprint*. 2019; arXiv:1906.02762. DOI: 10.48550/arXiv.1906.02762

ИНФОРМАЦИЯ ОБ АВТОРАХ / ABOUT THE AUTHORS

Владимир Борисович Гисин — кандидат физико-математических наук, профессор, профессор кафедры математики и анализа данных факультета информационных технологий и анализа больших данных, Финансовый университет при Правительстве Российской Федерации, Москва, Российская Федерация
Vladimir B. Gisin — Cand. Sci. (Phys. and Math.), Professor of the Mathematics and Data Analysis Department of the Faculty of Information Technology and Big Data Analysis, Financial University under the Government of the Russian Federation, Moscow, Russian Federation
https://orcid.org/0000-0002-7269-0587
vgisin@fa.ru

Конфликт интересов: автор заявляет об отсутствии конфликта интересов.
Conflicts of Interest Statement: The author has no conflicts of interest to declare.

Статья поступила 13.10.2025; принята к публикации 24.11.2025.
Автор прочитал и одобрил окончательный вариант рукописи.
The article was received on 13.10.2025; accepted for publication on 24.11.2025.
The author read and approved the final version of the manuscript.