ВЕСТНИК
НОВОСИБИРСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА

Научный журнал
Основан в ноябре 1999 года

Серия: Информационные технологии 2025. Том 23, № 4

СОДЕРЖАНИЕ

ISSN 1818-7900 (Print). ISSN 2410-0420 (Online)
Вестник НГУ. Серия: Информационные технологии. 2025. Том 23, № 4
Vestnik NSU. Series: Information Technologies, 2025, vol. 23, no. 4

V E S T N I K
NOVOSIBIRSK STATE UNIVERSITY

Scientific Journal
Since 1999, November
In Russian

Series: Information Technologies 2025. Volume 23, № 4

CONTENTS

ISSN 1818-7900 (Print). ISSN 2410-0420 (Online)
Вестник НГУ. Серия: Информационные технологии. 2025. Том 23, № 4
Vestnik NSU. Series: Information Technologies, 2025, vol. 23, no. 4

The journal is published quarterly in Russian since 1999
by Novosibirsk State University Press

The address for correspondence
Faculty of Information Technologies, Novosibirsk State University
1 Pirogov Street, Novosibirsk, 630090, Russia
Tel. +7 (383) 363 42 46

E-mail address: inftech@vestnik.nsu.ru
On-line version: http://elibrary.ru

ISSN 1818-7900 (Print). ISSN 2410-0420 (Online)
Вестник НГУ. Серия: Информационные технологии. 2025. Том 23, № 4
Vestnik NSU. Series: Information Technologies, 2025, vol. 23, no. 4

Научная статья

УДК 519.6, 517.95
DOI 10.25205/1818-7900-2025-23-4-5-22

Численное решение коэффициентной обратной задачи электроимпедансной томографии с использованием лабораторных измерений

Анна Александровна Афанасьева
Александр Васильевич Старченко

Томский государственный университет
Томск, Россия

anna.afanaseva@stud.tsu.ru
starch@math.tsu.ru

Аннотация

Представлен итерационный численный метод решения обратной коэффициентной задачи для однородного эллиптического уравнения с интегро-дифференциальными граничными условиями в замкнутой области. Метод опирается на конечно-объемные аппроксимации дифференциальных и интегральных операторов на неструктурированных сетках, численное решение последовательности прямых задач при известном кусочно-постоянном распределении коэффициентов разностного эллиптического уравнения и сходящийся итеративно регуляризованный метод Гаусса – Ньютона. Разработанный метод решения обратных задач электроимпедансной томографии прошел тестирование на измерениях электрического напряжения, выполненных на экспериментальном стенде KIT в университете Восточной Финляндии. Получены близкие к реальным результатам реконструкции электрической проводимости внутри области исследования.

Ключевые слова

коэффициентная обратная задача, уравнение эллиптического типа с кусочно-постоянными коэффициентами, интегро-дифференциальное граничное условие, метод конечного объема, неструктурированные сетки, полная электродная модель, реконструкция проводимости, итеративно регуляризованный метод Гаусса – Ньютона

Финансирование

Исследование выполнено при финансовой поддержке Министерства науки и высшего образования РФ (проект развития региональных математических центров).

Для цитирования

Афанасьева А. А., Старченко А. В. Численное решение коэффициентной обратной задачи электроимпедансной томографии с использованием лабораторных измерений // Вестник НГУ. Серия: Информационные технологии. 2025. Т. 23, № 3. С. 5–22. DOI 10.25205/1818-7900-2025-23-4-5-22

Numerical solution of the coefficient inverse problem of electrical impedance tomography using laboratory measurements

Anna A. Afanaseva, Alexander V. Starchenko

Tomsk State University
Tomsk, Russian Federation

anna.afanaseva@stud.tsu.ru
starch@math.tsu.ru

Abstract

An iterative numerical method for solving the inverse coefficient problem for a uniform elliptic equation with integro-differential boundary conditions in a closed domain is presented. The method relies on finite-volume approximations of differential and integral operators on unstructured grids, numerical solution of a sequence of direct problems with a known piecewise constant distribution of coefficients of a difference elliptic equation, and the convergent iteratively regularizable Gauss-Newton method. The developed method for solving inverse problems of electrical impedance tomography has been tested on electrical voltage measurements performed at the KIT experimental stand at the University of Eastern Finland. The results of reconstruction of electrical conductivity within the research area are close to the real ones.

Keywords

coefficient inverse problem, elliptic equation with piecewise constant coefficients, integro-differential boundary condition, finite volume method, unstructured grids, complete electrode model, conduction reconstruction, iteratively regularized Gauss-Newton method

Funding

The work was carried out with the financial support of the Ministry of Science and Higher Education of the Russian Federation (Project for the development of regional mathematical centers).

For citation

Afanaseva A. A., Starchenko A. V. Numerical solution of the coefficient inverse problem of electrical impedance tomography using laboratory measurements. *Vestnik NSU.* Series: *Information Technologies*, 2025, vol. 23, no. 4, pp. 5–22 (in Russ.) DOI 10.25205/1818-7900-2025-23-4-5-22

Введение

Электроимпедансная томография (ЭИТ) – это метод компьютерной визуализации внутренней структуры изучаемого объекта по полученному распределению значений коэффициента электропроводности объекта на основе измерений тока и напряжения на поверхности объекта с помощью прикрепленных электродов [1–3]. Механизм работы ЭИТ включает следующие шаги (рис. 1):

1. На поверхность исследуемого объекта (например, тело человека) накладывается специальная сетка из электродов. Количество электродов может варьироваться в зависимости от размера исследуемой области и требуемой разрешающей способности изображения внутренней структуры. Электроды должны обеспечивать хороший контакт с кожей для обеспечения точности измерений. Для улучшения контакта часто используют электропроводящий гель.

2. Через электроды пропускается слабый, безопасный для человека электрический ток низкой частоты. Сила тока тщательно контролируется и находится в пределах, безопасных для организма. Пациент не ощущает никакого дискомфорта во время прохождения тока.

3. При прохождении тока через ткани объекта сопротивление измеряется между парами электродов. Различные ткани обладают разным электрическим сопротивлением (импедансом), что зависит от их электропроводности, которая, в свою очередь, определяется содержанием воды, ионов и других компонентов. Например, легкие, заполненные воздухом, имеют высокое сопротивление (низкие значения электрической проводимости), а ткани органов, богатых водой, обладают низким сопротивлением (высокие значения электрической проводимости).

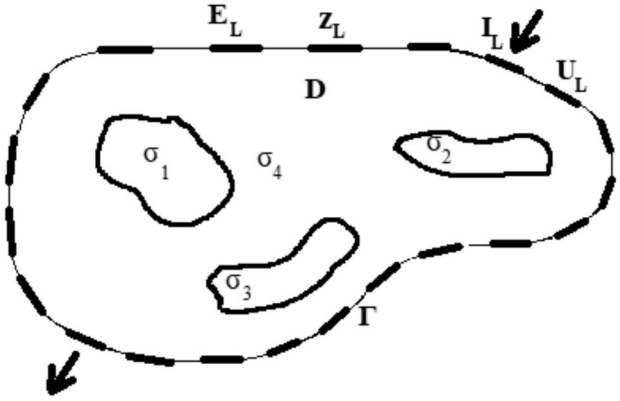


Рис. 1. Модель объекта, на которую нанесены электроды, с внутренними неоднородностями, σ – электрическая проводимость, $\{z_L\}$ – сопротивление на электродах, $\{I_L\}$ – электрический ток, $\{U_L\}$ – измеренное напряжение, E_L – размер электрода, L – количество электродов

Fig. 1. Model of the object on which the electrodes are applied, with internal inhomogeneities, σ – electrical conductivity, $\{z_L\}$ – resistance at the electrodes, $\{I_L\}$ – electrical current, $\{U_L\}$ – measured voltage, E_L – size of the electrode, L – number of electrodes

4. Измеренные на электродах значения электрического тока и напряжения обрабатываются с помощью специальных численных алгоритмов. Эти алгоритмы позволяют реконструировать изображение исследуемой области, отображающее распределение электрического импеданса тканей. Полученное изображение может быть визуализировано на экране монитора, подобно изображениям, получаемым с помощью других методов компьютерной томографии или магнитно-резонансной томографии [2; 3].

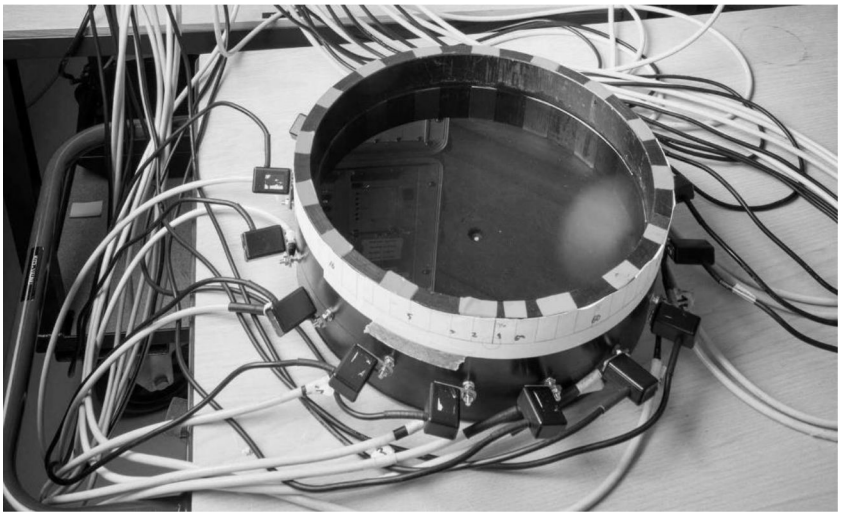


Рис. 2. Система измерения KIT [6]

Fig. 2. The KIT4 measurement system [6]

Реконструкция распределения электрической проводимости внутри области исследования обычно осуществляется путем решения коэффициентной обратной задачи ЭИТ, которая является сложной и многогранной проблемой теоретической и прикладной математики. Многие исследователи для ее решения обращаются к приближенным аналитическим и численным методам, однако для тестирования разработанных вычислительных технологий зачастую приме-

няются так называемые «синтетические данные», полученные в результате решения прямых задач ЭИТ для различных способов подключения электрического тока через электроды и даже накладывая псевдослучайные возмущения на вычисленные значения электродных напряжений [4; 5].

В настоящее время находятся в свободном доступе данные измерений напряжения на электродах при различных токовых конфигурациях, полученных на измерительной системе KIT (Kuopio Impedance Tomography) в университете Восточной Финляндии [6].

В измерительной системе KIT исследования проводились в резервуаре цилиндрической формы с диаметром 28 см (рис. 2). Внутренняя поверхность резервуара была оснащена шестнадцатью прямоугольными электродами из нержавеющей стали, каждый из которых имел размеры 7 см в высоту и 2,5 см в ширину. Электроды были расположены на равных расстояниях друг от друга, что способствовало симметричному распределению электрических полей внутри резервуара. Для удобства идентификации электроды были пронумерованы по часовой стрелке, начиная с верхнего, который обозначается как электрод 1. Исследования включали измерения разности электрического потенциала на электродах в различных условиях. Базовым экспериментом служили измерения с заполненным только водой резервуаром, что позволяет определить фоновое распределение электрической проводимости и оценить свойства самой измерительной системы, например, уровень шумов и паразитных емкостей. Далее проводились эксперименты с различными объектами, помещаемыми внутри заполненного водой резервуара: металлическое кольцо и пластиковые цилиндры разной формы. Металлическое кольцо, обладая высокой электропроводностью, должно существенно искажать электрическое поле, что позволяет оценить чувствительность системы к присутствию высокопроводящих объектов. Пластиковый цилиндр, наоборот, характеризуется низкой проводимостью, и его присутствие в резервуаре вызовет менее заметные, но все же регистрируемые изменения в распределении потенциала. В дополнение к этому исследовались более сложные конфигурации, такие как комбинация нескольких металлических колец или совокупность пластикового цилиндра и металлического кольца. Это позволяет изучить влияние взаимного расположения объектов на искажение электрического поля и оценить возможности системы KIT для решения обратной задачи – реконструкции формы и положения объектов внутри среды по измеренным данным [6].

Многие исследователи в своих работах использовали данную систему KIT для тестирования различных численных методов реконструкции изображений. В исследовании [4] рассматривается подход, основанный на глубоком обучении. Авторы предлагают обучать нейронную сеть для прямой реконструкции распределения проводимости по измеренным электрическим потенциалам на электродах. Обучение нейронной сети в работе [4] проводилось с использованием как синтетических данных, так и реальных экспериментальных данных, причем количество синтетических данных было существенно больше (>10000). Экспериментальные данные были собраны с помощью системы KIT [6], которая измеряла электрические потенциалы на электродах, расположенных по периметру резервуара. В качестве объектов исследования использовались различные фантомы – модели, имитирующие реальные объекты с различными электрическими свойствами и геометрическими формами. Рассматривались объекты из твердого пластика и полые металлические кольца, отличающиеся высокой электрической проводимостью. Дискретизация моделируемой области проводилась с высоким разрешением – на 1 696 треугольных ячейках, что обеспечивало достаточно детальное описание распределения проводимости, но, с другой стороны, требовало большого количества необходимых для реконструкции данных. Авторы [4] подчеркивают, что их подход обеспечивает быструю, устойчивую и качественную визуализацию распределения проводимости в исследуемой области. Заслуживает внимания тот факт, что, несмотря на обучение нейронной сети на данных, содержащих только круглые включения, она продемонстрировала способность к обобщению

и давала удовлетворительные результаты реконструкции для объектов более сложной формы, таких как треугольники и прямоугольники.

В работе [7] представлен итерационный метод решения обратной задачи ЭИТ с применением регуляризации Тихонова. Этот метод отличается своей универсальностью и может применяться к задачам, которые являются некорректными, т. е. к задачам, когда входных данных недостаточно или они существенно искажены. Одной из ключевых особенностей метода является влияние параметра регуляризации, который позволяет «сглаживать» решение и делать процесс его сходимости более устойчивым, что выражается в последовательном уменьшении нормы целевой функции и ее градиента. В [7] метод регуляризации Тихонова основывается на минимизации специального функционала, который включает в себя два основных компонента: меру близости рассчитанных напряжений к исходным (измеренным) данным и слагаемое регуляризации, пропорциональное L2-норме градиента решения. Это означает, что при выборе параметра регуляризации α необходимо найти баланс между точностью аппроксимации данных и гладкостью полученного решения. В работе [7] параметр регуляризации подбирался вручную для каждого эксперимента, что, хотя и позволило получить приемлемые результаты, не гарантирует оптимальность и объективность оценки. Авторы приводят визуальные сравнения. Полученные реконструированные изображения, несмотря на уверенное обнаружение включений/неоднородностей, страдали от размытости, что затрудняло точную оценку размеров и формы обнаруженных объектов. Также в [7] использовалась и l1-регуляризация, которая в некоторой степени повышает четкость изображений.

В исследовании [8] применяется метод быстрого приближенного вывода, основанный на распространении математического ожидания, для изучения апостериорного распределения вероятностей, возникающего в результате байесовской постановки нелинейных обратных задач. Этот метод применяется к решению обратной задачи электроимпедансной томографии в полной электродной постановке, используя измерения системы KIT [6]. Для численного решения задачи используется метод конечных элементов с кусочно-линейной аппроксимацией на сетке из 424 узлов и 750 треугольных элементов, локально уплотненной вблизи электродов, чтобы более точно описать измерения электрической проводимости в этой области. Проверка точности и эффективности метода проводилась на реальных данных, полученных при погружении пластиковых и металлических стержней в резервуар с водой [6]. Полученные результаты численного моделирования представлены в виде таблицы с относительными ошибками среднего значения и стандартного отклонения параметров.

Стоит подчеркнуть, что рассматриваемые подходы – итеративный метод и метод глубокого обучения – имеют свои преимущества и недостатки. Итеративные методы, как правило, более интерпретируемы и позволяют лучше понимать физические процессы, лежащие в основе решения обратной задачи. Однако они могут быть вычислительно более затратными и медленными, особенно для больших объемов данных. Методы глубокого обучения, напротив, могут быть более быстрыми и эффективными, но их «черный ящик» может затруднять понимание причин получаемых результатов и оценку их достоверности. Кроме того, качество реконструкции в методах глубокого обучения сильно зависит от качества и количества обучающих данных. Будущие исследования могут быть направлены на гибридные подходы, которые объединяют преимущества обоих методов, например, использование глубокого обучения для ускорения сходимости итеративных методов или использования итеративных методов для улучшения качества обучения нейронных сетей. Также важны исследования, направленные на повышение устойчивости методов к шуму и артефактам, которые неизбежно присутствуют в реальных экспериментальных данных [1–8].

Целью данной работы является применение разрабатываемого численного итеративно регуляризованного метода для решения обратной коэффициентной задачи ЭИТ на основе лабораторных измерений, выполненных с использованием измерительной системы KIT [6].

Остальная часть статьи организована следующим образом. Сначала подробно рассматривается главная составная часть метода решения обратной задачи ЭИТ – математическая постановка и численный метод решения прямой задачи ЭИТ на неструктурированных сетках. Приводятся результаты его тестирования на измерениях KIT. Затем описывается математическая постановка и численный метод решения обратной задачи ЭИТ, который опирается на итеративно регуляризованный метод Гаусса – Ньютона. Приводится блок-схема полного вычислительного процесса. Завершает статью раздел, в котором представлены результаты сравнения изображений численно реконструированной внутренней структуры с фотографиями реальных объектов и условий измерений. В заключении сформулированы основные результаты работы.

Математическая постановка и численный метод решения прямой задачи ЭИТ

Рассматривается двумерная область D с гладкой границей Г, которая отчасти либо соединяется с электродами для пропускания слабого электрического тока небольшой частоты Γ_l (l= 1, …, L – количество электродов), либо имеет общую границу Γ_{air} с непропускающей ток окружающей средой (см. рис. 1). В области $\bar{D} = D + \Gamma$ известно распределение электрической проводимости $\sigma(x, y) > 0$. Также известна используемая на электродах токовая конфигурация $\{I_l\}$ – способ подачи и приема электрического тока и значения силы тока на электродах. Внутренние источники тока отсутствуют. Магнитная напряженность пренебрежимо мала. Процесс стационарный.

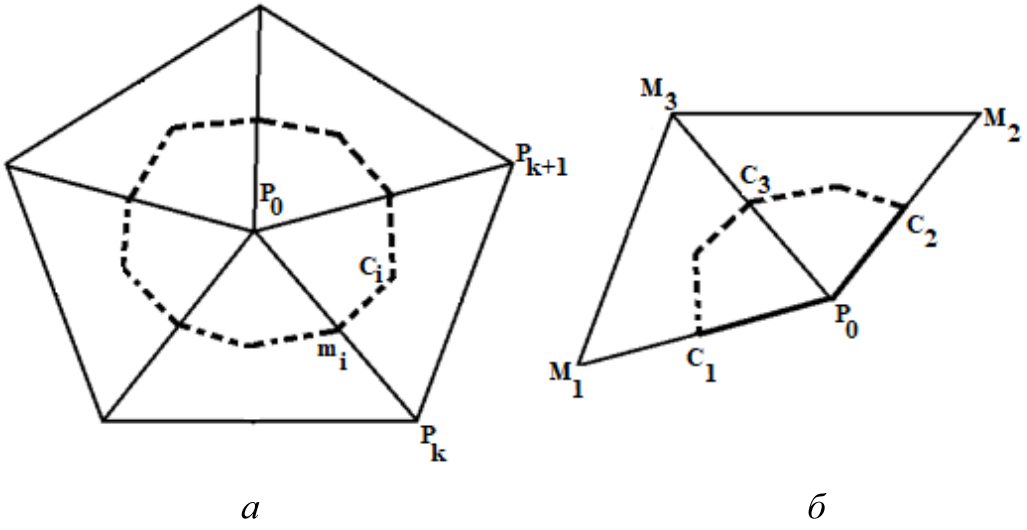
Для рассматриваемых условий из уравнений Максвелла и закона Ома для проводников математическая постановка задачи нахождения распределения электрического потенциала $u(x, y)$ в $\bar{D} = D + \Gamma$ и значений электрического напряжения $\{U_l\}$ на электродах может быть записана следующим образом [9; 10]:

$$\begin{cases} div\left(\sigma \cdot grad(u)\right) = 0, (x, y) \in D; \\ \left.\dfrac{\partial u}{\partial n}\right|_{\tilde{A}_{air}} = 0; \\ \left.\left(u + z_l \sigma \dfrac{\partial u}{\partial n}\right)\right|_{\Gamma_l} = \dfrac{1}{E_l}\left(\int\limits_{\Gamma_l} u\, ds + z_l I_l\right), l = 1, ..., L; \\ U_l = \dfrac{1}{E_l}\left(\int\limits_{\Gamma_l} u\, ds + z_l I_l\right); \int\limits_{\Gamma_l} \sigma \dfrac{\partial u}{\partial n} ds = I_l,\ l = 1, ..., L; \\ \sum\limits_{l=1}^{L} I_l = 0; \sum\limits_{l=1}^{L} U_l = 0; \end{cases} \quad (1)$$

В (1) $E_l > 0, z_l > 0$ – размеры и сопротивление (импеданс) l-го электрода. В работе [11] доказано, что такая постановка задачи обеспечивает получение единственного решения при определенных условиях на $u(x, y)$. На сегодняшний день такая постановка прямой задачи ЭИТ рассматривается как наиболее полно (с минимальной погрешностью) представляющая физические процессы передачи тока через биологические объекты [11; 12].

В данной работе численное решение прямой задачи ЭИТ (1) ищется с помощью метода конечного объема на неструктурированных сетках, покрывающих область $\bar{D}$. Построение неструктурированных сеток выполняется с помощью пакета Gambit, при этом особое внимание уделяется сгущению узлов сетки вблизи поверхности электродов Гl и границы контакта с окру-

жающей средой Γ_{air}. В качестве конечного объема рассматриваются барицентрические ячейки, окружающие каждый внутренний узел сетки (рис. 3).



а *б*

Рис. 3. Конечные объемы, использующиеся при получении разностной схемы: *а* – для внутреннего узла сетки; *б* – для граничного узла сетки

Fig. 3. The final volumes used in obtaining the difference scheme: *a* – for the inner node of the grid, *б* – for the boundary node of the grid

При построении разностной схемы предполагается, что в каждой треугольной ячейке сетки значение электрической проводимости $\sigma(x, y)$ постоянно, а приближенное распределение электрического потенциала можно представить билинейной функцией следующего вида: $v(x.y) = a_k + b_k x + c_k y$, где $k = 1, \ldots, NT$ – число треугольников неструктурированной сетки [13]. Значения постоянных величин $\{a_k, b_k, c_k\}$ могут быть выражены через приближенные значения потенциала в вершинах треугольника $P_0 P_k P_{k+1}$:

$$v^{(k)}(x,y) = v_{P_0}\Psi^{(k)}_{P_0}(x,y) + v_{P_k}\Psi^{(k)}_{P_k}(x,y) + v_{P_{k+1}}\Psi^{(k)}_{P_{k+1}}(x,y), \tag{2}$$

где $\Psi^{(k)}_{P_0}(x,y), \Psi^{(k)}_{P_k}(x,y), \Psi^{(k)}_{P_{k+1}}(x,y)$ – линейные базисные функции, причем они равны 1 только для своей вершины k-го треугольника, для остальных двух они равны 0. В связи с этим получается

$$\Psi^{(k)}_{P_0}(x,y) = \frac{1}{2S_k}\begin{vmatrix} 1 & x & y \\ 1 & x_{P_k} & y_{P_k} \\ 1 & x_{P_{k+1}} & y_{P_{k+1}} \end{vmatrix}, \qquad \Psi^{(k)}_{P_k}(x,y) = \frac{1}{2S_k}\begin{vmatrix} 1 & x_{P_0} & y_{P_0} \\ 1 & x & y \\ 1 & x_{P_{k+1}} & y_{P_{k+1}} \end{vmatrix},$$

$$\Psi^{(k)}_{P_{k+1}}(x,y) = \frac{1}{2S_k}\begin{vmatrix} 1 & x_{P_0} & y_{P_0} \\ 1 & x_{P_k} & y_{P_k} \\ 1 & x & y \end{vmatrix}, \qquad S_k = \frac{1}{2}\begin{vmatrix} 1 & x_{P_0} & y_{P_0} \\ 1 & x_{P_k} & y_{P_k} \\ 1 & x_{P_{k+1}} & y_{P_{k+1}} \end{vmatrix},$$

где S_k – площадь треугольника $\Delta P_0 P_k P_{k+1}$.

После интегрирования по конечному объему, внутри которого находится внутренний узел сетки P_0 (см. рис. 3), использования формулы Грина [14] и приближенного вычисления частных производных по x и по y с учетом (2) будет получена разностная схема следующего вида [13]:

$$\sum_{k=1}^{NT_0}\frac{\sigma_k}{4S_k}\Big[v_{P_0}\Big(\big(y_{P_k}-y_{P_{k+1}}\big)^2+\big(x_{P_{k+1}}-x_{P_k}\big)^2\Big)+$$
$$+v_{P_k}\Big(\big(y_{P_{k+1}}-y_{P_0}\big)\big(y_{P_k}-y_{P_{k+1}}\big)+\big(x_{P_0}-x_{P_{k+1}}\big)\big(x_{P_{k+1}}-x_{P_k}\big)\Big)+$$
$$+v_{P_{k+1}}\Big(\big(y_{P_0}-y_{P_k}\big)\big(y_{P_k}-y_{P_{k+1}}\big)+\big(x_{P_k}-x_{P_0}\big)\big(x_{P_{k+1}}-x_{P_k}\big)\Big)\Big]=0,\qquad P_0\in\omega_h. \tag{3}$$

NT_0 – количество треугольников в барицентрической ячейке с общей вершиной P_0. Суммирование выполняется по всем треугольным элементам сетки с общей вершиной P_0, находящейся внутри области D, причем, когда значение индекса $k+1$ становится больше NT_0, то нужно его взять равным 1.

Для узлов P_0, лежащих на границе, также строится конечный объем (см. рис. 3). Причем границы этого объема, находящиеся внутри области D, обрабатываются аналогично рассмотренным выше, а на криволинейных границах, совпадающих с Γ_l или Γ_{air}, напрямую используются граничные условия из интегро-дифференциальной постановки (1). Для приближенного вычисления интегралов по поверхности электродов от электрического потенциала используется формула трапеций, обеспечивающая второй порядок аппроксимации. В итоге для граничных узлов разностная схема будет выглядеть следующим образом [13]:

$$\sum_{k=1}^{NT_0}\frac{\sigma_k}{4S_k}\Big[v_{P_0}\Big(\big(y_{P_k}-y_{P_{k+1}}\big)^2+\big(x_{P_{k+1}}-x_{P_k}\big)^2\Big)+$$
$$+v_{P_k}\Big(\big(y_{P_{k+1}}-y_{P_0}\big)\big(y_{P_k}-y_{P_{k+1}}\big)+\big(x_{P_0}-x_{P_{k+1}}\big)\big(x_{P_{k+1}}-x_{P_k}\big)\Big)+$$
$$+v_{P_{k+1}}\Big(\big(y_{P_0}-y_{P_k}\big)\big(y_{P_k}-y_{P_{k+1}}\big)+\big(x_{P_k}-x_{P_0}\big)\big(x_{P_{k+1}}-x_{P_k}\big)\Big)\Big]+ \tag{4}$$
$$+\frac{j_1\big|C_1P_0\big|}{z_l}\left(\frac{1}{2E_l}\sum_{n=1}^{N_l-1}\big(v_{P_n}+v_{P_{n+1}}\big)\big|P_nP_{n+1}\big|+\frac{z_lI_l}{E_l}-\frac{v_{P_{k-1}}+3v_{P_0}}{4}\right)+$$
$$+\frac{j_2\big|P_0C_2\big|}{z_l}\left(\frac{1}{2E_l}\sum_{n=1}^{N_l-1}\big(v_{P_n}+v_{P_{n+1}}\big)\big|P_nP_{n+1}\big|+\frac{z_lI_l}{E_l}-\frac{3v_{P_0}+v_{P_{k+1}}}{4}\right)=0,P_0\in\gamma_h.$$

Здесь $j_1=1$, если $P_{k-1}P_0\in E_l$, иначе $j_l=0$; $j_2=1$, если $P_0P_{k+1}\in E_l$, иначе $j_2=0$; N_l – количество узлов сетки на электроде с номером l.

Для того чтобы строки системы линейных уравнений (3)–(4) были линейно независимы, необходимо одно из уравнений системы заменить на условие единственности получения решения [11] из (1):

$$\sum_{l=1}^{L}U_l=\sum_{l=1}^{L}\frac{1}{E_l}\left(\int_{\Gamma_l}u\,ds+z_lI_l\right)=0. \tag{5}$$

Приближенное представление (5) при использовании формулы трапеций для вычисления интегралов имеет вид

$$\sum_{l=1}^{L}\frac{1}{E_l}\left(\frac{1}{2}\sum_{n=1}^{N_l-1}\left(v_{P_n}+v_{P_{n+1}}\right)\left|P_nP_{n+1}\right|+z_lI_l\right)=0. \qquad (6)$$

Таким образом, чтобы найти приближенное распределение электрического потенциала, нужно решить систему (3), (4), в которой одно уравнение заменено на (6).

Для решения полученной системы линейных уравнений, имеющей несимметричную матрицу без диагонального преобладания, может быть использован метод Гаусса с выбором главного элемента [15] или итерационный метод BiCGstab [16].

Для тестирования рассмотренного метода численного решения прямой задачи ЭИТ были проведены расчеты для условий заполненной только водой емкости KIT для случая, когда через 5-й электрод подавался ток силой 2 мА, а 4-й электрод использовался для заземления. Для этого случая в исследовании [6] представлены результаты измерений напряжений на электродах (Case1-0 [6]). Значение электрической проводимости принято 0,0013 $Ом^{-1}\cdot м^{-1}$. Была построена треугольная сетка (рис. 4, *а*), состоящая из NT = 1200 треугольных элементов и N = 657 узлов. Сетка была уплотнена в областях, близких к границе, для повышения точности расчетов.

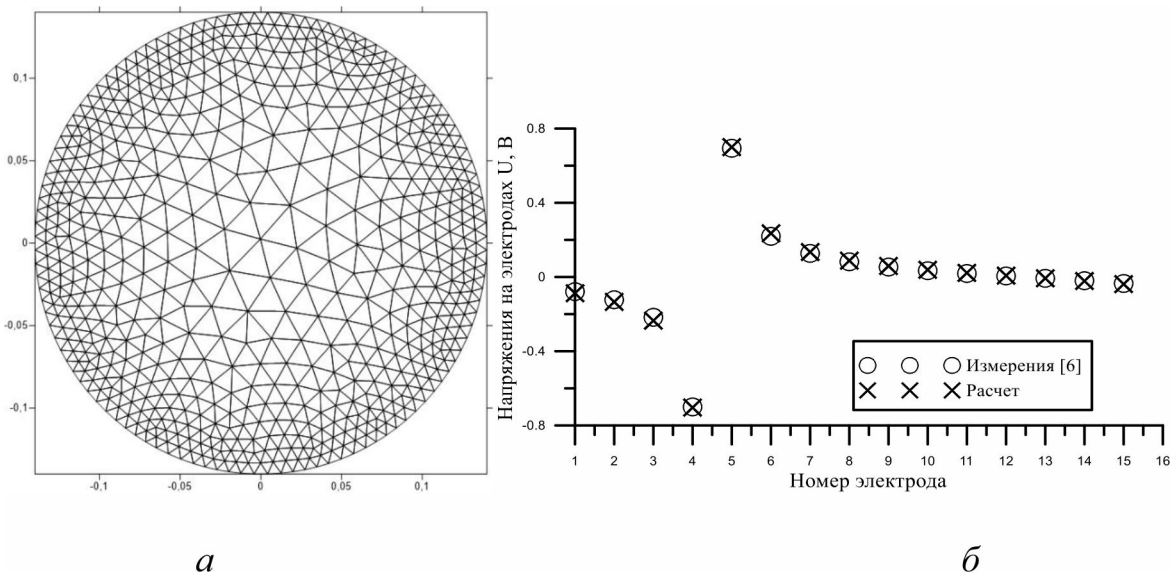


Рис. 4. Неструктурированная сетка, содержащая NT = 1200 треугольных ячеек и N = 657 узлов (*а*). Результаты сопоставления численного решения прямой задачи ЭИТ с измерениями (*б*) [6]
Fig. 4. An unstructured grid containing NT = 1200 triangular cells and N = 657 nodes (*a*). The results of comparing the numerical solution of the direct EIT problem with measurements (*б*) [6]

На рис. 4, *б* представлены вычисленные и измеренные значения напряжений на электродах. Средняя абсолютная ошибка составила 0,018 В. Как видно из рисунков, имеет место хорошее согласование расчетов и измерений. Применяя подход, аналогичный методу Рунге оценки главного члена погрешности на последовательности вложенных сеток [15], можно установить, что порядок аппроксимации построенной разностной схемы близок ко второму.

Математическая постановка и численный метод решения обратной задачи ЭИТ

Обратная задача ЭИТ предполагает, что известны размеры области $\overline{D}$, места крепления, размеры и сопротивления электродов, измеренные на электродах значения силы тока $\{I_l\}$ и на-

пряжений $\{U_l\}$, и требуется для постановки (1) или ее разностного аналога найти распределение электрической проводимости внутри области D. То есть на основе измерений осуществить визуализацию внутренней структуры исследуемого объекта. Эта обратная коэффициентная задача относится к некорректно поставленным задачам, потому что получаемые распределения $\sigma(x, y)$ в значительной степени зависят от погрешностей измеряемых величин. Для решения такого сорта нелинейных плохо обусловленных задач часто применяется метод регуляризации А. Н. Тихонова [17–19], суть которого заключается в нахождении приближенного решения системы нелинейных уравнений за счет добавления к условию минимизации невязки условия минимизации нормы решения.

В данной работе с помощью регуляризации А. Н. Тихонова ищется решение обратной задачи σ^*, которое представляется в виде минимума функционала [7; 20]:

$$\sigma^* = \arg\min_{\sigma\in\Sigma}\left\{\Phi_\alpha(\sigma) = \frac{1}{2}\sum_{\mu=1}^{M}\left\|\vec{U}^\mu(\sigma) - \vec{\tilde{U}}^\mu\right\|_2^2 + \frac{\alpha}{2}\left\|\sigma - \sigma_0\right\|_2^2\right\}, \tag{7}$$

где $\frac{\alpha}{2}\left\|\sigma - \sigma_0\right\|_2^2$ – регуляризирующий функционал; σ_0 – некоторое известное фоновое распределение проводимости; $\vec{\tilde{U}}^\mu = \left(\tilde{U}_1^\mu, \tilde{U}_2^\mu, ..., \tilde{U}_L^\mu\right)$ – измеренные с погрешностью значения напряжения для μ-й токовой конфигурации $\vec{I}^\mu = \left(I_1^\mu, I_2^\mu, ..., I_L^\mu\right)$, M – количество таких конфигураций; $\vec{U}^\mu(\sigma) = \left(U_1^\mu(\sigma), U_2^\mu(\sigma), ..., U_L^\mu(\sigma)\right)$ – рассчитанные из решения прямых задач (3) – (4) при некотором распределении σ для μ-й токовой конфигурации; $\Sigma = \left\{\sigma \in L^\infty(D) : c_0 \le \sigma \le c_1\right\}$ – допустимый набор электрической проводимости, где c_0, c_1 – известны.

Чтобы рассчитать минимум функционала, воспользуемся методом Гаусса – Ньютона, основанным на линеаризации. Представим $\vec{U}^\mu(\sigma) \approx \vec{U}^\mu(\sigma^0) + \vec{J}^\mu(\sigma^0)(\sigma - \sigma^0)$, где $\vec{J}^\mu(\sigma^0) = \nabla_\sigma \vec{U}^\mu(\sigma^0)$ – якобиан от $\vec{U}^\mu(\sigma)$ по σ при некотором, близком к искомому решению, начальном приближении σ^0. Подставляя это разложение в (7), получим линеаризованную задачу квадратичной минимизации:

$$\sigma^* = \arg\min_{\sigma\in\Sigma}\left\{\frac{1}{2}\sum_{\mu=1}^{M}\left\|\vec{J}^\mu(\sigma - \sigma^0) - \left(\vec{\tilde{U}}^\mu - \vec{U}^\mu(\sigma^0)\right)\right\|_2^2 + \frac{\alpha}{2}\left\|\sigma - \sigma_0\right\|_2^2\right\},$$

решение которой представляется следующим образом:

$$\left[\sum_{\mu=1}^{M}\vec{J}^{\mu T}(\sigma^0)\vec{J}^\mu(\sigma^0) + \alpha I\right](\sigma - \sigma^0) = -\left[\sum_{\mu=1}^{M}\vec{J}^{\mu T}(\vec{U}^\mu(\sigma^0) - \vec{\tilde{U}}^\mu) + \alpha(\sigma^0 - \sigma_0)\right]. \tag{8}$$

Система линейных уравнений (8) может быть решена прямым методом Холецкого [15], чтобы получить новую оценку для σ. Затем итеративно выполняется обновление реконструкции, принимая полученное решение в качестве первоначального предположения. На практике такая итеративная процедура обеспечивает сходимость в течение нескольких итераций [20], если дополнительно использовать монотонно уменьшающийся параметр регуляризации $\alpha_k > \alpha_{k+1} > 0$; $\lim_{k\to\infty}\alpha_k = 0$:

$$\sigma^{(k+1)} = \sigma^{(k)} - \left[\sum_{\mu=1}^{M}\vec{J}^{\mu T}(\sigma^{(k)})\vec{J}^\mu(\sigma^{(k)}) + \alpha_k I\right]^{-1}\left[\sum_{\mu=1}^{M}\vec{J}^{\mu T}(\vec{U}^\mu(\sigma^{(k)}) - \vec{\tilde{U}}^\mu) + \alpha_k(\sigma^{(k)} - \sigma_0)\right]. \tag{9}$$

Критерий остановки может быть основан на мониторинге относительного изменения таких величин, как $\Phi(\sigma^{(k)})$, $\sqrt{\frac{1}{M}\sum_{k=1}^{M}\left\|\vec{J}^{\mu}(\sigma^{(k)})\right\|_2^2}$ между последовательными итерациями.

На рис. 5 представлена блок-схема реализации итеративно регуляризованного метода Гаусса – Ньютона (9).

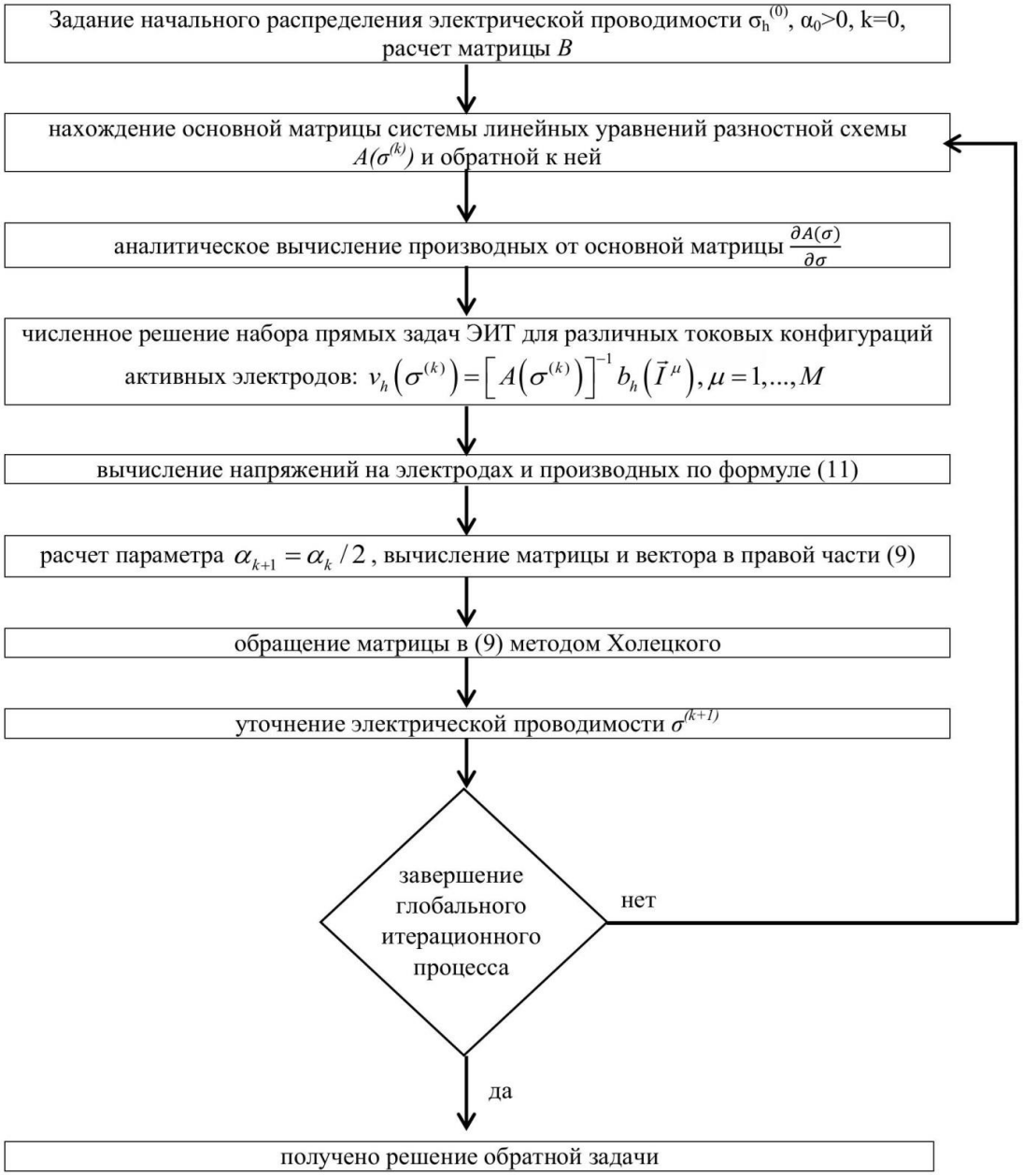


Рис. 5. Блок-схема итерационного процесса
Fig. 5. Block diagram of the iterative process

Заметим, что при вычислении значений якобиана можно воспользоваться следующим приемом [21]. В предыдущем разделе было показано, что при использовании выбранных вычис-

лительных технологий (метод конечного объема, барицентрические ячейки, кусочно-постоянные значения электрической проводимости и т. д.) получается система линейных уравнений (3), (4), (6), коэффициенты матрицы которой линейно зависят от электрической проводимости. Обозначим эту систему следующим образом:

$$A(\sigma)v_h = b_h\left(\vec{I}\right). \tag{10}$$

Искомые значения напряжений на электродах могут быть вычислены из формул (5) или (6) $\vec{U} = Bv_h + \vec{d}$, где вектор $\vec{d} = \left(\frac{z_1 I_1}{E_1}, \frac{z_2 I_2}{E_2}, ..., \frac{z_L I_L}{E_L}\right)$, а B – это матрица размером $L \times N$, в которой ненулевые элементы суть коэффициенты квадратурной формулы трапеций. Предполагая обратимость матрицы $A(\sigma)$, можно записать:

$$\vec{U}(\sigma) = B\left[A(\sigma)\right]^{-1} b_h\left(\vec{I}\right) + \vec{d}. \tag{11}$$

Тогда

$$\vec{J}(\sigma) = \frac{\partial \vec{U}(\sigma)}{\partial \sigma} = B\frac{\partial\left[A(\sigma)\right]^{-1}}{\partial \sigma} b_h\left(\vec{I}\right) = B\frac{\partial\left[A(\sigma)\right]^{-1}}{\partial \sigma} A(\sigma)v_h = -B\frac{\partial A(\sigma)}{\partial \sigma}\left[A(\sigma)\right]^{-1} v_h. \tag{12}$$

То есть в (12) производные от коэффициентов матрицы $A(\sigma)$ могут быть вычислены аналитически, а вторые производные и производные более высокого порядка будут равны нулю, что позволяет говорить о точной линеаризации $\vec{U}^{\mu}(\sigma) = \vec{U}^{\mu}(\sigma^0) + \vec{J}^{\mu}(\sigma^0)(\sigma - \sigma^0)$, $\mu = 1, \ldots, M$.

Результаты решения обратной задачи по лабораторным данным KIT

В этом разделе приводится реализация рассмотренного выше итеративно регуляризованного метода Гаусса – Ньютона с экспериментальными данными, которые были получены в университете Восточной Финляндии (Куопио) с использованием измерительной системы KIT [6] с $L = 16$ электродами. Для проведения измерений ЭИТ было использовано в общей сложности 79 парных токовый конфигураций. Они были разделены на пять наборов [6]:

Набор 1: Смежные токовые конфигурации. Инъекции с помощью электродов 1–2, 2–3, …, 15–16, 16–1.

Набор 2: Пропуск одного электрода. Инъекции с помощью электродов 1–3, 2–4, …, 14–16, 15–1, 16–2.

Набор 3: Пропуск двух электродов. Инъекции с помощью электродов 1–4, 2–5, …, 13–16, 14–1, …, 16–3.

Набор 4: Пропуск трех электродов. Инъекции с помощью электродов 1–5, 2–6, …, 12–16, 13–1, …, 16–4.

Набор 5: Все против одного. Инъекции с помощью электрода №1 и каждым из остальных l, где $l = 2, \ldots, 16$.

В каждой токовой конфигурации один электрод использовался для подачи тока, а другой – в качестве заземления. Амплитуда тока составляла 2 мА, т. е. действующее значение тока будет 1,41 мА. В соответствии с каждой инъекцией тока измерялись напряжения между всеми соседними электродами: 1–2, 2–3, …, 15–16, 16–1, что в итоге дало $79 \cdot 16 = 1\,264$ измерений.

Резервуар измерительной системы был наполнен водопроводной водой, и внутри него размещались цилиндрические предметы различных форм и материалов, таких как сталь и пластик. В данной работе рассматривалось несколько случаев [6]: случай 1 (Case1_0 [6]) – резервуар, заполненный исключительно водой; случай 2 (Case1_1 [6]) – в резервуар добавлялся один цилиндр из пластика; случай 3 (Case1_2 [6]) – добавлялся один полый стальной цилиндр; случай 4 (Case4_3 [6]) – один пластиковый цилиндр и один полый стальной цилиндр; случай 5 (Case3_4 [6]) – с тремя полыми стальными цилиндрами различных размеров.

Для решения обратной задачи необходимо знать значение коэффициента электрической проводимости воды σ_0 и значения сопротивлений электродов $\{z_l\}$. Для определения этих данных использовались измерения электродных напряжений для случая 1, когда резервуар заполнен только водой. Специально разработанная итерационная процедура для определения σ_0 и $\{z_l\}$ позволила с высокой точностью установить их значения: $\sigma_0 = 0{,}001295$ Ом-**1** $\cdot$ м$^{-1}$ и $\{z_l\}$ = {2.68e-4, 3.04e-4, 2.79e-4, 4.35e-4, 3.56e-4, 4.39e-4, 3.91e-4, 2.37e-4, 2.03e-4, 2.23e-4, 2.05e-4, 1.44e-4, 3.01e-4, 2.81e-4, 2.94e-4, 3.45e-4} Ом на подробной треугольной сетке, состоящей из 5 854 треугольных элементов. Следует отметить, что полученные значения электрической проводимости и сопротивления электродов неплохо согласуются со значениями этих величин, представленных в работах [7; 8]: $\sigma_0 = 0{,}00141$ Ом$^{-1} \cdot$ м$^{-1}$ и $\{z_l\}$ = {2.64e-4, 3e-4, 2.76e-4, 4.27e-4, 3.5e-4, 4.30e-4, 3.91e-4, 2.35e-4, 2.01e-4, 2.21e-4, 2.04e-4, 1.43e-4, 2.98e-4, 2.78e-4, 2.92e-4, 3.4e-4} Ом. Для количественной оценки точности полученных результатов были рассчитаны метрические оценки средней абсолютной ошибки (MAE = $3{,}736 \cdot 10^{-7}$) и среднеквадратичной ошибки (RMSE = $1{,}494 \cdot 10^{-6}$). Низкие значения MAE и RMSE свидетельствуют о высокой точности полученных данных. Эти значения демонстрируют хорошее соответствие модели экспериментальным данным.

При численной реконструкции распределения электрической проводимости для случаев 2–5 использовалась сетка, состоящая из 1 200 треугольных элементов (рис. 4, *а*). В качестве начального приближения в итеративно регуляризованном методе Гаусса – Ньютона использовалось значение фоновой проводимости σ_0. Для параметра регуляризации стартовое значение α_0 выбиралось равным единице.

На рис. 6–9 представлены результаты сравнительного анализа реальных измерений и результатов численной реконструкции распределения электрической проводимости в исследуемых объектах. Левая колонка каждого рисунка демонстрирует фотографии реальных экспериментальных установок, на которых проводились измерения потенциала на электродах. Подробные значения измеренных напряжений можно найти в источнике [6]. Правая колонка иллюстрирует результаты вычислительной реконструкции, полученные с помощью итеративно регуляризованного метода Гаусса – Ньютона, подробно описанного в предыдущих разделах. Этот метод, как показано, эффективно решает обратную задачу ЭИТ, позволяя определить расположение и размеры неоднородностей в исследуемом объекте. При графической демонстрации полученных решений обратной задачи ЭИТ использовалась интерполяция значений электрической проводимости в узел сетки, являющийся центром барицентрической ячейки, в соответствии со значением площади.

Важно отметить, что метод успешно различает неоднородности с существенно различной электрической проводимостью по сравнению с фоновой средой. Например, метод точно определяет местоположение как диэлектрических включений (например, пластиковых элементов, где проводимость σ значительно меньше фоновой проводимости σ_0), так и высокопроводящих объектов (например, металлических колец, где $\sigma >> \sigma_0$). Это демонстрирует высокую чувствительность метода к широкому диапазону изменений электрической проводимости. Разница в проводимости является ключевым фактором, влияющим на точность локализации неоднородности; чем больше разница между проводимостью неоднородности и окружающего материала, тем точнее определяется ее положение.

Однако, как уже упоминалось, метод Гаусса – Ньютона, несмотря на свою эффективность, демонстрирует определенное размытие границ неоднородностей на реконструированных изображениях, поэтому актуальным является проведение дальнейших исследований по оптимизации метода и возможного его улучшения [5; 7] с точки зрения уменьшения размытости границ неоднородностей.

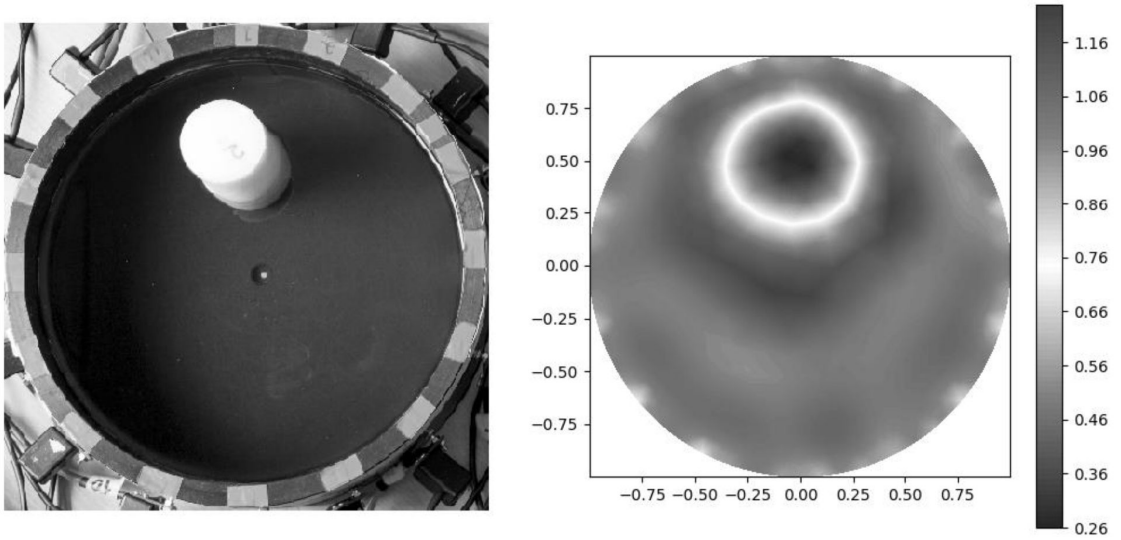


Рис. 6. Фотография условий проведения измерений и результат численной реконструкции для случая 2 (Case1_1 [6]) с пластиковым цилиндром. Справа шкала относительных значений σ/σ_0
Fig 6. A photograph of the measurement conditions and the result of numerical reconstruction for case 2 (Case 1_1 [6]) with a plastic cylinder. On the right is a scale of relative values σ/σ_0

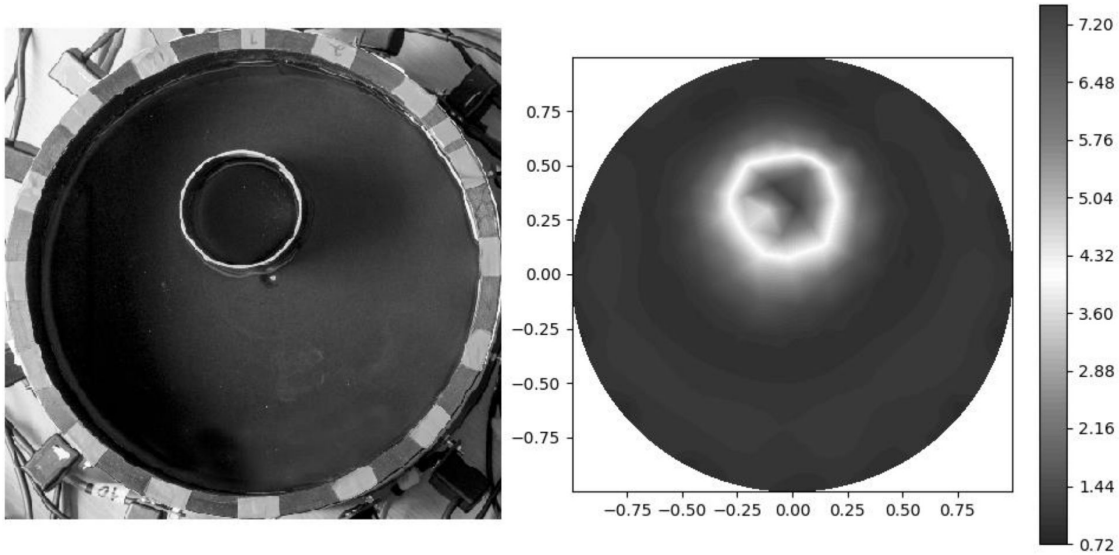


Рис. 7. Фотография условий проведения измерений и результат численной реконструкции для случая 3 (Case1_2 [6]) с полым металлическим цилиндром. Справа шкала относительных значений σ/σ_0
Fig. 7. A photograph of the measurement conditions and the result of numerical reconstruction for case 3 (Case 1_2 [6]) with a hollow metal cylinder. On the right is a scale of relative values σ/σ_0

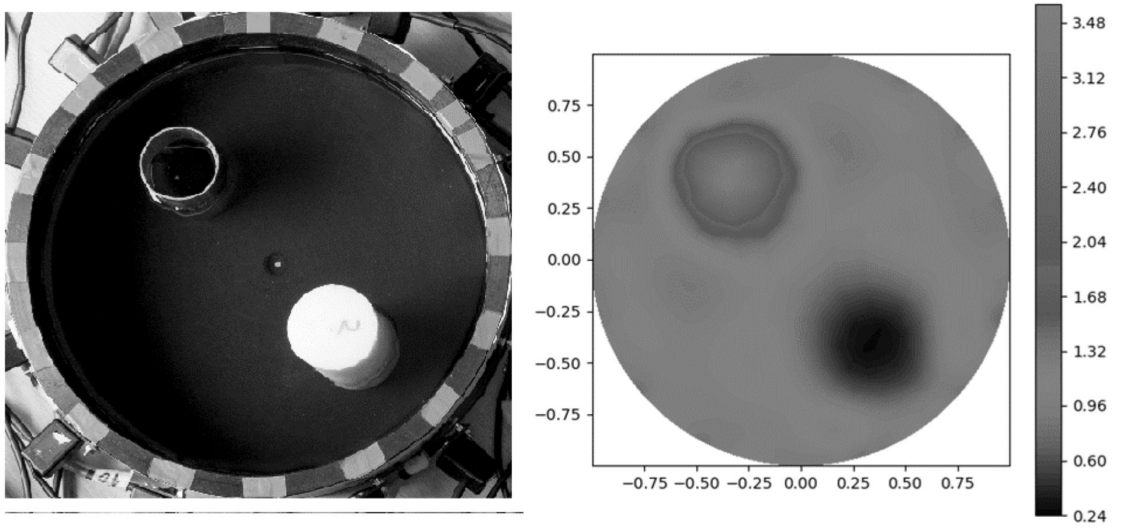


Рис. 8. Фотография условий проведения измерений и результат численной реконструкции для случая 4 (Case4_3 [6]) с пластиковым и полым стальным цилиндрами.
Справа шкала относительных значений σ/σ_0.
Fig. 8. A photograph of the measurement conditions and the result of numerical reconstruction for Case 4 (Case 4_3 [6]) with plastic and hollow steel cylinders. On the right is a scale of relative values σ/σ_0.

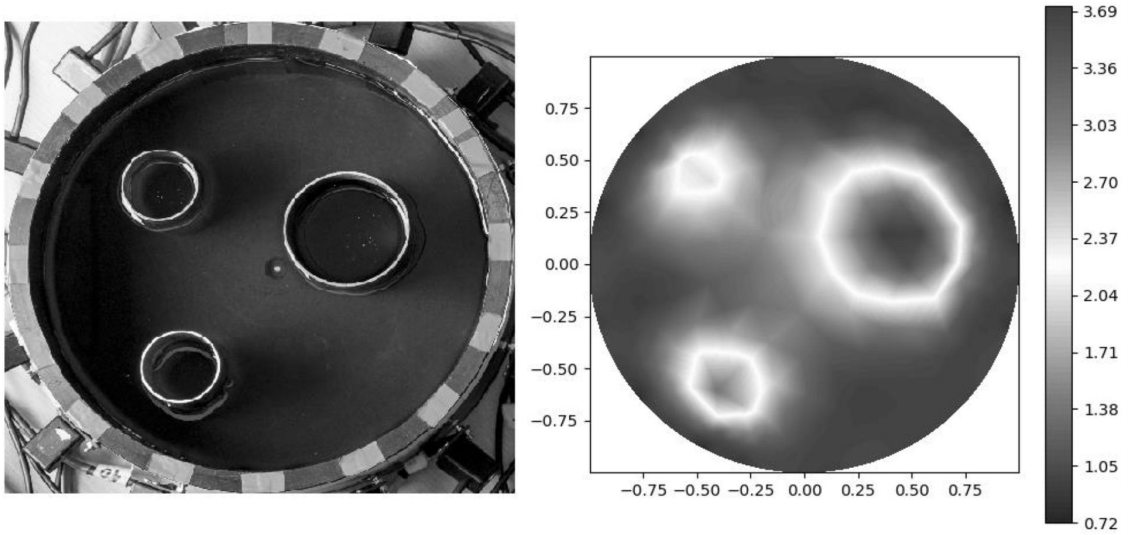


Рис. 9. Фотография условий проведения измерений и результат численной реконструкции для случая 5 (Case3_4 [6]) с тремя полыми стальными цилиндрами.
Справа шкала относительных значений σ/σ_0
Fig. 9. A photograph of the measurement conditions and the result of numerical reconstruction for case 5 (Case 3_4 [6]) with three hollow steel cylinders. On the right is a scale of relative values σ/σ_0

Заключение

Представлен итерационный численный метод решения обратной коэффициентной задачи для однородного эллиптического уравнения с кусочно-постоянными коэффициентами и интегро-дифференциальными граничными условиями в замкнутой области. Метод опирается на конечно-объемные аппроксимации дифференциальных и интегральных операторов на неструктурированных сетках, численное решение последовательности прямых задач при извест-

ном кусочно-постоянном распределении коэффициентов разностного эллиптического уравнения и сходящийся итеративно регуляризованный метод Гаусса – Ньютона. Разработанный метод решения обратных задач электроимпедансной томографии прошел тестирование на измерениях электрического напряжения, выполненных на экспериментальном стенде KIT в университете Восточной Финляндии. Получены близкие к реальным результатам реконструкции электрической проводимости внутри области исследования.

Список литературы

1. **Barber D. C., Brown B. H.** Applied potential tomography // J. Phys. E: Sci. Instrum. 1984. Vol. 17. P. 723–733.
2. **Корженевский А. В.** Электроимпедансная томография: исследования, медицинские приложения, коммерциализация // Альманах клинической медицины. 2006. № 12.
3. **Пеккер Я. С., Бразовский К. С., Усов В. Ю. и др.** Электроимпедансная томография. Томск: НТЛ, 2004. 192 с.
4. **Wei Z, Liu D, Chen X.** Dominant-Current Deep Learning Scheme for Electrical Impedance Tomography // IEEE Trans Biomed Eng. 2019. Vol. 66(9). P. 2546-2555. DOI: 10.1109/TBME.2019.2891676. Epub 2019 Jan 9. PMID: 30629486.
5. **Wang J.** A two-step accelerated Landweber-type iteration regularization algorithm for sparse reconstruction of electrical impedance tomography // Math Meth Appl Sci. 2024, Vol. 47. P. 3261–3272.
6. **Hauptmann A., Kolehmainen V., Mach N. M., Savolainen T., Seppänen A., Siltanen S.** Open 2D electrical impedance tomography data archive. 2017. P. 1–15. http://arxiv.org/abs/1704.01178.
7. **Gehre M., Kluth T., Lipponen A., Jin B., Seppänen A., Kaipio J. P., Maass P.** Sparsity reconstruction in electrical impedance tomography: An experimental evaluation // Journal of Computational and Applied Mathematics, 2012, Vol. 236, Issue 8, P. 2126–2136. https://DOI.org/10.1016/j.cam.2011.09.035
8. **Gehre M., Jin B.** Expectation Propagation for Nonlinear Inverse Problems – with an Application to Electrical Impedance Tomography // Numerical Analysis (math.NA). 2013. P. 1–35. DOI:10.1016/j.jcp.2013.12.010
9. **Borcea L.** Electrical impedance tomography, topical review // Inverse Problems. 2002. Vol. 18. R99–R136.
10. **Cheney M., Isaacson D., Newell J. C.** Electrical impedance tomography // SIAM review. 1999. Vol. 41 (1). P. 85–101.
11. **Somersalo E., Cheney M., Isaacson D.** Existence and uniqueness for electrode models for electric current computed tomography // SIAM Journal on Applied Mathematics. 1992. Vol. 52(4). 1023–1040.
12. **Gu D., Liu D., Smyl D., Deng J., Du J.** Supershape recovery from electrical impedance tomography data // IEEE Transactions on Instrumentation and Measurement. 2021. Vol. 70. P. 1–11.
13. **Афанасьева А. А., Старченко А. В.** Численное решение прямой задачи электроимпеданс-ной томографии в полной электродной постановке // Вестн. Томск. гос. ун-та. Матем. и мех. 2022. № 78. С. 5–21; DOI: 10.17223/19988621/78/1
14. **Тихонов А. Н., Самарский А. А.** Уравнения математической физики. М.: Наука, 1977. 735 с.
15. **Бахвалов Н. С., Жидков Н. П., Кобельков Г. М.** Численные методы: учеб. пособие для студентов физ.-мат. спец. вузов, 8-е изд. М.: Физматлит, 2000. 624 с.
16. **Саад Ю.** Итерационные методы для разреженных линейных систем: учеб. пособие. В 2 т. 2-е изд. М.: Изд-во Моск. ун-та, 2013. Т. 1. 344 с.

17. **Тихонов А. Н., Арсенин В. Я.** Методы решения некорректных задач: учеб. пособие для вузов по спец. «Прикл. Математика», 3-е изд., испр. М.: Наука, 1986. 286 с.
18. **Лаврентьев М. М., Романов В. Г., Шишатский С. П.** Некорректные задачи математической физики и анализа. М.: Наука, 1980. 286 с.
19. **Кабанихин С. И.** Обратные и некорректные задачи. Новосибирск: Сибирское научное изд-во, 2009.
20. **Бакушинский А. Б.** К проблеме сходимости интеративно регуляризованного метода Гаусса – Ньютона // Ж. вычисл. матем. и матем. физ. 1992. Т. 32, № 9. С. 1503–1509.
21. **Li J., Yuan Y.** Numerical simulation and analysis of generalized difference method on triangular networks for electrical impedance tomography // Appl. Math. Model. 2009. Vol. 3. No. 5. P. 2175–2186; DOI: 10.1016/j.apm.2008.05.025

References

1. **Barber D. C., Brown B. H.** Applied potential tomography. *J. Phys. E: Sci. Instrum*. 1984, vol. 17, pp. 723–733.
2. **Korzhenevsky A. V.** Elektroimpedansnaya tomografiya: issledovaniya, medicinskie prilozheniya, kommercializaciya [Electrical impedance tomography: research, medical applications, commercialization]. *Almanac of Clinical Medicine*. 2006, no. 12. (In Russ.)
3. **Pekker Ya. S., Brazovsky K. S., Usov V. Yu. et al.** Elektroimpedansnaya tomografiya [Electrical impedance tomography]. Publisher: Tomsk: NTL, 2004. 192 c. (In Russ.)
4. **Wei Z, Liu D, Chen X.** Dominant-Current Deep Learning Scheme for Electrical Impedance Tomography. *IEEE Trans Biomed Eng*. 2019 Sep;66(9):2546–2555. DOI: 10.1109/TBME.2019.2891676. Epub 2019 Jan 9. PMID: 30629486.
5. **Wang J.** A two-step accelerated Landweber-type iteration regularization algorithm for sparse reconstruction of electrical impedance tomography. *Math Meth Appl Sci*. 2024, vol. 47, pp. 3261–3272.
6. **Hauptmann A., Kolehmainen V., Mach N. M., Savolainen T., Seppänen A., Siltanen S.** Open 2D electrical impedance tomography data archive, 2017, pp. 1–15. http://arxiv.org/abs/1704.01178.
7. **Gehre M., Kluth T., Lipponen A., Jin B., Seppänen A., Kaipio J. P., Maass P.** Sparsity reconstruction in electrical impedance tomography: An experimental evaluation. *Journal of Computational and Applied Mathematics*, 2012, vol. 236, iss. 8, pp. 2126–2136. https://doi.org/10.1016/j.cam.2011.09.035
8. **Gehre M., Jin B.** Expectation Propagation for Nonlinear Inverse Problems – with an Application to Electrical Impedance Tomography. *Numerical Analysis (math. NA)*. 2013, pp. 1–35. DOI: 10.1016/j.jcp.2013.12.010
9. **Borcea L.** Electrical impedance tomography, topical review. Inverse Problems. 2002, vol. 18, pp. R99–R136.
10. **Cheney M., Isaacson D., Newell J. C.** Electrical impedance tomography, *SIAM review*, 1999, 41(1), pp. 85–101.
11. **Somersalo E., Cheney M., Isaacson D.** Existence and uniqueness for electrode models for electric current computed tomography. *SIAM Journal on Applied Mathematics*, 1992, 52(4), pp. 1023–1040.
12. **Gu D., Liu D., Smyl D., Deng J., Du J.** Supershape recovery from electrical impedance tomography data. *IEEE Transactions on Instrumentation and Measurement*, 2021, 70, pp. 1–11.
13. **Afanaseva A. A., Starchenko A. V.** Chislennoe reshenie pryamoj zadachi elektroimpedansnoj tomografii v polnoj elektrodnoj postanovke [Numerical solution of the direct problem of electrical

impedance tomography in the full electrode formulation]. *Vestn. Tomsk. state University. Mat. and Mech*. 2022, no. 78, pp. 5–21. DOI: 10.17223/19988621/78/1 (In Russ.)
14. **Tikhonov A. N., Samarsky A. A.** Uravneniya matematicheskoj fiziki [Equations of Mathematical Physics]. Moscow: Nauka, 1977. 735 c. (In Russ.)
15. **Bakhvalov N. S., Zhidkov N. P., Kobelkov G. M.** Chislennye metody: uchebnoe posobie dlya studentov fiziko-matematicheskih special'nostej vuzov [Numerical methods: a textbook for students of physics and mathematics specialties of higher education institutions]. 8th ed. Moscow [et al.]: Fizmatlit [et al.], 2000. 624 c. (In Russ.)
16. **Saad Yu.** Iteracionnye metody dlya razrezhennyh linejnyh sistem [Iterative methods for sparse linear systems]. In 2 volumes, vol. 1. Study guide, vol. 1, 2nd ed. // M.: Publishing house of Moscow University. 2013. 344 c. (In Russ.)
17. **Tikhonov A. N., Arsenin V. Ya.** Metody resheniya nekorrektnyh zadach [Methods for solving ill-posed problems]: [Textbook for universities in the specialty "Applied Mathematics"]. 3rd ed., corrected. Moscow: Nauka, 1986. 286 c. (In Russ.)
18. **Lavrentiev M. M., Romanov V. G., Shishatsky S. P.** Nekorrektnye zadachi matematicheskoj fiziki i analiza [Ill-posed problems of mathematical physics and analysis]. Moscow: Nauka, 1980. 286 c. (In Russ.)
19. **Kabanikhin S. I.** Obratnye i nekorrektnye zadachi [Inverse and ill-posed problems]. Novosibirsk: Siberian scientific publishing house, 2009. (In Russ.)
20. **Bakushinsky A. B.** K probleme skhodimosti interativno-regulyarizovannogo metoda Gaussa-N'yutona [On the problem of convergence of the iteratively regularized Gauss-Newton method]. *Journal of Computational Mathematics and Mathematical Physics*. 1992, vol. 32, no. 9, pp. 1503–1509. (In Russ.)
21. **Li J., Yuan Y.** Numerical simulation and analysis of generalized difference method on triangular networks for electrical impedance tomography. *Appl. Math. Model*. 2009, vol. 3, no. 5, pp. 2175–2186. DOI: 10.1016/j.apm.2008.05.025

Сведения об авторах

Афанасьева Анна Александровна, аспирант кафедры вычислительной математики и компьютерного моделирования Томского государственного университета

Старченко Александр Васильевич, профессор, доктор физико-математических наук, заведующий кафедрой вычислительной математики и компьютерного моделирования Томского государственного университета, научный сотрудник Регионального научно-образовательного математического центра Томского государственного университета
Researcher ID B-2354-2014

Information about the Authors

Anna A. Afanaseva, Graduate Student of Department of Computational Mathematics and Computer Modelling of National Research Tomsk State University, Tomsk, Russian Federation.

Alexander V. Starchenko, Professor, Doctor of Physical and Mathematical Sciences, Head of Department of Computational Mathematics and Computer Modelling of National Research Tomsk State University, Scientific Researcher, Regional Scientific Educational Mathematical Center of Tomsk State University, Tomsk, Russian Federation
Researcher ID B-2354-2014

Статья поступила в редакцию 13.02.2025;
одобрена после рецензирования 17.08.2025; принята к публикации 17.08.2025

The article was submitted 13.02.2025;
approved after reviewing 17.08.2025; accepted for publication 17.08.2025

Научная статья

УДК 004.09
DOI 10.25205/1818-7900-2025-23-4-23-43

Проблемы методов сжатия медицинских изображений

Андрей Васильевич Гаврилов [1]
Денис Владиславович Краюшкин [2]
Андрей Михайлович Чеповский [2,3]

[1]Московский государственный университет им. М. В. Ломоносова
Москва, Россия

[2]Национальный исследовательский университет «Высшая школа экономики»
Москва, Россия

[3]Российский экономический университета им. Г. В. Плеханова
Москва, Россия

agavrilov49@gmail.com; https://orcid.org/0000-0002-7838-584X
KrayushkinDenV@yandex.ru; https://orcid.org/0009-0004-5474-1397
achepovskiy@hse.ru; https://orcid.org/0000-0001-8959-6119

Аннотация

Автоматизация службы лучевой диагностики существенно повысила доступность радиологических исследований для точной диагностики заболеваний и травм. Вместе с тем расширение парка рентгенологического оборудования, внедрение телемедицины и сервисов поддержки врачебных решений на основе искусственного интеллекта требуют модернизации систем хранения и обработки изображений в уже существующих системах.

В данной статье представлен обзор современных методов сжатия радиологических изображений, которые обеспечивают более высокий коэффициент сжатия, улучшенное качество изображения и меньшее время кодирования/декодирования по сравнению со стандартами, предусмотренными спецификацией DICOM. Обзор научных публикаций позволяет заключить, что рентгенологические изображения обладают рядом особенностей, учет которых в алгоритмах сжатия позволяет улучшить показатели сжатия изображений. К таким особенностям относятся: высокая зашумленность, наличие локально симметричных областей (схожих участков), а также присутствие множества последовательных кадров в рамках одного исследования.

Применение современных подходов к сжатию данных способно повысить отказоустойчивость высоконагруженных медицинских систем и сократить затраты на хранение, передачу и обработку диагностических исследований.

Ключевые слова

сжатие изображений, медицинские данные, обзор

Финансирование

Исследование выполнено в рамках государственного задания МГУ им. М. В. Ломоносова.

Для цитирования

Гаврилов А. В., Краюшкин Д. В., Чеповский А. М. Проблемы методов сжатия медицинских изображений // Вестник НГУ. Серия: Информационные технологии. 2025. Т. 23, № 4. С. 23–43. DOI 10.25205/1818-7900-2025-23-4-23-43

Problems of the state of the art in medical images compression

Andrey V. Gavrilov [1], Denis V. Krayushkin [2]
Andrey M. Chepovskiy [2,3]

[1]Lomonosov Moscow State University
Moscow, Russian Federation

[2]HSE University,
Moscow, Russian Federation

[3]Plekhanov Russian University of Economics
Moscow, Russian Federation

agavrilov49@gmail.com; https://orcid.org/0000-0002-7838-584X
KrayushkinDenV@yandex.ru; https://orcid.org/0009-0004-5474-1397
achepovskiy@hse.ru; https://orcid.org/0000-0001-8959-6119

Abstract

The automation of radiology services has significantly improved access to radiological imaging for accurate diagnosis of diseases and injuries. However, the expansion of radiological equipment, the adoption of telemedicine, and the integration of AI-powered clinical decision support systems necessitate upgrades to existing medical image storage and processing solutions.

This article reviews modern compression methods for radiological images, which offer higher compression ratios, improved image quality, and faster encoding/decoding times compared to the standards defined by the DICOM specification. It is established that radiological images possess unique characteristics—such as high noise levels, locally symmetric regions (similar patches), and the presence of multiple sequential frames in a single study—which, when accounted for in compression algorithms, can enhance compression efficiency.

Implementing advanced data compression approaches can increase the fault tolerance of high-load medical systems and reduce costs associated with the storage, transmission, and processing of diagnostic studies.

Keywords

image compression, medical data, review, state of art

Financing

The research was carried out within the framework of the state assignment of the Lomonosov Moscow State University.

For citation

Gavrilov A. V., Krayushkin D. V., Chepovskiy A. M. Problems of the state of the art in medical images compression. *Vestnik NSU. Series: Information Technologies*, 2025, vol. 23, no. 4, pp. 23–43 (in Russ.) DOI 10.25205/1818-7900-2025-23-4-23-43

Введение

В последние годы использование радиологических изображений (КТ, МРТ, рентген и т. п.) и электрофизиологических данных (ЭКГ, ЭЭГ) для постановки медицинских заключений получило широкое распространение благодаря внедрению медицинских информационных систем, организующих хранение, обработку и передачу данных в соответствии со стандартом DICOM (Digital Imaging and Communications in Medicine). Все чаще такие системы интегрируют сервисы на основе искусственного интеллекта для автоматизации анализа и повышения точности диагностики, что вместе с развитием телемедицины и необходимостью долгосрочного архивирования исследований создает огромные потоки медицинских данных, особенно для объемных 3D-КТ, 3D-МРТ, высоко детализированных изображений маммографии и рентген-ангиографии, длительных записей УЗ-эхокардиографии, ЭКГ и т. п. [1]. Эти данные характеризуются большими размерами (до нескольких ГБ для отдельных исследований), требованием сохранения диагностически значимых деталей и необходимостью быстрого доступа, что делает эффективное сжатие критически важной технологией, позволяющей сократить затраты на хранение, ускорить передачу для телеконсультаций, обеспечить быстрый доступ

к архивам и поддерживать работу алгоритмов поддержки врачебных решений без потери диагностической ценности, при этом современные подходы к сжатию должны учитывать особенности медицинских изображений (высокую битовую глубину, шумы) и необходимость поддержки существующих медицинских стандартов.

Эффективные алгоритмы сжатия позволяют не только сократить затраты на хранение и передачу (особенно актуально для телемедицины), но и сохранить все клинически значимые особенности сигналов – от минимальных изменений зубцов на ЭКГ [2; 3] до минимальных изменений на сериях КТ-снимков, что критически важно для постановки точного диагноза и последующего лечения.

Наиболее широкое распространение сжатие изображений получило в зондировании земли, где применение методов сокращения представляемой информации обусловлено ограничением пропускного канала, высоким разрешением детализированных изображений, а также скоростью их создания. Такие изображения необходимо архивировать для последующей обработки и анализа. Повсеместная цифровизация медицинских учреждений, в том числе активное интегрирование в практическую работу врачей систем PACS/RIS (Picture archiving and communication system / Radiology Information System), а также внедрение современных высокопроизводительных радиологических приборов, предусматривающих возможность проведения исследований с высокой битовой глубиной (16 бит и выше), ставит новые вызовы, которые требуют эффективного сжатия медицинских данных подобно космической отрасли. Научная мысль современных исследований направлена на поиск особенностей и закономерностей в представлении информации медицинских изображений, результаты которых приводят к сокращению коммерческих издержек для хранения и повторного использования таких данных.

1. Исследование методов сжатия

В классической литературе преобразования информации называют отображениями. Если информацию можно восстановить в полном объеме после преобразования, то они являются обратимыми, в противном случае – необратимыми [4]. Таким образом, методы сжатия изображений разделяют на две большие группы: с потерями и без потерь.

Каждый метод содержит в себе ряд отображений, который приводит к сокращению представляемой информации путем устранения информационной избыточности изображения. Избыточность подразделяется на два вида: кодовая и пространственная/временная. Кодовая избыточность характеризуется разницей количества битов, в которых предоставляется информация, и количеством битов, в котором информация может передаваться. Главный принцип устранения кодовой избыточности заключается в преобразовании кодовых слов, в случае изображений значений яркости пикселя, так, чтобы более вероятным кодовым словам распределялось меньшее количество битов, а с меньшей вероятностью – большее. Пространственная/временная избыточность включает корреляцию значений яркости пикселей в одном и том же положении для последовательных кадров. Таким образом, эффективное сжатие медицинских исследований, содержащих несколько изображений или временных серий изображений, должно включать устранение как кодовой избыточности, так и пространственно-временной.

Оценка нижней границы среднего количества битов, необходимого для предоставления информации, производится по значению энтропии. Из теории информации известно, что источник информации может быть описан вероятностным процессом. Количество информации $I(E)$, которая содержится в случайном событии E, зависит от вероятности его возникновения $P(E)$ и может быть описана соотношением

$$I(E) = \log \frac{1}{P(E)} = -\log P(E). \quad (1)$$

Значение яркости пикселей можно рассматривать как источник статистически независимых случайных событий из дискретного набора случайных значений $\{a_1, a_2, \ldots, a_j\}$ с вероятностью их появления $\{P(a_1), P(a_2), \ldots, P(a_j)\}$. Каждое случайное значение является символом источника, а их набор – алфавитом источника. Энтропия источника, или неопределенность источника, определяется формулой

$$H = -\sum_{j=1}^{J} P\left(a_j\right)\log P\left(a_j\right). \tag{2}$$

В показателе энтропии заключено среднее количество информации на один символ источника. Именно энтропия источника показывает минимальную границу для кодирования символа алфавита. Если изображение считать источником без памяти, т. е. появляющиеся символы статистически независимы, а значения яркости лежат на интервале [0, L – 1], то энтропия будет определяться формулой

$$\widetilde{H} = -\sum_{k=0}^{L-1} p_r\left(r_k\right)\log P\left(a_j\right), \tag{3}$$

где $p_r(r_k)$ – значение вероятности появления дискретного значения.

Для уменьшения значения энтропии, которое представляет оценку минимального объема представления информации, необходимо учитывать коррелированность пикселей между собой. В этом случае источник обладает конечной памятью (такие источники называются марковскими). Чем ниже энтропия, тем эффективнее можно сжать изображение. Адаптивные модели динамически вычисляют вероятности появления символов на основе уже обработанных данных, постепенно уточняя статистику в процессе сжатия. В отличие от них, контекстно-зависимые модели определяют вероятности символов, анализируя структуру окрестности соседних пикселей.

Далее приведены различные основополагающие подходы к сжатию медицинских изображений, включая их математическое описание и стандарты (при наличии), а также обзор публикаций по разработке и применению современных методов сжатия в соответствии с базовым подходом.

1.1. Кодирование с предсказанием

Методы, основанные на кодировании ошибок предсказания, вычисляются как разница между истинным значением пикселя и предсказанным. Ранее было сказано, что энтропия источника показывает минимальную границу для кодирования символа алфавита. Оценка энтропии показывает, что сжатие исходного изображения имеет коэффициент энтропии значительно больше, чем при сжатии ошибки предсказания. Дело в том, что плотность распределения ошибок предсказания имеет пик в нуле в отличие от плотности распределения вероятностей яркостей. Это свойство позволяет сжимать изображение с меньшим количеством памяти, устраняя межэлементную избыточность. На основе распределения Лапласа с нулевым средним строится плотность распределения вероятностей ошибок $p_e(e)$:

$$p_e\left(e\right) = \frac{1}{\sqrt{2}\sigma_e} e^{-\frac{\sqrt{2}|e|}{\sigma_e}}, \tag{4}$$

где σ_e – величина стандартного отклонения e.

Рассмотрим наиболее часто используемый стандарт сжатия изображений на основе кодирования ошибки предсказания, который применяется в практике разработчиками алгоритмов

и программ обработки и анализа медицинских радиологических исследований. Таким стандартом является LossLess JPEG (JPEG – Joint Photographic Experts Group), использующий метод адаптивного предсказания на основе ближайших соседей (контекста). В его основе лежит алгоритм LOCO-I (Low Complexity Lossless Compression for Images). Один из этапов работы алгоритма включает достаточно простой и эффективный метод предсказания медианного детектора края (англ. Median Edge Detector, сокр. MED). Данный метод позволяет рассчитать предполагаемое значение для каждого пикселя на основе ближайших пикселей, размещенных в горизонтальной и вертикальной направлениях относительно рассматриваемого. Квантование в LossLess JPEG происходит на основе кодов Голомба. Для кодирования вычисленных значений ошибок предсказания используются кодирование Хаффмана или арифметическое кодирование.

Известны работы, где авторы стремятся уменьшить ошибку предсказания каждого пикселя путем увеличения рассматриваемого контекста, т. е. анализировать гораздо больший набор соседних вертикальных и горизонтальных пикселей [5]. Оценка в данной работе производится в сравнении с предсказателями MED, GAP (Golomb-Power Approximate Prediction), FLIF (Free Lossless Image Format), LBP (Local Binary Patterns) в показателях значений энтропии, среднего количества бит на один пиксель (англ. bits per pixel, сокр. Bpp), времени вычисления.

В работе [6] описан подход к сжатию изображения без потерь после устранения межэлементной избыточности на основе кодирования с предсказанием. Кодирование модулей ошибок предсказаний происходит на основе контекстно-зависимой модели. Такой подход позволяет получить вероятности множества кодируемых символов на основе окрестности пикселей. Полученные вероятности кодируются с помощью арифметического кодирования. Авторы отмечают, что такой способ является наиболее предпочтительным по причине простого алгоритма работы, который подстраивается к изменению статистически кодируемых данных. Было продемонстрировано на общедоступных банках изображений с высоким разрешением, что предложенный метод является наиболее эффективным по сравнению с распространенными форматами сжатия без потерь JPEG-2000 и JPEG-LS в сравнении по значениям Bpp. В работе [7] предложен метод сжатия изображений также с помощью арифметического кодирования ошибок предсказания, где ключевую роль играют условные кодовые распределения вероятностей элементов алфавита. Авторы связывают их со статической моделью, называемой «модель с вычисляемой последовательностью состояний». Элементы алфавита или последовательности формируются источником данных, который в каждый момент времени имеет состояние из множества, которые связаны с условным кодовым распределением вероятностей появления элементов алфавита. Кодер и декодер вычисляют текущее состояние источника и применяют к нему кодовое распределение. В рамках работы была произведена адаптация свободных параметров для изображений компьютерной томографии. В качестве целевого набора для тестирования использовались изображения брюшной полости, легких и головного мозга. Авторы вводят понятия устойчивых и неустойчивых параметров. Устойчивые параметры определяют эмпирическую энтропию, влияющую на кодовую избыточность. Неустойчивыми параметрами определяется сложность реализации алгоритма. Отмечено, что все параметры можно считать устойчивыми, хотя при этом увеличивается избыточность кодирования. В дальнейших работах этого направления [8–10] произведены оценки избыточности кодирования и минимальной скорости кодирования, а также приведен адаптивный метод сжатия со статической моделью источника и вычисляемым кодовым распределением на основе дискретного вейвлет-преобразования [8].

В работе [11] авторы демонстрируют уменьшение вычислительной сложности и увеличение коэффициентов сжатия комбинированием методов кодирования. Изображение разделяется на битовые плоскости, к каждой из которых применяется дифференциальная импульсно-кодовая модуляция (англ. Differential Pulse-Code Modulation, сокр. DPCM). Далее, в зависимости от контекста, поток битов кодируется с помощью метода кодирования длин серий (англ. Run-

Length Encoding, сокр. RLE) или метода арифметического кодирования. Авторы формируют математические модели распределения битов в битовых плоскостях на основе выбора структуры кодера с минимальным объемом кода.

МРТ и КТ-изображения многих органов и структур имеют симметрию, учитывание которой в модели предсказания существенно увеличивает коэффициент сжатия изображения [12]. Высокая степень самоподобия, аффинная симметричность и естественная избыточность [13] позволяют использовать эти свойства для сокращения размера исходного изображения.

Стоит отметить, что недостатком использования пространственной корреляции соседних пикселей является наличие в изображении текстур и резких перепадов яркости (контуры). В этом случае для методов прогнозирования возникает проблема коллинеарности, когда предсказатели линейно зависимы друг от друга. Ошибка предсказания, которая дополнительно масштабируется по всему изображению, возрастает на текстурах и контурах. В работе [14] предлагают линейный предсказатель, который выполняет переключения для нелинейных структур изображения или, как еще их называют, локальные особенности изображения. Результаты сравниваются со следующими методами прогнозирования: MED, GAP, DARC (Differentiable Architecture Compression), CALIC (Context-based, Adaptive, Lossless Image Codec).

Отдельным современным направлением применения алгоритмов на основе предсказателя, в том числе с использованием методов машинного обучения, является сжатие радиологических изображений высокого разрешения, содержащих несколько проекций (3D) [15].

1.2. Вейвлет-преобразования

Применение вейвлет-преобразований получило широкое распространение в сжатии изображений в том числе из-за локализации в нуле распределения вероятностей [16]. Ключевым стандартом, который работает с данным типом преобразования изображений, является JPEG 2000.

Любой сигнал может быть представлен в виде некоторого набора функций и коэффициентов разложения по формуле

$$f(x) = \sum_k \alpha_k \varphi_k(x). \tag{5}$$

Такие функции являются кусочно-постоянными и их называют масштабирующими.

Система функций образует ортонормированное пространство или базис V, если результат равен 0 при скалярном умножения пар различных функций (ортогональные), а также если результат равен 1 при скалярном умножении функции саму на себя:

$$\left\langle \varphi_n(x) \,|\, \varphi_k(x) \right\rangle = \delta_{nk} = \begin{cases} 0, & n \neq k \\ 1, & n = k \end{cases}. \tag{6}$$

Система является биортогональной в случае невыполнения условия ортогональности. В пространстве есть функция, которая не является ортогональной ко всем другим функциям системы. Это значит, что имеется более одного набора коэффициентов разложения. В этом случае существует набор функций, которые будут отвечать требованию

$$\left\langle \varphi_n(x) \,|\, \tilde{\varphi}_k(x) \right\rangle = \delta_{nk} = \begin{cases} 0, & n \neq k \\ 1, & n = k \end{cases}. \tag{7}$$

Коэффициенты разложения определяются формулой

$$\alpha_k = \left\langle f(x) \,|\, \tilde{\varphi}(x) \right\rangle = \int f(x) \tilde{\varphi}^*(x) dx. \tag{8}$$

Вейвлет-преобразование является частью кратномасштабного анализа (КМА). Пространство функций V_j может быть уточняемо путем представления в виде суммы подпространства V_{j-1} и его ортогонального дополнения W_{j-1}, где W – набор вейвлет-функций, которые описывают разность между пространствами V_j и V_{j-1}. Таким образом, с помощью масштабирующих функций формируется последовательность приближений исходного сигнала. Разность между соседними приближениями описывается вейвлет-функциями, описывающими целые сдвиги и изменение масштаба:

$$V_j = V_{j-1} \oplus W_{j-1}. \tag{9}$$

где $\oplus$ – сумма пространств.

Разложение сигнала соответственно определяется пространством:

$$f(x) = \ldots = \sum_k \alpha_k^{j-1} \varphi_k^{j-1} = \sum_k \alpha_k^j \varphi_k^j = \ldots. \tag{10}$$

Ортогональным базисом является базис, в котором масштабирующие функции ортогональны друг другу, вейвлеты ортогональны друг другу, а также каждый вейвлет ортогонален каждой масштабирующей функции предыдущего уровня:

$$\begin{cases} \left\langle \varphi_k^j(x) \mid \varphi_n^j(x) \right\rangle = \delta_{k,n} \\ \left\langle \psi_k^j(x) \mid \psi_n^j(x) \right\rangle = \delta_{k,n} \\ \left\langle \varphi_k^j(x) \mid \psi_n^j(x) \right\rangle = 0 \end{cases} \text{ для всех } j, k, n. \tag{11}$$

Ключевая особенность биортогонального базиса заключается в том, что масштабирующие функции *основного базиса* ортогональны *двойственным* функциям разложения, а *двойственные* масштабирующие функции ортогональны функциям разложения *основного* базиса:

$$\begin{cases} \left\langle \varphi_k^j(x) \mid \tilde{\psi}_n^j(x) \right\rangle = 0 \\ \left\langle \psi_k^j(x) \mid \tilde{\varphi}_n^j(x) \right\rangle = 0 \end{cases} \text{ для всех } j, k, n. \tag{12}$$

Связь масштабирующих функций и вейвлет-функций представлена соотношением

$$\left(f(x), \psi_n\right)\left(\sum_{k=1}^{\infty} \alpha_k \varphi_k, \psi_n\right) = \sum_{k=1}^{\infty} \alpha_k (\psi_n, \varphi_n) = \sum_{k=1}^{\infty} \alpha_k \delta_n^k = \alpha_n. \tag{13}$$

Результатом преобразования является получение горизонтальных ($d^H(m, n)$), вертикальных ($d^V(m, n)$) и диагональных коэффициентов ($d^D(m, n)$), коэффициентов приближения ($a(m, n)$). Высокочастотный фильтр обеспечивает получение детализированной исходной последовательности, низкочастотный – приближение.

Комплексирование методов вейвлет-преобразования и Хаффмана является эффективным решением сжатия изображений, на основе которых предлагаются решения для повышения коэффициентов сжатия [17; 18]. Группа исследователей [19] предложила применять векторное квантование (VQ) к коэффициентам вейвлет-преобразования (Добеша 9/7) за исключением поддиапазона LL, который несет основную информацию о контурах в изображении. Метод VQ использует древовидную структуру для квантования вектора коэффициентов вейвлет-преобразования. Удаление шумовой составляющей исходного изображения также осуществляется с помощью медианного фильтра. Результат кодируется с помощью метода Хаффмана.

Исследователи применяют данные подход к сжатию разнообразных радиологических медицинских изображений и демонстрируют эффективность с HEVC-RA (High Efficiency Video Coding – Random Access) по показателям PSNR (Peak Signal-to-Noise Ratio) и Bpp. Для увеличения коэффициента сжатия вейвлет-преобразования используют совместно с методами прогнозирования [20; 21]. В работе [21] к указанной связке было предложено добавить нелинейный предсказатель на основе дифференциальной импульсно-кодовой модуляции (DPCM) для дальнейшего преобразования матрицы ошибок предсказания и для последующего кодирования. Отмечено в результате тестирования, что коэффициент сжатия увеличивается, но время исполнения алгоритма сжатия возрастает.

1.3. Фурье-преобразования

Преобразования Фурье нашли широкое применение в области сжатия в стандарте JPEG. Этот стандарт является достаточно популярным в сжатии 8-битных полутоновых изображений из-за низкой вычислительной сложности. На основе дискретного преобразования был разработан формат JPEG XT, для которого многие исследователи занимаются его улучшением и усовершенствованием. Его основное применение в медицине – это сжатие изображений с высокой битовой глубиной. Стандарт является совместимым с классическим JPEG, потому что алгоритм сжатия включает в себя разделение на базовый слой и слой расширения. Последний слой является предметом исследования для применения и поиска эффективных методов сжатия [22].

Любой непрерывный сигнал на отрезке $\{0, T\}$ может быть представлен в виде набора гармонических функций или ряда Фурье в соответствии с формулой

$$f(t) = \sum_{n=-\infty}^{\infty} X_k e^{i2\pi k\frac{t}{T}}. \tag{14}$$

Коэффициенты разложения определяются по формуле

$$X_k = \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} f(t) e^{-i2\pi k\frac{t}{T}} \quad \text{для } k = \pm 0, 1, 2, \ldots \tag{15}$$

Коэффициенты разложения для дискретного сигнала извлекаются путем применения к непрерывному сигналу операций квантования и дискретизации. В этом случае коэффициенты разложения для дискретного сигнала размерностью N определяются соотношением

$$X_k = \sum_{N}^{N-1} f(n) e^{-i2\pi\frac{n}{N}k} \quad \text{для } k = \pm 0, 1, 2, \ldots \tag{16}$$

Применяя формулу Эйлера в комплексной форме, уравнение (соотношение) получает вид

$$X_k = \sum_{N}^{N-1} f(n)\left(\cos 2\pi k\frac{n}{N} - i\sin 2\pi k\frac{n}{N}\right). \tag{17}$$

Фурье-преобразование является обратимым. Таким образом, получение исходной дискретной последовательности из последовательности коэффициентов разложения определяется как

$$f(n) = \frac{1}{N}\sum_{k=0}^{N-1} X_k e^{i2\pi\frac{n}{N}k} = \frac{1}{N}\sum_{k=0}^{N-1} X_k\left(\cos 2\pi k\frac{n}{N} + i\sin 2\pi k\frac{n}{N}\right). \tag{18}$$

Математическая задача преобразования Фурье заключается в приведении функции к новому ортонормированному базису, который состоит только из ортогональных косинусных функций.

Например, коэффициенты, получаемые в результате дискретного преобразования Фурье, состоят из целой и десятичной части. В общем случае, во время квантования отбрасывается десятичная часть, а также небольшие целые значения. Кроме того, коэффициенты переменной составляющей преобразования округляются во время квантования. Это значительно увеличивает коэффициент сжатия, но ухудшает качество текстурной составляющей изображения, которая зарегистрирована с высоким разрешением, например, изображения тканей [22]. Одно из решений предлагает дополнительное усиление этих частей коэффициентов, чтобы в процессе квантования информация была более детальной [23]. В работе [24] предлагают усовершенствованный алгоритм быстрых преобразований Фурье для сжатия медицинских изображений.

1.4. Фрактальные преобразования

Фрактальные методы сжатия не входят в перечень допустимых стандартов сжатия DICOM, хотя многие исследовательские работы демонстрируют, что фрактальные алгоритмы очень перспективны для сжатия медицинских изображений. Дело в том, что фрактальные преобразования чрезвычайно эффективны на изображениях, которые обладают высокой степенью самоподобия. Главное преимущество фрактальных алгоритмов заключается в сохранении качества контуров при сжатии с потерями, которые особенно важны при передаче медицинских изображений. В общем случае изображение разбивается на доменные и ранговые области. Каждый домен приближают к размеру ранговой области путем преобразования яркости. После преобразования всех доменов происходит сопоставление всех ранговых областей к наиболее подходящим доменным. В основе фрактального метода используется сжатие коэффициентов преобразования при сопоставлении областей [25]. Достаточно просто отражается принцип фрактального представления изображения в математическом описании фрактальной размерности D, которая определяется соотношением

$$D = \frac{\ln N}{\ln S}, \tag{19}$$

где N – число самоподобных областей, на которые можно разделить изображение, а S – коэффициент масштабирования, необходимый для наблюдения N частей.

В работе [26] представлены результаты сжатия МРТ головы с помощью различных методов фрактального кодирования без потерь (классическое, фрактальное и квази-). Известны работы, где применяется фрактальное сжатие, в основе которого лежит BC (Box-counting) алгоритм [27; 28]. Метод BC основывается на разделении изображения на области (boxes) с равными сторонами, площадь которых изменяется. Классический поиск фрактальной размерности подразумевает изменения масштаба изображения. Реализация алгоритма BC применительно к медицинским данным представлена в работе [29].

Авторами [28] предложено увеличить быстродействие сжатия фрактальным методом путем уменьшения емкости пула доменов, в котором происходит поиск для текущей части изображения. Уменьшение пула доменов происходит за счет классификации их по пространственно-временному сходству. Дополнительно, для увеличения коэффициента сжатия, используется метод остаточной компенсации, который обеспечивается большей степенью корреляции между исходным изображением и восстановленным.

2. Критерии для оценки качества сжатия

Для оценки степени сжатия изображений и сохранения качества изображения при сжатии используют следующие характеристики.

Коэффициент сжатия (англ. *Compression Ratio*, сокр. *CR*) – отношение количества требуемой памяти для хранения исходного изображения $I(x, y)$ к количеству требуемой памяти для хранения сжатого изображения $I(x, y)'$:

$$CR = \frac{size(I(x,y))}{size(I(x,y)')}. \tag{20}$$

Среднее количество бит на один пиксель (англ. *Bits Per Pixel*, сокр. Bpp). Показатель отражает среднее количество бит, которое необходимо для представления одного пикселя:

$$Bpp = \frac{\text{size}(I(x,y)')}{\text{number of pixels}}. \tag{21}$$

Максимальная абсолютная ошибка (англ. *Maximum Absolute Error*, сокр. *MAE*). Показатель максимальной разницы между исходным и сжатым изображениями:

$$MAE = \max\left(\left|I(x,y) - I(x,y)'\right|\right). \tag{22}$$

Средний квадрат ошибки (англ. *Mean Square Error*, сокр. *MSE*). Показатель средней квадратичной разницы между исходным и сжатым изображениями:

$$MSE = \frac{1}{MN}\sum_{y=1}^{M}\sum_{x=1}^{N}\left[I(x,y) - I(x,y)'\right]^2. \tag{23}$$

Отношения пикового сигнала к шуму (англ. *Peak Signal-to-Noise Ratio*, сокр. *PSNR*). Отражает зависимость между максимально возможной интенсивностью пикселя к среднему квадрату ошибки. Чем выше это значение, тем лучшее качество изображения можно получить при сжатии:

$$PSNR = 20\log\frac{\left(2^n - 1\right)}{\sqrt{MSE}} = [\text{дБ}]. \tag{24}$$

Отношение сигнала к шуму (англ. *Signal-to-Noise Ratio*, сокр. *SNR*). Отношение мощности сигнала к мощности шума:

$$SNR = 10\log\frac{\sum_{y=1}^{M}\sum_{x=1}^{N}\left[I(x,y)\right]^2}{\sum_{y=1}^{M}\sum_{x=1}^{N}\left[I(x,y) - I(x,y)'\right]^2} = [\text{дБ}]. \tag{25}$$

Структурность содержимого (англ. *Structural Content*, сокр. *SC*). Показатель позволяет оценить качество изображения и детализированность структуры информации после сжатия (26). Чем выше значение, тем ниже качество восстановленного изображения:

$$SC = \frac{\sum_{y=1}^{M}\sum_{x=1}^{N}\left[I(x,y)\right]^2}{\sum_{y=1}^{M}\sum_{x=1}^{N}\left[I(x,y)'\right]^2}. \tag{26}$$

Индекс структурного сходства (англ. *Structural Similarity Index Measure*, сокр. *SSIM*). Показатель позволяет оценить ухудшение качества между исходным изображением и сжатым (27). В *SSIM* заключены особенности восприятия человеческого глаза, которые не учитывают другие показатели:

$$SSIM = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} = \left[\frac{2\mu_x\mu_y}{\mu_x^2 + \mu_x^2}\right]\left[\frac{2\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2}\right]\left[\frac{\sigma_{xy}}{\sigma_x\sigma_y}\right], \quad (27)$$

где C_1 и C_2 – константы; μ – локальное среднее значение и σ – стандартное отклонение, такие что (28) – (31):

$$\mu_x = \frac{1}{K}\sum_{i=1}^{K} x_i, \quad (28)$$

$$\mu_y = \frac{1}{K}\sum_{i=1}^{K} y_i, \quad (29)$$

$$\sigma_x = \left[\frac{1}{K-1}\sum_{i=1}^{K}\left(x_i - \mu_x\right)\right]^{\frac{1}{2}}, \quad (30)$$

$$\sigma_y = \left[\frac{1}{K-1}\sum_{i=1}^{K}\left(y_i - \mu_y\right)\right]^{\frac{1}{2}}. \quad (31)$$

Коэффициент корреляции (англ. *Correlation Coefficient*, сокр. *CC*). Показатель позволяет оценить корреляцию между исходным и сжатым изображением:

$$CC = \frac{\sum_{y=1}^{M}\sum_{x=1}^{N} I\left(x,y\right) I(x,y)'}{\sqrt{\sum_{y=1}^{M}\sum_{x=1}^{N}\left[I(x,y)'\right]^2}\sqrt{\sum_{y=1}^{M}\sum_{x=1}^{N}\left[I(x,y)\right]^2}}. \quad (32)$$

В табл. 1, приведен обзор современных алгоритмов сжатия медицинских изображений с указанием используемых методов, видов изображений и критериев, по которым проводилась оценка.

Таблица 1

Обзор исследований сжатия медицинских изображений

Table 1

An overview in medical images compression

Источник	Методы	Тип изображений	Оценка
1	2	3	4
[19]	1. Медианный фильтр 2. Прямое вейвлет-преобразование (Добеши 9/7) 3. Вектор квантования (древовидная структура) 4. Кодироване Хаффмана	МРТ, КТ, ультразвук Легкие, головной мозг, брюшная полость, позвоночник	1. CR 2. Bpp 3. MSE 4. PSNR 5. SSIM

Пролодження табл. 1

1	2	3	4
[17]	1. Дискретное косинусное преобразование 2. Кодирование Хаффмана	–	–
[14]	1. Лифтинг-схема на основе целочисленного вейвлет-преобразования (англ. Integer Wavelet Transform, сокр. IWT) 3. Адаптивный предсказатель 2. Энтропийное кодирование Хаффмана	МРТ, КТ Головной мозг, легкие, позвоночник, брюшная полость, суставы	1. CR 2. MAE 3. PSNR
[30]	1. SVD для извлечения ROI 2. Комбинированный метод (вейвлет Хаара и сверточная нейронная сеть) 3. EZW-кодирование	МРТ, КТ Головной мозг	1. MSE 2. PSNR 3. SSIM 5. Время исполнения (Runtime)
[11]	1. DPCM 2. Комбинированный метод (кодирование длин серий и арифметическое кодирование)	Медицинские изображения	1. CR 2. Удельная ошибка прогнозирования
[5]	1. LOCO-I (Low Complexity Lossless Compression) 2. Предсказатель на основе ближайшего соседа с расширенным контекстом 3. Энтропийное кодирование	–	1. Значение энтропии 2. Bpp 3. Время исполнения (Runtime/ Estimated time)
[31]	1. Алгоритм предсказания ALCM и адаптивный нелинейный фильтр 2. Кодирование Binary Layers Scaning (BLS)	3D КТ Головной мозг	1. Bpp
[27]	Фрактальные преобразования	3D МРТ Головной мозг	1. CR 2. PSNR 3. MSE 4. Время исполнения (Runtime/ Estimated time)
[21]	1. Предсказатель на DPCM 2. Вейвлет-преобразование Хаара 3. Кодирование Хаффмана	КТ Головной мозг, грудная клетка	1. CR 2. PSNR 3. MSE
[32]	1. Сегментация для извлечения ROI 2. Целочисленное вейвлет-преобразование 3. Кодирование Хаффмана	X-Ray	1. PSNR 2. SSIM 3. MSE 4. Время исполнения (Runtime/ Estimated time)

Окончание табл. 1

1	2	3	4
[33]	1. Логарифмическое преобразование 2. Вейвлет-преобразование 3. Кодирование Хаффмана	Ультразвук Брюшная полость	1. CR 2. PSNR 3. SSIM 4. CC
[34]	1. Биортогональный вейвлет / вейвлет Хаара / стационарное вейвлет-преобразование (англ. Sationary Wavelet Transform, сокр. SWT) 2. Кодирование Хаффмана / арифметическое / блочное	КТ, МРТ Голова	1. CR 2. PSNR 3. Bpp
[35]	1. Линейный предсказатель 2. Дискретное вейвлет-преобразование (DWT) 3. Кодирование Хаффмана	МРТ Голова	1. CR 2. MSE 3. PSNR 4. SNR 5. Время исполнения (Runtime/ Estimated time)
[24]	Быстрое преобразование Фурье (FFT)	КТ Головной мозг, грудная клетка	1. CR 2. PSNR 3. MSE 4. SC
[15]	1. LTSM (Long short-term memory) 2. Арифметическое кодирование	3D КТ и МРТ высокого разрешения Голова, шея, торс	1. Bpp 2. Время исполнения (Runtime/ Estimated time)
[23]	JPEG-XT 1. Двумерное прямое преобразование Фурье. 2. DC коэффициенты – DPCM кодирование AC коэффициенты – кодирование Хаффмана и RLE	3D МРТ высокого разрешения Голова	1. CR 2. MSE 3. PSNR
[36]	1. Дискретное косинусное преобразование 2. SVD (Singular Value Decomposition) 3. SPIHT (Set Partitioning In Hierarchical trees)	МРТ, КТ Головной мозг, позвоночник	1. CR 2. PSNR
[37]	Дискретное косинусное преобразование	МРТ Кости	1. CR 2. PSNR 3. SSIM

2. Обсуждение результатов

Обзор публикаций показал, что наиболее активно работа исследователей ведется в направлениях сжатия полутоновых изображений, а также изображений с высокой битовой глубиной. На сегодняшний момент первый тип изображений является наиболее распространенным при обмене и хранении в существующих системах PACS/RIS. Для решения текущих вызовов исследователи пытаются найти методы с более высоким коэффициентом сжатия и/или скорости кодирования/декодирования исходных/сжатых изображений. С точки зрения последующей перспективы применения в прикладных задачах чрезвычайно важное значение имеет разработка методов для хранения и передачи изображений с высокой битовой глубиной.

Алгоритмы сжатия разделяются на две категории – это методы, анализирующие частотную (вейвлет-функции, разложение Фурье) и пространственную области (прогнозирование, фракталы). Первые осуществляют сжатие благодаря поиску корреляций в спектральном составе исходного изображения и в последующем снижение с локализацией пространственного и частотного представлений. Вторые осуществляют сжатие благодаря поиску зависимостей значения пикселей от контекста или областей.

В случае применения вейвлет-преобразований можно сказать, что задача сводится к поиску вейвлет-функций, с помощью которых при разложении достигается наибольшее количество коэффициентов, равных нулю. Данное свойство позволяет более эффективно сжимать последовательность бит. Стандарт сжатия, основанный на вейвлет-функциях – JPEG 2000.

При использовании методов, основанных на предсказании, главным показателем является минимизация ошибки предсказания. С одной стороны, это обеспечивается захватом более широкого контекста пикселей, а с другой стороны, минимизацией этого количества путем устранения из рассмотрения пикселей, не оказывающих влияние или оказывающих ложное. Стандартом сжатия, базирующимся на поиске ошибки предсказания на основе контекста пикселей, является LossLess JPEG.

Особенностью применения Фурье-преобразования для входной последовательности в задачах сжатия является его эффективное представление областей изображений с высоким уровнем корреляции, а также с низкими значениями коэффициентов высокочастотных составляющих. Последнее позволяет приближать значения к нулю путем применения различных цифровых фильтров, рабочая частотная область которых отсекает шумы. Стандарты сжатия, основанные на Фурье-преобразованиях – JPEG и JPEG XR. JPEG XR является совместимыми с JPEG и активно развивается для сжатия медицинских изображений с высокой битовой глубиной.

Фрактальные методы по сравнению с известными стандартами способны сжимать изображения с высокими коэффициентами сжатия. Недостатком является возникновение потерь при сжатии, которые заложены в основе поиска соответствия между ранговыми и доменными областями. Однако отмечается, что при таком методе сжатия контуры костей и органов на изображениях хорошо различимы.

Использование различных методов прогнозирования показывает, что ошибка предсказания возрастает на контурах и текстурах. Для томографических данных головного мозга или костей снижение ошибки предсказания может быть осуществлено благодаря выравниванию исходного изображения к шаблону, а затем введения дополнительного контекста, который включает пиксели из соседней симметричной части изображения. Если говорить о многокадровых исследованиях, то дополнительный контекст может быть введен на основе предыдущих кадров. В этом случае устраняется межкадровая избыточность, которая в свою очередь снижает необходимое количество памяти для хранения. Способы ее устранения могут основываться на преобразованиях подобия или алгоритмах компенсации движения. Следует отметить, что КТ- и МРТ-изображения обладают высоким уровнем зашумленности. В этом случае в спектральном составе входной последовательности обнаруживается высокочастотная составляющая. Большинство

походов к сжатию изображений требует использование цифровых фильтров в составе преобразований. Например, наличие шумов существенно увеличивает ошибку предсказания, поэтому использование цифровых фильтров является важным при предсказании пикселей на основе соседей или схожих областей.

Отметим, что при сжатии изображений с потерями ключевой задачей является не только эффективное уменьшение объема занимаемой памяти, но и обеспечение максимально возможно точного восстановления исходного сигнала. Использование фундаментальных математических исследований в этой области позволяет принципиально улучшить существующие методы сжатия, предлагая алгоритмы, которые восстанавливают сигнал почти без потерь даже при сильной неполноте исходных данных. Например, теорема Котельникова – Найквиста устанавливает нижнюю границу частоты дискретизации, необходимую для однозначного восстановления сигнала. Однако существует подход «сжатого измерения», ключевая особенность которого заключается в восстановлении сигнала из наиболее значимых коэффициентов нового базиса. В работе [38] авторы восстанавливают исходный сигнал из набора скалярных произведений за полиномиальное время и с меньшим количеством измерений, чем это требуется по теореме Котельникова – Найквиста. Как было сказано ранее, существуют методы сжатия, которые преобразуют исходное изображение, представляя его в новом базисе. В случае преобразования Фурье и некоторых вейвлет-функций – это ортонормированный базис. Применение подходов, аналогичных описанным в [38; 39], позволит либо быстрее восстанавливать сигнал, либо увеличивать коэффициент сжатия, благодаря обнулению большего количества коэффициентов разложения.

Стандарт DICOM не предусматривает свободное изменение стандартов сжатия. Это обусловлено сложностью и многокомпонентностью систем PACS/РИС (Picture Archiving and Communication System – Радиологическая информационная система), МИС (Медицинская информационная система), ЦАМИ (Центральный архив медицинских изображений) и требованием к их отказоустойчивости и совместимости. Их наполненность, взаимосвязь и возможность масштабирования требуют строгой стандартизации данных, которые в них поступают, и методов их обработки. Однако при использовании отрытых или собственных архитектур и разработок появляется возможность внедрения новых методов сжатия для оптимизации и повышения быстродействия систем. Внедрение в отдельные компоненты таких систем новых разработок эффективных алгоритмов сжатия позволяет доказать их практическую значимость для последующего внедрения в стандарт и дальнейшего широкого применения.

Кроме того, получение показателей влияния внедрения новых методов сжатия в существующие высоконагруженные медицинские системы может быть произведено с использованием инструмента администрирования PacsMan [40]. Он позволяет проанализировать весь обмен данных и все события в информационной системе, начиная от генерации исследования на приборе и его архивирования в ЦАМИ и заканчивая запросами исходных исследований для обработки на станции врача или сервисами поддержки врачебных решений [41], например [42; 43]. Таким образом, можно получить наиболее реальную оценку при внедрении новых методов сжатия в рабочую систему.

Заключение

Требование к повышению эффективности хранения, передачи и обработки радиологических исследований связано с увеличением потока данных в последние годы. Существенное увеличение обрабатываемых исследований связано с активным распространением телемедицины, а также повышением доступности к диагностированию заболеваний на раннем этапе для населения с помощью рентгенологии. Стоит отметить, что дальнейшее повсеместное вне-

дрение систем поддержки врачебных решений, в том числе с использованием искусственного интеллекта, еще острее ставит проблему эффективного использования ресурсов.

В настоящей работе был проведен обзор публикаций по методам сжатия радиологических исследований, который показал большое количество разнообразных подходов в зависимости от требований. К требованиям относятся улучшение показателей сжатия в сравнении с существующими стандартами, предусмотренными спецификацией DICOM.

Анализ литературных источников показал, что использование высокочастотных цифровых фильтров в большинстве подходов повышает коэффициент сжатия из-за высокой зашумленности изображения, хотя уменьшает быстродействие. Некоторым типам изображений, например КТ и МРТ головы и костей, свойственно наличие симметрии схожих областей относительно центра изображения, что возможно использовать при сжатии, используя алгоритмы для выравнивания исходного изображения. Кроме того, каждое исследование состоит из множества последовательных кадров, поэтому использование алгоритмов, анализирующих перемещение отдельных блоков от кадра к кадру (преобразования подобия или алгоритмы компенсации движения), сокращает межкадровую избыточность исследования. Таким образом, использование вышеперечисленных свойств, а также разработка и применение более эффективных цифровых фильтров позволяет оптимизировать использование ресурсов для хранения, передачи и обработки медицинских исследований.

Список литературы

1. **Краюшкин Д. В., Чеповский А. М.** Проблемы сопровождения медицинских информационных систем // Инжиниринг предприятий и управление знаниями (ИП&УЗ-2024), М.: РЭУ им. Г. В. Плеханова. 2024. С. 172–177.
2. **Mu L., Liu H.** Noninvasive electrocardiographic imaging with low-rank and non-local total variation regularization // Pattern Recognition Letters, 2020, vol. 138, pp. 106–114. DOI 10.1016/j.patrec.2020.07.007
3. **Данилкина Ю. С., Крамм М. Н., Чыонг Т. Л. Н., Бодин А. Ю., Краюшкин Д. В.** Многоканальная регистрация ЭКГ с поверхности женского торса и визуализация характеристик сердца // Научная визуализация. 2024. Т. 16, № 3. С. 97–105. DOI 10.26583/sv.16.3.10
4. **Gonzalez R. C., Woods R. E.** Digital Image Processing. 4th ed. Pearson, 2018.
5. **Amin M. S., Jabeen S., Wang C. et al.** Improved Median Edge Detection (iMED) for Lossless Image Compression // Image Analysis and Stereology, 2023, vol. 42(1), pp. 25–35. DOI 10.5566/ias.2786
6. **Егоров Н. Д., Новиков Д. В., Гильмутдинов М. Р.** Метод сжатия изображений без потерь с помощью контекстного кодирования по двоичным уровням // Информационно-управляющие системы. 2017. Т. 6(91). С. 96–106. DOI: 10.15217/issn1684-8853.2017.6.96
7. **Сушко Д. В., Штарьков Ю. М.** О сжатии томографических данных // Информационные процессы. 2008. Т. 8(4) С. 240–255.
8. **Стефанович А. И., Сушко Д. В.** Обратимое сжатие данных посредством универсального арифметического кодирования // Информатика и ее применения. 2017. Т. 11(1). С. 20–45. DOI: 10.14357/19922264170103
9. **Сушко Д. В.** Алгоритмы сжатия данных массивов силовых кривых I: кодирование ошибок предсказания // Информатика и ее применения, 2021, Т. 15(2). С. 82–88. DOI 10.14357/19922264210212
10. **Сушко Д. В.** Алгоритмы сжатия данных массивов силовых кривых II: кодирование компонент вейвлет-преобразования // Информатика и ее применения. 2021. Т. 15(3). С. 16–23. DOI: 10.14357/19922264210303

11. **Садик Б. Д. С., Цветков В. Ю., Бобов М. Н.** Адаптивное комбинированное кодирование изображений с прогнозированием объема арифметического кода // Доклады БГУИР. 2021. Т. 19(2). С. 31–39. DOI 10.35596/1729-7648-2021-19-2-31-39
12. **Karimi N., Samavi S., Amraee S. et al.** Use of symmetry in prediction-error field for lossless compression of 3D MRI images // Multimedia Tools and Applications, 2014, vol. 74. DOI 10.1007/s11042-014-2214-9
13. **Bhalerao A., Wilson R.** Warplets: an image-dependent wavelet representation // IEEE ICIP, 2005. DOI 10.1109/ICIP.2005.1530099
14. **Pathak K. C., Sarvaiya J. N.** Lossless medical image compression using transform domain adaptive prediction for telemedicine // IEEE WiSPNET, 2017. DOI 10.1109/WiSPNET.2017.8299918
15. **Nagoor O. H., Whittle J., Deng J. et al.** Sampling strategies for learning-based 3D medical image compression // Machine Learning with Applications, 2022, vol. 8, pp. 100273. DOI 10.1016/j.mlwa.2022.100273
16. **Antonini M., Barlaud M., Mathieu P. et al.** Image coding using wavelet transform // IEEE Transactions on Image Processing, 1992, vol. 1(2), pp. 205–220.
17. **Chaudhary A. K., Mehrotra R., Ansari M. A. et al.** A Novel Scheme for Medical Image Compression Using Huffman and DCT Techniques // Advances in Smart Communication and Imaging Systems, 2021. DOI: 10.1007/978-981-15-9938-5_28
18. **Agrawal R., Singh K., Goyal A.** Advances in Smart Communication and Imaging Systems // Lecture Notes in Electrical Engineering, 2021, vol. 721. DOI 10.1007/978-981-15-9938-5
19. **Ammah P. N. T., Owusu E.** Robust medical image compression based on wavelet transform and vector quantization // Informatics in Medicine Unlocked, 2019, vol. 15, pp. 100183. DOI 10.1016/j.imu.2019.100183
20. **Bairagi V.** Lossless Medical Image Compression by Integer Wavelet and Predictive Coding // Biomedical Engineering, 2013. DOI 10.1155/2013/832527
21. **Abo-Zahhad M., Gharieb R., Ahmed S., Abd-Ellah M.** Huffman Image Compression Incorporating DPCM and DWT // Journal of Signal and Information Processing, 2015, vol. 6. pp. 123–135. DOI 10.4236/jsip.2015.62012.
22. **Richter T., Bruylants T., Schelkens P. et al.** The JPEG XT suite of standards: status and future plans // Proc. SPIE, 2015. DOI 10.1117/12.2189873
23. **Li Z., Ramos A., Li Z. et al.** An optimized JPEG-XT-based algorithm for the lossy and lossless compression of 16-bit depth medical image // Biomedical Signal Processing and Control, 2021, vol. 64. DOI 10.1016/j.bspc.2020.102306
24. **Karthikeyan T., Thirumoorthi C.** A Hybrid Medical Image Compression Techniques for Lung Cancer // Indian Journal of Science and Technology, 2016, vol. 9(39), pp. 1–6. DOI 10.17485/ijst/2016/v9i39/91500
25. **Илюшин С. В.** Разработка алгоритмов быстрого фрактального сжатия цифровых изображений: дис. … канд. техн. наук, М., 2012.
26. **Bhavani S., Thanushkodi K. G.** Comparison of fractal coding methods for medical image compression // IET Image Process, 2013, vol. 7(7), pp. 686–693. DOI 10.1049/iet-ipr.2012.0041
27. **Liu S., Bai W., Zeng N. et al.** A fast fractal based compression for MRI images // IEEE Access, 2019, vol. 7 pp. 62412–62420. DOI 10.1109/ACCESS.2019.2916934
28. **Srimal A., Peters J. F., Ramanna S. et al.** Quaternionic views of rs-fMRI hierarchical brain activation regions // Chaos, Solitons & Fractals, 2021, vol. 152, pp. 111351. DOI 10.1016/j.chaos.2021.111351
29. **Miras J. R., Posadas M. A., Ibañez-Molina A. J. et al.** Fast Computation of Fractal Dimension for 2D, 3D and 4D Data // Journal of Computational Science, 2023, vol. 66, pp. 101908. DOI 10.1016/j.jocs.2022.101908
30. **Li S., Lu J., Hu Y. et al.** Towards scalable medical image compression using hybrid model analysis // Journal of Big Data, 2025, vol. 12. DOI: 10.1186/s40537-025-01073-1

31. **Гильмутдинов М. Р., Веселов А. И.** Способ устранения межкадровой избыточности при сжатии медицинских 3D-изображений // сб. докладов 18-й Междунар. конф. «Цифровая обработка сигналов и ее применение – DSPA-2016». С. 832–838.
32. **Vamsikrishna M., Sudhakar O., Bugge B.P. et al.** Region based lossless compression for digital images using entropy coding // Indonesian Journal of Electrical Engineering and Computer Science. 2025, vol. 38(3), pp. 1870–1879. DOI 10.11591/ijeecs.v38.i3
33. **Kim K., Pak M., Ri Y.-W. et al.** An image compression method for improving noise robustness of ultrasonic medical image compression in wavelet domain // Multimedia Tools and Applications, 2025. DOI 10.1007/s11042-025-20943-7
34. **Anusuya V., Stency V., Srividhya G. et al.** Performance Comparison of Wavelet Transforms based Medical Image Compression // Journal of Cybersecurity and Information Management, 2025, vol. 16, pp. 1–12. DOI 10.54216/JCIM.160201
35. **Mofreh A., Barakat T. M., Refaat A. M.** A New Lossless Medical Image Compression Technique using Hybrid Prediction Model // SPIJ, 2016, vol. 10(3).
36. **Reddy M. R., Ravichandran K. S. et al.** Improved pharma education system in the field of medical images using compression techniques // Cluster Computing, 2019, vol. 22(2). DOI 10.1007/s10586-018-2496-1
37. **Pandey A., Yadav P., Chaudhary J. et al.** Compression of 99mTc Methylene Diphosphonate Bone Scan Images using Discrete Cosine Transformation // Indian Journal of Nuclear Medicine, 2022, vol. 37(4), pp. 337. DOI 10.4103/ijnm.ijnm_45
38. **Кашин В. С., Темляков В. Н.** Замечание о задаче сжатого измерения // Математические заметки. 2007. Т. 82, № 6. С. 829–837.
39. **Kosov E., Temlyakov V.** Sampling discretization of the uniform norm and applications // Journal of Mathematical Analysis and Applications, 2024, vol. 538(2), pp. 128431. DOI 10.1016/j.jmaa.2024.128431
40. **Гаврилов А. В., Куликов И. В., Краюшкин Д. В.** Программа графического пользовательского интерфейса для сопровождения информационной системы PACS/RIS. Свидетельство о гос. регистрации программы для ЭВМ № 2025615636, 2025.
41. **Гаврилов А. В., Краюшкин Д. В., Куликов И. В., Соломинов М. В., Чеповский А. М.** Современные подходы к сопровождению медицинских информационных систем при внедрении новых технологий // Вопросы кибербезопасности. 2025. Т. 2(66). С. 41–51. DOI 10.21681/2311-3456-2025-2-41-51
42. **Гаврилов А. В., Долотова Д. Д., Парусников А. В., Благосклонова Е. Р., Соломинова Т. А., Акимова Е. А., Краюшкин Д. В.** Программа комплексного анализа DICOM-изображений компьютерной томографии головного мозга при острых нарушениях мозгового кровообращения «Multivox AI Stroke». Свидетельство о гос. регистрации программы для ЭВМ № 2024690602, 2024.
43. **Сандриков В. А., Кулагина Т. Ю., Гаврилов А. В., Благосклонова Е. Р., Соломинов М. А., Краюшкин Д. В.** Программа регистрации, визуализации, обработки и архивирования ультразвуковых исследований движения стенок аорты методом спекл-трекинга. Свидетельство о гос. регистрации программы для ЭВМ № 2025668216, 2025.

References

1. **Krayushkin D. V., Chepovskiy A. M.** Problems of supporting medical information data systems. *Inzhiniring predpriyatiy i upravlenie znaniyami [Enterprise Engineering and Knowledge Management] Moscow: Plekhanov Russian Academy of Economics*, 2024, pp. 172–177.
2. **Mu L., Liu H.** Noninvasive electrocardiographic imaging with low-rank and non-local total variation regularization. *Pattern Recognition Letters*, 2020, vol. 138, pp. 106–114. DOI 10.1016/j.patrec.2020.07.007

3. **Danilkina Y. S., Kramm M. N., TTruong T. L. N., Bodin A. Y., Krayushkin D. V.** Multichannel ECG ecording from the surface of the female torso and visualization of heart characteristics. *Scientific Visualization*, 2024, vol. 16(3), pp. 97–105, DOI: 10.26583/sv.16.3.10
4. **Gonzalez R. C., Woods R. E.** Digital Image Processing. 4th ed. Pearson, 2018.
5. **Amin M. S., Jabeen S., Wang C. et al.** Improved Median Edge Detection (iMED) for Lossless Image Compression. *Image Analysis and Stereology*, 2023, vol. 42(1), pp. 25–35. DOI 10.5566/ias.2786
6. **Egorov N. D., Novikov D. V., Gilmutdinov M. R.** Lossless Image Compression using Binary Layers Scanning Data Encoding. *Information and Control Systems*. 2017, vol. 6(91), pp. 96–106. DOI: 10.15217/issn1684-8853.2017.6.96
7. **Sushko D.V., Shtar'kov Y.M.** O szhatii tomograficheskikh dannykh [On tomography data compression]. *Information Processes*. 2008, vol. 8(4), pp. 240–255. (in Russ.)
8. **Stefanovich A. I., Sushko D. V.** Reversible data compression by universal arithmetic coding. *Informatics and Applications*, 2017, vol. 11(1), pp. 20–45. DOI10.14357/19922264170103
9. **Sushko D. V.** Compression algorithms for force volume data I: Prediction errors coding. *Informatics and Applications*, 2021, vol. 15(2), pp. 82–88. DOI 10.14357/19922264210212
10. **Sushko D. V.** Compression algorithms for force volume data ii: coding of wavelet transform components. *Informatics and Applications*, 2021, vol, 15(3), pp. 16–23. DOI: 10.14357/19922264210303
11. **Sadiq B. J., Tsviatkou V. Yu., Bobov M. N.** Adaptive combined image coding with prediction of arithmetic code volume. *Doklady BGUIR*, 2021, vol. 19(2), pp. 31–39. DOI 10.35596/1729-7648-2021-19-2-31-39
12. **Karimi N., Samavi S., Amraee S. et al.** Use of symmetry in prediction-error field for lossless compression of 3D MRI images. *Multimedia Tools and Applications*, 2014, vol. 74. DOI 10.1007/s11042-014-2214-9
13. **Bhalerao A., Wilson R.** Warplets: an image-dependent wavelet representation. *IEEE ICIP*, 2005. DOI: 10.1109/ICIP.2005.1530099
14. **Pathak K. C., Sarvaiya J. N.** Lossless medical image compression using transform domain adaptive prediction for telemedicine. *IEEE WiSPNET*, 2017. DOI: 10.1109/WiSPNET.2017.8299918
15. **Nagoor O. H., Whittle J., Deng J. et al.** Sampling strategies for learning-based 3D medical image compression. *Machine Learning with Applications*, 2022, vol. 8, pp. 100273. DOI 10.1016/j.mlwa.2022.100273
16. **Antonini M., Barlaud M., Mathieu P. et al.** Image coding using wavelet transform. *IEEE Transactions on Image Processing*, 1992, vol. 1(2), pp. 205–220.
17. **Chaudhary A. K., Mehrotra R., Ansari M. A. et al.** A Novel Scheme for Medical Image Compression Using Huffman and DCT Techniques. *Advances in Smart Communication and Imaging Systems*, 2021. DOI: 10.1007/978-981-15-9938-5_28
18. **Agrawal R., Singh K., Goyal A.** Advances in Smart Communication and Imaging Systems. *Lecture Notes in Electrical Engineering, Springer*, 2021, vol. 721. DOI 10.1007/978-981-15-9938-5
19. **Ammah P. N. T., Owusu E.** Robust medical image compression based on wavelet transform and vector quantization. *Informatics in Medicine Unlocked*, 2019, vol. 15, pp. 100183. DOI 10.1016/j.imu.2019.100183
20. **Bairagi V.** Lossless Medical Image Compression by Integer Wavelet and Predictive Coding. *Biomedical Engineering*, 2013. DOI 10.1155/2013/832527
21. **Abo-Zahhad M., Gharieb R., Ahmed S., Abd-Ellah M.** Huffman Image Compression Incorporating DPCM and DWT. *Journal of Signal and Information Processing*, 2015, vol. 6, pp. 123–135. DOI 10.4236/jsip.2015.62012.
22. **Richter T., Bruylants T., Schelkens P. et al.** The JPEG XT suite of standards: status and future plans. *Proc*. SPIE, 2015. DOI: 10.1117/12.2189873
23. **Li Z., Ramos A., Li Z. et al.** An optimized JPEG-XT-based algorithm for the lossy and lossless compression of 16-bit depth medical image. *Biomedical Signal Processing and Control*, 2021. vol. 64. DOI 10.1016/j.bspc.2020.102306

24. **Karthikeyan T., Thirumoorthi C.** A Hybrid Medical Image Compression Techniques for Lung Cancer. *Indian Journal of Science and Technology*, 2016, vol. 9(39), pp. 1–6. DOI 10.17485/ijst/2016/v9i39/91500
25. **Ilyushin S. V.** Development of algorithms for fast fractal compression of digital images: dis. ... Candidate of Technical Sciences, Moscow, 2012.
26. **Bhavani S., Thanushkodi K. G.** Comparison of fractal coding methods for medical image compression. *IET Image Process*, 2013, vol. 7(7), pp. 686–693. DOI 10.1049/iet-ipr.2012.0041
27. **Liu S., Bai W., Zeng N. et al.** A fast fractal based compression for MRI images. *IEEE Access*, 2019, vol. 7 pp. 62412–62420. DOI 10.1109/ACCESS.2019.2916934
28. **Srimal A., Peters J. F., Ramanna S. et al.** Quaternionic views of rs-fMRI hierarchical brain activation regions. *Chaos, Solitons & Fractals*, 2021, vol. 152, pp. 111351. DOI 10.1016/j.chaos.2021.111351
29. **Miras J. R., Posadas M. A., Ibañez-Molina A.J . et al.** Fast Computation of Fractal Dimension for 2D, 3D and 4D Data. *Journal of Computational Science*, 2023, vol. 66, pp. 101908. DOI 10.1016/j.jocs.2022.101908
30. **Li S., Lu J., Hu Y. et al.** Towards scalable medical image compression using hybrid model analysis. *Journal of Big Data*, 2025, vol. 12. DOI: 10.1186/s40537-025-01073-1
31. **Gilmutdinov M. R., Veselov A. I.** A way to eliminate inter-frame redundancy when compressing 3D medical images. *Collection of reports of the 18th International conference "Digital Signal processing and its application – DSPA-2016"*. pp. 832-838.
32. **Vamsikrishna M., Sudhakar O., Bugge B. P. et al.** Region based lossless compression for digital images using entropy coding. *Indonesian Journal of Electrical Engineering and Computer Science*. 2025, vol. 38(3), pp. 1870–1879. DOI: 10.11591/ijeecs.v38.i3
33. **Kim K., Pak M., Ri Y.-W. et al.** An image compression method for improving noise robustness of ultrasonic medical image compression in wavelet domain. *Multimedia Tools and Applications*, 2025. DOI: 10.1007/s11042-025-20943-7
34. **Anusuya V., Stency V., Srividhya G. et al.** Performance Comparison of Wavelet Transforms based Medical Image Compression. *Journal of Cybersecurity and Information Management*, 2025, vol. 16, pp. 1–12. DOI 10.54216/JCIM.160201
35. **Mofreh A., Barakat T. M., Refaat A. M.** A New Lossless Medical Image Compression Technique using Hybrid Prediction Model. SPIJ, 2016, vol. 10(3).
36. **Reddy M. R., Ravichandran K. S. et al.** Improved pharma education system in the field of medical images using compression techniques. *Cluster Computing*, 2019, vol. 22(2). DOI 10.1007/s10586-018-2496-1
37. **Pandey A., Yadav P., Chaudhary J. et al.** Compression of 99mTc Methylene Diphosphonate Bone Scan Images using Discrete Cosine Transformation. *Indian Journal of Nuclear Medicine*, 2022, vol. 37(4), pp. 337. DOI 10.4103/ijnm.ijnm_45
38. **Kashin B. S., Temlyakov V. N.** A remark on Compressed Sensing. *Mathematical Notes*, 2007, vol. 82, pp. 748–755. DOI 10.1134/S0001434607110193
39. **Kosov E., Temlyakov V.** Sampling discretization of the uniform norm and applications. *Journal of Mathematical Analysis and Applications*, 2024, vol. 538(2), pp. 128431. DOI 10.1016/j.jmaa.2024.128431.
40. **Gavrilov A. V., Kulikov I. V., Krayushkin D. V.** Programma graficheskogo polzovatel'skogo interfeysa dlya soprovozhdeniya informatsionnoy sistemy PACS/RIS [A GUI Tool for PACS/RIS Information System Support]. The Certificate on Official Registration of the Computer Program in Russia. No. 2025615636. 2025. (In Russ.)
41. **Gavrilov A. V., Krayshkin D. V., Kulikov I. V., Solominov M. V., Chepovsky A. M.** Modern approaches to supporting and improving medical information systems. Voprosy kiberbezopasnosti, 2025, vol. 2(66), pp. 41–51. DOI 10.21681/2311-3456-2025-2-41-51

42. **Gavrilov A. V., Dolotova D. D., Parusnikov A. V., Blagosklonova E. R., Solominova T. A., Akimova E. A., Krayushkin D. V.** Programma kompleksnogo analiza DICOM-izobrazheniy komp'yuternoy tomografii golovnogo mozga pri ostrykh narusheniyakh mozgovogo krovoobrashcheniya «Multivox AI Stroke» [A Software for Comprehensive Analysis of DICOM Images from Brain Computed Tomography in Acute Cerebral Circulation Disorders «Multivox AI Stroke»]. The Certificate on Official Registration of the Computer Program in Russia. No. 2024690602. 2024. (In Russ.)
43. **Sandrikov V. A., Kulagina T. Y., Gavrilov A. V., Blagosklonova E. R., Solominov M. A., Krayushkin D. V.** Programma registracii, vizualizacii, obrabotki i arhivirovaniya ul'trazvukovyh issledovanij dvizheniya stenok aorty metodom spekl-trekinga [A Software for acquisition, visualization, processing, and archiving of ultrasound-based aortic wall motion studies using speckle tracking]. The Certificate on Official Registration of the Computer Program in Russia. No. 2025668216. 2025. (In Russ.)

Сведения об авторах

Гаврилов Андрей Васильевич, кандидат технических наук, заведующий лабораторией медицинских компьютерных систем Научно-исследовательского института ядерной физики им. Д. В. Скобельцына Московского государственного университета им. М. В. Ломоносова

Краюшкин Денис Владиславович, аспирант Национального исследовательского университета «Высшая школа экономики»

Чеповский Андрей Михайлович, доктор технических наук, профессор; профессор Национального исследовательского университета «Высшая школа экономики»; профессор Российского экономического университета им. Г. В. Плеханова

Information about the Authors

Andrey V. Gavrilov, Candidate of Engineering Sciences, Head of the laboratory at Lomonosov Moscow State University

Denis V. Krayushkin, Postgraduate student at the HSE University

Andrey M. Chepovskiy, Doctor of Engineering Sciences, Professor at the HSE University, Moscow, Russia; Professor at the Plekhanov Russian University of Economics

Статья поступила в редакцию 30.09.2025;
одобрена после рецензирования 15.10.2025; принята к публикации 15.10.2025

The article was submitted 30.09.2025;
approved after reviewing 15.10.2025; accepted for publication 15.10.2025

Научная статья

УДК 004.89
DOI 10.25205/1818-7900-2025-23-4-44-61

Исследование методов оптимизации скорости исполнения больших языковых моделей для задачи распознавания команд

Александр Игоревич Гончаренко[1]
Максим Иванович Чупров[2]
Евгений Семенович Нежевенко[3]

[1]Институт интеллектуальной робототехники НГУ
Новосибирск, Россия

[2]ООО «Экспасофт»
Новосибирск, Россия

[3]Институт автоматики и электрометрии СО РАН
Новосибирск, Россия

a.goncharenko@expasoft.tech; https://orcid.org/0009-0000-5087-8506
m.chuprov@expasoft.tech
nedj@iae.nsk.su

Аннотация

Целью данной работы являлось исследование и реализация методов оптимизации (особенно методов прунинга) больших языковых моделей для задачи function calling, а также сравнение точности и скорости работы полученных моделей.

В качестве базовой модели была выбрана модель Mistral-7B. Для эффективной тренировки модели использовался датасет glaive-function-calling-v2, предназначенный для задачи function calling. Для обучения базовой модели использовалось квантование до 4 бит в формате nf4 и двойное квантование в сочетании с методом QLoRA (Quantized Low-Rank Adaptation).

Оптимизация модели проводилась несколькими способами: (1) с использованием метода ShortGPT, (2) с помощью критерия Тейлора для послойного прунинга, (3) методом LLM-Pruner, который отбрасывает параметры модели поканально, оставляя при этом количество слоев модели неизменным, и (4) методом PowerInfer, который использует свойство контекстуальной разреженности в больших языковых моделях. Для всех перечисленных способов оптимизации были построены оптимизированные модели, и проведено сравнение точности и скорости работы полученных моделей.

Результаты экспериментов показали, что наибольшая точность была достигнута на модели, которая была оптимизирована с помощью метода послойного прунинга по критерию Тейлора важности слоя. Для данного метода был проведен ряд экспериментов, в которых исследовалась разная расстановка гейтов внутри слоя декодера, а также различные способы агрегирования важности слоя на гейтах. По итогам экспериментов можно сделать вывод, что расстановка гейтов после блоков Multi-Head Attention и использование агрегирования важности с помощью L2-нормы вектора градиентов дают наибольшую точность по сравнению с другими возможными вариантами.

Научная значимость работы состоит в сравнении передовых методов прунинга, исходя из соотношения качество/скорость модели, и получении ускоренной версии модели для задачи function calling.

Ключевые слова

прунинг, квантование, ряд Тейлора, большие языковые модели, механизм внимания, function calling, PowerInfer

Благодарности

Авторы выражают благодарность Чеблаковой Елене Анатольевне за помощь в оформлении статьи.

Для цитирования

Гончаренко А. И., Чупров М. И., Нежевенко Е. С. Исследование методов оптимизации скорости исполнения больших языковых моделей для задачи распознавания команд // Вестник НГУ. Серия: Информационные технологии. 2025. Т. 23, № 4. С. 44–61. DOI 10.25205/1818-7900-2025-23-4-44-61

Research of inference speed optimization methods of large language models for function calling task

Alexander I. Goncharenko[1], Maxim I. Chuprov[2]
Evgeniy S. Nejevenko[3]

[1]Institute of Intelligent Robotics of Novosibirsk State University
Novosibirsk, Russian Federation

[2]Expasoft LLC
Novosibirsk, Russian Federation

[3]Institute of Automation and Electrometry of the Siberian Branch of the Russian Academy of Sciences
Novosibirsk, Russian Federation

a.goncharenko@expasoft.tech; https://orcid.org/0009-0000-5087-8506
m.chuprov@expasoft.tech
nedj@iae.nsk.su

Abstract

This work is devoted to study and practical implementation of optimization methods (especially pruning) for large language models (LLM) in the context of function calling task, as well as comparison of accuracy and speed of the obtained models.

Authors chose Mistral-7B as the basic model; glaive-function-calling-v2 – as dataset for training. 4-bit quantization in nf4 format and double quantization were used in combination with QLoRA (Quantized Low-Rank Adaptation) method. Four different pruning methods were applied for model optimization. The first method, ShortGPT, focuses on reducing the model size by trimming less significant parts. The second method is based on Taylor's criterion for layer-by-layer pruning. The third method, LLM-Pruner, removes parameters channel-by-channel maintaining the total number of layers. The fourth method, PowerInfer, uses contextual sparsity of large language models. Optimized models were implemented for all these methods; the accuracy and speed of resulting models were compared.

Results of experiments show that the highest accuracy was achieved using the layer-by-layer pruning according to Taylor's criterion of layer importance. This method was tested with different placement of gates within the decoder layer and different ways of aggregation of layer importance on the gates. Experiments show that best results were achieved by placing the gates after Multi-Head Attention blocks and using the L2 norm of the gradient vector to aggregate layer importance.

Scholarly importance of the work includes comparison of advanced pruning methods in the context of quality/speed ratio and obtaining a speed up version model for the function calling task.

Keywords

pruning, quantization, Taylor series, large language models, attention mechanism, function calling, PowerInfer.

Acknowledgements

The authors thank Elena A. Cheblakova for help in drafting the article.

For citation

Goncharenko A. I., Chuprov M. I., Nejevenko E. S. Research of inference speed optimization methods of large language models for function calling task. *Vestnik NSU. Series: Information Technologies*, 2025, vol. 23, no. 4, pp. 44–61 (in Russ.) DOI 10.25205/1818-7900-2025-23-4-44-61

Введение

В последнее время большую популярность приобрели большие языковые модели (Large Language Models, LLM). Они стали ключевым инструментом в различных областях, начиная с обработки естественного языка и генерации текста и заканчивая агентными системами и база-

ми знаний. Такие модели, как GPT (Generative Pretrained Transformer) [1] и BERT (Bidirectional Encoder Representations from Transformers) [2], показали впечатляющие результаты в области понимания и обработки человеческого языка.

Отличительно чертой больших языковых моделей является использование огромных объемов данных и вычислительных ресурсов для достижения высокой производительности и универсальности. Целью данной работы являлось исследование и реализация методов оптимизации больших языковых моделей для повышения их эффективности и доступности в использовании.

1. Анализ предметной области

1.1. Языковое моделирование и большие языковые модели

На данный момент в области обработки естественного языка (Natural Language Processing, NLP) одной из самых распространенных задач является задача языкового моделирования. Языковое моделирование является основой для многих приложений NLP, таких как машинный перевод, суммаризация, ответы на вопросы (Question answering, QA) и др.

Языковое моделирование представляет собой предсказание следующего слова или последовательности слов в заданном контексте на основе статистического анализа языка. Для решения данной задачи существуют различные языковые модели, начиная от ранних Word2Vec [3] и заканчивая современными большими языковыми моделями.

В основе больших языковых моделей обычно лежит архитектура «трансформер» [4], точнее, ее авторегрессионная часть для генерации последовательностей в виде текста, также называемая декодером. Трансформер – это архитектура глубокой нейронной сети, предназначенная для обработки последовательностей, таких как тексты или временные ряды. Она основана на механизме внимания (self-attention), который позволяет модели фокусироваться на различных частях входных данных в зависимости от их важности для конкретной задачи. Механизм внимания в трансформере реализуется блоками Multi-Head Attention. Иллюстрация работы этих блоков приведена на рис. 1.

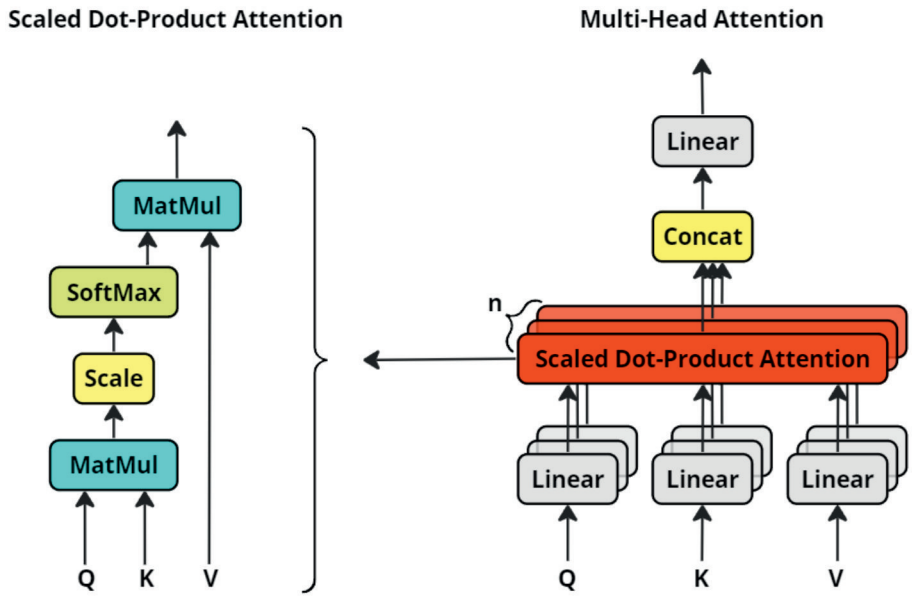


Рис. 1. Блок Multi-Head Attention
Fig. 1. Multi-Head Attention block

1.2. Методы оптимизации нейронных сетей

Основными методами оптимизации нейронных сетей являются прунинг и квантование.

1.2.1. Прунинг

Прунинг – это метод оптимизации, используемый в искусственных нейронных сетях для удаления отдельных параметров или групп параметров из существующей сети, чтобы сохранить точность сети и повысить ее эффективность. Существует два вида прунинга: структурированный и неструктурированный.

Структурированный прунинг предполагает удаление параметров из сети с определенной структурой, такой как удаление целых слоев или блоков параметров. Например, можно удалять целые сверточные фильтры в сверточных слоях или целые нейроны в полносвязных слоях. Этот подход обычно более прост в реализации и может обеспечить более стабильные результаты при сохранении производительности сети.

Неструктурированный прунинг, напротив, удаляет отдельные параметры независимо от их структуры, что может привести к более разреженным моделям. Например, это может быть удаление отдельных весов внутри сверточного фильтра или отдельных весов в полносвязных слоях. Хотя неструктурированный прунинг позволяет обеспечить большую степень сжатия модели, он может быть более сложен в реализации и требователен к вычислительным ресурсам.

Пример обоих видов прунинга изображен на рис. 2.

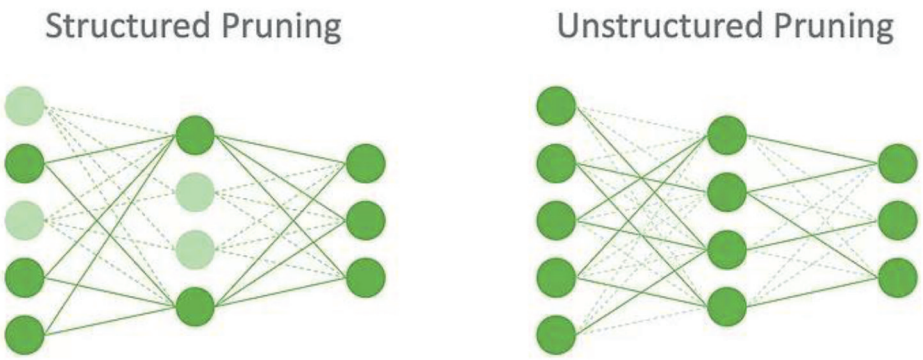


Рис. 2. Структурированный и неструктурированный прунинг
Fig. 2. Structured and unstructured pruning

Помимо этого, существуют различные критерии прунинга нейронных сетей, например, на основе информации об L2-норме параметра или их градиенте.

Так, при использовании L2-критерия, важность параметра определяется как

$$I_w = w^2,$$

где w – значение параметра; I_w – важность параметра w, определяемая как квадрат значения данного параметра. Важность группы параметров определяется как сумма важностей параметров:

$$I_W = \sum_{w \in W} w^2,$$

где W – группа параметров; I_W – важность группы параметров.

Критерий на основе градиента основывается на том, что важность параметра может быть оценена по ошибке, вызванной его удалением. Эта ошибка может быть измерена как разность значений функций потерь с параметром и без него следующим образом:

$$I_w = \left| L_w - L_{w=0} \right|,$$

где L_w – значение функции потерь с параметром w; $L_{w=0}$ – значение функции потерь без параметра w; I_w – важность параметра w. Если представить функции потерь в виде разложения Тейлора, то получим следующее:

$$I_w = \left| L_{w=0} + \frac{\partial L_w}{\partial w} w - L_{w=0} - \frac{\partial L_{w=0}}{\partial \left(w = 0 \right)} \cdot 0 \right| = \left| \frac{\partial L_w}{\partial w} w \right| = \left| g_w w \right|,$$

где g_w – значение градиента для веса w; I_w – важность веса w, которая считается как модуль произведения веса на значение градиента этого веса.

Приведенные примеры критериев для прунинга имеют свои достоинства и недостатки, которые следует учитывать при выборе конкретного метода. Так, к достоинствам критерия на основе L2-нормы весов можно отнести простоту применения и интуитивную интерпретацию. Недостатком критерия является то, что он не учитывает взаимосвязь между параметрами, что может привести к потере важной информации при прунинге.

У критерия на основе градиента наоборот, благодаря использованию градиента, появляется возможность оценить, насколько параметр влияет на функцию потерь, что позволяет сохранить наиболее важные параметры, но при этом сам критерий является вычислительно затратным из-за необходимости вычислять этот градиент.

1.2.2. Квантование

Квантование – это процесс уменьшения размера весов, смещений и активаций, обычно с 32-битных значений с плавающей точкой до более низких битов, например 16 или 8. Такое снижение точности позволяет получить более компактное представление модели, что приводит к снижению потребления памяти и повышению скорости вычислений.

В общем виде операция квантования определяется следующим образом:

$$Z = \left[q_{\min} - \frac{r_{\min}}{S} \right],$$

$$S = \frac{r_{\max} - r_{\min}}{2^b - 1},$$

$$X_q = \left[\frac{X}{S} + Z \right],$$

где Z – константа квантования, соответствующая нулевому значению; S – константа квантования, отвечающая за масштаб преобразования; $[r_{\min}, r_{\max}]$ – вещественный диапазон значений во входных данных; b – количество бит в квантованном типе данных; $q_{\min}$ – минимальное значение в квантованном типе данных; X – входные данные; X_q – квантованные данные.

Операция деквантования определяется как

$$X = S(X_q - Z).$$

Квантование имеет важное значение для развертывания больших нейронных сетей на устройствах с ограниченными ресурсами, таких как микроконтроллеры или одноплатные компьютеры, без значительного снижения точности.

1.3. Особенности оптимизации больших языковых моделей

При оптимизации больших языковых моделей следует учитывать особенности, связанные с их работой. Прежде всего это авторегрессионная природа больших языковых моделей. Авторегрессионные модели, такие как GPT, генерируют текст последовательно, токен за токеном, основываясь на ранее сгенерированных токенах. Это усложняет задачу параллелизации процесса генерации.

Другой важной особенностью является то, что для предварительного обучения больших языковых моделей требуются огромные вычислительные мощности и большие корпуса текста. Поскольку такие модели состоят из миллиардов параметров, предварительное обучение может стать трудоемкой или даже невозможной задачей в рамках оптимизации модели. Это происходит потому, что модели данного типа являются универсальными и предназначены для решения большинства типов задач без предварительного обучения.

1.4. Обзор существующих методов оптимизации

1.4.1. Метод LLM-Pruner

LLM-Pruner [5] представляет собой метод статического структурированного прунинга больших языковых моделей. Метод состоит из трех основных этапов: обнаружение зависимостей, оценка важности весов и восстановление качества работы.

В рамках этапа обнаружения зависимостей необходимо разбить языковую модель на независимые группы нейронов, а чтобы сгруппировать их, необходимо определить, как одни нейроны зависят от других. Авторы [5] вводят два вида зависимостей:

1) если нейрон N_j исходит только от нейрона N_i, то N_j зависит от N_i;

2) если нейрон N_i входит только в нейрон N_j, то N_i зависит от N_j.

Принцип зависимости заключается в том, что если текущий нейрон зависит исключительно от другого нейрона и этот другой нейрон подвергается прунингу, то и текущий нейрон также должен быть подвергнут прунингу. Используя такое определение зависимости, появляется возможность автоматически анализировать связанные структуры в языковой модели и затем группировать их для последующего прунинга.

После обнаружения всех зависимостей в модели наступает этап оценки важности весов. В рамках данного этапа из языковой модели удаляются те группы, которые имеют наименьшую важность по некоторому заданному критерию. Метод поддерживает множество критериев важности весов, однако основными являются критерии на основе L2-нормы весов и на основе градиентов. Определив важность каждого отдельного веса, появляется возможность определить важность группы, состоящей из этих весов. Авторы предлагают агрегировать информацию о важности группы четырьмя разными способами: суммированием важности весов в группе, произведением важности весов в группе, взятием максимума важности весов внутри группы и взятием важности последней структуры в группе. После оценки важности каждой группы авторы ранжируют группы по их важности, а затем прунят группы с более низкой важностью на основе заранее заданного коэффициента прунинга.

Для восстановления качества работы модели используется метод LoRA или QLoRA для минимизации количества обучаемых параметров, что позволяет сократить сложность обучения.

1.4.2. Метод ShortGPT

В отличие от метода LLM-Pruner, авторы метода ShortGPT [6] предложили метод структурированного прунинга по слоям, а не по каналам большой языковой модели. Делается это в два этапа:

1. Расчет меры важности каждого слоя.
2. Удаление определенной доли наименее важных слоев.

Авторы метода определяют меру важности с помощью косинусной близости между скрытыми состояниями на входе и выходе слоя. Чем меньше косинусная близость, тем более важен данный слой в большой языковой модели. Как показано на рис. 3, важность i-го слоя может быть посчитана как

$$score_i = 1 - \frac{1}{n}\sum_{t=1}^{n}\frac{X_{i,t}{}^{T} X_{i+1,t}}{\left\|X_{i,t}\right\|_2 \left\|X_{i+1,t}\right\|_2},$$

где $X_{i,t}$ – вектор скрытого состояния для t-го токена в последовательности на слое i; $X_{i+1,t}$ – вектор скрытого состояния для t-го токена в последовательности на слое $i + 1$; n – количество токенов в последовательности.

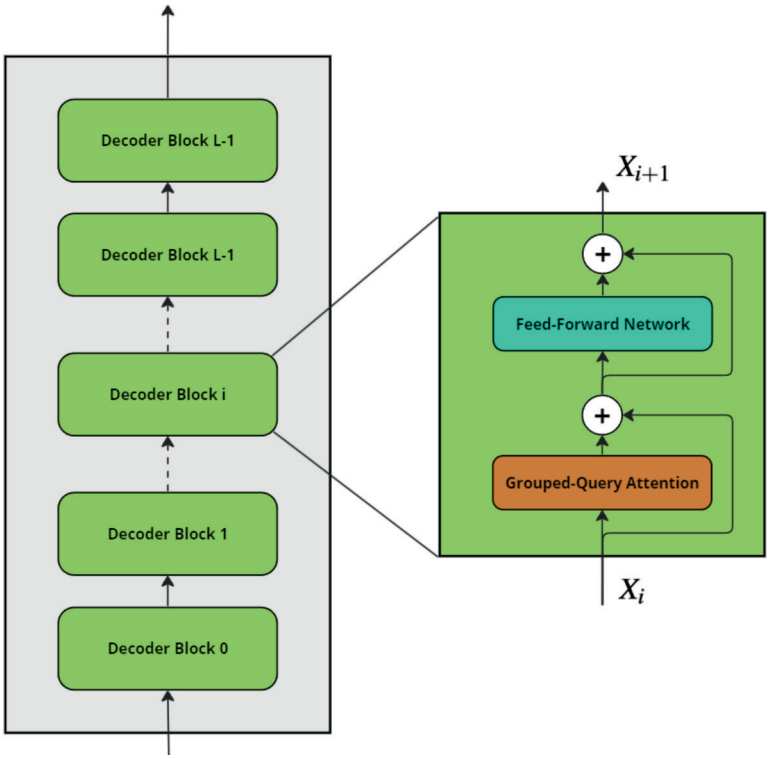


Рис. 3. Иллюстрация для пояснения формулы расчета важности
Fig. 3. Illustration to explain the importance calculation formula

Такой подход позволяет более эффективно сжимать модели, так как в нем, в отличие от LLM-Pruner, сокращается не только количество параметров и вычислений, но и количество последовательных операций.

1.4.3. Метод SparseGPT

SparseGPT [7] предлагает новый метод статического неструктурированного прунинга для моделей типа GPT, позволяющий провести быструю оптимизацию модели без дополнительного дообучения даже для моделей с сотнями миллиардов параметров. Это стало возмож-

ным из-за разреженности активаций в больших языковых моделях, когда только небольшая часть нейронов в каждом слое имеет влияние на активацию слоя.

Достигается такой результат благодаря применению подхода Mask Selection & Weight Reconstruction, где после прунинга части весов обновляются оставшиеся веса, чтобы компенсировать запруненные. В своем методе авторы используют несколько нововведений, таких как аппроксимация гессиана и итеративный выбор маски прунинга, которые снижают вычислительную сложность реконструкции весов без сильной потери качества запруненной модели.

Однако SparseGPT имеет несколько недостатков. Во-первых, он использует неструктурированный прунинг, в котором присутствуют разреженные вычисления для достижения ускорения нейросети. Даже при прунинге 50 % весов удается добиться ускорения не более чем в 1,6–1,7 раза по сравнению с исходной моделью. При большем проценте прунинга качество модели начинает существенно ухудшаться. Во-вторых, не все устройства эффективно поддерживают разреженные вычисления, из-за чего возникает ограниченность использования данного метода, особенно на встраиваемых устройствах.

1.4.4. Метод DejaVu

Авторы метода DejaVu [8] используют особенность LLM, связанную с разреженностью активаций. Они обнаружили, что активность нейронов на каждом слое напрямую зависит от входных значений на этих слоях. Эту особенность авторы назвали контекстуальной разреженностью, и в своей работе они исследуют контекстуальную разреженность в LLM для достижения ускорения без ухудшения качества модели.

Суть предложенного метода состоит в том, чтобы на основании входных данных предсказывать, какую малую часть нейронов в MLP и MHSA блоках необходимо активировать для достижения ускорения за счет уменьшения итоговых вычислений. Делается это за счет создания небольших MLP-предикторов поверх каждого блока в LLM, которые во время инференса модели предсказывают, какие нейроны нужно активировать, благодаря чему нет необходимости использовать все нейроны для предсказания модели. Это позволяет добиться значительного ускорения модели без существенного снижения качества ее работы.

1.4.5. Метод PowerInfer

PowerInfer развивает идеи предыдущего метода и также основывается на использовании контекстуальной разреженности для оптимизации модели. Однако, помимо оптимизации скорости инференса модели, PowerInfer [9] снижает потребление GPU, делая представленный фреймворк для инференса больших языковых моделей наиболее предпочтительным для запуска на устройствах с ограниченными ресурсами среди всех представленных выше методов.

PowerInfer представляет собой гибридный GPU/CPU-метод для инференса больших языковых моделей, использующий разреженное распределение активаций нейронов для достижения более быстрой скорости инференса на одном графическом процессоре. Предварительное размещение наиболее активных нейронов на графическом процессоре и менее активных нейронов на центральном процессоре, а также использование онлайн-предикторов для выбора нейронов для активации на графическом процессоре и центральном процессоре обеспечивает эффективную работу даже на пользовательских GPU, таких как RTX 4090 или RTX 2080Ti.

Такое решение позволяет запускать достаточно большие языковые модели (более 70 миллиардов параметров), добиваясь ускорения в 8–12 раз по сравнению с исходными моделями.

2. Материалы и методы

2.1. Выбор задачи и модели

Для дальнейших экспериментов в качестве задачи языкового моделирования была выбрана задача function calling, которая в контексте больших языковых моделей означает возможность модели вызывать определенные функции или API в зависимости от контекста запроса. Это позволяет большим языковым моделям предлагать вызов функций, что расширяет их возможности взаимодействия на естественном языке.

Function calling – это задача, приближенная к реальному практическому использованию больших языковых моделей в отличие от популярных бенчмарков с выбором ответа или генерацией односложного ответа. Function calling требует не только хорошего понимания пользовательского запроса, но и генерации сложного и точного ответа.

В качестве модели была выбрана Mistral-7B [10], которая является сильной базовой моделью для многих задач языкового моделирования, включая задачу function calling.

2.2. Тренировочные и тестовые данные

Для эффективной тренировки модели для задачи function calling использовался датасет glaive-function-calling-v2[1]. Он включает в себя 113 тысяч примеров и содержит обширный набор функций общего назначения. Помимо примеров с вызовом функции, в датасете также есть примеры, где либо отсутствует доступ к внешним функциям, либо отсутствуют подходящие для запроса функции.

Разбиение датасета на валидационную и тренировочную выборку проводилось следующим образом: выбиралось случайное подмножество функций, участвующих в вызовах в датасете, и все примеры с вызовами данных функций отбирались для валидационной выборки. Остальные примеры использовались для непосредственной тренировки модели. Такое разбиение позволяет избежать «утечки» данных, когда модель тренируется также на примерах из валидационной выборки – в данном случае, на примерах, где в тренировочной и валидационной выборках вызываются одни и те же функции.

2.3. Предобработка данных

Для выбранного датасета были выполнены определенные шаги предобработки данных, чтобы привести его к необходимому формату для последующего использования в обучении модели.

Предобработка glaive-function-calling-v2 производилась в несколько шагов:

– фильтрация примеров, где отсутствуют вызовы внешних функций, чтобы обеспечить наличие только релевантных данных для обучения;

– удаление текста после вызова функции. Данный шаг представляет возможный ответ системы в виде вызова функционального API;

– преобразование из формата мессенджера в формат mistral-instruct, чтобы стандартизировать представление данных.

После создания датасета и разбиения его на тренировочную и валидационную выборки была произведена синтетическая генерация примеров с множественными вызовами функций. Это было сделано для того, чтобы модель обрела возможность по единственному пользовательскому запросу вызывать несколько функций одновременно. При этом учитывалось, что в тренировочной выборке отсутствовали примеры с множественными вызовами функций.

[1] https://huggingface.co/datasets/glaiveai/glaive-function-calling-v2 (дата обращения: 15.03.2024).

Аналогичная ситуация была в валидационной выборке, где также наблюдалось недостаточное количество таких примеров.

2.4. Метрики качества модели

В качестве метрик, по которым будет оцениваться качество модели, были выбраны следующие метрики:

1. Exact match. Это метрика, которая сопоставляет один в один ответ модели с правильным ответом.
2. Время генерации. Эта метрика показывает, сколько времени в миллисекундах уходит в среднем на генерацию одного токена моделью.

2.5. Обучение базовой модели

Обучение больших языковых моделей напрямую сопряжено с трудностями из-за их огромного размера. Например, чтобы обучить модель с 7 миллиардами параметров, потребуется приблизительно 28 гигабайт видеопамяти на графическом процессоре, что существенно превышает возможности современных пользовательских видеокарт.

Для преодоления этих трудностей в данной работе использовался метод квантования в сочетании с QLoRA (Quantized Low-Rank Adaptation). Квантование позволяет существенно сократить размер модели путем уменьшения количества байт, используемых для представления весов и активаций. В данной работе использовалось квантование до 4 бит в формате nf4 и двойное квантование. Формат nf4 – это тип данных, при котором значения представляют собой квантили стандартного нормального распределения в предположении, что веса модели распределены соответствующим образом. Квантование весов выполняется блоками по 64 элемента, и для каждого элемента хранятся константы в формате fp32 для последующего деквантования весов. Суть двойного квантования состоит в том, что дополнительно квантуются хранимые константы блоками по 256 элементов, и тем самым сокращается потребление памяти еще на 0,4 бита на параметр.

QLoRA, в свою очередь, позволяет дообучать квантованные модели с помощью обучаемой низкоранговой добавки к весам модели. Таким образом, основные веса остаются квантованными, что помогает сократить размер модели, а обучаемая добавка к весам позволяет дообучать модель на тренировочных данных. QLoRA применялась ко всем линейным слоям модели с рангом 8 и alpha, равным 16.

Иллюстрация работы метода QLoRA приведена на рис. 4.

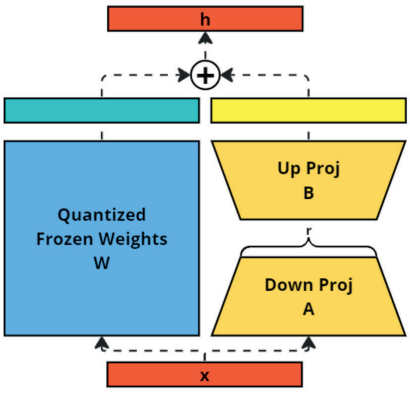


Рис. 4. Иллюстрация работы метода QloRA
Fig. 4. Illustration of QloRA method work

Для обучения модели использовались следующие гиперпараметры:

– batch_size: 1 (с накоплением градиента в 32 шага)
– steps: 4000
– warmup steps: 2000
– learning rate: 7e-5
– weight decay: 5e-5
– scheduler: линейный
– optimizer: paged AdamW 8bit

Процесс обучения базовой модели включал несколько шагов, а именно:

1. Загрузка предобученной модели Mistral с оптимизированными весами.
2. Квантование с использованием библиотеки bits&bytes, которая сжимает веса модели до 4 бит.
3. Добавление QLoRA-слоев на каждый линейный слой модели.
4. Загрузка подготовленного датасета, описанного в пункте «Предобработка данных».
5. Создание класса Trainer из библиотеки transformers для обучения модели.
6. Обучение QLoRA-слоев по заданным гиперпараметрам для минимизации ошибки во время обучения.
7. Сохранение обученной модели, в которой веса QLoRA объединены с весами модели.

2.6. Оптимизация методом ShortGPT

Оптимизация модели с использованием метода ShortGPT проходит в несколько шагов, начиная с калибровки на тренировочном датасете. Этот процесс начинается с прогона каждого примера из тренировочного датасета через модель, что позволяет получить скрытые состояния на входе и выходе каждого слоя модели для каждого примера. Затем для каждого примера и каждого слоя вычисляется косинусное расстояние между выходным и выходным скрытым состоянием, что позволяет оценить степень изменения состояний на разных слоях модели. После этого находится среднее значение между всеми примерами для каждого слоя, что позволяет оценить важность каждого слоя модели.

Вторым этапом оптимизации является прунинг, который основан на статистике важности каждого слоя, накопленной во время калибровки. Этот процесс начинается с сортировки слоев по степени их важности от наименее значимого до наиболее значимого. Затем выбираются 10 наименее важных слоев, которые удаляются из модели, тем самым сокращаются вычислительные и аппаратные требования к инференсу модели.

Последний этап оптимизации – дообучение. На этом этапе берется запруненная модель, из которой были удалены наименее важные слои, и проводится дообучение стандартной процедурой дообучения, тем самым восстанавливая потерянную способность к решению задачи function calling.

2.7. Оптимизация с помощью критерия Тейлора

Метод ShortGPT неплохо показал себя в качестве метода оптимизации моделей, используя в качестве критерия для прунинга косинусное расстояние между скрытыми состояниями на входе и выходе слоев. Однако такой критерий не учитывает информацию о влиянии удаленных слоев на итоговое значение функции потерь. В связи с этим была предложена идея использовать критерий Тейлора для послойного прунинга.

Суть метода заключается в добавлении «гейтов» [11] после блока GQA в каждом слое модели перед skip connection. Гейт – это слой в нейронной сети, представляющий собой вектор, инициализированный единицами, веса которого умножаются на входные значения и переда-

ются дальше. Такой подход позволяет накапливать градиенты на этих весах для дальнейшей калибровки модели.

Выясним, почему градиент, накопленный на гейте, действительно показывает важность той группы весов, которая с ним связана. Определим следующие функции:

$$y = \sum_{i=1}^{n} w_i x_i,$$
$$z = hy,$$

где y – выход группы весов; n – количество нейронов в группе; w_i – i-й вес в группе весов, x_i – вход для i-го веса; z – выход на гейте; h – вес на гейте, равный единице.

Найдем производные:

$$\frac{\partial L}{\partial h} = \frac{\partial L}{\partial z} y = \frac{\partial L}{\partial z} \sum_{i=1}^{n} w_i x_i,$$

$$\frac{\partial L}{\partial w_i} = \frac{\partial L}{\partial z} \frac{\partial z}{\partial y} \frac{\partial y}{w_i} = \frac{\partial L}{\partial z} x_i h = \frac{\partial L}{\partial z} x_i,$$

где L – значение функции потерь.

Важность параметра определяется как

$$I_w = g_w w.$$

Важность группы параметров определяется как сумма важностей всех параметров:

$$I_L = \sum_{i=1}^{n} w_i \frac{\partial L}{\partial w_i} = \sum_{i=1}^{n} w_i \frac{\partial L}{\partial z} x_i = \frac{\partial L}{\partial z} \sum_{i=1}^{n} w_i x_i = \frac{\partial L}{\partial h},$$

где I_L – важность данной группы. Таким образом, градиент, накопленный на гейте, равен сумме важностей параметров, входящих в данную группу.

При использовании гейтов возможно их различное расположение внутри слоя декодера. Всего можно представить четыре варианта расстановки: гейт после блока Multi-Head Attention, гейт после блока Feed-Forward Network, два разных гейта после обоих блоков, один и тот же гейт после обоих блоков (рис. 5–8).

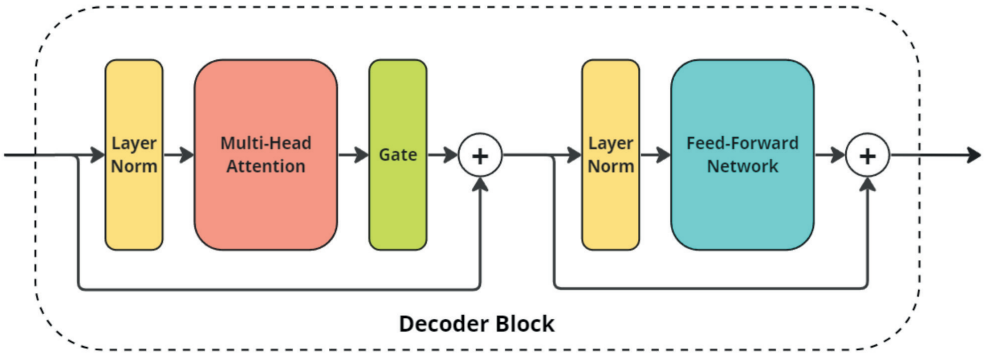


Рис. 5. Расстановка гейта после блока Multi-Head Attention
Fig. 5. Placing the gate after Multi-Head Attention block

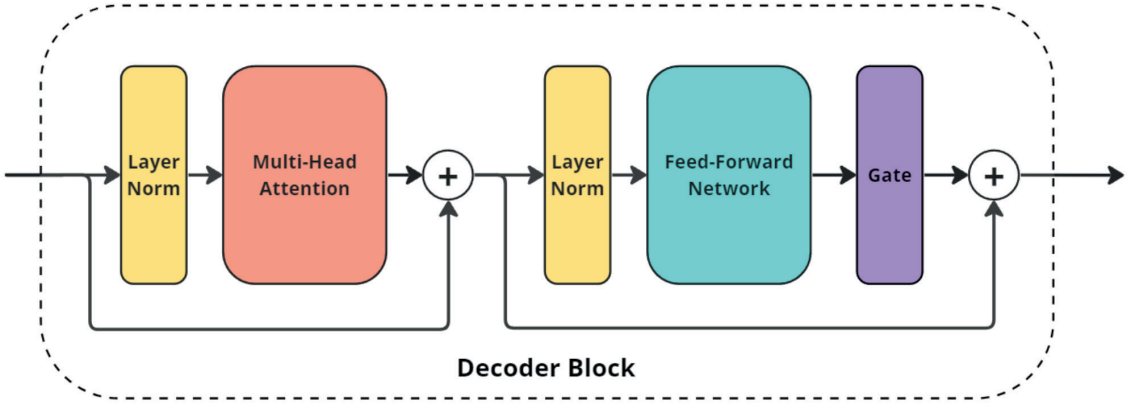


Рис. 6. Расстановка гейта после блока Feed-Forward Network
Fig. 6. Placing the gate after Feed-Forward Network block

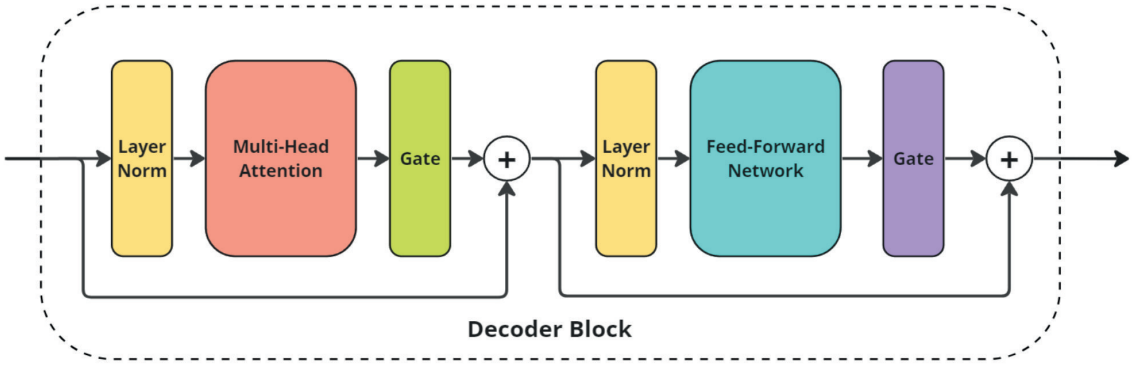


Рис. 7. Расстановка двух разных гейтов после блоков Multi-Head Attention и Feed-Forward Network
Fig. 7. Placing two different gates after Multi-Head Attention and Feed-Forward Network blocks

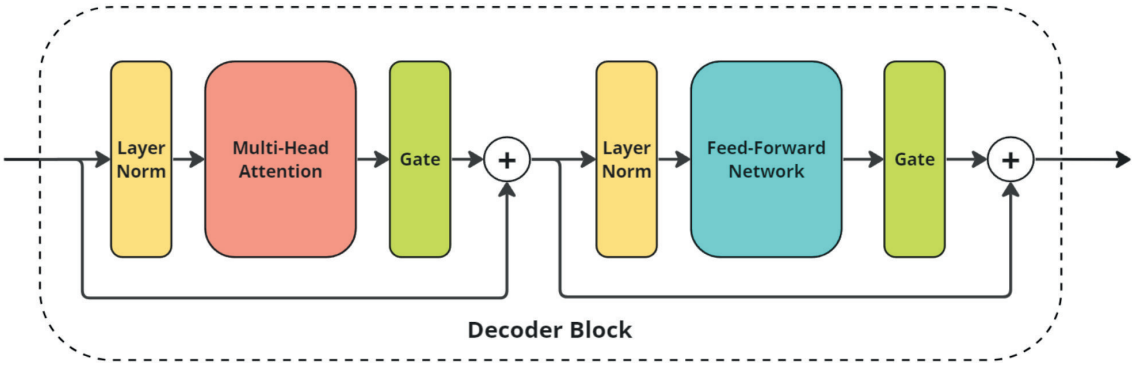


Рис. 8. Расстановка одного и того же гейта после блоков Multi-Head Attention и Feed-Forward Network
Fig. 8. Placing the same gate after Multi-Head Attention and Feed-Forward Network blocks

Важность слоя можно определить двумя разными способами: как модуль суммы градиентов на гейте и как L2-норму вектора градиентов. В первом случае важность слоя будет меньше зависеть от важности каждого канала по отдельности, во втором случае – наоборот, больше учитывается наибольший вклад каждого отдельного канала.

Калибровка модели производится следующим образом: на тренировочном датасете прогоняется каждый пример через модель, вычисляется ошибка с помощью функции потерь, применяется обратное распространение ошибки, и сохраняются градиенты на каждом гейте. В самом конце, когда все градиенты посчитаны, определяется важность каждого слоя.

Процедура прунинга и дообучения повторяет процедуру прунинга методом ShortGPT, однако использует статистику важности, накопленную в процессе калибровки критерием Тейлора. Данный метод представляет собой способ оптимизации больших языковых моделей, учитывающий информацию о градиенте.

2.8. Оптимизация методом LLM-Pruner

Помимо послойного прунинга, также был опробован метод LLM-Pruner, который отбрасывает параметры модели поканально, оставляя при этом количество слоев модели неизменным. Метод был выбран для эксперимента, чтобы узнать, будет ли он, имея меньшую эффективность в плане ускорения модели, обладать большей точностью на задаче function calling благодаря сохранению глубины оригинальной модели.

Первым шагом был выбор критерия прунинга. Было решено использовать критерий важности на основе L2-нормы весов, состоящий в том, что чем меньше значение веса, тем меньше его важность. Важно отметить, что существует и альтернативный критерий Тейлора, который оценивает важность на основе градиента веса. Однако было решено не использовать этот критерий из-за его повышенных требований к памяти видеокарты, на которой запускается и оптимизируется модель.

Следующим шагом было выявление и удаление наименее важных каналов с использованием выбранного критерия прунинга. Количество запруненных каналов было рассчитано таким образом, чтобы общее количество параметров примерно совпадало с количеством параметров моделей, оптимизированных послойными методами оптимизации.

После прунинга модели была применена стандартная процедура дообучения, которая применялась ранее для других методов оптимизации.

2.9. Оптимизация методом PowerInfer

Помимо привычных методов структурированного прунинга, когда из модели удаляются группы параметров, был опробован метод PowerInfer, использующий свойство контекстуальной разреженности в больших языковых моделях.

Прежде всего было решено использовать предварительно обученные веса модели, предоставленные авторами метода PowerInfer [9]. Это решение позволило сократить время и ресурсы, которые бы потребовались на самостоятельное дообучение модели после конвертации в совместимый с методом формат. Далее, уже на основе этих весов, модель дообучалась под искомую задачу. Процедура дообучения повторяла процедуру дообучения базовой модели для остальных методов оптимизации. Точность полученной модели оказалась ниже, чем исходной базовой модели. Это объясняется тем, что модель для PowerInfer имеет гораздо большую разреженность активаций, из-за чего обучение такой модели представляет собой более сложную задачу.

Далее необходимо было обучить онлайн-предикторы, которые играют ключевую роль в методе PowerInfer. В ходе экспериментов были подобраны процедуры обучения и параметры предикторов, которые обеспечили допустимое качество работы модели с онлайн-предикторами. В качестве гиперпараметров предикторов были выбраны следующие значения:

– epochs: 10
– learning rate: 0.02

– weight decay: 0.0001
– optimizer: RAdam
– predictor hidden size: 2048
– criterion: weighted binary cross entropy

После обучения модели и онлайн-предикторов все необходимые компоненты метода были готовы к запуску на представленном авторами метода ПО для инференса.

2.10. Измерение скорости инференса

Для измерения скорости моделей использовался фреймворк для инференса больших языковых моделей – llama.cpp [12]. Этот инструмент предоставляет удобный и понятный интерфейс для запуска больших языковых моделей, а также вывод основных характеристик работы модели, в том числе скорости работы.

Для измерения скорости моделей был использован пример из валидационной выборки длиной 322 входных токена и 27 токенов на генерацию. Все вычисления производились на CPU в однопоточном режиме и на GPU.

Методология измерения времени генерации моделей представляет собой использование двух прогонов модели для «прогрева» и пять прогонов для тестирования. В конце результаты усредняются для получения средних показателей скорости работы моделей.

Инференс моделей происходил с весами в формате fp16 с использованием метода декодирования greedy search.

3. Результаты

Тестирование точности exact match моделей проводилось на валидационной выборке датасета. В качестве целевого устройства для замеров скорости работы моделей был выбран персональный компьютер со следующими характеристиками:

– процессор: Intel Core i7-11700K 3.60 GHz
– видеокарта: Nvidia RTX 4090 24GB
– оперативная память: 60 GB
– операционная система: Linux

Для метода оптимизации с помощью критерия Тейлора был проведен ряд экспериментов, где рассматривалась разная расстановка гейтов внутри слоя декодера, а также различные способы агрегирования важности слоя на гейтах. Результаты приведены в табл. 1.

Таблица 1

Результаты экспериментов с критерием Тейлора

Table 1

Results of experiments with Taylor criterion

Модель	**Параметры**	**Способ агрегирования критерия**	**Расстановка гейтов**	**Точность, %**
Базовая модель	7241732096	–	–	81,83
L2 + FFN	5060612096	L2 norm	после FFN	76,83
L2 + ATTN	5060612096	L2 norm	после Attention	80,34
sum + FFN + ATTN	5060612096	sum	после FFN и Attention	78,30
L2 + FFN + ATTN	5060612096	L2 norm	после FFN и Attention	77,34

По итогам экспериментов (см. табл. 1) было выявлено, что расстановка гейтов после блоков Multi-Head Attention и использование агрегирования важности с помощью L2-нормы вектора градиентов дают наибольшую точность среди остальных опробованных вариантов.

Также было проведено сравнение опробованных методов оптимизации. В табл. 2 представлены результаты тестирования базовой и оптимизированных моделей по разным метрикам качества.

Таблица 2

Результаты тестирования моделей

Table 2

Results of model testing

Метод оптимизации	**Кол-во параметров**	**Точность, exact match, %**	**Скорость генерации токенов на CPU, ms**	**Скорость генерации токенов на GPU, ms**
Базовая модель	7241732096	81,83	491,73 (+–32,75)	21,334 (+–0,14)
ShortGPT	5060612096	79,21	326,15 (+–11,36)	12,06 (+–0,03)
Критерий Тейлора (L2 + ATTN)	5060612096	80,34	326,15 (+–11,36)	12,06 (+–0,03)
LLM-Pruner L2	5035266848	79,13	306,87 (+–4,76)	16,12 (+–0,12)
PowerInfer	8449691648	67,69	342,81 (+–11,12)	69,51 (+–3,12)

Из результатов экспериментов можно сделать следующие выводы.

1. Наибольшая точность для оптимизированной модели была достигнута с помощью метода послойного прунинга по критерию Тейлора важности слоя. При этом потери составляли всего 1,49 % точности. Этому же методу, вместе с методом ShortGPT, удалось достичь наилучшей скорости на GPU – ускорение модели составило 43,47 %. Таким образом, можно сказать, что на ускорение языковых моделей на GPU больше влияет их глубина, нежели количество каналов.

2. При работе на CPU глубина модели влияет на ее ускорение не так сильно, как общее количество параметров модели. Это видно по ускорению моделей методами LLM-Pruner и ShortGPT. При схожем количестве параметров модели имеют сопоставимые значения ускорения.

3. Метод PowerInfer показал себя худшим образом из всех использованных методов оптимизации, уступая остальным методам из табл. 2 как в скорости, так и в точности.

Таким образом, можно сделать вывод, что методы послойного прунинга оказались лучшими по соотношению точность/скорость генерации среди всех опробованных методов.

Заключение

В результате исследования были изучены и применены на практике различные методы оптимизации больших языковых моделей для задачи function calling. Был разработан программный код, который позволяет воспользоваться представленными в работе методами оптимизации. Код написан на языке Python с использованием современных инструментов для создания, обучения и инференса больших языковых моделей.

В качестве результатов применения методов оптимизации были получены оптимизированные модели и составлена таблица сравнения точности и скорости работы полученных моделей. Сделаны выводы о применимости использованных методов оптимизации для каждого конкретного случая.

Список литературы / References

1. **Radford A. et al.** Improving language understanding by generative pre-training. 2018.
2. **Devlin J. et al.** Bert: Pre-training of deep bidirectional transformers for language understanding // arXiv preprint arXiv: 1810.04805. 2018. DOI: 10.18653/V1/N19-1423
3. **Mikolov T. et al.** Efficient estimation of word representations in vector space // arXiv preprint arXiv: 1301.3781. 2013. https://doi.org/10.48550/arXiv.1301.3781
4. **Vaswani A. et al.** Attention is all you need // Advances in neural information processing systems. 2017. T. 30. DOI/10.5555/3295222.3295349
5. **Ma X., Fang G., Wang X.** Llm-pruner: On the structural pruning of large language models // Advances in neural information processing systems. 2023, vol. 36, pp. 21702–21720. https://doi.org/10.48550/arXiv.2305.11627
6. **Men X. et al.** Shortgpt: Layers in large language models are more redundant than you expect // arXiv preprint arXiv: 2403.03853. 2024. https://doi.org/10.48550/arXiv.2403.03853
7. **Frantar E., Alistarh D.** Sparsegpt: Massive language models can be accurately pruned in one-shot // International Conference on Machine Learning. PMLR, 2023. P. 10323–10337. https://doi.org/10.48550/arXiv.2301.00774
8. **Liu Z. et al.** Deja vu: Contextual sparsity for efficient llms at inference time // International Conference on Machine Learning. PMLR, 2023. P. 22137–22176. DOI/10.5555/3618408.3619327
9. **Song Y. et al.** Powerinfer: Fast large language model serving with a consumer-grade gpu // arXiv preprint arXiv: 2312.12456. 2023. https://doi.org/10.1145/3694715.3695964
10. **Jiang A. Q. et al.** Mistral 7B // arXiv preprint arXiv: 2310.06825. 2023. https://doi.org/10.48550/arXiv.2310.06825
11. **Molchanov P. et al.** Importance estimation for neural network pruning // Proceedings of the IEEE / CVF conference on computer vision and pattern recognition. 2019. P. 11264-11272. DOI: 10.1109/CVPR.2019.01152
12. **Gerganov G.** GitHub – ggerganov/llama.cpp: Port of Facebook's LLaMA model in C/C++ – github.com. 2023.

Сведения об авторах

Гончаренко Александр Игоревич, старший преподаватель Института интеллектуальной робототехники НГУ

Чупров Максим Иванович, разработчик-исследователь систем искусственного интеллекта компании ООО «Экспасофт»

Нежевенко Евгений Семенович, доктор технических наук, ведущий научный сотрудник тематической группы оптико-электронных специализированных процессоров Института автоматики и электрометрии Сибирского отделения Российской академии наук

Information about the Authors

Alexander I. Goncharenko, Senior lecturer, Institute of Intelligent Robotics of Novosibirsk State University

Maxim I. Chuprov, Artificial intelligence systems developer/researcher, Expasoft LLC

Evgeniy S. Nezhevenko, PhD., Leading researcher of the subject group of optical-electronic specialized processors, Institute of Automation and Electrometry of the Siberian Branch of the Russian Academy of Sciences

Статья поступила в редакцию 09.06.2025;
одобрена после рецензирования 23.08.2025; принята к публикации 23.08.2025

The article was submitted 09.06.2025;
approved after reviewing 23.08.2025; accepted for publication 23.08.2025

Научная статья

УДК 519.766.48
DOI 10.25205/1818-7900-2025-23-4-62-73

Оценка качества перевода художественного текста с амхарского на английский язык с использованием методов сжатия данных

Йешевас Гетачеу Лулу

Новосибирский государственный университет
Новосибирск, Россия

j.lulu@g.nsu.ru; https://orcid.org/0009-0006-8054-9846

Аннотация

Оценка качества перевода является важной задачей в области компьютерной лингвистики. В данном исследовании рассматривается использование методов сжатия данных для оценки точности перевода путем выявления характерных языковых закономерностей. Традиционные методы оценки перевода основаны на анализе стилистических показателей и машинном обучении, однако на эти подходы часто влияют длина текста и предопределенные лингвистические особенности. Чтобы устранить эти ограничения, мы используем теоретико-информационный метод, основанный на сжатии данных.

Наша методология использует алгоритмы сжатия для анализа перевода с целью оценки качества. Мы оцениваем неосознанный стилистический вклад переводчиков, сравнивая несколько переводов одних и тех же литературных произведений. Кроме того, мы применяем классификацию на основе сжатия, чтобы различать оригинальные тексты на амхарском языке, тексты, переведенные человеком с амхарского на английский, и тексты, переведенные компьютером. В наших экспериментах мы использовали шесть оригинальных романов на амхарском языке для анализа авторских стилей, а для оценки качества перевода – известные произведения, переведенные как переводчиками-людьми, так и компьютерными переводчиками. Среди различных алгоритмов сжатия данных без потерь были протестированы следующие: Prediction by Partial Matching (PPM), кодирование Хаффмана, преобразование Барроуза – Уилера (BWT) и алгоритм Лемпеля – Зива – Маркова (LZMA) с целью оценки их эффективности. Согласно коэффициенту V Крамера, рассчитанному по результатам различных экспериментов, алгоритм Prediction by Partial Matching (PPM) показал наивысшую стабильность и поэтому был выбран для всех последующих анализов.

Результаты показывают, что алгоритм PPM достигает наивысшей точности классификации: коэффициент Крамера (V) составил 0,89 для авторских текстов на амхарском языке, 0,762 и 1 для текстов, переведенных человеком с английского на амхарский, 0,91 для текстов, переведенных компьютером с амхарского на английский, и 0,53 для задач компьютерного перевода с английского на амхарский.

Исследование демонстрирует, что методы сжатия данных обеспечивают жизнеспособный, не зависящий от языка подход к оценке качества перевода, особенно для языков с ограниченными ресурсами, таких как амхарский. Эти результаты подчеркивают потенциал теоретико-информационных методов в лингвистическом анализе и компьютерных исследованиях перевода.

Ключевые слова

сжатие данных, перевод текста на амхарский язык, лингвистический анализ, оценка качества перевода, коэффициент Крамера

Благодарность

Автор выражает искреннюю благодарность научному руководителю, профессору Борису Рябко за ценное время, опыт и содержательные отзывы об этой исследовательской работе. Его конструктивные комментарии и продуманные предложения сыграли ключевую роль в повышении качества и ясности статьи.

Для цитирования

Лулу Й. Г. Оценка качества художественного перевода с амхарского на английский язык с использованием методов сжатия данных // Вестник НГУ. Серия: Информационные технологии. 2025. Т. 23, № 4. С. 62–73. DOI 10.25205/1818-7900-2025-23-4-62-73

Assessment of Amharic-English Literary Translation Quality Through Data Compression Techniques

Yeshewas Getachew Lulu

Novosibirsk State University
Novosibirsk, Russian Federation

j.lulu@g.nsu.ru; https://orcid.org/0009-0006-8054-9846

Abstract

Translation quality assessment are crucial challenges in computational linguistics. This study explores the use of data compression techniques to evaluate translation accuracy by identifying distinct linguistic patterns. Traditional methods for translation evaluation rely on style metric analysis and machine learning; however, these approaches are often influenced by text length and predefined linguistic features. To address these limitations, we employ an information-theoretic method based on data compression.

Our methodology utilizes compression algorithms to analyze translation of quality assessment. We assess the unconscious stylistic contribution of translators by comparing multiple translations of the same literary works. Additionally, we apply compression-based classification to distinguish between original Amharic texts, human-translated Amharic-to-English texts, and computer-translated texts. In our Experiments were conducted using six original Amharic novels for authorship styles and for translation quality assessment we utilize well-known translated works by human translator and computer translators. Among various lossless data compression algorithms, the following were tested: Prediction by Partial Matching (PPM), Huffman coding, Barrows-Wheeler Transform (BWT) and Lempel-Ziv-Markov Algorithm (LZMA), in order to evaluate their performance. According to the Cramer's V coefficient calculated from different experiments, the Prediction by Partial Matching (PPM) algorithm showed the highest stability and was therefore selected for all subsequent analyses.

Results indicate that PPM achieves the highest classification accuracy, with a Cramer coefficient (V) of 0.89 for Amharic authorship works, 0.762 and 1 for human-translated English-to-Amharic texts,0.91 for computer based translated Amharic-to-English texts and 0.53 for English Amharic computer translated tasks.

The study demonstrates that data compression techniques provide a viable, language-independent approach for translation quality assessment, particularly for low-resource languages like Amharic. These findings highlight the potential of information-theoretic methods in linguistic analysis and computational translation studies.

Keywords

Data compression, Amharic text translation, Linguistic analysis, Translation quality assessment, Cramer coefficient

Acknowledgements

We would like to express our sincere gratitude to the supervisor Professor Boris Ryabko for their valuable time, expertise, and insightful feedback on this research work. Their constructive comments and thoughtful suggestions have played a pivotal role in enhancing the quality and clarity of the manuscript.

For citation

Yeshewas Getachew Lulu. Assessment of Amharic-English Literary Translation Quality Through Data Compression Techniques. *Vestnik NSU. Series: Information Technologies*, 2025, vol. 23, no. 4, pp. 62–73 (in Russ.) DOI 10.25205/1818-7900-2025-23-4-62-73

1. Introduction

In an increasingly interconnected global landscape, translation functions as a critical conduit for cross-linguistic communication, enabling the dissemination of information across domains such as literature, journalism, film, and digital social platforms. The escalating demand for high-quality

translations has catalyzed substantial developments in both human and machine translation methodologies, thereby stimulating scholarly inquiry into their relative efficacy and fidelity [1].

Traditionally, literary scholars have evaluated translations using established analytical methods, accumulating extensive insights into both individual works and broader translation principles. In recent years, mathematical and computational approaches have emerged, offering new ways to assess translation quality [2; 3]. However, defining what constitutes a high-quality translation remains a challenge due to the subjective nature of style and interpretation.

From a computational perspective, translation quality is formulated as a classification problem that relies on distinctive linguistic features to capture an author's writing style [1; 3]. Style metric analysis, which includes vocabulary richness, word frequency distributions, and lexical repetition, is widely used for this purpose. However, as noted by Madigan et al [4], many of these metrics are highly dependent on text length, making them difficult to apply reliably. Researchers have explored alternative stylistic markers, such as word class frequencies, syntactic structures, word collocations, grammatical errors, and document structure (e.g., sentence and paragraph length) [5–8]. Additionally, machine learning techniques, including Support Vector Machines [9], Neural Networks [10], and Decision Trees [11], have been employed for authorship classification and translation quality evaluations.

To avoid the need for predefined linguistic features, some researchers have proposed using Data compression models for authorship attribution and translation quality assessment [1–3], as well as related tasks such as text categorization [12], language identification [13], genome classification, and clustering [14].

Amharic is a Semitic language spoken in North Central Ethiopia by the Amhara. It is the second most spoken Semitic language after Arabic, and the official language of Ethiopia. Amharic is also the official or working language of several of the states, including Amhara Region and the multi-ethnic Southern Nations, Nationalities, and People's Region [15]. Amharic has been spoken in Ethiopia since the late 12th century in various industries including the legal system, commerce, communications, the military and religion [16].

an additional challenge arises in translation between low-resource languages, such as Amharic to English. Unlike widely spoken languages, these translations suffer from limited linguistic data, inadequate machine translation models, and a lack of comprehensive studies on quality assessment. The scarcity of research in this area highlights the need for more systematic approaches to improve translation accuracy and preserve the nuances of both languages.

2. Methodology

In this section, we will briefly describe Ryabko et al.'s [2; 3] approach to literary text attribution and the Information-Theoretic Method for assessing translation quality. We will also explain how we use the data compression method for evaluating the quality of translations between Amharic and English in both directions, applying our analytical approach. Here below in figure 1 outline the steps to conduct this research.

2.1 Data collection and parameter selection

We began by selecting six well known Amharic novels, "*Oromai*", "Fiker Eskemekaber", "YeHelina Dewel", "Shotelay", "Keadmase Bashager", and "Emego" written by Dr. Hadis Alemayehu, Bealu Girma, Mamo Wodneh, and Alemneh Wassie respectively for the first experiment.

For our experiment, we prepared our dataset by selecting a 64kb training sample from each book. Each sample was then divided into 32 slices, each with a size of 2 KB.

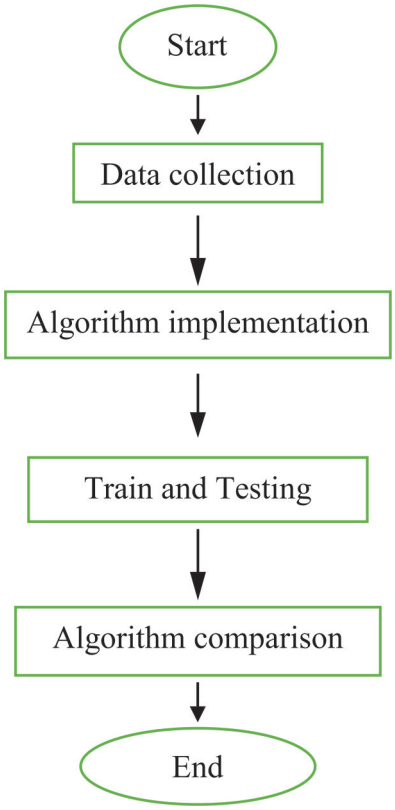


Fig. 1. Flow chart of the research stage

In the second experiment, we selected two well-known English books that had been translated into Amharic by two different translators. Specifically: the first book entitled "*Les Miserable*" by Victor Hugo, which was translated into Amharic as "*Menduban*" by Yohannes G. Meskel and "*Mekergunchu*" by Sehale Selassie Berhan and the second book "*Crime and Punishment*" by Fyodor Dostoevsky, which was translated into Amharic as "*Wenjel Ena Ketat*" by Muluberhan and "*Wenjel Ena Ferde*" by Kassa

For the third experiment, due to unavailability of one books translated from Amharic to English by several human translators we utilized three computer translator that support Amharic language two of them the common language translator google translator [17], Yandex Translator [18] and one of them is translator that support Artificial intelligence large language model and neural machines translation model that is lingvanex translator [19]. we select famous novel in Amharic entitled "*Fiker Esk mekaber*" (Love up to death) written by Dr. Hadis alemayhu for this work. For the last experiment we select an English book entitled "Born a crime" by Trevor Noah and translated to Amharic by applying the above mentioned computer translators.

2.2 Algorithm implementation and selection

In our investigation of we utilized the following modern lossless data compression algorithm: GNU ZIP (Gzip) [20], BZIP2[21], Lempel–Ziv–Markov chain algorithm (LZMA) [22], Brotli[23], Zstandard (Zstd) [24], and Prediction by partial matching (PPM)[25].

We cheek all mentioned potential modern lossless Data compressor and calculating their chi-square and crammer coefficient V in order to select the efficient algorithm for our experiment.

The Cramer coefficient V varies from 0 (corresponding to no association between the variables) to +1 (complete association). It is based on Pearson's chi-squared statistic. Here's below the steps followed to calculate Cramer's V for each contingency table for Data compressors used in the test:

1. Construct the Contingency Table:
   - Create a contingency table that shows the frequency distribution of the variables.
2. Calculate the Chi-Squared Statistic (χ^2):
   - Use the formula for Pearson's chi-squared test:

$$x^2 = \frac{\sum (Oij - Eij) 2}{Eij}.$$

where O_{ij} is the observed frequency and E_{ij} is the expected frequency under the assumption of independence.

3. Determine the Total Sample Size (N):
   - Sum all the observed frequencies in the table to get N.
4. Find the Minimum Dimension (k):
   - Determine the smaller value between the number of rows (r) and the number of columns (c) in the table:

$$k = \min(\mathrm{r} - 1, \mathrm{c} - 1)$$

5. Compute Cramér's V:
   - Use the formula

$$V = \sqrt{\frac{x^2}{N * K}}.$$

2.3. Training and Test the model

In this section we use the proposed method is based on the approach developed by [2; 3] for the comparative analysis of two characteristics of translation quality, based on a comparative quantitative analysis of the unconscious style of translation texts. First, we quantify the contribution of the unconscious author's style by comparing different Amharic original work of different Authors. Secondly, we indirectly compare the contribution of the author of the work and the translator by analyzing translations of the same work by different translators.

For our investigation we split our Data set in two exclusive sets which is the training 64kb and the test set from each book. Then we split the test set into 32 slices with size 2kB to Test the model with the given training set.

Modern digital systems employ data compression techniques mentioned [19–25], known as archivers, which rely on principles from information theory and formal grammars. These tools reduce text size by detecting patterns in character frequency, enabling efficient data storage and retrieval [16]. This study based on an information-theoretic method for determining translation quality assessment from Amharic to English and vice versa.

3.0. experimental setup and Results

We prototype a python program for our experiment because Python is easy to import data compression libraries and mathematical manipulation using numpy library and data visualization mod-

ules. You can find the research implementation and Data in GitHub repository, https://github.com/yeshewas/Information-Theoretic-Method-for-Assessing-the-Quality-of-Translations.

Here below in the subsection 3.1 we investigate the authorship style of Amharic works, in subsection 3.2 we show results of English to Amharic translation of two books each with two different human translators and four different English books by a single Amharic translator, in section 3.3 we show the results of one Amharic work to English by three computer translators and finally in section 3.4 we investigate one English work to Amharic by computer translator's.

3.1 Authorship style of Amharic original texts

For authorship identifications experiment, we used original Amharic works by Dr. Hadis Alemayehu, Bealu Girma, Mamo Wodneh, and Alemneh Wassie, with N = 6 and m = 32. From these authors, we created six training samples, X1, X2, ..., X6, each with a size of 64kb. Additionally, we extracted 32 test samples from each author's work—Y1j (j = 1, ..., 32) from Hadis Alemayhu, Y2j from Bealu Girma, and so on, up to Y6j from alemneh Wassie. Each test sample was 2 KB in size.

Table 1: Results of the experiments. The data obtained for the PPM archiver, training set 64kb, and 32 Test slice each with 2 kB size. For All Potential Data compressor results calculated there Cramer coefficient in the same fashion and the result attached in Appendix 1.

PPM Cramer coefficient V=0.89

	Hadis-alemayhu	Bealu-Girma	Mamo-wedneh	Tsegaye-G/medhin	Michel-kebede	Alemayhu-waasie
Hadis-alemayhu	32	0	0	0	0	0
Bealu-Girma	0	30	0	0	0	2
Mamo-wedneh	0	0	31	1	0	0
Tsegaye-G/medhin	0	0	0	32	0	0
Michel-kebede	0	0	0	0	32	0
Alemayhu-waasie	0	2	0	0	0	30

From Table 1 presents a confusion matrix showing the attribution of texts to six different authors based on stylistic features. The results demonstrate a high degree of accuracy in distinguishing between the authors' styles.

Most notably, *Hadis Alemayhu*, T*segaye G/Medhin*, and *Michel Kebede* each achieved perfect classification, with all 32 samples correctly attributed to them. *Mamo Wedneh* also shows very strong consistency, with 31 texts correctly identified and only one misclassified as *Tsegaye G/Medhin*.

Slight confusion appears between *Bealu Girma* and *Alemayhu Waasie*, where each had 30 texts correctly attributed to them, but two samples were misclassified as each other. This suggests some stylistic overlap or similarity between these two authors.

Overall, the matrix indicates that the stylistic signatures of each author are distinct and reliably identifiable, with only minor exceptions.

3.2. English to Amharic translated books by two different real translators

In this experiment we utilize two English books translated to Amharic by real Amharic writers, firstly, a Book from vector ego known as "lie measurable" translated two Amharic novels "menduban" and "mekergonchu" written by Yohans G/mesekel and salhalselasi berhan respectively. Secondly a book from crime and punishment by devestoky to Amharic novels the books entitled "wenjel ena keta " and "wenjel ena ferde" by two translator muluberhan and kassa respectively, Each books with Training size of 64kb and Test slice of 32 with Slice size 2kB.

Table 2 presents a contingency matrix showing the attribution of Amharic translations of the English work "*Lie Measurable*" to two novels—"*Menduban*" by Yohans and "*Mekerognochu*" by Sehaleselasi—using a Partial Prediction Matching (PPM) compression-based method. The Cramer's V value of **0.762** indicates a **strong association** between the predicted and actual sources, suggesting the method is effective at distinguishing between the stylistic features of the two translations.

PPM V=0.762

	Menduban	**Mekerognochu**
Menduban by Yohans	31	1
Mekrognchu by sehaleselasi	7	25

From table 2 The high Cramer's V value (0.762) reflects a substantial level of distinguishability between the two translated texts. While *"Menduban"* displays a highly distinctive and consistent style, *"Mekerognochu"* appears to share certain stylistic traits with *"Menduban"*, leading to some misclassifications. This overlap may point to either subtle stylistic similarities between the authors or influences in translation that blur the boundaries between their styles.

Overall, the PPM compression method shows strong performance in identifying authorial style, particularly for Yohans's *"Menduban"*, though some refinement may be needed to improve accuracy for Sehaleselasi's *"Mekerognochu"*.

Table 3: contingency table of translation of English work "crime and punishment" to Amharic Novels "Wenjel ena ketat" and "Wenjel ena ferde" by Muluberhan and Kassa respectively Cramer V=1 for predication partial matching(PPM) compressor.

PPM v=1

	Wenjl ena ferde	**Wenjel ena ketat**
Wenjel ena ferde	32	0
Wenjel ena ketat	0	32

From Table 3 The results demonstrate that the PPM method can perfectly distinguish between the translation styles of Muluberhan and Kassa. This suggests that each translator imposed a highly distinct stylistic signature on their respective Amharic versions of *Crime and Punishment*. The absence of any misclassification indicates no stylistic overlap, making this an ideal case of translation style differentiation.

Table 4: For this experiment, I analyzed three literary works translated from English into Amharic by the renowned Ethiopian translator and author Sahle Sellassie Berhane Mariam, along with one of his original novels. The translated works included:

- Victor Hugo's *Les Misérables* (Amharic title: *Mekergochu*)
- Charles Dickens' *A Tale of Two Cities* (Amharic title: *Ye Hulet Ketemawech Weg*)
- Pearl Buck's *The Mother* (Amharic title: *Emye*)
- Sahle Sellassie's original work *Basha-Ketaw*

was included in the analysis. The experiment utilized a training corpus of 64 kB, with 32 test slices of 2 kB each for textual analysis.

Writers	**Sahle Sellassie**	**Victor Hugo**	**Charles Dickens**	**Pearl Buck**
Sahle Sellassie	32	0	0	0
Victor Hugo	0	32	0	0
Charles Dickens	0	0	29	3
Pearl Buck	2	8	1	21

From the table 4, we observe that the original style of Sahle Sellassie's novel was perfectly preserved in all 32 instances, as his translations consistently matched his own stylistic traits. Similarly, Victor Hugo's translations were distinctly different from those of other writers, with all 32 samples closely aligning with his unique style. However, Charles Dickens' style was less accurately preserved—three of his translated excerpts bore a closer resemblance to Pearl Buck's works.

Pearl Buck's style proved the most challenging to retain, with eleven of her translated slices leaning more toward other authors' styles. Overall, Sahle Sellassie's translations exhibited the worse preservation of the original author's style. This observation is supported by the Cramer's V coefficient of 0.77, indicating a strong but imperfect association between authors and their translated styles.

Notably, certain writers present exceptional difficulties for translators. Pearl Buck stands out as one such author—her distinctive style remains, among the most difficult to translate faithfully.

3.3. Amharic to English translated work by computer translators

In third experiment We used three computer translator that support Amharic language two of them the common language translator google translator, Yandex and one of them is translator that support Artificial intelligence large language model and neural machines translation model that is lingvanex translator [25].for this task our Data source is a famous novel in Amharic entitled "Fiker esk mekaber" (Love up to death) written by Dr. Hadis alemayhu with training size 64 kB and 32 Test slices each with 2 kB.

Based on the above experimental setup draw the contingency table and calculate the Cramer coefficient for all potential compressor (Gzip, LZMA, BZIP2, PPM, brotli, zstad) to select the efficient one. So, the result shows that almost all compressor scores approximate Cramer coefficient value which is 0.78 except the PPM V=0.83.

Table 5: contingency table of translation of Amharic work "Feker Esk mekaber" to English by computer translators. Cramer V=0.91 for predication partial matching (PPM) compressor.

PPM V=0.91

	Google-translators	**Yandex**	**lingvanex**
Google-translators	29	3	0
Yandex	4	28	0
Lingvanex	0	0	32

From Table 5 demonstrates that the translations produced by Google, Yandex, and Lingvanex differ significantly in style. This indicates that each translator retains elements of their own unconscious stylistic patterns. The Cramer's V coefficient of 0.91 further confirms that the distinctive styles of the translators are strongly expressed—while the original style of the classic novel is not as clearly preserved.

3.4. English to Amharic translated work by computer translators

Lastly, we used three computer translator that support Amharic language two of them the common language translator google translator, Yandex and one of them is translator that support Artificial intelligence large language model and neural machines translation model that is lingvanex translator [25]. for this task our Data source is a famous novel in English novel entitled "Born a crime" written by Trevor Noah with training size 64 kB and 32 Test slices each with 2 kB.

Table 6: contingency table of translation of English work "Born a crime" to Amharic by computer translators. Cramer V=53 for predication partial matching (PPM) compressor.

PPM V=0.53

	Google	Yandex	lingvanex
Google-translators	16	4	12
Yandex	1	27	4
Lingvanex-AI	0	3	29

From Table 6, we observe that the style of Lingvanex translations stands out significantly from the styles of other translators. Specifically, 29 out of 32 slices are most similar to Lingvanex's own translations. In comparison, Yandex's translation style is less distinct—four of its translation slices resemble Lingvanex's style more closely, and one aligns with Google's. Finally, Google's translation style is the least consistently conveyed, with 16 of its slices being more similar to those of other translators.

To assess high-quality translation, one essential condition must be met: the translator's influence on the translation style should be minimal—or ideally, absent altogether. The data in this table suggest a near-ideal scenario, as indicated by the Cramer's V coefficient value of 0.53. This supports the conclusion that the translator's impact on the translated text is relatively small.

4.0 Conclusion

This study demonstrates the potential of data compression techniques in assessing translation quality and authorship style analysis. By leveraging information-theoretic methods, we successfully analyzed linguistic patterns and stylistic similarities across original and translated texts. The results confirm that lossless data compressors, particularly PPM, effectively capture differences in authorship styles and translation styles of Amharic to English literary texts.

Our experiments revealed that translations by different human translators exhibit distinguishable linguistic patterns, whereas machine-translated texts show varying degrees of similarity depending on the algorithm used. The application of Cramer's V metric allowed us to quantify these differences, reinforcing the validity of data compression as a tool for evaluating translation fidelity and authorial influence.

The findings also highlight the challenges associated with translations between low-resource languages like Amharic and English. The variability observed among machine-generated translations underscores the need for improved natural language processing models that preserve linguistic degrees more effectively.

Future research should explore expanding the dataset to include more language pairs and integrating hybrid models that combine traditional linguistic features with data compression techniques. Additionally, refining machine translation algorithms based on compression-based quality metrics could lead to significant advancements in translation accuracy and authorship identification.

Appendix 1:

Calculating Cramer coefficient of All compressor:

LZMA V=0.86

	Hadis-alemayhu	Bealu-Girma	Mamo-wedneh	Tsegaye-G/medhin	Michel-kebede	Alemayhu-waasie
Hadis-alemayhu	32	0	0	0	0	0
Bealu-Girma	0	32	0	0	0	0

	Hadis-alemayhu	Bealu-Girma	Mamo-wedneh	Tsegaye-G/medhin	Michel-kebede	Alemayhu-waasie
Mamo-wedneh	0	0	32	3	0	0
Tsegaye-G/medhin	0	0	0	32	0	0
Michel-kebede	0	0	0	5	27	0
Alemayhu-waasie	0	4	0	0	0	28

BZIP2 V=0.75

	Hadis-alemayhu	Bealu-Girma	Mamo-wedneh	Tsegaye-G/medhin	Michel-kebede	Alemayhu-waasie
Hadis-alemayhu	19	0	0	13	0	0
Bealu-Girma	0	23	0	0	8	1
Mamo-wedneh	0	0	29	3	0	0
Tsegaye-G/medhin	0	0	0	32	0	0
Michel-kebede	0	0	0	8	24	0
Alemayhu-waasie	0	1	0	3	0	28

Zstand V= 0.87

	Hadis-alemayhu	Bealu-Girma	Mamo-wedneh	Tsegaye-G/medhin	Michel-kebede	Alemayhu-waasie
Hadis-alemayhu	32	0	0	0	0	0
Bealu-Girma	0	30	0	0	0	2
Mamo-wedneh	0	0	32	0	0	0
Tsegaye-G/medhin	0	0	0	32	0	0
Michel-kebede	0	0	0	3	29	0
Alemayhu-waasie	0	2	0	3	0	30

Brotli V=0.88

	Hadis-alemayhu	**Bealu-Girma**	**Mamo-wedneh**	**Tsegaye-G/medhin**	**Michel-kebede**	**Alemayhu-waasie**
Hadis-alemayhu	32	0	0	0	0	0
Bealu-Girma	0	28	0	0	0	4
Mamo-wedneh	0	0	32	0	0	0
Tsegaye-G/medhin	0	0	0	32	2	0
Michel-kebede	0	0	0	1	31	0
Alemayhu-waasie	0	0	0	0	0	32

References

1. **Malyutov M.B.** Authorship Attribution of Texts: A Review General Theory of Information Transfer and Combinatory. Lecture Notes in Computer Science, vol. 4123. Springer, Berlin, Heidelberg. DOI.org/10.1007/11889342_20.
2. **Ryabko B.Y., Savina N.** Information-Theoretic Method for Assessing the Quality of Translations. Entropy 2022, 24, 1739. DOI.org/10.3390/e24121739
3. **Ryabko B.Y., Savina N.** Using Data Compression to Build a Method for Statistically Verified Attribution of Literary Texts. Entropy 2021, 23, 1302. DOI.org/10.3390/e23101302
4. **Madigan D; Genkin A; Lewis D. D; Argamon, S; Fradkin, D; Ye L.** Author identification on the large scale. 2005, June. In *Proceedings of the 2005 Meeting of the Classification Society of North America (CSNA*).
5. **Argamon S; Šarić M; Stein S.S.** Style mining of electronic messages for multiple authorship discrimination: first results. 2003, August. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 475–480.
6. **Forsyth R.S; David I.H.** Feature-finding for text classification. Literary and Linguistic Computing, 1996, vol. 11, no. 4. pp. 163–174.
7. **Juola P.** Authorship attribution. Foundations and Trends® in Information Retrieval. 2008 Mar 6, vol. 1, no. 3. pp. 233–334.
8. **Katirai Hooman; Waterloo Ontario; Dale Schuurmans.** Filtering junk e-mail. 1999. Department of Electrical & Computer Engineering, University of Waterloo.
9. **Cover T. M.** Elements of information theory. 1999. John Wiley & Sons.
10. **Williams C. B.** Mendenhall's studies of word-length distribution in the works of Shakespeare and Bacon.1975. Biometrika, pp. 207–212.
11. **Thompson J. W; Padover S. K.** Secret diplomacy: espionage and cryptography. 1963, pp. 1500–1815.
12. **Mendenhall T. C.** The characteristic curves of composition. Science. 11 Mar 1887, no. 9, issue 214, pp. 237–246. DOI: 10.1126/science. ns-9.214S.237.
13. **Bahtin M.** Problemi poetike Dostojevskoga. Zadar: Sveučilište u Zadru; 2020.
14. **Brinegar C. S.** Mark Twain and the Quintus Curtius Snodgrass letters: A statistical test of authorship. 1963. Journal of the American Statistical Association. vol. 58, no. 301, pp. 85–96.

15. **Bende M. L.** The origin of Amharic. *Ethiopian Journal of Languages and Literature*. 1983, vol. 1, pp. 41–52.
16. **Adugna, Gabe.** Research: Language Learning - Amharic: Home. library.bu.edu. Retrieved 2025-05-10.
17. https://translate.google.ru/?hl=en&sl=it&tl=ru&op=translate
18. https://translate.yandex.com/en/
19. https://lingvanex.com/translate/
20. **Abdelfattah M.S; Hagiescu A; Singh D.** Gzip on a chip: High performance lossless data compression on fpgas using opencl. In Proceedings of the international workshop on openCL. 2013 & 2014, pp. 1–9.
21. **Seward J.** bzip2 and libbzip2. avaliable at http://www. bzip. Org. 1996, pp. 8–18.
22. **Onuma Y., Terashima Y., Kiyohara R.** Compression method for ECU software updates. IEEE Tenth International Conference on Mobile Computing and Ubiquitous Network (ICMU). 2017 Oct 3, pp. 1–6.
23. **Alakuijala J., Farruggia A., Ferragina P., Kliuchnikov E., Obryk R., Szabadka Z., Vandevenne L.** Brotli: A general-purpose data compressor. ACM Transactions on Information Systems (TOIS).2018, vol. 37, pp. 1–30.
24. **Collet Y., Kucherawy M.** Z-standard Compression and the application/zstd Media Type. 2018, no. rfc8478.
25. **Shkarin D.** PPM: one step to practicality. In: Proceedings of the Data Compression Conference. vol. DDC ’02, pp. 202. IEEE Computer Society (2002)

Information about the Author

Yeshewas Getachew Lulu, PhD Student in specialty of Mathematical and software of computer systems, complexes and computer network in Faculty of information Technology, Novosibirsk state university, Russian Federation.

Сведения об авторе

Йешевас Гетачеу Лулу, аспирант факультета информационных технологий Новосибирского государственного университета

Статья поступила в редакцию 09.06.2025;
одобрена после рецензирования 31.07.2025; принята к публикации 31.07.2025

The article was submitted 09.06.2025;
approved after reviewing 31.07.2025; accepted for publication 31.07.2025

Научная статья

УДК 004.056.55
DOI 10.25205/1818-7900-2025-23-4-74-93

Комбинированный матрично-блочный алгоритм шифрования с использованием эллиптических кривых

Ольга Алексеевна Сергеева
Анастасия Сергеевна Кутовая
Владислав Сергеевич Сергеев

Кемеровский государственный университет
Кемерово, Россия

okoin@yandex.ru; https://orcid.org/0000-0002-9204-6318
askutovaia@mail.ru
v1a05@ya.ru

Аннотация

В статье рассматривается блочный криптографический алгоритм с использованием двухкомпонентного общего секретного ключа, полученного по принципу ключевого обмена Диффи – Хеллмана на точках эллиптической кривой над полем Z_p. Цель алгоритма – устранить недостатки отдельных классических алгоритмов и за счет их комбинирования повысить общую стойкость системы. Генерация и обмен ключами между пользователями осуществляются по типу эллиптических криптографических систем с открытым ключом. При этом предлагается два способа генерации общих секретных ключей для взаимодействующих пользователей: применение криптографического протокола Диффи – Хеллмана на нескольких точках эллиптической кривой или дополнительное использование рекуррентной формулы. Элементы шифрования в алгоритме представлены блоками в виде квадратных матриц, построенных на координатах точек эллиптической кривой. Собственно шифрование проходит в два этапа, на первом из которых используется поточное гаммирование с операцией вычисления кратной точки эллиптической кривой, а на втором проводится формирование матричных блоков и выполняется их матричное преобразование Хилла с использованием обратной связи. Каждый этап шифрования задействует соответствующий ему компонент общего секретного ключа пользователей: числовую гамма-последовательность или квадратную ключ-матрицу. Криптографическая стойкость алгоритма базируется на трудоемкости решения задачи дискретного логарифмирования на эллиптических кривых и защищенности сервиса совместного доступа с безопасной аутентификацией взаимодействующих пользователей. Блочная реализация второго этапа шифрования обеспечивает стойкость системы к частотному анализу. В качестве иллюстрации работы приведенного алгоритма в статье пошагово разбирается пример шифрования/дешифрования текстового сообщения.

Ключевые слова

шифрование, криптографический алгоритм, протокол Диффи – Хеллмана, эллиптические кривые, кратная точка, гаммирование, преобразование Хилла, матричные блоки

Для цитирования

Сергеева О. А., Кутовая А. С., Сергеев В. С. Комбинированный матрично-блочный алгоритм шифрования с использованием эллиптических кривых // Вестник НГУ. Серия: Информационные технологии. 2025. Т. 23, № 4. С. 74–93. DOI 10.25205/1818-7900-2025-23-4-74-93

Combined Matrix-Block Encryption Algorithm Using Elliptic Curves

Olga A. Sergeeva, Anastasia S. Kutovaya, Vladislav S. Sergeev

Kemerovo State University
Kemerovo, Russian Federation

okoin@yandex.ru; https://orcid.org/0000-0002-9204-6318
askutovaia@mail.ru
v1a05@ya.ru

Abstract

The article examines a block cryptographic algorithm using a two-component shared secret key obtained according to the Diffie-Hellman key exchange principle on elliptic curve points over the field Z_p. The goal is to eliminate shortcomings of individual classical algorithms and, through their combination, increase overall system strength. Key generation and exchange between users are carried out using elliptic curve cryptographic systems with public key. Two methods are proposed for generating shared secret keys for interacting users: applying the Diffie-Hellman cryptographic protocol on multiple elliptic curve points or additionally using a recurrence formula. Encryption elements are represented by blocks as square matrices constructed on elliptic curve point coordinates. Encryption proceeds in two stages: the first uses stream cipher with scalar multiplication of elliptic curve points, and the second involves forming matrix blocks and performing Hill matrix transformation with feedback. Each encryption stage utilizes its corresponding component of the users' shared secret key: a numerical gamma sequence or a square key matrix. The cryptographic strength is based on the computational complexity of solving the discrete logarithm problem on elliptic curves and the security of the sharing service with secure authentication of interacting users. The block implementation of the second encryption stage ensures the system's resistance to frequency analysis. As an illustration of the presented algorithm's operation, the article provides a step-by-step example of encrypting/decrypting a text message.

Keywords

Encryption, cryptographic algorithm, Diffie-Hellman protocol, elliptic curves, scalar multiple, gamma encryption, Hill transformation, matrix blocks

For citation

Sergeeva O. A., Kutovaya A. S., Sergeev V. S. Combined matrix-block encryption algorithm using elliptic curves. *Vestnik NSU. Series: Information Technologies*, 2025, vol. 23, no. 4, pp. 74–93 (in Russ.) DOI 10.25205/1818-7900-2025-23-4-74-93

Введение

Поточное шифрование при непосредственном кодировании символов алфавита в классических криптосистемах имеет своим главным недостатком нестойкость к статистическому частотному анализу текста [1]. Для устранения подобных уязвимостей в криптосистемах, использующих числовое кодирование элементов шифрования, отдают предпочтение блочным видам реализации, при которых символы открытого текста объединяются в биграммы или числовые блоки фиксированной длины [2–4]. Использование числовых блоков наделяет шифр свойством перемешивания, при котором скрываются статистические зависимости между исходным и зашифрованным текстами. Такой подход не был рассмотрен ранее для эллиптических криптосистем ввиду особенности операций в группе точек эллиптической кривой [5–7]. Однако использование числовых координат точек кривой по отдельности с сохранением их порядка позволяет обойти это ограничение и рассмотреть блочную реализацию эллиптической криптосистемы [8].

Результаты

В качестве варианта блочной реализации эллиптической криптографической системы в работе рассматривается матрично-блочный криптографический алгоритм с использованием

двухкомпонентного общего секретного ключа, полученного по принципу ключевого обмена Диффи – Хеллмана на точках эллиптической кривой [9]. Для успешного сочетания сильных сторон асимметричных и симметричных криптосистем в алгоритме совместно применяются особенности ключевого обмена в эллиптических криптосистемах и классические методы симметричного шифрования, такие как гаммирование [10], матричное шифрование Хилла [11; 12] и шифрование с обратной связью блоков [13]. Алгоритм включает пять последовательных этапов:

I. Генерация и обмен открытыми ключами.

II. Генерация двух компонентов общего секретного ключа.

III. Кодирование открытого текста точками эллиптической кривой.

IV. Гаммирование кодированного текста с использованием операции вычисления кратной точки.

V. Формирование блоков и их матричное преобразование Хилла с обратной связью.

Опишем подробно основные этапы такого блочного алгоритма и приведем наглядный пример его работы.

1. Генерация двухкомпонентного общего секретного ключа на основе ключевого обмена Диффи – Хеллмана

Обязательным начальным этапом любой криптографической системы является этап генерации ключей пользователей [14]. Данная система организуется по принципу криптосистем с открытым ключом, в которых каждый пользователь имеет два типа личных ключей: секретный и открытый. В соответствии с протоколом ключевого обмена Диффи – Хеллмана на основе личных ключей двух взаимодействующих пользователей системы можно сгенерировать их общий секретный ключ [15].

Для согласования общих входных параметров криптографической системы и реализации ключевого обмена между пользователями в сервисе совместного доступа и использования файлов публикуется информация о следующих используемых в системе данных:

1) конечное поле вычетов Z_p по простому модулю p;

2) эллиптическая группа $E_p(a, b)$, состоящая из точек эллиптической кривой с уравнением $y^2 = x^3 + ax + b$, определенной над полем Z_p;

3) генерирующая точка $G = (x, y)$, $G \in E_p(a, b)$;

4) порядок N эллиптической группы $E_p(a, b)$, равный общему количеству элементов в эллиптической группе, включая бесконечно удаленную точку;

5) порядок n квадратной ключ-матрицы K – четное натуральное число;

6) инструкция преобразования числовых значений (матрицы) секретного ключа при нарушении условий их допустимости: число (первый элемент матрицы) увеличивается на единицу до первого подходящего значения;

7) рекуррентная формула для последовательности точек эллиптической группы (при необходимости);

8) используемый алфавит (множество символов), закодированный конечными точками эллиптической группы $E_p(a, b)$;

9) открытые ключи пользователей в виде конечной последовательности точек эллиптической группы $E_p(a, b)$.

Размещение перечисленной выше информации на сервисе совместного доступа позволяет организовать бесконтактную работу взаимодействующих пользователей криптосистемы, без необходимости предварительных встреч, переговоров и передачи данных по защищенным каналам связи. Данная информация не является конфиденциальной и доступна к ознакомлению и использованию всеми заинтересованными лицами. В пункте 6 речь идет о дистанци-

онной договоренности между взаимодействующими пользователями о конкретных действиях с целью устранения недопустимых числовых значений, в частности коэффициентов кратности для точек эллиптической кривой, приводящих к появлению точки в бесконечности. Принятая инструкция числовых преобразований может пригодиться пользователям на этапе генерации их общих секретных ключей при анализе невырожденности матричного ключа и числовых элементов гамма-последовательности, а также при вычислении точек ключа по рекуррентной формуле.

Далее на основе открытых ключей двух взаимодействующих пользователей по аналогии с ключевым обменом Диффи – Хеллмана генерируется их общий секретный ключ, состоящий в данном случае из двух компонентов: квадратной ключ-матрицы K фиксированного четного порядка n и числовой гамма-последовательности Г из n^2 чисел, построенных из координат выбранных точек эллиптической группы $E_p(a, b)$. Рассмотрим подробнее два возможных варианта генерации общего секретного ключа для двух воображаемых взаимодействующих пользователей данной криптосистемы Алисы и Боба.

1.1. Вариант с заданной числовой последовательностью секретного ключа

Генерация личных секретных ключей

Каждый из пользователей генерирует свою личную числовую последовательность S, состоящую из $n^2/2$ целых чисел, не кратных порядку N эллиптической группы $E_p(a, b)$. Числовая последовательность S и будет представлять секретный ключ данного пользователя.

Генерация открытого ключа

Каждый из пользователей, используя свой секретный ключ S и операцию вычисления кратной точки в эллиптической группе, вычисляет свою последовательность точек эллиптической группы $E_p(a, b)$. Полученная последовательность точек и будет выполнять роль открытого ключа данного пользователя. Рассмотрим генерацию открытого ключа на примере одного из пользователей, например Алисы. Пусть Алиса имеет своим секретным ключом числовую последовательность $S^A = \left(s_1^A, s_2^A, \ldots, s_{n^2/2}^A\right)$. Тогда она вычисляет последовательность точек кривой $E_p(a, b)$, перемножая элементы числовой последовательности S^A на известную генерирующую точку G:

$$s_1^A \cdot G = Q_1^A = \left(u_1, v_1\right),$$
$$s_2^A \cdot G = Q_2^A = \left(u_2, v_2\right),$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$s_{n^2/2}^A \cdot G = Q_{n^2/2}^A = \left(u_{n^2/2}, v_{n^2/2}\right).$$

В итоге будет сформирован открытый ключ Алисы в виде последовательности точек кривой $E_p(a, b)$:

$$Q^A = \left(Q_1^A, Q_2^A, \ldots, Q_{n^2/2}^A\right).$$

Свои открытые ключи пользователи также выкладывают в сервис совместного доступа.

Генерация общего двухкомпонентного секретного ключа

Имея доступ к открытому ключу своего напарника, пользователи находят общий секретный ключ в виде ключ-матрицы K. Рассмотрим процесс генерации общего секретного ключа на примере двух воображаемых пользователей-напарников Алисы и Боба. Алиса узнает от-

крытый ключ Боба $Q^B = \left(Q_1^B, Q_2^B, \ldots, Q_{n^2/2}^B\right)$, выложенный им на открытый сервис совместного доступа. Используя точки открытой последовательности Q^B и числа своей личной секретной последовательности S^A, Алиса находит последовательность точек общего секретного ключа как результат вычисления кратных точек в эллиптической группе $E_p(a, b)$:

$$s_1^A \cdot Q_1^B = P_1 = \left(\xi_1, \tau_1\right),$$
$$s_2^A \cdot Q_2^B = P_2 = \left(\xi_2, \tau_1\right),$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$s_{n^2/2}^A \cdot Q_{n^2/2}^B = P_{n^2/2} = \left(\xi_{n^2/2}, \tau_{n^2/2}\right).$$

Из координат полученных точек Алиса формирует квадратную ключ-матрицу K, построчно записывая в нее координаты точек. Чтобы понять, как именно это происходит, предположим что $n = 4$ и Алиса указанным выше способом получила общую с Бобом секретную последовательность из $(4 \times 4)/2 = 8$ точек $\left(P_1, P_2, P_3, P_4, P_5, P_6, P_7, P_8\right)$, где $P_i = \left(\xi_i, \tau_i\right), i = 1, \ldots, 8$. Тогда ключ-матрица K получает размерность 4×4 и формируется по правилу

$$K = \begin{pmatrix} \xi_1 & \tau_1 & \xi_2 & \tau_2 \\ \xi_3 & \tau_3 & \xi_4 & \tau_4 \\ \xi_5 & \tau_5 & \xi_6 & \tau_6 \\ \xi_7 & \tau_7 & \xi_8 & \tau_8 \end{pmatrix}.$$

Точно такая же ключ-матрица получится у Боба – напарника Алисы, на основе его личной секретной числовой последовательности S^B и открытого ключа Алисы $Q^A = \left(Q_1^A, Q_2^A, \ldots, Q_{n^2/2}^A\right)$, так как

$$s_i^A \cdot Q_i^B = s_i^A \cdot (s_i^B \cdot G) = s_i^B \cdot (s_i^A \cdot G) = s_i^B \cdot Q_i^A, \; i = 1, \ldots, n^2/2.$$

Ключ-матрица K должна быть невырожденной, т. е. должна быть обратимой по модулю. Нарушение этого условия не позволит использовать этот матричный ключ и потребует его изменения в соответствии с инструкцией.

Кроме ключ-матрицы K, на основе полученной совместной последовательности точек P_i, $i = 1, \ldots, n^2/2$ ормируется числовая гамма-последовательность $\Gamma = \left\{x_1, y_1, x_2, y_2, \ldots, x_{\frac{n^2}{2}}, y_{\frac{n^2}{2}}\right\} = \left\{\gamma_i\right\}_{i=1}^{n^2}$ длиною n^2. Все числа этой последовательности независимо друг от друга будут использованы при шифровании как числовые коэффициенты кратности при вычислении кратных точек в эллиптической группе $E_p(a, b)$. Поэтому они должны соответствовать обязательному условию обратимости этой операции, а именно должны быть взаимно простыми с порядком эллиптической группы N, т. е. должно выполняться равенство $\text{НОД}\left(N, \gamma_i\right) = 1, i = 1, \ldots, n^2$. В противном случае при нарушении этого условия, согласно опубликованной инструкции, недопустимое числовое значение γ_i этой последовательности преобразуется путем последовательного увеличения на 1 до первого подходящего значения.

Таким образом, на основе общей секретной последовательности точек эллиптической кривой $\left\{P_i\right\}_{i=1}^{n^2/2}$, $P_i = (x_i, y_i) \in E_p(a, b)$ будет сгенерирован общий двухкомпонентный секретный ключ для взаимодействующих пользователей Алисы и Боба, представленный:

1) квадратной ключ-матрицей K размером $n \times n$;

2) числовой гамма-последовательностью $\Gamma = \left\{\gamma_i\right\}_{i=1}^{n^2}$.

1.2. Вариант с использованием рекуррентной формулы

Если в списке общих параметров системы порядок n ключ-матрицы K будет выбран достаточно большим, то, чтобы избежать сложности организации работы с большим объемом входных личных секретных данных, удобнее использовать рекуррентную формулу для вычисления последовательности точек общего секретного ключа пользователей.

В этом случае числовую последовательность личных секретных ключей пользователей достаточно задать двумя случайными целыми числами, не кратными порядку эллиптической группы N. С их помощью по тем же формулам, что и в предыдущем варианте, каждый пользователь находит свой открытый ключ, представляющий собой последовательность из двух точек эллиптической кривой, и публикует его на платформе сервиса совместного доступа.

После обмена открытыми ключами пользователи Алиса и Боб независимо друг от друга находят общий секретный ключ, состоящий из двух точек (P_1, P_2) эллиптической группы $E_p(a, b)$. Чтобы найти оставшиеся $\left(\frac{n^2}{2}-2\right)$ точек, необходимые для формирования ключ-матрицы K, пользователи используют рекуррентную формулу, опубликованную на открытом сервисе совместного доступа вместе с другими общими входными данными системы. Если порядок n ключ-матрицы не кратен порядку N эллиптической группы, то в качестве рекуррентной формулы может быть использована, например, такая формула по двум точкам:

$$P_{i+2} = n \cdot P_i + P_{i+1}, \tag{1}$$

где $i = 1, \ldots, n^2/2 - 2$; $P_1 \in E_p(a,b)$ – первая точка общего секретного ключа; $P_2 \in E_p(a,b)$ – вторая точка общего секретного ключа.

Если при вычислении по формуле (1) на каком-то i-м шаге $(i = 1, \ldots, n^2/2 - 2)$ кратная точка $n \cdot P_i$ окажется противоположной точкой для точки P_{i+1}, то, чтобы избежать в качестве результата их сложения точки в бесконечности, коэффициент кратности n в этом случае увеличивают на 1 до первого подходящего значения, т. е. значения, не кратного N, при котором точки $n \cdot P_i$ и P_{i+1} не будут являться противоположными. Договоренность о таких действиях также публикуется заранее на используемом сервисе совместного доступа для взаимодействующих пользователей.

Таким образом, на основе двух общих секретных входных точек P_1 и P_2, полученных по принципу ключевого обмена Диффи – Хеллмана, и известной рекуррентной формулы пользователями Алисой и Бобом бесконтактно будет сгенерирована общая секретная последовательность из $n^2/2$ точек эллиптической группы $E_p(a, b)$ и далее будут сформированы их общие секретные ключ-матрица K и гамма-последовательность Г. Стоит отметить, что открытая информация о значении порядка n ключ-матрицы K и о самой рекуррентной формуле без известных входных точек P_1 и P_2 не позволит несанкционированным пользователям сгенерировать последовательность точек общего секретного ключа Алисы и Боба, а значит, и определить их общую ключ-матрицу K. Информация об общих секретных входных точках P_1 и P_2 защищена высокой трудоемкостью решения задачи дискретного логарифмирования.

Формирование ключ-матрицы K и гамма-последовательности Г осуществляется из координат полученных точек $P_i, i = 1, \ldots, n^2/2,$ по той же схеме, как показано в предыдущем варианте.

Если в качестве входных секретных личных ключей пользователей взять последовательность из k целых чисел, где $k < n^2/2$, то открытые ключи пользователей будут представлены упорядоченным набором из k точек эллиптической группы. Тогда построение общего секретного ключа будет включать вычисление по принципу Диффи – Хеллмана последовательности

из k общих точек P_1, P_2, ..., P_k и определение оставшихся ($n^2/2 - k$) точек по рекуррентной формуле вида

$$P_{k+i} = n^{k-1} \cdot P_i + n^{k-2} \cdot P_{i+1} + \ldots + n \cdot P_{i+k-2} + P_{i+k-1}, \quad (2)$$

2. Шифрование/дешифрование с использованием двухкомпонентного общего секретного ключа

Любая криптографическая система состоит из двух взаимообратных алгоритмов: шифрования и дешифрования. Алгоритм шифрования с использованием двухкомпонентного общего секретного ключа состоит из двух последовательных этапов:

Этап 1: гаммирование

После кодирования открытого текста точками эллиптической кривой осуществляется его гаммирование с операцией умножения точек на числовой коэффициент, т. е. вычисления кратной точки, где в качестве коэффициентов кратности выступают числовые значения секретной гамма-последовательности. При этом сама гамма повторяется столько раз, сколько необходимо для покрытия всего открытого текста. В итоге будет получен промежуточный шифр, представляющий собой последовательность полученных кратных точек.

Этап 2: матричное преобразование

Координаты точек промежуточного шифра, полученного после гаммирования, объединяются в общую числовую последовательность с сохранением порядка следования точек и порядка их координат. Далее числовая последовательность разбивается на n^2-блоки, т. е. блоки последовательности длиною n^2. Если целое деление длины последовательности на n^2 невозможно и в итоге остается неполный блок, то в его конец добавляются значения, взятые с начала первого блока.

Из каждого полученного числового n^2-блока построчными срезами длиной n формируется квадратная матрица и промежуточный блочный шифр принимает вид последовательности матриц $\{T_i\}$. Далее каждая матрица из полученной последовательности умножается слева на ключ-матрицу K по модулю p с использованием обратной связи с предыдущей матрицей-блоком по правилу

$$C_1 = K \cdot T_1 \,(mod\ p),\ C_i = \left(K \cdot T_i + T_{i-1}\right)(mod\ p), i > 1. \quad (3)$$

Полученная последовательность матриц $\{C_i\}$ и представляет шифр по данному алгоритму.

Дешифрование также проходит в два этапа, но в обратном порядке. Общий алгоритм дешифрования имеет вид:

1. Определяется обратная матрица K^{-1} (*mod N*) для секретной ключ-матрицы K.
2. Вычисляются обратные значения по модулю N для числовых элементов гамма-последовательности Г: γ^{-1} (*mod N*), из которых составляется дешифрующая гамма-последовательность $Г^{-1}$ (*mod N*).
3. Каждая шифр-матрица C_i преобразуется по формуле матричного дешифрования с использованием обратной связи:

$$T_1 = K^{-1} \cdot C_1 \,(mod\ p),\ T_i = K^{-1} \cdot \left(C_i - T_{i-1}\right)(mod\ p),\ i > 1. \quad (4)$$

4. Из матриц T_i извлекаются построчные срезы, из элементов которых составляется последовательность числовых значений. Далее с сохранением порядка следования значения объединяются в упорядоченные пары, представляющие точки эллиптической группы $E_p(a, b)$.

5. Каждая точка из полученной последовательности точек эллиптической группы умножается на соответствующий элемент дешифрующей гамма-последовательности.

6. Полученный набор точек эллиптической группы декодируется в соответствующие символы алфавита и открытый текст восстанавливается.

3. Пример работы алгоритма с рекуррентной формулой для открытых ключей

Предположим, что для организации работы взаимодействующих пользователей на каком-либо сервисе совместного доступа и использования файлов была предварительно опубликована информация о следующих используемых в системе данных:

1) конечное поле вычетов Z_{43} по простому модулю $p = 43$;

2) эллиптическая группа $Z_{43}(17,24)$ состоящая из точек эллиптической кривой $y^2 = x^3 + 17x + 24$, определенной над полем Z_{43};

3) генерирующая точка $G = (16, 36)$, $G \in E_{43}(17, 24)$;

4) порядок $N = 51$ эллиптической группы $E_{43}(17, 24)$;

5) порядок $n = 4$ квадратной ключ-матрицы K – четное натуральное число;

6) инструкция преобразования числовых значений (матрицы) секретного ключа при нарушении условий их допустимости: число (первый элемент матрицы) увеличивается на единицу до первого подходящего значения;

7) рекуррентная формула (1) для последовательности точек эллиптической группы;

8) кодированный русский алфавит на точках эллиптической кривой $E_{43}(17, 24)$ (табл. 1).

Таблица 1

Кодированный алфавит

Table 1

Encoded alphabet

№	символ	точка	№	символ	точка	№	символ	точка
1	**А**	(0, 14)	18	**Р**	(18, 23)	35	**.**	(30, 10)
2	**Б**	(0, 29)	19	**С**	(19, 5)	36	**,**	(30, 33)
3	**В**	(2, 18)	20	**Т**	(19, 38)	37	**!**	(32, 21)
4	**Г**	(2, 25)	21	**У**	(21, 15)	38	**?**	(32, 22)
5	**Д**	(3, 4)	22	**Ф**	(21, 28)	39	**:**	(33, 12)
6	**Е**	(3, 39)	23	**Х**	(22, 9)	40	*****	(33, 31)
7	**Ё**	(6, 16)	24	**Ц**	(22, 34)	41	**0**	(35, 8)
8	**Ж**	(6, 27)	25	**Ч**	(24, 18)	42	**1**	(35, 35)
9	**З**	(7, 20)	26	**Ш**	(24, 25)	43	**2**	(36, 11)
10	**И**	(7, 23)	27	**Щ**	(25, 11)	44	**3**	(36, 32)
11	**Й**	(12, 8)	28	**Ъ**	(25, 32)	45	**4**	(39, 8)
12	**К**	(12, 35)	29	**Ы**	(26, 5)	46	**5**	(39, 35)
13	**Л**	(16, 7)	30	**Ь**	(26, 38)	47	**6**	(41, 5)
14	**М**	(16, 36)	31	**Э**	(28, 7)	48	**7**	(41, 38)
15	**Н**	(17, 18)	32	**Ю**	(28, 36)	49	**8**	(42, 7)
16	**О**	(17, 25)	33	**Я**	(29, 3)	50	**9**	(42, 36)
17	**П**	(18, 20)	34	**_**	(29, 40)			

На основании этой информации взаимодействующие пользователи Алиса и Боб генерируют сначала свои личные ключи – секретный и открытый, размещают информацию о своих открытых ключах на совместном сервере и затем приступают к генерации компонентов общего секретного ключа.

Генерация двухкомпонентного общего секретного ключа

Предположим, что Алиса своим секретным ключом выбрала упорядоченную числовую пару $S^A = (13,45)$, а Боб $S^B = (9,21)$.

Пользователи по отдельности находят свои открытые ключи в виде двух упорядоченных точек эллиптической группы $E_{43}(17, 24)$ (табл. 2).

Таблица 2

Генерация открытых ключей

Table 2

Public key generation

Алиса	**Боб**
$Q_1^A = 13 \cdot G = 13 \cdot (16,36) = (2,18);$ $Q_2^A = 45 \cdot G = 45 \cdot (16,36) = (29,3).$	$Q_1^B = 9 \cdot G = 9 \cdot (16,36) = (21,15);$ $Q_1^B = 21 \cdot G = 21(16,36) = (30,10).$

Полученные открытые ключи пользователей Алисы и Боба представлены в таблице, опубликованной на сервисе общего пользования (табл. 3).

Таблица 3

Публикация открытых ключей

Table 3

Public key publication

Алиса	**Боб**
$Q^A = ((2,18), (29,3))$	$Q^B = ((21,15), (30,10))$

Используя открытый ключ напарника и числовые значения своего личного секретного ключа, каждый из пользователей формирует ключ-матрицу K и гамма-последовательность Г. Для этого сначала вычисляются две начальные общие точки, из которых потом по рекуррентной формуле (1) вычисляются оставшиеся общие точки пользователей (табл. 4).

Таблица 4

Генерация общих секретных точек

Table 4

Shared secret point generation

Алиса	**Боб**
$P_1 = 13 \cdot (21,15) = (33,12);$ $P_2 = 45 \cdot (30,10) = (24,25).$	$P_1 = 9 \cdot (2,18) = (33,12);$ $P_2 = 21 \cdot (29,3) = (24,25).$
$P_3 = 4 \cdot P_1 + P_2 = 4 \cdot (33,15) + (24,25) = (33,31),$ $P_4 = 4 \cdot P_2 + P_3 = 4 \cdot (24,25) + (33,31) = (21,28),$ $P_5 = 4 \cdot P_3 + P_4 = 4 \cdot (33,31) + (21,28) = (42,7),$ $P_6 = 4 \cdot P_4 + P_5 = 4 \cdot (21,28) + (42,7) = (6,16),$ $P_7 = 4 \cdot P_5 + P_6 = 4 \cdot (42,7) + (6,16) = (24,25),$ $P_8 = 4 \cdot P_6 + P_7 = 4 \cdot (6,16) + (24,25) = (33,12).$	

В итоге из координат полученных точек в соответствии с их порядком каждый пользователь формирует секретную ключ-матрицу – первый компонент общего секретного ключа:

$$K = \begin{pmatrix} 33 & 12 & 24 & 25 \\ 33 & 31 & 21 & 28 \\ 42 & 7 & 6 & 16 \\ 24 & 25 & 33 & 12 \end{pmatrix}$$

и числовую последовательность:

$$\Gamma = (33,12,24,25,33,31,21,28,42,7,6,16,24,25,33,12).$$

Все элементы числовой последовательности проверяются на допустимость, и в случае нарушения условия допустимости для какого-либо элемента его числовое значение будет увеличиваться на единицу до тех пор, пока новое значение и порядок кривой $N = 51$ не станут взаимно простыми числами. В итоге этих действий Алиса и Боб независимо друг от друга получают гамма-последовательность – второй компонент общего секретного ключа:

$$\Gamma = (33 + 2, 12 + 1, 24 + 1, 25, 33 + 2, 31, 21 + 1, 28, 42 + 1,$$
$$7, 6 + 1, 16, 24 + 1, 25, 33 + 2, 12, + 1) =$$
$$= (35, 13, 25, 25, 35, 31, 22, 28, 43, 7, 7, 16, 25, 25, 35, 13).$$

Шифрование с использованием двухкомпонентного общего секретного ключа

Для наглядной демонстрации работы предложенного алгоритма шифрования предположим, что Алиса шифрует исходное сообщение, а Боб его расшифровывает. В качестве примера текста исходного сообщения возьмем тематическое предложение:

«АЛИСА_И_БОБ_ЯВЛЯЮТСЯ_ВЗАИМОДЕЙСТВУЮЩИМИ_ АГЕНТАМИ_КРИПТОГРАФИЧЕСКИХ_СИСТЕМ.»

Числовое значение длины текста сообщения, равное 74, не кратно 16, поэтому Алиса добавляет в конец текста 6 его первых символов и получает расширенный текст сообщения из 80 символов, которое не меняет своего первоначального смысла. Далее Алиса кодирует полученный текст точками эллиптической кривой E_{43} (17, 24) в соответствии с алфавитной таблицей (табл. 5).

Таблица 5

Кодирование открытого текста

Table 5

Plaintext encoding

А	**Л**	**И**	**С**	**А**	**_**	**И**	**_**	**Б**
(0,14)	(16,7)	(7,23)	(19,5)	(0,14)	(29,40)	(7,23)	(29,40)	(0,29)
О	**Б**	**_**	**Я**	**В**	**Л**	**Я**	**Ю**	**Т**
(17,25)	(0,29)	(29,40)	(29,3)	(2,18)	(16,7)	(29,3)	(28,36)	(19,38)
С	**Я**	**_**	**В**	**З**	**А**	**И**	**М**	**О**
(19,5)	(29,3)	(29,40)	(2,18)	(7,20)	(0,14)	(7,23)	(16,36)	(17,25)
Д	**Е**	**Й**	**С**	**Т**	**В**	**У**	**Ю**	**Щ**
(3,4)	(3,39)	(12,8)	(19,5)	(19,38)	(2,18)	(21,15)	(28,36)	(25,11)

Окончсание табл. 5

И	**М**	**И**	**_**	**А**	**Г**	**Е**	**Н**	**Т**
(7,23)	(16,36)	(7,23)	(29,40)	(0,14)	(2,25)	(3,39)	(17,18)	(19,38)

А	**М**	**И**	**_**	**К**	**Р**	**И**	**П**	**Т**
(0,14)	(16,36)	(7,23)	(29,40)	(12,35)	(18,23)	(7,23)	(18,20)	(19,38)

О	**Г**	**Р**	**А**	**Ф**	**И**	**Ч**	**Е**	**С**
(17,25)	(2,25)	(18,23)	(0,14)	(21,28)	(7,23)	(24,18)	(3,39)	(19,5)

К	**И**	**Х**	**_**	**С**	**И**	**С**	**Т**	**Е**
(12,35)	(7,23)	(22,9)	(29,40)	(19,5)	(7,23)	(19,5)	(19,38)	(3,39)

М	**.**	**А**	**Л**	**И**	**С**	**А**	**_**
(16,36)	(30,10)	(0,14)	(16,7)	(7,23)	(19,5)	(0,14)	(29,40)

После кодирования текста Алиса поэлементно накладывает на него гамма-последовательность Г и производит его гаммирование с операцией вычисления кратной точки на эллиптической кривой (табл. 6).

Таблица 6

Гаммирование

Table 6

Gamma encryption

(0,14)	**(16,7)**	**(7,23)**	**(19,5)**	**(0,14)**	**(29,40)**	**(7,23)**	**(29,40)**	**(0,29)**
35	13	25	25	35	31	22	28	43
(12,8)	(2,25)	(18,25)	(17,18)	(12,8)	(42,7)	(16,36)	(33,12)	(36,11)

(17,25)	**(0,29)**	**(29,40)**	**(29,3)**	**(2,18)**	**(16,7)**	**(29,3)**	**(28,36)**	**(19,38)**
7	7	16	25	25	35	13	35	13
(36,32)	(18,20)	(29,3)	(16,27)	(28,7)	(32,22)	(24,18)	(25,32)	(7,23)

(19,5)	**(29,3)**	**(29,40)**	**(2,18)**	**(7,20)**	**(0,14)**	**(7,23)**	**(16,36)**	**(17,25)**
25	25	35	31	22	28	43	7	7
(17,18)	(6,27)	(29,40)	(22,9)	(16,7)	(17,25)	(22,9)	(7,23)	(36,32)

(3,4)	**(3,39)**	**(12,8)**	**(19,5)**	**(19,38)**	**(2,18)**	**(21,15)**	**(28,36)**	**(25,11)**
16	25	25	35	13	35	13	25	25
(7,23)	(22,34)	(36,11)	(41,38)	(7,23)	(36,11)	(21,15)	(28,36)	(16,7)

(7,23)	**(16,36)**	**(7,23)**	**(29,40)**	**(0,14)**	**(2,25)**	**(3,39)**	**(17,18)**	**(19,38)**
35	31	22	28	43	7	7	16	25
(3,39)	(35,8)	(16,36)	(33,12)	(36,32)	(41,5)	(28,36)	(35,35)	(17,25)

(0,14)	**(16,36)**	**(7,23)**	**(29,40)**	**(12,35)**	**(18,23)**	**(7,23)**	**(18,20)**	**(19,38)**
25	35	13	35	13	25	25	35	31
(2,18)	(32,21)	(41,38)	(29,40)	(25,32)	(41,38)	(18,23)	(22,9)	(16,36)

Окончание табл. 6

(17,25)	**(2,25)**	**(18,23)**	**(0,14)**	**(21,28)**	**(7,23)**	**(24,18)**	**(3,39)**	**(19,5)**
22	28	43	7	7	16	25	25	35
(25,32)	(7,20)	(19,38)	(18,23)	(39,25)	(3,4)	(39,35)	(22,34)	(41,38)

(12,35)	**(7,23)**	**(22,9)**	**(29,40)**	**(19,5)**	**(7,23)**	**(19,5)**	**(19,38)**	**(3,39)**
13	35	13	25	25	35	31	22	28
(2,18)	(3,39)	(17,25)	(6,16)	(17,18)	(3,39)	(16,7)	(36,32)	(0,29)

(16,36)	**(30,10)**	**(0,14)**	**(16,7)**	**(7,23)**	**(19,5)**	**(0,14)**	**(29,40)**
43	7	7	16	25	25	35	13
(12,35)	(29,3)	(18,23)	(32,21)	(18,23)	(17,18)	(12,8)	(24,25)

В результате гаммирования Алиса получает последовательность точек эллиптической кривой, из координат которой построчно формирует квадратные матрицы размером 4 × 4 и в итоге получает промежуточный шифр в виде последовательности матриц $T_{ii=1}^{10}$:

$$T_1 = \begin{pmatrix} 12 & 8 & 2 & 25 \\ 18 & 23 & 17 & 18 \\ 12 & 8 & 42 & 7 \\ 16 & 26 & 33 & 12 \end{pmatrix} \qquad T_6 = \begin{pmatrix} 36 & 32 & 41 & 5 \\ 28 & 36 & 35 & 35 \\ 17 & 25 & 2 & 18 \\ 32 & 21 & 41 & 38 \end{pmatrix}$$

$$T_2 = \begin{pmatrix} 36 & 11 & 36 & 32 \\ 18 & 20 & 29 & 3 \\ 6 & 27 & 28 & 7 \\ 32 & 22 & 24 & 18 \end{pmatrix} \qquad T_7 = \begin{pmatrix} 29 & 40 & 25 & 32 \\ 41 & 38 & 18 & 23 \\ 22 & 9 & 16 & 36 \\ 25 & 32 & 7 & 20 \end{pmatrix}$$

$$T_3 = \begin{pmatrix} 25 & 32 & 7 & 23 \\ 17 & 18 & 6 & 27 \\ 29 & 40 & 22 & 9 \\ 16 & 7 & 17 & 25 \end{pmatrix} \qquad T_8 = \begin{pmatrix} 19 & 38 & 18 & 23 \\ 39 & 35 & 3 & 4 \\ 39 & 35 & 22 & 38 \\ 41 & 38 & 2 & 18 \end{pmatrix}$$

$$T_4 = \begin{pmatrix} 22 & 9 & 7 & 23 \\ 36 & 32 & 7 & 23 \\ 22 & 34 & 36 & 11 \\ 41 & 38 & 7 & 23 \end{pmatrix} \qquad T_9 = \begin{pmatrix} 3 & 39 & 17 & 25 \\ 6 & 16 & 17 & 18 \\ 3 & 39 & 16 & 7 \\ 36 & 32 & 0 & 29 \end{pmatrix}$$

$$T_5 = \begin{pmatrix} 36 & 11 & 21 & 15 \\ 28 & 36 & 16 & 7 \\ 3 & 39 & 35 & 8 \\ 16 & 36 & 33 & 12 \end{pmatrix} \qquad T_{10} = \begin{pmatrix} 12 & 35 & 29 & 3 \\ 18 & 23 & 32 & 21 \\ 18 & 23 & 17 & 18 \\ 12 & 8 & 24 & 25 \end{pmatrix}$$

Далее каждая матрица из полученной последовательности умножается слева на ключ-матрицу K по модулю $p = 43$ с использованием обратной связи по правилу (3):

$$C_1 = K \cdot T_1 (mod\ 43) = \begin{pmatrix} 10 & 41 & 39 & 4 \\ 20 & 3 & 34 & 17 \\ 12 & 3 & 37 & 34 \\ 36 & 1 & 19 & 6 \end{pmatrix} (mod\ 43),$$

$$C_2 = K \cdot T_2 + T_1 (mod\ 43) = \begin{pmatrix} 38 & 3 & 15 & 15 \\ 34 & 39 & 10 & 12 \\ 5 & 6 & 30 & 25 \\ 20 & 20 & 39 & 12 \end{pmatrix} (mod\ 43),$$

$$C_3 = K \cdot T_3 + T_2 (mod\ 43) = \begin{pmatrix} 11 & 10 & 2 & 21 \\ 12 & 4 & 8 & 37 \\ 14 & 0 & 37 & 25 \\ 12 & 21 & 25 & 36 \end{pmatrix} (mod\ 43),$$

$$C_4 = K \cdot T_4 + T_3 (mod\ 43) = \begin{pmatrix} 27 & 28 & 28 & 5 \\ 29 & 32 & 30 & 9 \\ 15 & 35 & 5 & 22 \\ 39 & 21 & 41 & 28 \end{pmatrix} (mod\ 43),$$

$$C_5 = K \cdot T_5 + T_4 (mod\ 43) = \begin{pmatrix} 40 & 17 & 20 & 19 \\ 23 & 27 & 17 & 35 \\ 26 & 10 & 5 & 27 \\ 4 & 40 & 11 & 20 \end{pmatrix} (mod\ 43),$$

$$C_6 = K \cdot T_6 + T_5 (mod\ 43) = \begin{pmatrix} 16 & 1 & 29 & 4 \\ 26 & 10 & 32 & 33 \\ 3 & 14 & 4 & 18 \\ 31 & 29 & 42 & 36 \end{pmatrix} (mod\ 43),$$

$$C_7 = K \cdot T_7 + T_6 (mod\ 43) = \begin{pmatrix} 15 & 29 & 7 & 35 \\ 21 & 7 & 15 & 24 \\ 33 & 0 & 10 & 38 \\ 27 & 32 & 26 & 14 \end{pmatrix} (mod\ 43),$$

$$C_8 = K \cdot T_8 + T_7 (mod\ 43) = \begin{pmatrix} 32 & 21 & 29 & 41 \\ 17 & 5 & 19 & 17 \\ 5 & 2 & 11 & 17 \\ 10 & 33 & 17 & 32 \end{pmatrix} (mod\ 43),$$

$$C_9 = K \cdot T_9 + T_8 (mod\ 43) = \begin{pmatrix} 1 & 28 & 6 & 22 \\ 19 & 7 & 8 & 24 \\ 27 & 37 & 5 & 39 \\ 20 & 35 & 30 & 13 \end{pmatrix} (mod\ 43),$$

$$C_{10} = K \cdot T_{10} + T_9 (mod\ 43) = \begin{pmatrix} 14 & 29 & 1 & 14 \\ 40 & 11 & 28 & 40 \\ 30 & 1 & 9 & 14 \\ 7 & 23 & 23 & 15 \end{pmatrix} (mod\ 43),$$

Полученная последовательность матриц $\{C_i\}_{i=1}^{10}$ представляет окончательный шифр, который Алиса отправляет Бобу.

Дешифрование

Боб, получив шифр в виде последовательности матриц $\{C_i\}_{i=1}^{10}$, приступает к его дешифрованию. Для этого он находит обратную матрицу по модулю $p = 43$ к ключ-матрице K:

$$K^{-1} (mod\ 43) = \begin{pmatrix} 33 & 12 & 24 & 25 \\ 33 & 31 & 21 & 28 \\ 42 & 7 & 6 & 16 \\ 24 & 25 & 33 & 12 \end{pmatrix}^{-1} (mod\ 43) = \begin{pmatrix} 29 & 9 & 40 & 5 \\ 8 & 36 & 19 & 3 \\ 30 & 34 & 26 & 17 \\ 22 & 7 & 6 & 41 \end{pmatrix}$$

и также обратные значения по модулю $N = 51$ к элементам гамма-последовательности Г:

$$35^{-1} = 35(mod\ 51) \qquad 31^{-1} = 28(mod\ 51) \qquad 43^{-1} = 19(mod\ 51)$$

$$13^{-1} = 4(mod\ 51) \qquad 22^{-1} = 7(mod\ 51) \qquad 7^{-1} = 22(mod\ 51)$$

$$25^{-1} = 49(mod\ 51) \qquad 28^{-1} = 31(mod\ 51) \qquad 16^{-1} = 16(mod\ 51)$$

В итоге обратная гамма-последовательность принимает вид:

$$\Gamma^{-1}\ (mod\ 51) = (35, 4, 49, 49, 35, 28, 7, 31, 19, 22, 22, 16, 49, 49, 35, 4).$$

После определения ключей дешифрования $K^{-1}\ (mod\ 51)$ и $\Gamma^{-1}\ (mod\ 51)$ Боб на первом этапе по дешифрующим формулам (4) умножает слева полученные от Алисы матрицы из последовательности $\{C_i\}_{i=1}^{10}$ на обратную ключ-матрицу с использованием связи с предыдущей матрицей и находит промежуточный шифр – последовательность матриц $\{T_i\}_{i=1}^{10}$:

$$T_1 = K^{-1} (mod\ 43) = \begin{pmatrix} 33 & 12 & 24 & 25 \\ 33 & 31 & 21 & 28 \\ 42 & 7 & 6 & 16 \\ 24 & 25 & 33 & 12 \end{pmatrix} (mod\ 43),$$

$$T_2 = K^{-1} \cdot \left(C_2 - T_1\right)(mod\ 43) = \begin{pmatrix} 36 & 11 & 36 & 32 \\ 18 & 20 & 29 & 3 \\ 6 & 27 & 28 & 7 \\ 32 & 22 & 24 & 18 \end{pmatrix} (mod\ 43),$$

$$T_3 = K^{-1} \cdot \left(C_3 - T_2\right)(mod\ 43) = \begin{pmatrix} 25 & 32 & 7 & 23 \\ 17 & 18 & 6 & 27 \\ 29 & 40 & 22 & 9 \\ 16 & 7 & 17 & 25 \end{pmatrix} (mod\ 43),$$

$$T_4 = K^{-1} \cdot \left(C_4 - T_3\right)(mod\ 43) = \begin{pmatrix} 22 & 9 & 7 & 23 \\ 36 & 32 & 7 & 23 \\ 22 & 34 & 36 & 11 \\ 41 & 38 & 7 & 23 \end{pmatrix} (mod\ 43),$$

$$T_5 = K^{-1} \cdot \left(C_5 - T_4\right)(mod\ 43) = \begin{pmatrix} 36 & 11 & 21 & 15 \\ 28 & 36 & 16 & 7 \\ 3 & 39 & 35 & 8 \\ 16 & 36 & 33 & 12 \end{pmatrix} (mod\ 43),$$

$$T_6 = K^{-1} \cdot \left(C_6 - T_5\right)(mod\ 43) = \begin{pmatrix} 36 & 32 & 41 & 5 \\ 28 & 36 & 35 & 35 \\ 17 & 25 & 2 & 18 \\ 32 & 21 & 41 & 38 \end{pmatrix} (mod\ 43),$$

$$T_7 = K^{-1} \cdot \left(C_7 - T_6\right)(mod\ 43) = \begin{pmatrix} 29 & 40 & 25 & 32 \\ 41 & 38 & 18 & 23 \\ 22 & 9 & 16 & 36 \\ 25 & 32 & 7 & 20 \end{pmatrix} (mod\ 43),$$

$$T_8 = K^{-1} \cdot \left(C_8 - T_7\right)(mod\ 43) = \begin{pmatrix} 19 & 38 & 18 & 23 \\ 39 & 35 & 3 & 4 \\ 39 & 35 & 22 & 38 \\ 41 & 38 & 2 & 18 \end{pmatrix} (mod\ 43),$$

$$T_9 = K^{-1} \cdot \left(C_9 - T_8\right)(mod\ 43) = \begin{pmatrix} 3 & 39 & 17 & 25 \\ 6 & 16 & 17 & 18 \\ 3 & 39 & 16 & 7 \\ 36 & 32 & 0 & 29 \end{pmatrix} (mod\ 43),$$

$$T_{10} = K^{-1} \cdot \left(C_{10} - T_9\right)(mod\ 43) = \begin{pmatrix} 12 & 35 & 29 & 3 \\ 18 & 23 & 32 & 21 \\ 18 & 23 & 17 & 18 \\ 12 & 8 & 24 & 25 \end{pmatrix} (mod\ 43),$$

На втором этапе дешифрования из упорядоченных матриц промежуточного шифра T_i (i = 1,…,10) Боб извлекает построчные срезы, составляет из их элементов последовательность числовых значений, которую затем разбивает на упорядоченные пары – точки эллиптической кривой. Далее для полученной последовательности точек Боб производит поэлементное наложение обратной гамма-последовательности Γ^{-1} (*mod* 51) и выполняет гаммирование с операцией умножения точки эллиптической кривой на числовой элемент обратной гамма-последовательности (табл. 7).

Таблица 7

Обратное гаммирование

Table 7

Inverse gamma encryption

(12,8)	**(2,25)**	**(18,25)**	**(17,18)**	**(12,8)**	**(42,7)**	**(16,36)**	**(33,12)**	**(36,11)**
35	4	49	49	35	28	7	31	19
(0,14)	(16,7)	(7,23)	(19,5)	(0,14)	(29,40)	(7,23)	(29,40)	(0,29)

(36,32)	**(18,20)**	**(29,3)**	**(16,27)**	**(28,7)**	**(32,22)**	**(24,18)**	**(25,32)**	**(7,23)**
22	22	16	49	49	35	4	35	4
(17,25)	(0,29)	(29,40)	(29,3)	(2,18)	(16,7)	(29,3)	(28,36)	(19,38)

(17,18)	**(6,27)**	**(29,40)**	**(22,9)**	**(16,7)**	**(17,25)**	**(22,9)**	**(7,23)**	**(36,32)**
49	49	35	28	7	31	19	22	22
(19,5)	(29,3)	(29,40)	(2,18)	(7,20)	(0,14)	(7,23)	(16,36)	(17,25)

(7,23)	**(22,34)**	**(36,11)**	**(41,38)**	**(7,23)**	**(36,11)**	**(21,15)**	**(28,36)**	**(16,7)**
16	49	49	35	4	35	4	49	49
(3,4)	(3,39)	(12,8)	(19,5)	(19,38)	(2,18)	(21,15)	(28,36)	(25,11)

(3,39)	**(35,8)**	**(16,36)**	**(33,12)**	**(36,32)**	**(41,5)**	**(28,36)**	**(35,35)**	**(17,25)**
35	28	7	31	19	22	22	16	49
(7,23)	(16,36)	(7,23)	(29,40)	(0,14)	(2,25)	(3,39)	(17,18)	(19,38)

(2,18)	**(32,21)**	**(41,38)**	**(29,40)**	**(25,32)**	**(41,38)**	**(18,23)**	**(22,9)**	**(16,36)**
49	35	4	35	4	49	49	35	28
(0,14)	(16,36)	(7,23)	(29,40)	(12,35)	(18,23)	(7,23)	(18,20)	(19,38)

(25,32)	**(7,20)**	**(19,38)**	**(18,23)**	**(39,25)**	**(3,4)**	**(39,35)**	**(22,34)**	**(41,38)**
7	31	19	22	22	16	49	49	35
(17,25)	(2,25)	(18,23)	(0,14)	(21,28)	(7,23)	(24,18)	(3,39)	(19,5)

Окончание табл. 7

(2,18)	**(3,39)**	**(17,25)**	**(6,16)**	**(17,18)**	**(3,39)**	**(16,7)**	**(36,32)**	**(0,29)**
4	35	4	49	49	35	28	7	31
(12,35)	(7,23)	(22,9)	(29,40)	(19,5)	(7,23)	(19,5)	(19,38)	(3,39)

(12,35)	**(29,3)**	**(18,23)**	**(32,21)**	**(18,23)**	**(17,18)**	**(12,8)**	**(24,25)**
19	22	22	16	49	49	35	4
(16,36)	(30,10)	(0,14)	(16,7)	(7,23)	(19,5)	(0,14)	(29,40)

В итоге описанных выше действий Боб получает набор точек эллиптической кривой, декодирует их в соответствующие символы алфавита по табл. 5 и тем самым восстанавливает открытый текст:

«АЛИСА_И_БОБ_ЯВЛЯЮТСЯ_ВЗАИМОДЕЙСТВУЮЩИМИ_
АГЕНТАМИ_КРИПТОГРАФИЧЕСКИХ_СИСТЕМ.АЛИСА_»

Последние 6 символов «АЛИСА_» дублируют начало текста, они были использованы для технической необходимости шифрования и не нарушают смысла исходного сообщения.

Замечание. Для повышения криптостойкости приведенного алгоритма для двух разных компонентов общего секретного ключа целесообразно использовать различные общие точки эллиптической кривой, сгенерированные по рекуррентной формуле (2).

Заключение

Рассмотренный в статье алгоритм двухэтапного шифрования имеет своей особенностью комплексный подход к защите текстовой информации, а именно: алгоритм комбинирует в себе симметричные и асимметричные методы классической криптографии с использованием точек эллиптической кривой, при этом на разных этапах шифрования использует различную математическую структуру шифруемых элементов. Как и любой комбинированный алгоритм шифрования, он сочетает в себе преимущества асимметричных эллиптических криптосистем с производительностью симметричных.

Криптографическая стойкость приведенного алгоритма основывается на трудоемкости решения задачи дискретного логарифмирования на эллиптических кривых (ECDLP) [16] и защищенности сервиса совместного доступа с безопасной аутентификацией взаимодействующих пользователей [17]. Блочная реализация второго этапа шифрования гарантирует стойкость шифра к статистическому анализу текста. Но, как и любая криптографическая система, базирующаяся на ключевом обмене Диффи – Хеллмана, предложенный алгоритм будет уязвим к атакам «человек посередине», так как не обеспечивает аутентификацию взаимодействующих пользователей. Использование данного алгоритма в комплексе с защищенными сервисами совместного доступа, предоставляющими доступ к общим данным только для уполномоченных пользователей через проверку их электронных цифровых подписей, позволяет минимизировать уязвимость к таким атакам и предоставляет надежную защиту конфиденциальности.

Список литературы

1. **Свечников С. Н.** Частотный анализ при использование классических криптоалгоритмов // Инновации в науке и практике. Уфа: Вестник науки, 2020. С. 57–62.
2. **Сергеева О. А., Кутовая А. С.** Основы криптографии: учеб.-метод. пособие [Электронный ресурс]; Кемеров. гос. ун-т. Электрон. дан. (2,26 Мб). Кемерово: КемГУ, 2024. Систем. требования: Intel Pentium (или аналогичный процессор других производителей), 1,2 ГГц; 512 Мб оперативной памяти; видеокарта SVGA, 1280x1024 High Color (32 bit); 2Мб свободного дискового пространства; операционная система Windows XP и выше; Adobe Reader. Загл. с экрана, № госрегистрации: 0322400576, 26.02.2024.
3. **Земор Ж.** Курс криптографии. М.-Ижевск: Регулярная и хаотическая динамика, 2019. 256 с.
4. **Гулевич С. А.** Общие сведения о современной криптографии и подходах к ее изучению // RATIO ET NATURA. 2022. № 2 (6).
5. **Кузнецов А. В., Шишкина, Э. Л.** Методы алгебраической геометрии в криптографии: учеб. пособие. Воронеж: Издательский дом ВГУ, 2023. 125 с.
6. **Жданов О. Н., Чалкин Т. А.** Эллиптические кривые и их применение в криптографии: учеб. пособие. Красноярск: СибГАУ, 2011. 65 с.
7. **Жданов О. Н., Чалкин Т. А.** Эллиптические кривые. Основы теории и криптографические приложения. URSS: Либроком, 2020. 200 с.
8. **Кутовая А. С., Сергеева О. А.** Комбинированный блочный алгоритм шифрования с открытым ключом на базе эллиптической кривой: в сб.: Фундаментальные и прикладные исследования в физике, химии, математике и информатике // Материалы симпозиума XIX (LI) Междунар. науч. конф. студентов, аспирантов и молодых ученых, г. Кемерово, 2024. Кемеров. гос. ун-ет. С. 151–154.
9. **Стрельцова А. С., Ухваркин С. П., Филимонов В. В.** Применение эллиптических кривых в алгоритме Диффи – Хеллмана // Научный альманах. 2019. № 1-3 (51). С. 62–64.
10. **Гущин А. В., Осипов М. Н.** Криптографические методы защиты информации: учеб. пособие. Самара, 2024. Самарский ун-т. 126 с.
11. **Кутовая А. С.** Матричные алгоритмы криптографической защиты информации // Фундаментальные и прикладные исследования в физике, математике и информатике. Кемерово: Кемеровский гос. ун-т, 2022. С. 171–174.
12. **Мунерман В. И.** Реализация алгоритма шифрования Хилла на основе алгебры многомерных матриц // Системы высокой доступности. 2019. Т. 15, № 1. С. 21–27.
13. **Жуков А. Е.** Системы блочного шифрования: учеб. пособие по курсу «Криптографические методы защиты информации». М.: Изд-во МГТУ им. Н.Э. Баумана, 2013. 79 с.
14. **Карпов А. В., Ишмуратов Р. А.** Введение в криптографию: учеб. пособие. Казань: Казан. ун-т, 2024. 128 с.
15. **Молдовян А. А., Молдовян Д. Н., Левина А. Б.** Протоколы аутентификации с нулевым разглашением секрета: учеб. пособие. СПб: Университет ИТМО, 2016. 55 с.
16. **Семаев И. А.** О вычислении логарифмов на эллиптических кривых. Дискрет. матем., 1996. Т. 8, вып. 1. С. 65–71.
17. 13 сервисов для безопасного общего доступа и обмена файлами. URL: https://www.securitylab.ru/blog/company/PandaSecurityRus/351556.php (дата обращения: 23.09.2025).

References

1. **Svechnikov S. N.** Frequency analysis in the application of classical cryptographic algorithms. *Innovations in Science and Practice*. Ufa: LLC "Scientific Publishing Center "Vestnik Nauki"", 2020, p. 57–62.
2. **Sergeeva O. A., Kutovaya A. S.** Fundamentals of cryptography: educational and methodological guide [Electronic resource]; Kemerovo State University. Electronic data (2.26 MB). Kemerovo: KemSU, 2024. System requirements: Intel Pentium (or similar processor from other manufacturers), 1.2 GHz; 512 MB RAM; SVGA video card, 1280x1024 High Color (32 bit); 2 MB free disk space; Windows XP operating system or higher; Adobe Reader. Title from screen, State registration No.: 0322400576, 26.02.2024.
3. **Zémor G.** A course in cryptography. Moscow-Izhevsk: SRC "Regular and Chaotic Dynamics", 2019. 256 p.
4. **Gulevich S. A.** General information about modern cryptography and approaches to its study. *RATIO ET NATURA*. 2022. No. 2 (6).
5. **Kuznetsov A. V., Shishkina E. L.** Methods of algebraic geometry in cryptography: textbook. Voronezh: VSU Publishing House, 2023. 125 p.
6. **Zhdanov O. N., Chalkin V. A.** Elliptic curves and their application in cryptography: textbook. Krasnoyarsk: SibSAU, 2011. 65 p.
7. **Zhdanov O. N., Chalkin T. A.** Elliptic curves. Fundamentals of theory and cryptographic applications. Moscow: URSS: Librokom, 2020. 200 p.
8. **Kutovaya A. S., Sergeeva O. A.** Combined block encryption algorithm with public key based on elliptic curve. *In: Fundamental and Applied Research in Physics, Chemistry, Mathematics and Computer Science. Proceedings of the Symposium of the XIX (LI) International Scientific Conference of Students, Graduate Students and Young Scientists*, Kemerovo, 2024. Kemerovo: Kemerovo State University, 2024, p. 151–154.
9. **Streltsova A. S., Ukhvarkin S. P., Filimonov V. V.** Application of elliptic curves in the Diffie-Hellman algorithm. *Scientific Almanac*. 2019, no. 1-3 (51), p. 62–64.
10. **Gushchin A. V., Osipov M. N.** Cryptographic methods of information protection: textbook / A. V. Gushchin, M. N. Osipov. Samara: Samara University Publishing House, 2024. 126 p.
11. **Kutovaya A. S.** Matrix algorithms for cryptographic information protection. *Fundamental and Applied Research in Physics, Mathematics and Computer Science*. Kemerovo: Kemerovo State University, 2022, p. 171–174.
12. **Munerman V. I.** Implementation of the Hill encryption algorithm based on the algebra of multidimensional matrices. *Highly Available Systems*. 2019, vol. 15, no. 1, pp. 21–27.
13. **Zhukov A. E.** Block cipher systems: textbook for the course "Cryptographic methods of information protection". Moscow: Bauman MSTU Publishing House, 2013. 79 p.
14. **Karpov A. V., Ishmuratov R. A.** Introduction to cryptography: textbook. Kazan: Kazan University, 2024. 128 p.
15. **Moldovyan A. A., Moldovyan D. N., Levina A. B.** Zero-knowledge authentication protocols: textbook. St. Petersburg: ITMO University, 2016. 55 p.
16. **Semaev I. A.** On the computation of logarithms on elliptic curves. *Discrete Mathematics and Applications*. 1996, vol. 8, no. 1, pp. 65–71.
17. 13 services for secure shared access and file URL: https://www.securitylab.ru/blog/company/PandaSecurityRus/351556.php (accessed: 23.09.2025).

Сведения об авторах

Сергеева Ольга Алексеевна, кандидат физико-математических наук, доцент кафедры фундаментальной математики Кемеровского государственного университета

Кутовая Анастасия Сергеевна, магистр, учитель математики и информатики СОШ № 31 им. В. Д. Мартемьянова

Сергеев Владислав Сергеевич, магистрант 2-го курса кафедры фундаментальной математики Кемеровского государственного университета

Information about the Authors

Olga A. Sergeeva, Candidate of Physical and Mathematical Sciences, Associate Professor, Department of Fundamental Mathematics Kemerovo State University, Russia, Kemerovo

Anastasia S. Kutovaya, Master's degree holder, teacher of mathematics and informatics Secondary General Education School No. 31 named after V. D. Martemyanov

Vladislav S. Sergeev, Second-year master's student at the Department of Fundamental Mathematics Kemerovo State University

Статья поступила в редакцию 02.09.2025;
одобрена после рецензирования 15.11.2025; принята к публикации 15.11.2025

The article was submitted 02.09.2025;
approved after reviewing 15.11.2025; accepted for publication 15.11.2025

Правила оформления текста рукописи

Авторы представляют статьи на русском или английском языке объемом от 0,5 авторского листа (20 тыс. знаков) до 1 авторского листа (40 тыс. знаков), включая иллюстрации (1 иллюстрация форматом 190 × 270 мм = 1/6 авторского листа, или 6,7 тыс. знаков). Публикации, превышающие указанный объем, допускаются к рассмотрению только после индивидуального согласования с редакцией журнала.

Текст рукописи должен быть представлен в редколлегию в виде файла MS Word (.doc, .docx). Гарнитура Times New Roman, размер шрифта 11, межстрочный интервал 1, размеры полей – стандартные значения текстового редактора. Форматирование – выравнивание по ширине страницы, переносы слов включены, каждый новый абзац начинается с красной строки. Не допускается ручное форматирование абзацев (пробелами, лишними переводами строк, разрывами страниц).

Структура статьи

- Индекс УДК (универсальной десятичной классификации). Выравнивание по левому краю
- Название статьи. Выравнивание по центру, полужирный шрифт
- ФИО авторов (полностью). Выравнивание по центру, полужирный шрифт
- Места работы всех авторов. Выравнивание по центру, курсив
- Адреса электронной почты, ORCID авторов
- Аннотация статьи
- Ключевые слова, не более 10
- Благодарности, сведения о финансовой поддержке
- Название статьи **на английском языке**. Выравнивание по центру, полужирный шрифт
- ФИО авторов **на английском языке** (полностью). Выравнивание по центру, полужирный шрифт
- Места работы авторов **на английском язык**е. Выравнивание по центру, курсив
- Аннотация статьи **на английском языке (Abstract)**, 200–250 слов
- Ключевые слова **на английском языке (Keywords)**, не более 10
- Благодарности, сведения о финансовой поддержке **на английском язык**е, если есть соответствующий раздел на русском языке **(Acknowledgements)**
- Основной текст
- Список литературы / **References**
- Сведения об авторах

Требования к оформлению основного текста и иллюстративных материалов

Основной текст должен быть представлен в структурированном виде, рекомендуется использовать подзаголовки – например: Введение, Методика..., Выводы, Результаты, Заключение.

Подзаголовки отделяются и набираются полужирным шрифтом. В целях выделения частей текста и отдельных слов и словосочетаний допускается использование курсива или полужирного шрифта. Подчеркивание, разрядка, изменение основного кегля и выделение цветом не используются.

Иллюстрации к рукописи статьи должны быть приложены в виде отдельных файлов. При этом в тексте должно содержаться включенное изображение с указанием имени файла. Все иллюстрации, содержащие схемы, графики, алгоритмы и т. п., должны быть представлены в векторном виде (.ai, .eps, .cdr). Скриншоты и другие растровые изображения должны быть представлены в максимально высоком качестве, без каких-либо потерь и искажений (.jpg, .tif). Все иллюстрации должны иметь подрисуночную подпись – свое название. Надписи к таблицам и подписи к иллюстрациям приводятся **на двух языках (русском и английском)**.

Примеры:

Рис. 1. Диаграмма производительности...
Fig. 1. Performance diagram...

Таблица 1

Сравнение алгоритмов...

Table 1

Comparison of algorithms...

Нумерация последовательная и неразрывная от начала статьи. Не допускается использование других наименований, кроме «Рис.» / «Fig.», «Таблица» / «Table», и усложнение нумерации (например, «Рис. 3.2.»). Ссылка на иллюстрацию в тексте должна быть приведена в круглых скобках, например: (рис. 1), (табл. 1).

Формулы должны быть набраны с использованием редактора MathType либо встроенного редактора формул MS Word. Кегль основных символов – 11, греческие символы набираются прямым шрифтом, латинские – курсивом. Нумеруются только те формулы, на которые автор ссылается в тексте.

Abstract

Аннотация статьи на английском языке (Abstract) не должна быть дословным переводом русскоязычной аннотации. Раздел Abstract, как и основной текст, должен быть структурирован, в нем должно содержаться описание цели работы, методов исследования, научной значимости, выводов / результатов. Требуется качественный перевод на английский язык (при необходимости просим авторов обращаться к профессиональным переводчикам). **Объем Abstract 200–250 слов.**

Список литературы / References

Список литературы и список литературы на английском языке (References) размещаются в общем разделе. Рекомендуемое количество цитируемых в статье источников – не менее 10, в список желательно включать ссылки на актуальные работы по теме исследования, особенно в иностранных периодических изданиях.

В тексте статьи ссылки на литературу указываются цифрами в квадратных скобках, при необходимости указываются номера страниц, например: [2; 3. С. 15].

Список литературы нумеруется в порядке цитирования и оформляется в соответствии с ГОСТ Р 7.0.5-2008 на библиографическое описание (знаки тире в описании опускаются). Ссылки на неопубликованные работы, а также на интернет-ресурсы (кроме электронных изданий, поддающихся библиографическому описанию) оформляются в виде сноски.

В Список литературы ссылки на источники следует включать на оригинальном языке опубликования. Каждый источник должен быть также оформлен на английском языке (References) по международному стандарту для публикаций в области информатики IEEE Style со следующими отличиями:

- инициалы авторов указываются после фамилии;
- название статьи не берется в кавычки, отделяется точкой;

- отсутствует союз «and» перед фамилией последнего автора;
- в диапазоне страниц – удвоенная «p» (например, «pp. 2–9»);
- год издания указывается после места издания (для книг) и сразу после названия журнала (для периодики).
- Перевод источника на английский язык:
- если источник имеет выходные данные на английском языке, то для формирования References **следует использовать именно эти данные**;
- если оригинальная публикация не содержит выходных данных на английском языке, то допускается транслитерация названия материала на латинский алфавит в сочетании с переводом на английский язык в квадратных скобках. В конце описания указывается, на каком языке написана эта работа, например, (in Russ.). При транслитерации можно воспользоваться интернет-ресурсом http://ru.translit.ru/, рекомендуется выбрать стандарт BSI. Место издания не транслитерируется, указывается полностью на английском языке, например: Moscow. Название издательства / издателя, как правило, транслитерируется. Для журналов, у которых есть официальное название на английском языке, – использовать его (проверить на сайте журнала, или, например, в библиотеке WorldCat), если названия на английском языке нет, использовать транслитерацию по системе BSI. Не следует самостоятельно переводить названия журналов.

Если у цитируемого источника есть **цифровой идентификатор DOI** (https://search. crossref. org), его требуется обязательно указывать в конце библиографической ссылки.

Примеры оформления ссылок. Каждый источник в том же пункте дублируется на английском языке (References).

Источник на русском языке, перевод на английский доступен в метаданных статьи

1. **Журавлев С. С., Рудометов С. В., Окольнишников В. В., Шакиров С. Р.** Применение модельно-ориентированного проектирования к созданию АСУ ТП опасных промышленных объектов // Вестник НГУ. Серия: Информационные технологии. 2018. Т. 16, № 4. С. 56–67. DOI 10.25205/1818-7900-2018-16-4-56-67

Zhuravlev S. S., Rudometov S. V., Okolnishnikov V. V., Shakirov S. R. Model-Based Design Approach for Development Process Control Systems of Hazardous Industrial Facilities. *Vestnik NSU. Series: Information Technologies*, 2018, vol. 16, no. 4, pp. 56–67. (in Russ.) DOI 10.25205/ 1818-7900-2018-16-4-56-67

Источник на английском языке. Оформляем согласно требованиям для References. Приводим только 1 раз.

2. **Telnov V. I.** Optimization of the Beam Crossing Angle at the ILC for E + e- and yy Collisions. *Journal of Instrumentation*, 2018, vol. 13, no. 03, pp. P03020–P03020. DOI 10.1088/1748-0221/13/03/ p03020

Метаданные источника доступны только на русском языке

3. **Жижимов О. Л., Федотов А. М., Шокин Ю. И.** Технологическая платформа массовой интеграции гетерогенных данных // Вестник НГУ. Серия: Информационные технологии. 2013. Т. 11, вып. 1. С. 24–41.

Zhizhimov O. L., Fedotov A. M., Shokin Yu. I. Tekhnologicheskaya platforma massovoi integratsii geterogennykh dannykh [Technology Platform for the Mass Integration of Heterogeneous Data]. *Vestnik NSU. Series: Information Technologies*, 2013, vol. 11, no. 1, pp. 24–41. (in Russ.)

Сведения об авторах

Последний раздел статьи – информация об авторе / авторах **на русском и английском языках**:

- ФИО полностью, ученая степень, ученое звание;
- идентификаторы автора, такие как ResearcherID (всем авторам рекомендуется использовать данные сервисы для ведения актуального списка своих публикаций);
- контактный телефон (не публикуется).

Если статья представляется на английском языке, необходимо приложить перевод на русский язык названия, аннотации, ключевых слов, сведений об авторе.

Доставка материалов

Материалы предоставляются в редакцию по электронной почте inftech@vestnik.nsu.ru.

Порядок рецензирования

Все статьи сначала проходят проверку на заимствование и только после этого отправляются на рецензирование. Редакционный совет не допускает к публикации материал, если имеется достаточно оснований полагать, что он является плагиатом.

Тип рецензирования статей – двухуровневое, одностороннее анонимное («слепое»).

Для каждой статьи редколлегией выбираются рецензенты, научная деятельность которых связана с темой представленного материала. Ответственный секретарь журнала обращается к ним с просьбой дать экспертную оценку статье либо помочь организовать рецензирование.

Рецензии для журнала «Вестник НГУ. Серия: Информационные технологии» составляются по единой схеме и подразумевают оценку по следующим критериям: соответствие тематике журнала, оригинальность и значимость результатов, качество изложения материала.

Заполненный бланк рецензии высылается на электронный адрес редакции. В зависимости от экспертных заключений статья может быть принята редакционным советом к опубликованию, рекомендована автору к доработке (с последующим повторным рецензированием либо без него) или отклонена (с предоставлением автору мотивированного отказа). Автору на электронный адрес высылается текст рецензии без указания ФИО рецензента и его контактных данных.

Все рецензии хранятся в редакции журнала не менее 5 лет. Редколлегия журнала обязуется при поступлении соответствующего запроса направлять копии рецензий в Министерство науки и высшего образования Российской Федерации.