15. *Erokhin V.* Poisk vredonosnykh stsenariev powershell s ispol'zovaniem sintaksicheskikh derev'ev [Searching for malicious powershell scripts using syntax trees], *Bezopasnost' informatsionnykh tekhnologiy* [Information Technology Security], 30 (3), pp. 77-89. DOI: http://dx.doi.org/10.26583/bit.2023.3.05.
16. Follina Exploit Leads to Domain Compromise. Available at: https://thedfirreport.com/2022/10/31/follina-exploit-leads-to-domain-compromise/.
17. *Salitin M.A., Zolait A.H.* The role of User Entity Behavior Analytics to detect network attacks in real time, *2018 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), Sakhier, Bahrain, 2018*, pp. 1-5. DOI: 10.1109/3ICT.2018.8855782.
18. *Hutchins E.M., Cloppert M.J., Amin R.M.* Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains, *Lockheed Martin Corporation.* Available at: https://www.lockheedmartin.com/content/dam/lockheed-martin/rms/documents/cyber/LM-White-Paper-Intel-Driven-Defense.pdf.
19. MITRE ATT&CK. Available at: https://attack.mitre.org/.
20. Bank dannykh ugroz bezopasnosti informatsii [Data bank of information security threats], Federal'naya sluzhba po tekhnicheskomu i eksportnomu kontrolyu, Gosudarstvennyy nauchno-issledovatel'skiy ispytatel'nyy institut problem tekhnicheskoy zashchity informatsii [Federal Service for Technical and Export Control, State Research and Testing Institute of Problems of Technical Protection of Information]. Available at: https://bdu.fstec.ru/.

Статью рекомендовал к опубликованию д.т.н., профессор А.В. Боженюк.

Олейникова Анна Алексеевна – ООО «Интеллектуальная безопасность»; e-mail: ana.oleynikova@gmail.com, г. Москва, Россия; тел.: 83912227639.

Золотарев Вячеслав Владимирович – Сибирский государственный университет науки и технологий; e-mail: amida.2@yandex.ru; г. Красноярск, Россия; тел.: 83912227639; к.т.н.; доцент.

Oleynikova Anna Alekseevna – Intellectual Security LLC; e-mail: ana.oleynikova@gmail.com; Moscow, Russia; phone: +73912227639.

Zolotarev Vyacheslav Vladimirovich – Siberian State University of Science and Technology; e-mail: amida.2@yandex.ru; Krasnoyarsk, Russia; phone: +73912227639; cand. of eng. sc.; associate professor.

УДК 004.89 DOI 10.18522/2311-3103-2023-5-81-92

В.С. Усатюк, С.И. Егоров, А.П. Локтионов, Е.А. Титенко, И.Е. Чернецкая

АРХИТЕКТУРА НЕЙРОННЫХ СЕТЕЙ НА ОСНОВЕ КОДОВ НА ГРАФАХ

Одним из важных достижений теории помехоустойчивого кодирования является открытие кодов на графах и их важного подмножества низкоплотностных кодов (LDPC-кодов). Используя проверочную матрицу кода на графе, можно получить марковское случайное поле. LDPC-код может быть вложен в модель Изинга (разновидность марковского случайного поля) путем использования топологии тора с отрицательной кривизной. При этом кодовые слова соответствуют седловым точкам (экстремумам) в модели, а треппин-сеты соответствуют локальным минимумам. Использование LDPC-кодов с увеличенным кодовым расстоянием позволяет максимально разнести седловые точки, и таким образом повысить устойчивость нейронной сети к шуму и мощность представления. При этом блочная и разряженная структура, характерная для тора отрицательной кривизны, упрощает мультиплексирование и снижает число обучаемых параметров нейронной сети. Целью исследования являются снижение вычислительной сложности и увеличение точности нейронных сетей за счёт применения априорных структурных (квазициклических) разряженных графов для широкого класса задач машинного обучения на марковских случайных полях. В работе представлен новый подход, позволяющий осуществ-

лять синтез архитектур нейронных сетей на основе кодов на графах. Предложенный подход осуществляет эффективное представление марковских случайных полей за счёт применения разряженных блочных (квазициклических) матриц (тензоров). Предложенный подход позволяет снизить число обучаемых параметров и логарифмически снизить сложность мультиплексирования тензора. Полученная на основе предложенного подхода архитектура трансформера в задаче поиска пути (pathfinder) с конкурса трансформеров (long range arena) заняла пятое место по точности классификации изображений 94.95% (1.72% от первого места) при значительно меньшей сложности (число параметров (умножений) синтезированной сети меньше в более чем 5 раз). Применение предложенного подхода к задачам факторизации на плотных графах, сетевых задачах, поверхностных сетках, ковариационных матрицах позволило увеличить точность реконструкции по метрике Фробениуса (на отдельных задачах на 8 порядков) в сочетание с упрощением структуры мультиплексора в сравнение с методами усеченного сингулярного разложения TSVD и хордовой разряженной факторизации.

Нейронные сети; коды на графах; низкоплотностные коды; модель Изинга; матричная факторизация; вложение многообразия; торическая гиперболическая топология.

V.S. Usatyuk, S.I. Egorov, A.P. Loktionov, E.A. Titenko, I.E. Chernetskaya

NEURAL NETWORK ARCHITECTURE BASED ON GRAPH CODES

One of the important achievements of the theory of error-correcting coding is the discovery of graph codes and their important subset - low-density parity check codes (LDPC codes). Using the parity check matrix of the code on the graph, one can obtain a Markov random field. LDPC code can be embedded in an Ising model (a type of Markov random field) by using a torus topology with negative curvature. In this case, codewords correspond to saddle points (extrema) in the model, and trappin sets correspond to local minima. The use of LDPC codes with an increased code distance allows for maximum separation of saddle points, and thus increases the noise resistance of the neural network and the representation power. At the same time, the block and sparse structure, characteristic of a torus of negative curvature, simplifies multiplexing and reduces the number of trainable parameters of the neural network. The aim of the research is to reduce the computational complexity and increase the accuracy of neural networks through the use of a priori structural (quasi-cyclic) sparse graphs for a wide class of machine learning problems on Markov random fields. The paper presents a new approach that allows the synthesis of neural network architectures based on graph codes. The proposed approach provides an effective representation of Markov random fields through the use of QC-LDPC matrices and tensors. The proposed approach allows us to reduce the number of trainable parameters and logarithmically reduce the complexity of tensor multiplexing. The proposed approach provided an accuracy of 94.95% (1.72% to first place) of the binary classification problem "Pathfinder" of the "Long Range Arena" competition, with more than 5 times fewer parameters (multiplications). Application of the proposed approach to factorization problems on dense graphs, network problems, surface meshes, covariance matrices made it possible to increase the accuracy of reconstruction using the Frobenius metric in individual problems by more than 8 orders of magnitude in combination with simplifying the structure of the multiplexer.

Neural networks; codes on graph; LDPC codes; Ising model; matrix factorization; manifold embedding; toric hyperbolic topology.

Введение. Технологии искусственного интеллекта (ИИ) приобретают все большее значение в современном мире. Наибольший прогресс в области ИИ обеспечило применение нейронных сетей с глубоким обучением, которые в ряде приложений уже могут заменить человека. Для расширения области применения этих сетей необходимо уменьшить их сложность. В статье предлагается подход к построению нейронных сетей уменьшенной сложности на основе применения кодов на графах.

1. Коды на графах. Коды на графах были введены Таннером [1], который доказал оптимальность алгоритмов исправления ошибок методом распространения доверия (BP, belief-propagation) для кодов, определенных на графах без циклов. В классе кодов на графах особый практический интерес в силу наименьшей сложности декодирования имеют низкоплотностные коды.

Низкоплотностный (LDPC, Low Density Parity Check) код– это блочный линейный код размерностью k и длиной кодового слова n, задаваемый проверочной матрицей H размерностью $(n\text{-}k)\cdot n$, имеющей небольшую плотность отличных от нуля символов [2]. По определению проверочной матрицы для любого кодового слова v LDPC-кода справедливо следующее: $v\cdot H^T = 0$. Каждая строка проверочной матрицы H задает уравнение проверки на четность:

$$\sum_{l=0}^{n-1} v_l \cdot h_{j,l} = 0, \quad (1)$$

где $h_{j,l}$ – элемент проверочной матрицы, j – номер строки проверочной матрицы (номер проверочного уравнения), l – номер символа кодового слова.

Достоинством низкоплотностных кодов является возможность применения субоптимального алгоритма декодирования с мягкими решениями методом распространения доверия (BP), обладающего значительно большей помехоустойчивостью по сравнению с алгоритмами декодирования с жесткими решениями, сложность которого растет линейно относительно длины кода.

Алгоритм BP предусматривает представление LDPC-кода в виде двудольного графа Таннера (пример графа Таннера приведен на рис. 1).

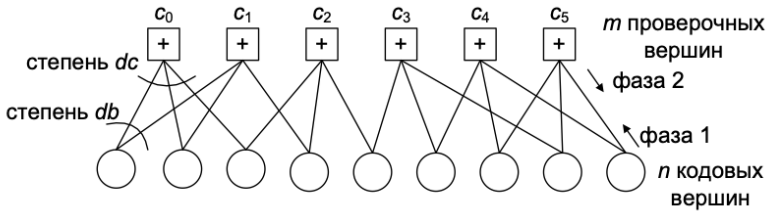

Рис. 1. Двудольный граф двоичного регулярного LDPC кода длины 9

Граф Таннера G имеет два множества вершин. Одно состоит из $m=n\text{-}k$ проверочных вершин $\{c_0, c_1, \dots, c_{m-1}\}$, соответствующих m строкам матрицы H**,** второе – из n кодовых вершин $\{v_0, v_1, \dots, v_{n-1}\}$, соответствующих n столбцам матрицы H. Кодовая вершина v_l соединяется ребром с проверочной вершиной c_i в том случае, если $h_{j,l} \neq 0$.

В соответствии с итеративным алгоритмом декодирования BP получение верных значений бит кодового слова осуществляется в результате многократного обмена сообщениями вершинами графа Таннера. Каждая итерация алгоритма содержит две фазы. В фазе 1 обновляются сообщения проверочных вершин на основе анализа сообщений кодовых вершин; в фазе 2 – сообщения кодовых вершин на основе анализа сообщений проверочных вершин.

На эффективность BP-декодирования отрицательно влияет наличие циклов в графе Таннера, образующих треппин-сеты (Trapping set, TS,) или (a,b)-подграфы (подграфы в графе Таннера, состоящие из a символьных узлов, b из которых инцидентны проверочным узлам с нечетнымий степенями). Эти подграфы обуславливают ошибку BP-декодирования. В случае если сообщения проверочных узлов изменит значения символьных узлов, инцидентных нечетному числу проверок, то, вследствие неправильного подсчета условных вероятностей, обусловлен-

ного циклами графа Таннера, на символьных узлах подграфа ошибка не будет скорректирована, даже если она является корректируемой в соответствие с дистантными свойствами кода.

Важной характеристикой LDPC-кода является приближенный **с**пектр связности (ACE Spectrum) графа Таннера. Спектр связности представляется в виде вектора:

$$ACE(H) = (ACE_{4,0}(VN), ACE_{4,1}(VN), ..., ACE_{i,j}(VN), ..., ACE_{k,m}(VN)),$$

где $ACE_{i,j}(VN)$- количество символьных узлов, содержащихся в цикле длины i со значением связности j, VN-подмножество символьных узлов графа Таннера, образующих циклы длины $i \in \{4,6,...\}$. Значение связности (ACE, Approximated Cycle Extrinsic Message Degree) вычисляется для символьных узлов, содержащихся в подграфе образованном циклом $v_i \in C$, $ACE(C) = \sum_{v \in C}(d(v) - 2)$, где $d(v)$- степень инцидентности символьного узла. Таким образом, код с лучшим спектром связности при одинаковом спектре кода при использовании декодирования методом распространения доверия обеспечивает лучшую помехоустойчивость. Этому коду не будут мешать псевдокодовые слова, обусловленные TS.

На практике для уменьшения сложности декодеров LDPC-кодов используются квазициклические коды. Квазициклический регулярный (*J*, *L*) LDPC-код (*J* – вес столбца (число единиц в столбце) матрицы, *L* – вес строки) задается проверочной матрицей [3]:

$$H = \begin{bmatrix} I(p_{0,0}) & I(p_{0,1}) & ... & I(p_{0,L-1}) \\ I(p_{1,0}) & I(p_{1,1}) & & I(p_{1,L-1}) \\ \vdots & \vdots & \ddots & \vdots \\ I(p_{J-1,0}) & I(p_{J-1,1}) & ... & I(p_{J-1,L-1}) \end{bmatrix}, \quad (2)$$

где $0 \le j \le J-1$, $0 \le l \le L-1$ и $I(p_{j,l})$ – подматрица перестановки размера $z \cdot z$ (циркулянт - единичная матрица, циклически сдвинутая вправо на $p_{j,l}$ символов). Проверочная матрица регулярного кода также может содержать нулевые подматрицы размера $z \cdot z$. Если веса строк (столбцов) проверочной матрицы LDPC-кода принимают различные значения, то такой код называют нерегулярным.

2. Марковские случайные поля, описываемые кодами на графах. Марковское случайное поле (МСП) это – графовая модель, в которой множество случайных величин обладает Марковским свойством, описанным неориентированным графом. МСП широко применяются в статистической физике и обработке изображений [4, 5]. Например, МСП применяются для сглаживания, сегментации, восстановления, регистрации изображений, синтеза текстур, повышения разрешения, согласования изображений в стереопарах, аннотации, извлечения информации и решения других задач. Физическим прототипом МСП является модель Изинга намагничивания материала в статистической физике. Рассмотрим гамильтонову модель Изинга Эдварда-Андерсона, [2]:

$$H_{EA} = -\sum_{i=1,...,n} \sum_{a=1,...,m} C_{ij} J_{ij} \sigma_i \sigma_j, \quad (3)$$

где C_{ij} – элемент матрицы связности, равный 1, если два спина взаимодействуют, или 0 в противном случае, J_{ij} – вес взаимодействия между *i*-ым и *j*-ым спинами , σ_i – спин, n – количество столбцов, m – количество строк. Величины J_{ij} определяют силу двухспинового взаимодействия и обычно рассматриваются как независимые случайные величины с известным распределением вероятностей, например с J_0 средним значением, ΔJ^2 дисперсией. Тогда гамильтониан H_{EA}, можно переписать в не локальную (infinite range) модель:

$$H_{ECC} = -\sum_p \sum_{i_{1,\dots,i_p}=1,\dots,n} C^{(p)}_{i_{1,\dots,i_p}} J_{i_{1,\dots,i_p}} \sigma_{i_1} \dots \sigma_{i_p}, \quad (4)$$

и $J_0^{(p)}$ и $\Delta\ J^2_{(p)}$ – математическое ожидание и дисперсия величины J_{ij}.

Рассмотрим модель Изинга, используя код на графе с заданной проверочной матрицей H (1). Проверочная матрица LDPC-кода задает матрицу C. Тогда n будет длиной кода, m – количеством проверочных уравнений, n бит кодового слова a_i определят значения спинов $\sigma_i = 2a_i - 1$. Закодированное сообщение соответствует элементу матрицы $J^0_{i_{1,\dots,i_p}} = \sigma_{i_1} \dots \sigma_{i_p}$. Декодирование кодового слова соответствует нахождению энергетического минимума гамильтониана (4). При этом Js – выход канала (парные спиновые взаимодействия с моментами $J_0^{(p)}$ и $\Delta\ J^2_{(p)}$) и J^0 – шум в канале.

Для каждого p в (4) номер координаты $z_i = \sum_{j_{2,\dots,j_p}} C_{i,j_{2,\dots,j_p}} = z$ не зависит от i и скорость кода равна $R = p/z$. При $p \to \infty$ получается модель случайной энергии, при $p = 2$ получаются спиновые стекла Шеррингтона-Киркпатрика, в которых каждый спин взаимодействует со всеми остальными спинами. Для уменьшения сложности оценки основного состояния в BP можно предположить, что каждая кодовая вершина имеет дерево «окрестностей». Такая модель короткодействующей корреляции в спиновом стекле используется в методе нарушения симметрии реплик [2] (replica symmetry breaking method).

В работе [6] было доказано, что LDPC-код может быть вложен в модель Изинга путем использования топологии тора с отрицательной кривизной (гиперболической). При этом кодовые слова (треппин-сеты TS(a,0)) соответствуют седловым точкам (экстремумам) в модели, а треппин-сеты (TS(a,b), b ≠ 0) соответствуют локальным минимумам. Использование LDPC-кодов с увеличенным кодовым расстоянием позволяет максимально разнести седловые точки, и таким образом повысить устойчивость нейронной сети к шуму и мощность представления (способностью нейронной сети присваивать правильные метки конкретному экземпляру и создавать четко определенные точные границы принятия решений для этого класса). При этом блочная и разряженная структура, характерная для тора отрицательной кривизны, упрощает мультиплексирование и снижает число обучаемых параметров (умножений на веса нейронной сети).

3. Приложение предложенного подхода для построения архитектуры глубокой нейронной сети (Трансформера) в задачах обработки изображений. Предложенный подход применялся для синтеза глубокой нейронной сети архитектуры «Трансформер» с 15 слоями Каждый из слоев описывается квазициклической проверочной матрицей размера 1024×1024, содержащей 3072 ненулевых позиций, представленной на рис. 2 (слева). Мультиплексирование осуществляется сдвиговым регистром размера 64. Глубокая нейронная сеть, состоящая из 15 слоев, представляет собой иерархическую систему с топологией вложенных торов (до 15 торов), которые могут совпадать, частично совпадать, либо различаться (рис. 2, справа), [6].

Применительно к задаче “Pathfinder” конкурса “Long Range Arena” [7] синтезированная глубокая сеть обеспечила точность бинарной классификации 94.95%, уступая только 1,72% сети, занявшей на конкурсе первое место, при значительно меньшей сложности. Число параметров (умножений) синтезированной сети меньше в более чем 5 раз.

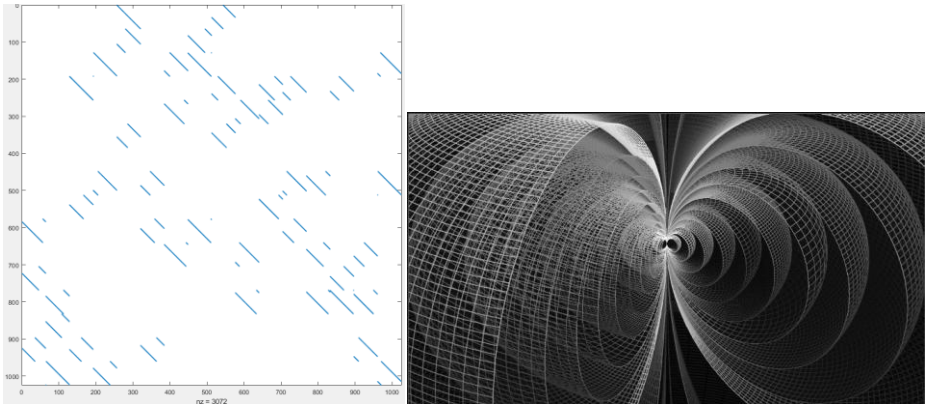

Рис. 2. Квазициклическая проверочная матрица (слева) и топология синтезированной нейронной сети (15 не пересекающихся вложенных торов (справа))

4. Приложение предложенного подхода в задачах факторизации. Факторизация матрицы представляет ее аппроксимацию путем разложения на произведение двух или более матриц. Обычно разложение выполняется на произведение матриц меньшей размерности или на произведение разряженных матриц.

Точность матричной аппроксимации (ее ошибку) можно оценить с помощью квадрата нормы Фробениуса. Квадрат нормы Фробениуса является общей метрикой, используемой для оценки разницы между двумя матрицами и определяется как сумма квадратов разностей между соответствующими элементами двух матриц [8].

$$D\left(X \parallel \hat{X}\right) = \left\|X - \hat{X}\right\|_F^2 = \sum_{ij} \left(X_{ij} - \hat{X}_{ij}\right)^2, \quad (5)$$

где X – исходная матрица и $\hat{X}$ – ее аппроксимация.

Путем минимизации квадрата нормы Фробениуса между исходной матрицей и ее аппроксимацией можно оптимизировать факторизацию для достижения наилучшего возможного приближения исходной матрицы.

Факторизация матрицы малого ранга размерности $N \times N$ предполагает ее аппроксимацию путем разложения на произведение двух или более матриц меньшей размерности [9]: $\approx \hat{X} = WH$ и $X \approx WSH$, где $W \in \mathrm{R}^{N \times r}$, $S \in \mathrm{R}^{r \times r}$, $H \in \mathrm{R}^{r \times N}$, r – ранг аппроксимации и $r \ll N$.

В работах [9, 10] показано, что минимизация квадрата нормы Фробениуса $\|X - WH\|_F^2$ относительно W и H может быть выполнена с помощью усеченного разложения по сингулярным значениям (TSVD). TSVD метод предполагает выбор диагональной матрицы Λ с r наибольшими сингулярными значениями и использование соответствующих левых и правых сингулярных векторов в качестве столбцов для формирования матриц U и V, соответственно. Тогда минимизация достигается путем выбора $W = U$ и $H = \Lambda V^T$. Масштабирующий множитель можно перенести из матрицы W в матрицу H.

Пример факторизации по методу TSVD приведен на рис. 3.

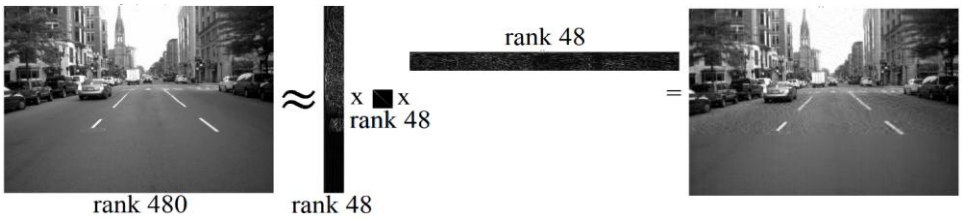

Рис. 3. Пример факторизации матрицы низкого ранга с использованием TSVD (функции svdsketch в Matlab)

Целью малоранговой факторизации является выявление основных закономерностей и структур в больших наборах данных за счет уменьшения сложности исходной матрицы. Матричную факторизацию можно рассматривать как частный случай вложения в модель МСП (спиновых стекол) проекций низкой размерности [11]. При этом матрица раскладывается на две или более матрицы, которые представляют собой скрытые факторы или признаки. Эти скрытые факторы можно рассматривать как малоранговое вложение, которое отражает основные отношения и структуру данных. Представляя данные в виде вложения, методы матричной факторизации позволяют, среди прочего, уменьшать размерность, извлекать шаблоны и применяться в рекомендательных системах. Таким образом, факторизацию матрицы можно рассматривать как особую форму кодов вложения модели Изинга на основе кодов на графах, где матрица раскладывается на представления меньшей размерности.

Альтернативой малоранговой факторизации является факторизация матрицы путем разложения ее на произведение разряженных матриц той же размерности (SF, sparse factorization).

Для разряженной факторизации SF решается следующая задача оптимизации:

$$\min_{W^{(1)},\dots,W^{(M)}} \left\| X - \prod_{m=1}^{M} W^{(M)} \right\|_F^2, \tag{6}$$

где $W^{(M)}$ – M разреженных квадратных матриц с небольшим числом ненулевыхи позиций. Ищутся разреженные матрицы с минимальным числом ненулевых элементов, обеспечивающие минимум квадрата нормы Фробениуса.

Пример разряженной факторизации SF Chord представлен в статье [12]. В работе для построения шаблонов разреженных матриц использовался метод хорд, основанный на протоколе P2P Chord [13]. Значения ошибки аппроксимации этого метода в сравнении с TSVD для разных типов квадратных матриц представлены в таблице.

Наш подход, использующий коды на графах, позволяет решить задачу разреженной факторизации более эффективно. В соответствии с этим подходом шаблоны разряженных матриц строятся как матрицы проверок на четность LDPC-кодов, оптимизированных с использованием PEG+ACE методов [14, 15] и методов улучшение спектра связности и кодового расстояния, предложенных в работах [16–19]. Этот метод факторизации назовем LDPC-методом.

Оптимизационная задача (6) при наличии шаблона в виде проверочной матрицы может быть решена любым методом неограниченной гладкой оптимизации. Нами использовался метод Бройдена-Флетчера-Гольдфарба-Шанно (BFGS). Минимизация ошибки аппроксимации дает в результате факторизующие матрицы $W^{(1)}, \dots W^{(M)}$.

5. Экспериментальное исследование метода факторизации с использованием LDPC-кодов. Экспериментальное исследование предложенного метода факторизации с использованием LDPC-кодов проводилось в системе Matlab.

В рамках исследования строились шаблоны разреженных матриц $W^{(M)}$ на основе проверочной матрицы LDPC-кода. Затем решалась оптимизационная задача (6), в результате которой получались сами факторизующие матрицы и значение ошибки аппроксимации.

Полученные результаты сравнивались по точности аппроксимации с результатами, полученными для методов TSVD и SF Chord [12]. Сравнение выполнялось для одинаковых квадратных матриц из [12]. Одинаковая вычислительная сложность факторизации обеспечивалась следующим выбором параметров: для TSVD $r = \lceil (\log_2 N)^2/2 \rceil$, $q = 2Nr + r$; для SF и LDPC $q = N(\log_2 N)^2$ (q - общее количество ненулевых элементов для шаблонов).

LDPC шаблоны строились с использованием PEG-метода с оптимизацией спектра связности (ACE), имитацией отжига с оптимизацией степени внешнего сообщения точного цикла (EMD) [16–19] для QC-LDPC кода.

Для решения задачи (6) использовался fminunc-оптимизатор Matlab. Ненулевые элементы факторизующих матриц были инициализированы случайными числами в диапазоне $[K^{-1}, K^{-1} + 10^{-2}], K = log_2 N$.

fminunc-оптимизатор Matlab выполнялся на рабочей станции с процессором AMD Ryzen 9 3950X @ 4,0 ГГц и 128 ГБ DDR4 2400 МГц (двухканальный Kingston SK Hynix) системной памяти. Задача непараметрической оптимизации ковариации «Mfeat» с использованием BFGS Matlab, для которой требуется 137 ГБ ОЗУ, решалась на сервере с 2-мя процессорами Intel Xeon E5-2696 V4 с 256 ГБ DDR4 2133 МГц (8 каналов Cisco SK Hynix). Исходные коды использованного программного обеспечения доступны в репозитории github [20].

В качестве аппроксимированных матриц использовались изображения в оттенках серого размером 256×256. Рис. 4 показывает шесть квадратных матриц типовых изображений, а также результирующие ошибки аппроксимации методов TSVD, SF Chord и LDPC под каждым изображением. Оценивая точность различных методов на различных типах наборов данных, можно получить представление о наиболее эффективных подходах к разреженной факторизации больших квадратных матриц.

Шахматный образ близок к низкоранговому, поскольку представляет собой черно-белую шахматную доску, имеющую всего два цвета. Поэтому TSVD работает лучше всех для этого изображения. Однако для других изображений, содержащих богатые высокочастотные детали, такие как линии и углы, TSVD не так хорош, как LDPC. Чтобы убедиться в этом, были вычислены величины градиента изображений (показано на рис. 4 справа). Матрицы градиентов, интерпретируемые как изображения, имеют нули в областях, соответствующих областям исходного изображения с постоянной интенсивностью, Ненулевые значения матриц градиентов представляют в основном высокочастотные детали исходных изображений.

Из рис. 4 видно, что TSVD дает большую ошибку аппроксимации для всех матриц градиентов, чем методы разряженной аппроксимации. Для 9 из 12 матриц изображений лучшую точность обеспечивает наш метод в сравнении с SF Chord и TSVD.

Ошибки аппроксимации по квадрату нормы Фробениуса с использованием TSVD и SF Chord [12], LDPC PEG+ACE [14,15], QC-LDPC SA+EMD [16–19] приведены в табл. 1.

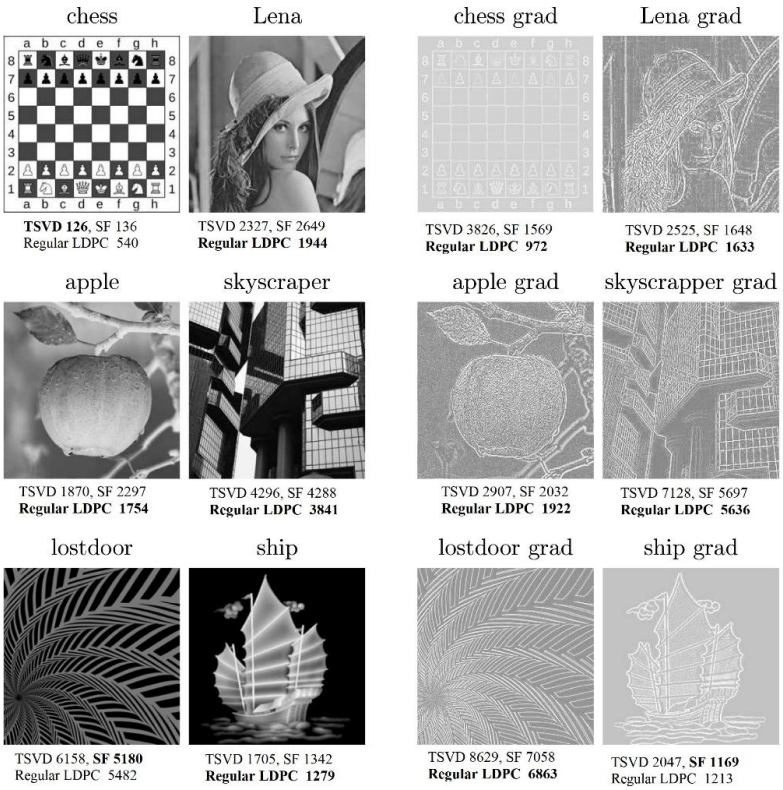

Рис. 4. Примеры квадратных матриц: исходные изображения (слева) и изображения величины градиента (справа) (отображаются после выравнивания гистограммы для лучшей наглядности)

Таблица 1

Величина ошибки аппроксимации в метрике Фробениуса различных методов факторизации квадратных матриц [6], Приложение А] (TSVD, SF Chord, LDPC)

Тип данных квадратной матрицы	Название матрицы	TSVD	SF Chord	LDPC PEG+ACE	QC-LDPC SA+EMD (MET)
Плотный граф	AuraSonar	8.54E+00	8.68E+00	**7.01E+00**	-
Плотный граф	Protein	1.17E+01	1.09E+01	**0.74E+01**	-
Плотный граф	Yeast	3.72E+01	3.61E+01	**3.15E+01**	-
Плотный граф	Voting	**8.07E-04**	1.71E+01	1.47E-02	2.1E-02
Сеть	Sawmill	3.24E+00	1.03E+00	**0.26E+00**	-
Сеть	Scotland	5.90E+00	3.76E+00	3.99E+00	**2.42E+00**
Сеть	A99 m	1.47E+01	1.01E+01	1.04E+01	**0.99E+01**
Сеть	Mexican Power	3.85E+00	1.71E+00	**0.51E+00**	-
Сеть	Strike	2.73E+00	1.04E+00	**0.20E+00**	-
Сеть	Webkb Cornell	6.98E+00	4.80E+00	5.48E+00	**4.36E+00**
Сеть	Worldtrade	8.65E+04	4.47E+04	**3.54E+04**	-
Поверхн. сетка	Mesh1e1	1.87E+01	9.82E+00	**8.45E+00**	-
Поверхн. сетка	Mesh2e1	2.48E+02	3.47E+02	**1.86E+02**	-
Поверхн. сетка	OrbitRaising	9.37E+01	8.53E+01	**7.83E+01**	-

Окончание табл. 1

Поверхн. сетка	Shuttle Entry	2.73E+03	1.86E+03	**1.85E+03**	-
Поверхн. сетка	AntiAngiogenesis	5.85E+01	3.29E+01	4.36E+01	**2.03E+01**
Ковар. матрица	Phoneme	2.80E+01	5.27E+01	**2.34E+01**	-
Ковар. матрица	MiniBooNE	**1.04E+00**	6.36E+03	4.81E+05	1.90E+03
Ковар. матрица	Covertype	8.22E-02	1.90E-02	**1.50E-02**	-
Ковар. матрица	Mfeat	1.11E+03	4.01E+05	3.36E+03	**0.72E+03**
Ковар. матрица	OptDigits	3.28E+01	7.01E+01	**1.23E+01**	-
Ковар. матрица	PenDigits	4.00E+02	1.87E+02	**6.00E-07**	-
Ковар. матрица	Acoustic	1.36E-02	1.11E-02	5.00E-03	**4.55E-03**
Ковар. матрица	IJCNN	5.24E-02	3.03E-02	**5.10E-03**	-
Ковар. матрица	Spam Ham	1.07E-01	4.97E-02	**4.70E-02**	-
Ковар. матрица	TIMIT	9.64E+01	1.56E+02	**8.85E+01**	-
Ковар. матрица	Votes	4.00E-01	1.70E-01	**2.12E-05**	-

Помимо квадратных матриц изображений было выполнено сравнение применения методов TSVD, SF Chord и LDPC для других типов квадратных матриц. В табл. 1 показаны результаты сравнения. Типы данных включают матрицы плотных графов (“Плотный граф” в таблице), матрицу разреженных сетей (“Сеть”), матрицу поверхностной сетки по трехмерным объектам (“Поверх. сетка”) и матрицу ковариации векторных данных (“Ковар. матрица”).

Типы данных описаны в приложение А статьи [6].

Ниже приведены проверочные матрицы H для задач «Антиангиогенез» (циркулянт z = 205) (7) и "Webkb Cornell" (циркулянт $z = 65$) (8), созданные методом QC-LDPC SA+EMD, [19]:

$$H_1 = (I^0 + I^2 + I^3 + I^{65} + I^{70} + I^{85} + I^{97} + I^{154}), \quad (7)$$

$$H_2 = \begin{pmatrix} I^1 + I^{26} + I^{50} & I^2 + I^{19} + I^{49} & I^5 + I^{13} + I^{42} \\ I^5 + I^{58} & I^5 + I^{60} & I^4 + I^{60} \\ I^4 + I^{18} + I^{48} & I^{23} + I^{28} + I^{61} & I^1 + I^4 + I^{53} \end{pmatrix}. \quad (8)$$

Из табл. 1 видно, что точность реконструкции в метрике Фробениуса метода LDPC выше почти для всех матриц (на отдельных задачах на 8 порядков) в сравнение с методами усеченного сингулярного разложения TSVD и хордовой разряженной факторизации SF Cord. Отметим, что использование квазициклических LDPC -кодов упрощает мультиплексирование.

Заключение. В работе представлен новый подход, позволяющий осуществлять синтез архитектур нейронных сетей на основе кодов на графах. Предложенный подход позволил синтезировать глубокую нейронную сеть (Transformer), обеспечивающую точность бинарной классификации 94.95% (1,72% до первого места) для задачи “Pathfinder” конкурса “Long Range Arena”, при более, чем 5-кратно меньшем числе параметров (умножений). Применение предложенного подхода к задачам факторизации на плотных графах, сетевых задачах, поверхностных сетках, ковариационных матрицах позволило увеличить точность реконструкции по метрике Фробениуса в отдельных задачах на более чем 8 порядков.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Tanner R.* A recursive approach to low complexity codes // IEEE Transactions on Information Theory. – 1981. – Vol. 27, No. 5. – P. 533-547.
2. *Mézard M., Montanari A.* Information, Physics, and Computation. – Oxford Graduate Texts, 2009. – 569 p.
3. *Tanner R., et al.* LDPC block and convolutional codes based on circulant matrices // IEEE Transactions on Information Theory. – 2004. – Vol. 50, No. 12. – P. 2966-2984.
4. *Сукар Л.Э.* Вероятностные графовые модели. Принципы и приложения. – М.: ДМК Пресс, 2021. – 338 с.

5. *Stan Z. Li*. Markov Random Field Modeling in Image Analysis. – Springer London, 2009. – 362 p.
6. *Usatyuk V., Sapozhnikov D., Egorov S.* Spherical and Hyperbolic Toric Topology-Based Codes on Graph Embedding for Ising MRF Models: Classical and Quantum Topology Machine Learning. – 2023. – 71 p. – https://arxiv.org/abs/2307.15778.
7. *Yi T., et al.* Long Range Arena: A Benchmark for Efficient Transformers. – ICLR, 2021. – 15 p.
8. *Коэн М.И.* Прикладная линейная алгебра для исследователей данных. – М.: ДМК Пресс, 2023. – 328 с.
9. *Eckart C., Young G.* The Approximation of One Matrix by Another of Lower Rank // Psychometrika. – 1936. – Vol. 1, No. 3. – P. 211-218.
10. *Fomin F. V., et al.* Approximation schemes for low-rank binary matrix approximation problems // ACM Transactions on Algor. – 2019. – Vol. 16, No. 1. – P. 12:1-12:39.
11. *Camilli F., Mezard M.* Matrix factorization with neural networks // American Physical Society, Phys. Rev. E. – 2023. – Vol. 107, No. 6.
12. *Khalitov R., et al.* Sparse factorization of square matrices with application to neural attention modeling // Neural Networks. – 2022. – Vol. 152. – P. 160-168.
13. *Stoica I., et al.* Chord: A scalable peer-to-peer lookup service for internet applications // ACM SIGCOMM Computer Communication Review. – 2001. – Vol. 31, No 4. – P. 149-160.
14. *Hu Xiao-Yu, Eleftheriou E., Arnold D. M.* Regular and irregular progressive edge-growth tanner graphs // IEEE Transactions on Information Theory. – 2005. – Vol. 51, No. 1. – P. 386-398.
15. *Tian Tao, Jones C. R., Villasenor J. D., Wesel R. D.* Selective avoidance of cycles in irregular LDPC code construction // IEEE Transactions on Communications. – 2004. – Vol. 52, No. 8. – P. 1242-1247.
16. *Усатюк В.С., Егоров С.И.* Построение квазициклических недвоичных низкоплотностных кодов на основе совместной оценки их дистантных свойств и спектров связности // Телекоммуникации. – 2016. – № 8. – С. 32-40.
17. *Усатюк В.С., Егоров С.И.* Устройство для оценки кодового расстояния линейного блочного кода методом геометрии чисел // Известия Юго-Западного государственного университета. Серия: Управление, вычислительная техника, информатика. Медицинское приборостроение. – 2017. – № 4 (25). – С. 24-33.
18. *Usatyuk V., Egorov S., Svistunov G.* Construction of Length and Rate Adaptive MET QC-LDPC Codes by Cyclic Group Decomposition // 2019 IEEE East-West Design & Test Symposium (EWDTS). – Batumi, Georgia, 2019. – P. 1-5.
19. *Usatyuk V., Vorobyev I.* Simulated Annealing Method for Construction of High-Girth QC-LDPC Codes // Intern. Conf. on Telecom. and Signal Proc. – 2018. – P. 1-5.
20. *Usatyuk V.S.* Matlab implementation of TSVD, SF Chord, LDPC PEG+ACE, QC-LDPC SA+EMD (MET) codes sparse factorization. – URL: https://github.com/Lcrypto/Classical-and-Quantum-Topology-ML-toric-spherical.

REFERENCES

1. *Tanner R.* A recursive approach to low complexity codes, *IEEE Transactions on Information Theory*, 1981, Vol. 27, No. 5, pp. 533-547.
2. *Mézard M., Montanari A.* Information, Physics, and Computation. Oxford Graduate Texts, 2009, 569 p.
3. *Tanner R., et al.* LDPC block and convolutional codes based on circulant matrices, *IEEE Transactions on Information Theory*, 2004, Vol. 50, No. 12, pp. 2966-2984.
4. *Sukar L.E.* Veroyatnostnye grafovye modeli. Printsipy i prilozheniya [Probabilistic graph models. Principles and applications]. Moscow: DMK Press, 2021, 338 p.
5. *Stan Z. Li*. Markov Random Field Modeling in Image Analysis. Springer London, 2009, 362 p.
6. *Usatyuk V., Sapozhnikov D., Egorov S.* Spherical and Hyperbolic Toric Topology-Based Codes on Graph Embedding for Ising MRF Models: Classical and Quantum Topology Machine Learning, 2023, 71 p. Available at: https://arxiv.org/abs/2307.15778.
7. *Yi T., et al.* Long Range Arena: A Benchmark for Efficient Transformers. ICLR, 2021, 15 p.
8. *Koen M.I.* Prikladnaya lineynaya algebra dlya issledovateley dannykh [Applied linear algebra for data scientists]. Moscow: DMK Press, 2023, 328 p.
9. *Eckart C., Young G.* The Approximation of One Matrix by Another of Lower Rank, *Psychometrika*, 1936, Vol. 1, No. 3, pp. 211-218.

10. *Fomin F. V., et al.* Approximation schemes for low-rank binary matrix approximation problems, *ACM Transactions on Algor*., 2019, Vol. 16, No. 1, pp. 12:1-12:39.
11. *Camilli F., Mezard M.* Matrix factorization with neural networks, *American Physical Society, Phys. Rev. E*., 2023, Vol. 107, No. 6.
12. *Khalitov R., et al.* Sparse factorization of square matrices with application to neural attention modeling, *Neural Networks*, 2022, Vol. 152, pp. 160-168.
13. *Stoica I., et al.* Chord: A scalable peer-to-peer lookup service for internet applications, *ACM SIGCOMM Computer Communication Review*, 2001, Vol. 31, No 4, pp. 149-160.
14. *Hu Xiao-Yu, Eleftheriou E., Arnold D. M.* Regular and irregular progressive edge-growth tanner graphs, *IEEE Transactions on Information Theory*, 2005, Vol. 51, No. 1, pp. 386-398.
15. *Tian Tao, Jones C. R., Villasenor J. D., Wesel R. D.* Selective avoidance of cycles in irregular LDPC code construction, *IEEE Transactions on Communications*, 2004, Vol. 52, No. 8, pp. 1242-1247.
16. *Usatyuk V.S., Egorov S.I.* Postroenie kvazitsiklicheskikh nedvoichnykh nizkoplotnostnykh kodov na osnove sovmestnoy otsenki ikh distantnykh svoystv i spektrov svyaznosti [Construction of quasi-cyclic non-binary low-density codes based on a joint assessment of their distant properties and connectivity spectra], *Telekommunikatsii* [Telecommunications], 2016, No. 8, pp. 32-40.
17. *Usatyuk V.S., Egorov S.I.* Ustroystvo dlya otsenki kodovogo rasstoyaniya lineynogo blochnogo koda metodom geometrii chisel [Device for estimating the code distance of a linear block code using the geometry of numbers method], *Izvestiya Yugo-Zapadnogo gosudarstvennogo universiteta. Seriya: Upravlenie, vychislitel'naya tekhnika, informatika. Meditsinskoe priborostroenie* [News of the South-West State University. Series: Management, computer technology, computer science. Medical instrumentation], 2017, No. 4 (25), pp. 24-33.
18. *Usatyuk V., Egorov S., Svistunov G.* Construction of Length and Rate Adaptive MET QC-LDPC Codes by Cyclic Group Decomposition, *2019 IEEE East-West Design & Test Symposium (EWDTS)*. Batumi, Georgia, 2019, pp. 1-5.
19. *Usatyuk V., Vorobyev I.* Simulated Annealing Method for Construction of High-Girth QC-LDPC Codes, *Intern. Conf. on Telecom. and Signal Proc*., 2018, pp. 1-5.
20. *Usatyuk V.S.* Matlab implementation of TSVD, SF Chord, LDPC PEG+ACE, QC-LDPC SA+EMD (MET) codes sparse factorization. Available at: https://github.com/Lcrypto/ Classical-and-Quantum-Topology-ML-toric-spherical.

Статью рекомендовал к опубликованию д.т.н., профессор В.М. Курейчик.

Усатюк Василий Станиславович – ООО «Т8»; e-mail: l@lcrypto.com; г. Москва, Россия; к.т.н.

Егоров Сергей Иванович – Юго-Западный государственный университет; e-mail: sie58@mail.ru; г. Курск, Россия; д.т.н.; доцент; профессор кафедры вычислительной техники.

Локтионов Аскольд Петрович – e-mail: loapa@mail.ru; д.т.н.; доцент; старший научный сотрудник кафедры уникальных зданий и сооружений.

Титенко Евгений Анатольевич – e-mail: johntit@mail.ru; тел.: +79051588904; кафедра программной инженерии; к.т.н.; доцент.

Чернецкая Ирина Евгеньевна – e-mail: white@mail.ru; кафедра вычислительной техники; зав. кафедрой; д.т.н.; доцент.

Usatyuk Vasily Stanislavovich – T8 LLC; e-mail: l@lcrypto.com; Moscow, Russia; cand. of eng. sc.

Egorov Sergey Ivanovich – South-West State University; e-mail: sie58@mail.ru; Kursk, Russia; dr. of eng. sc.; associate professor; professor of the department of computer science.

Loktionov Askold Petrovich – e-mail: loapa@mail.ru; dr. of eng. sc.; associate professor; senior researcher at the department of unique buildings and structures.

Titenko Evgeny Anatolievich – e-mail: johntit@mail.ru; phone: +79051588904; the department of computer science cand. of eng. sc.; associate professor.

Chernetskaya Irina Evgenievna – e-mail: white@mail.ru; the department of computer science; head of. department; dr. of eng. sc.; associate professor.