УДК 004.383.8.032.26

Применение имитационного моделирования для вычисления погрешности матрично-векторного умножения в мемристивных кроссбарах искусственных нейронных сетей

Борданов И.А., Антонов А.М., Королев Л.Я.

Аналоговая реализация искусственных нейронных сетей (ИНС) на базе мемристивных устройств (ИНСМ) позволяет повысить скорость работы ИНСМ при этом уменьшив их энергопотребление по сравнению с наиболее распространенным цифровым созданием ИНС на ЭВМ в архитектуре фон Неймана. Однако несмотря на данные преимущества мемристоры имеют ряд недостатков, которые приводят к снижению точности вычисления матрично-векторного умножения (МВУ) и соответственно негативно сказываются на качестве работы ИНСМ. В данной работе предложен подход для решения данной проблемы основанный на методологии имитационного моделирования. Реализация данного подхода продемонстрирована на примере решения задачи по вычислению погрешности МВУ для кроссбара ИНСМ, содержащего двадцать значений её весовых коэффициентов синапсов нейронов.

Ключевые слова: искусственные нейронные сети, мемристоры, матрично-векторное умножение, погрешность вычисления.

Введение

Объем данных и сложность алгоритмов их обработки растут с каждым годом, что сопровождается ростом количества и номенклатуры используемой электронной продукции и возрастанием нагрузки на существующую телекоммуникационную, вычислительную и энергетическую инфраструктуры. Дополнительно этот тренд усиливается за счёт увеличения доли применения систем искусственного интеллекта (ИИ) для анализа больших данных и применения сложных алгоритмов, таких как искусственные нейронные сети (ИНС) [1-4]. В связи с этим необходимо разрабатывать новые эффективные устройства, которые позволят организовать такие вычисления на новом более качественном уровне, с высоким быстродействием и низким энергопотреблением.

Мемристоры являются наиболее перспективными электронными компонентами для аппаратной реализации, как формальных ИНС со статическими весами [5], так и спайковых ИНС с изменяющимися в процессе работы весами [6]. Они представляют собой энергонезависимые электронные компоненты, которые изменяют своё сопротивление при протекании через них тока и могут сохранять его длительное время [7]. В ИНС они используются для выполнения операции матрично-векторного умножения (МВУ), которое происходит в соответствии с естественным физическим законом Ома [8], что позволяет повысить скорость работы ИНС и снизить потребление энергии.

Не смотря на существенные преимущества мемристоров при аппаратной реализации формальных ИНС, основным недостатком является не возможность задать с абсолютной точностью нужное значение сопротивления. Это приводит к возникновению погрешностей [9] при выполнении МВУ, которые в итоге могут привести к ошибкам в работе ИНС и в целом снизить надежность разрабатываемых систем.

Результаты исследований [10-15] показывают, что не смотря на погрешности при выполнении МВУ, ИНС способны работать без возникновения критических ошибок для достаточно широкого диапазона разброса весов. Это зависит от многих факторов: выбранной архитектуры и структуры, алгоритма обучения, конструктивной реализации весов и т.д. Диапазон допустимых погрешностей можно определить для каждого конкретного варианта сочетания вышеназванных факторов путем имитационного моделирования. Результат такого моделирования позволит разработчику ответить на вопрос – будет ли ИНС, аппаратно реализованная на базе мемристоров (ИНСМ), работать с той же точностью и

надежностью, как при симуляции модели на цифровых вычислительных устройствах?

С другой стороны, при аппаратной реализации ИНСМ матрицы весов разбиваются на участки размером, эквивалентным размеру кроссбара мемристивных устройств. Таким образом каждый кроссбар выполняет МВУ на определенном участке ИНС. На выходе строк каждого кроссбара в результате вычисления МВУ формируется определенное суммарное напряжение, которое затем подается на следующий кроссбар в качестве входного напряжения. Под значение этого напряжения рассчитываются параметры всех электронных компонентов системы, поэтому его нельзя превышать, чтобы не потерять информацию и не вызвать ошибки в работе системы. Это вторая важная причина, по которой на этапе проектирования необходимо определить допустимые разбросы сопротивлений и рассчитывать предельные значения для каждого режима работы.

В данной работе рассмотрено применение имитационного моделирования для решения задачи оценки погрешности выполнения МВУ в ИНСМ.

Метод

Имитационное моделирование является мощнейшим, а зачастую единственным средством исследования ИНСМ, так как они являются сложно формализуемыми объектами, в которых информация проходит поэтапную обработку с нелинейными преобразованиями. Зависимость точности работы ИНСМ от величины погрешности сопротивлений мемристоров невозможно рассчитать аналитически. Мы предлагаем следующий алгоритм оценки точности вычислений в ИНСМ, основанный на имитационном моделировании:

- для конкретного типа мемристивных устройств провести эксперименты, заключающиеся программировании нескольких резистивных состояний Rm в диапазоне минимального сопротивления мемристора R_{min} до максимального R_{max} и накопить статистику;
- для полученных экспериментальных данных построить плотности вероятности и определить закон распределения и значения его параметров, получив график зависимости погрешности от сопротивления мемристора Rm;
- аппроксимировать результаты погрешности на диапазоне от R_{min} до R_{max} и получить функцию зависимости параметров закона распределения от сопротивления мемристора Rm;
- полученные зависимости использовать для имитационного моделирования погрешностей весов ИНСМ, перейдя от уровня физических устройств на уровень обработки информации;
- на имитационной модели оценить точность вычислений в ИНСМ и сделать вывод о применимости рассматриваемого типа мемристивных устройств для аппаратной реализации сети.

Частным случаем применения данного алгоритма, является расчет погрешности МВУ в ИНСМ, состоящий из следующих шагов:

- Перед тем, как выполнять имитационное моделирование необходимо первоначально обучить ИНС решению требуемой задачи после чего нужно извлечь полученные веса из ИНС и разбить их на участки размером с кроссбар (рис. 1).
- Необходимо определиться со схемой, которая будет осуществлять умножение входного напряжения на весовой коэффициент сети, роль которого играет один или несколько мемристоров в зависимости от реализации. Понимание схемы и её реализации необходимо для выведения формулы для получения эквивалентных весу сопротивлений для мемристора.
- После того как требуемая формула была определена, переводим весовые коэффициенты в сопротивления.
- Теперь, зная сопротивления мемристоров, а также погрешность на их программирование, можно рассчитать погрешность конкретного веса при его переносе на аппаратную базу.

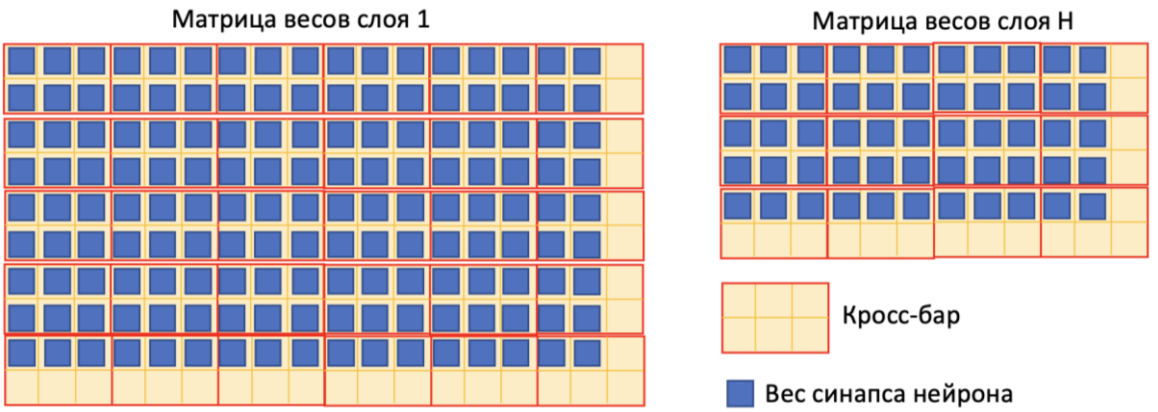

Рис. 1. – «Покрытие» матриц весов синапсов нейронов кроссбарами мемристивных устройств

• Полученные в ходе предыдущего пункта погрешности весов можно подставить в модель кроссбара и рассчитать погрешность выходных данных.

Для реализации данного алгоритма необходимо создание имитационной модели, которая будет специфична каждому конкретному конструкторскому варианту организации вычислений в ИНСМ. Общим для всех моделей является соблюдение концепции, в соответствии с которой погрешности, возникающие в ИНСМ на физическом уровне, вызывают погрешности в работе ИНСМ на уровне обработки информации, соответственно имитационная модель должна учитывать эту взаимосвязь. Для программной реализации предложенных алгоритмов мы используем язык программирования Python.

Эксперимент

Рассмотрим предложенный метод на примере одного участка ИНС размером с кроссбар 5 на 4 которая была обучена решению задачи распознавания рукописных цифр на датасете MNIST.

Из данной сети были извлечены 20 значений весовых коэффициентов синапсов, связывающих входной и первый скрытый слой ИНС.

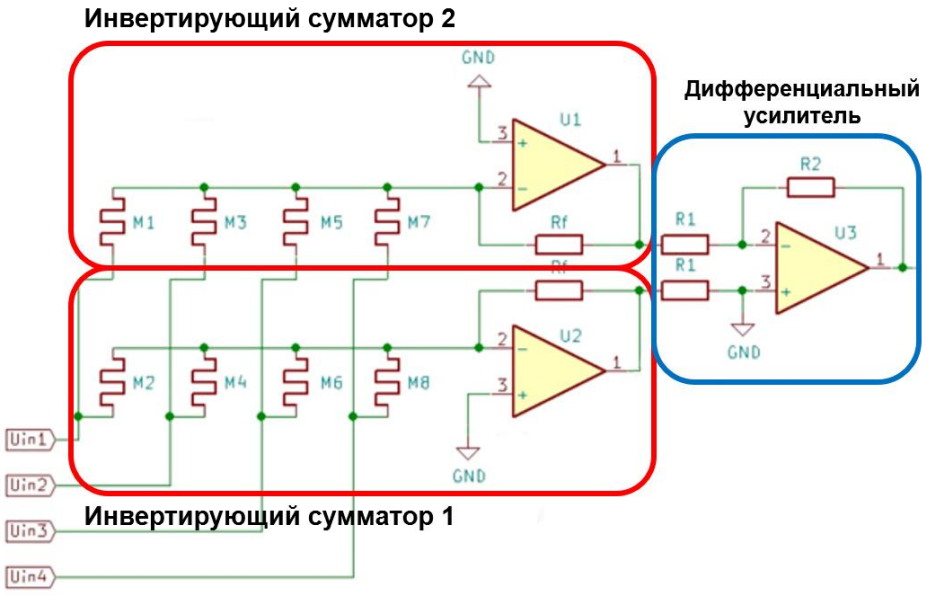

Рис. 2. – Схема ВМУ для одного нейрона с 4 синапсами. Красными контурами выделены схемы инвертирующих сумматоров вместе с мемристорами, которые подключены к их отрицательному входу. Синим контуром выделен дифференциальный усилитель, на вход которому подключается пара инвертирующих сумматоров для реализации функции разности входных сигналов

В качестве схемы аппаратной реализации ИНСМ была выбрана схема синапса, выполненная на основе двух мемристоров (рис. 2), что позволяет организовывать хранения не только положительных, но и отрицательных весов.

Умножение входного напряжения, поступающего на входы кроссбара, на вес осуществляется по средствам инвертирующего сумматора. При этом каждая пара выходных строк соединяется с дифференциальным усилителем, который выполняет функцию разности входных сигналов, благодаря чему достигается возможность получения и положительных и отрицательных весов для ИНСМ. Формула расчета весового коэффициента при такой реализации представлена ниже:

$$w = Rf \cdot (R1 - R2) / (R1 \cdot R2) \qquad (1)$$

где w – значение веса синапса нейрона; Rf – коэффициент масштабирования (Ом); $R1$, $R2$ – сопротивления мемристоров синапса.

Из формулы (1) можно вывести формулы для расчета сопротивлений мемристоров, которые необходимы для дальнейших расчетов.

Ниже представлена формула для расчета $R1$ при значении веса больше 0, при этом значение $R2 = R_{max}$:

$$R1 = R_{max} / (w / Rf \cdot (R_{max} + 1)), \qquad (2)$$

где R_{max} – максимальное значение сопротивления мемристора.

Ниже представлена формула для расчета $R2$ при значении веса меньше или равном 0, при этом значение $R1 = R_{max}$:

$$R1 = - R_{max} / (w / Rf \cdot (R_{max} - 1)). \qquad (3)$$

Затем, используя формулы (2) и (3), были рассчитаны значения сопротивлений мемристоров для всех весовых коэффициентов.

На полученные значения сопротивлений были накинуты погрешность программирования весов, которая была условно установлена в районе 20 % по нормальному закону распределения. Данное действие выполнялось 1000 раз в результате чего было получено по 1000

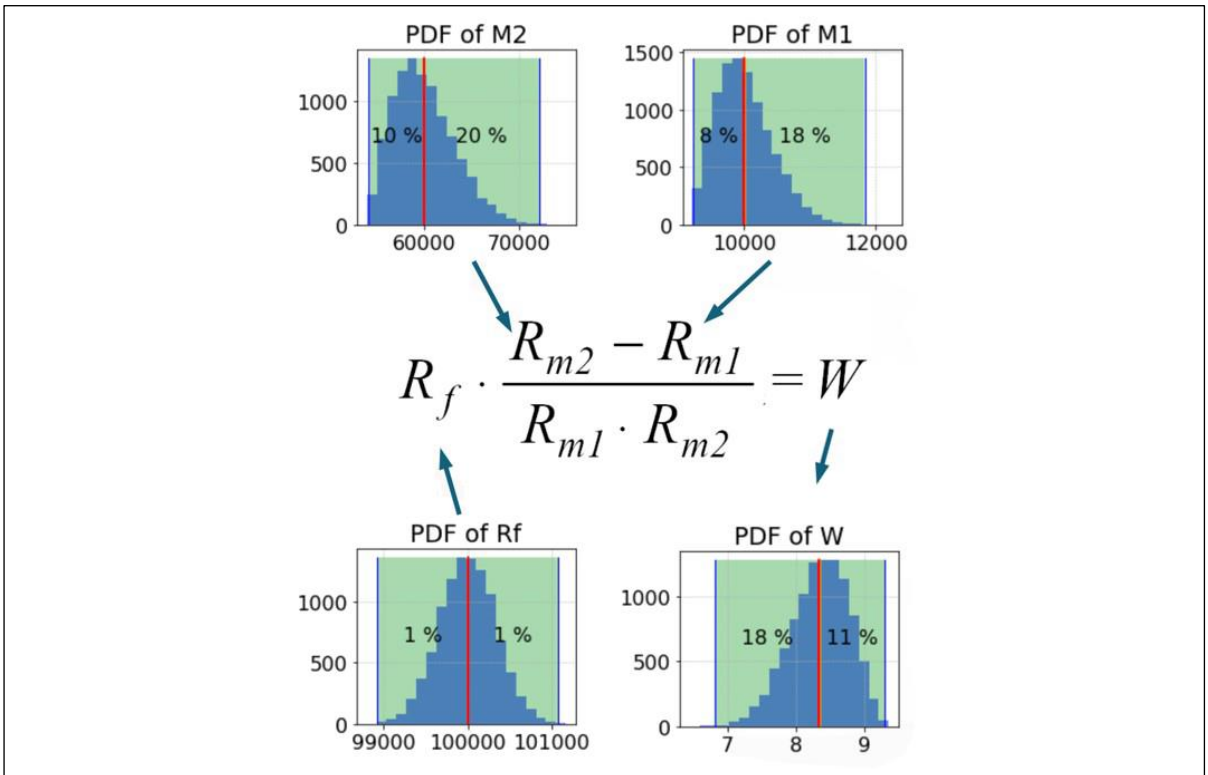

Рис. 3. – Влияние погрешностей мемристоров синапса на вес. Красная линия на гистограммах соответствует среднему значению сопротивления. Зеленые области, ограниченные синими линиями, соответствуют максимальному разбросу в отрицательную и в положительную сторону относительно среднего значения. По оси абсцисс на гистограммах PDF of M1, PDF of M2 и PDF of Rf представлено значение сопротивления в омах, а на PDF of M1 значение веса синапса нейрона. По оси ординат на всех гистограммах отражено количество данных значений

значений сопротивлений с погрешностями для каждого мемристора в синапсе.

Зная данные значения и подставив их в формулу (1), была получена 1000 матриц весов ИНС с погрешностями, проявившимися в результате погрешности программирования МВУ. Наглядная иллюстрация данной операции для весового коэффициента синапса нейрона представлена на рис. 3.

На основе полученных данных и формулы (4) было рассчитано значение погрешности для каждого веса кроссбара, которое теперь можно использовать для расчета погрешности ВМУ. Результаты данных расчетов были представлены в виде диаграммы на рис. 4.

$$\Delta w = 3 \cdot w / w_{mean} \cdot 100, \qquad (4)$$

где Δw – значение погрешности веса синапса нейрона; w – значение веса синапса нейрона с погрешностью; w_{mean} – среднее значение веса для всей 1000 матриц.

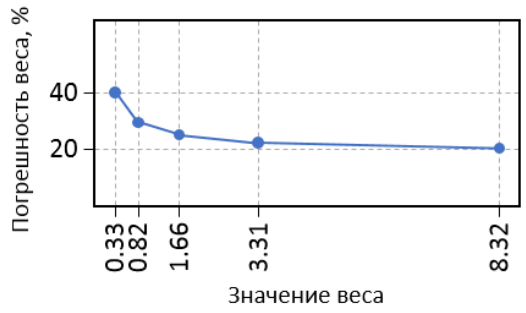

Рис. 4. – Влияние погрешностей мемристоров синапса на вес. По оси абсцисс здесь показано значение веса синапса нейрона, а по оси ординат величина погрешности соответствующего веса

Полученные значения весовых коэффициентов и соответствующие им погрешности были использованы при выполнении имитационного моделирования операции ВМУ на протяжении 1000 раз. При этом погрешности для каждого веса рассчитывались по нормальному. Результаты расчетов представлены на рис. 5.

Из рис. 5 видно, что из-за погрешностей сопротивлений мемристивных элементов, которые соответственно добавляют разбросы в значения весов синапсов нейронов, выходной вектор, полученный в результате ВМУ, может иметь погрешность более 50% процентов при 20% погрешности сопротивлений мемристивных элементов.

Полученные данные могут быть использованы для масштабирования выходного вектора перед подачей в следующий кроссбар. Это является важным моментом так как если диапазон рабочих сопротивлений к примеру, составляет от -6 до 6 V, а при 27% погрешности 5,03 V значение получается более 6 V, то часть информации теряется, что также вносит дополнительные погрешности и может стать причиной снижения точности работы ИНСМ до неудовлетворительных значений.

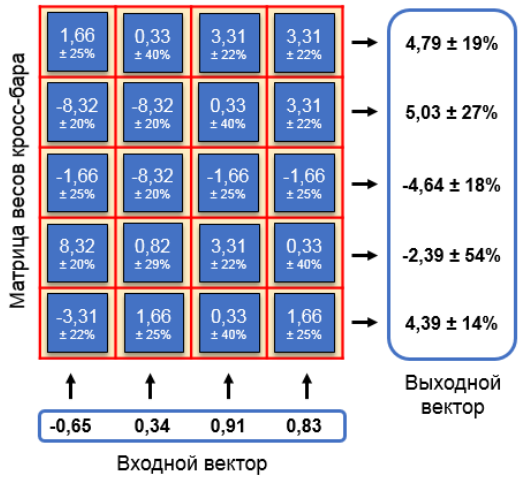

Рис. 5. – Расчет погрешности вычисления ВМУ. Входной вектор, состоящий из 4 чисел, умножается на матрицу весовых коэффициентов размером 5x4. При этом под каждым значением веса расположенном в синим квадрате написан разброс его значения. В выходном векторе представлено номинальное значения вычисления ВМУ, как если бы оно выполнялось без учета погрешностей весов, а также его разброс, полученный в результате имитационного моделирования

Заключение

Таким образов в результате выполнения данной работы были показаны возможности по определению на имитационной модели погрешности МВУ для заданной аппаратной реализации весов нейронной сети.

В дальнейшем этот подход может быть использован для расчета допустимых отклонений весов при заданном допуске на погрешность МВУ.

Литература

1. Privezentsev D.G. et al. Formation of the Structure and Training of a Multilayer Neural Network for the Analysis and Blood Glucose Level Prediction // J. Phys. Conf. Ser. IOP Publishing. 2021. Vol. 1828. № 1. P. 012014.

2. Shtanko A., Kulik S. Preliminary Experiment on Emotion Detection in Illustrations Using Convolutional Neural Network // Adv. Intell. Syst. Comput. Springer Science and Business Media Deutschland GmbH. 2021. Vol. 1310. P. 490–494.

3. Kulik SD. Neural Network Model Of Artificial Intelligence For Handwriting Recognition // J. Theor. Appl. Inf. Technol. 2015. Vol. 73, № 2.

4. Kulik S.D., Shtanko A.N. Experiments with Neural Net Object Detection System YOLO on Small Training Datasets for Intelligent Robotics // Mech. Mach. Sci. Springer Science and Business Media B.V. 2020. Vol. 80. P. 157–162.

5. Mikhaylov A.N. et al. One-Board Design and Simulation of Double-Layer Perceptron Based on Metal-Oxide Memristive Nanostructures // IEEE Trans. Emerg. Top. Comput. Intell. Institute of Electrical and Electronics Engineers Inc. 2018. Vol. 2. № 5. P. 371–379.

6. Sboev A. et al. Modeling the Dynamics of Spiking Networks with Memristor-Based STDP to Solve Classification Tasks // Math. 2021, Vol. 9, Page 3237. Multidisciplinary Digital Publishing Institute. 2021. Vol. 9. № 24. P. 3237.

7. Strukov D.B. et al. The missing memristor found // Nat. 2008 4537191. Nature Publishing Group. 2008. Vol. 453, № 7191. P. 80–83.

8. Amirsoleimani A. et al. In-Memory Vector-Matrix Multiplication in Monolithic Complementary Metal–Oxide–Semiconductor-Memristor Integrated Circuits: Design Choices, Challenges, and Perspectives // Adv. Intell. Syst. Wiley. 2020. Vol. 2. № 11. P. 2000115.

9. Mehonic A. et al. Simulation of inference accuracy using realistic rram devices // Front. Neurosci. Frontiers Media S.A. 2019. Vol. 13. P. 593.

10. Danilin S.N., Shchanikov S.A., Galushkin A.I. The research of memristor-based neural network components operation accuracy in control and communication systems // 2015 Int. Sib. Conf. Control Commun. SIBCON 2015 - Proc. Institute of Electrical and Electronics Engineers Inc. 2015.

11. Danilin S.N., Shchanikov S.A., Panteleev S.V. Determining Operation Tolerances of Memristor-Based Artificial Neural Networks. Institute of Electrical and Electronics Engineers (IEEE). 2017. P. 34–38.

12. Danilin S.N., Shchanikov S.A. Neural network control over operation accuracy of memristor-based hardware // Proc. 2015 Int. Conf. Mech. Eng. Autom. Control Syst. MEACS 2015. Institute of Electrical and Electronics Engineers Inc. 2016.

13. Danilin S.N., Makarov V. V., Shchanikov S.A. Design of Artificial Neural Networks with a Specified Quality of Functioning // Proc. - 2014 Int. Conf. Eng. Telecommun. EnT 2014. Institute of Electrical and Electronics Engineers Inc. 2014. P. 67–71.

14. Danilin S.N. et al. Determining the fault tolerance of memristorsbased neural network using simulation and design of experiments // Proc. - 5th Int. Conf. Eng. Telecommun. EnT-MIPT 2018. Institute of Electrical and Electronics Engineers Inc. 2018. P. 205–209.

15. Shchanikov S.A. Methodology for Hardware-in-the-Loop Simulation of Memristive Neuromorphic Systems // Nanobiotechnology Reports. Springer. 2022. Vol. 16, № 6. P. 782–789.

Работа выполнена при поддержке стипендии Президента РФ СП-3988.2022.5.

Поступила 14 сентября 2022 г.

The analog implementation of artificial neural networks (ANNs) based on memristive devices (ANNM) allows to increase the speed of the ANNM while reducing their power consumption compared to the most common digital creation of ANN on a computer in the von Neumann architecture. However, despite these advantages, memristors have a number of disadvantages that lead to a decrease in the accuracy of the calculation of matrix-vector multiplication (MVM) and, accordingly, negatively affect the quality of the ANNMs. This paper proposes an approach to solve this problem based on the simulation methodology. The implementation of this approach is demonstrated on the example of solving the problem of calculating the error of the MVM for the ANNM crossbar containing twenty values of its weight coefficients of neuron synapses.

Key words: artificial neural networks, memristors, matrix-vector multiplication, calculation error.

Борданов Илья Алексеевич – инженер-исследователь лаборатории разработки систем искусственного интеллекта Муромского института (филиала) ФГОУ ВО «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых».

E-mail: bordanov2011@yandex.ru.

Антонов Александр Михайлович – лаборант-исследователь лаборатории разработки систем искусственного интеллекта Муромского института (филиала) ФГОУ ВО «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых».

E-mail: cfifant@mail.ru.

Королев Леонид Ярославович – студент кафедры информационных систем Муромского института (филиала) ФГОУ ВО «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых».

E-mail: madimtor@ya.ru.

Адрес: 602264, Муром, ул. Орловская, д. 23.