Инженерный вестник Дона, №6 (2025)
ivdon.ru/ru/magazine/archive/n6y2025/10111

Улучшение сжатия данных: инновации и будущие перспективы

М.Е. Козлов, А.С. Попов, Д.М. Махмудов, И.С. Макаров

Поволжский государственный университет телекоммуникаций и информатики, Самара

Аннотация: Статья посвящена теме применения современных методов генеративного сжатия изображений с использованием вариационных автокодировщиков и нейросетевых архитектур. Особое внимание уделяется анализу существующих подходов к генерации и восстановлению изображений, а также сравнительной оценке качества сжатия с точки зрения визуального восприятия и метрических показателей. Целью исследования является систематизация методов глубокого сжатия изображений и выявление наиболее эффективных решений, основанных на вариационном байесовском подходе. В работе рассмотрены различные архитектуры, в том числе условные автокодировщики и модели с гиперсетями, а также методы оценки качества получаемых данных. В качестве основных методов исследования применялись анализ научной литературы, сравнительный эксперимент над архитектурами генеративных моделей и вычислительная оценка сжатия на основе метрик. Результаты исследования показали, что использование вариационных автокодировщиков в сочетании с рекуррентными и сверточными слоями позволяет добиться высокого качества восстановления изображений при значительном снижении объема данных. Сделан вывод о перспективности использования условных вариационных автокодировщиков в задачах сжатия изображений, особенно при наличии дополнительной информации (например, метаданных). Представленные подходы могут быть полезны для разработки эффективных систем хранения и передачи визуальных данных.
Ключевые слова: вариационные автокодировщики, генеративные модели, сжатие изображений, глубокое обучение, нейросетевые архитектуры, восстановление данных, условные модели.

Введение

Современный мир словно погружен в цифровую революцию, а объем данных, которые мы каждый день создаем, только растет. В таких условиях важно разработать эффективные методы сжатия данных, чтобы уменьшить их размер, но сохранить всю важную информацию. В этой статье мы рассмотрим инновационные методы и технологии, которые помогут улучшить сжатие данных, а также обсудим перспективы этого направления. Традиционные методы сжатия, такие как алгоритмы Хаффмана или Лемпеля-Зива-Уэлча (Lempel-Ziv-Welch – LZW) уже не являются в полной мере эффективными с учетом современных требований к информации. Одно из самых перспективных направлений в области сжатия данных - применение нейронных сетей. Последние исследования показали, что нейронные сети

могут быть очень эффективными при сжатии различных типов данных, включая изображения, видео и аудио. В рамках научной статьи особое внимание уделено таким основным архитектурам как классический автокодировщик (Autoencoder) и вариационный автокодировщик (variational autoencoder – VAE). На основе проведенного сравнения был выбран более эффективный метод сжатия изображений, основанный на глубоком обучении, который позволяет получить высокую степень сжатия без значительной потери качества [1]. Прогресс в области нейронных сетей открывает «новые возможности для создания адаптивных и эффективных алгоритмов сжатия данных». Эффективное сжатие данных будет играть ключевую роль в оптимизации передачи огромных объемов информации также и в нейронных сетях следующего поколения.

Целью научной статьи является сравнительный анализ эффективности Autoencoder и VAE в задаче сжатия изображений, а также сопоставление их производительности с традиционными методами сжатия данных. Исследование направлено на выявление преимуществ и ограничений нейросетевых подходов к сжатию данных, для оценки потенциал для практического применения в различных областях знаний.

Методология

Для научного исследования были применены такие методы как исторический анализ, метод научного абстрагирования, статические методы анализа, эмпирический метод и др. Были проведены эксперименты по сжатию изображений на основе дата сета MNIST [2] с применением разработанных нейросетевых моделей и алгоритма ZIP, проведена количественная оценка качества сжатия при помощи метрик среднеквадратическая ошибка и пиковых отношений сигнал/шум. Для оценки качества также проводилось визуальное сравнение оригинальных и реконструированных изображений. Результаты проведенных экспериментов

были обработаны статистически для обеспечен для достоверности полученных выводов.

Современные достижения в области искусственного интеллекта открывают новые возможности для технологий сжатия данных. Благодаря своей способности выявлять иерархические особенности в данных, нейронные сети демонстрируют значительный успех в сжатии изображений, видео и аудио без потерь [3].

Архитектуры нейронных сетей

Одна из самых перспективных архитектур - сверточная нейронная сеть (convolutional neural network – CNN), которая отлично справляется с визуальными данными [4]. CNN использует сверточные слои для автоматического и адаптивного извлечения характеристик, что делает их идеальными для сжатия изображений. Еще один тип нейронных сетей, широко используемый для сжатия данных - автоэнкодеры [5]. Они кодируют входные данные в компактное представление и затем восстанавливают их [6].

Сравнение с традиционными методами

Когда сравниваем нейронные сети с традиционными способами сжатия, такими как JPEG для изображений или MPEG для видео, становится ясно, что нейронные сети имеют преимущество в гибкости и качестве сжатия. В таблице 1 приведено сравнение, которое подтверждает этот факт [1, 7, 8]. Нейронные сети способны обучаться на различных типах данных и оптимизировать процесс сжатия специально для них, что позволяет достичь более высокого уровня сжатия без существенных потерь качества [3].

Примеры применения

Примером успешного применения нейронных сетей в сжатии данных может служить разработка компанией Google формата сжатия изображений WebP [4]. Этот формат использует алгоритмы машинного обучения для

улучшения сжатия и поддерживает как сжатие без потерь, так и сжатие с потерями [5].

Таблица №1

Сравнение нейронных сетей с традиционными методами машинного обучения

Свойство	Нейронные сети	Традиционные методы машинного обучения
Архитектура	Состоит из множества взаимосвязанных нейронов, организованных в слои	Основаны на математических и статистических моделях
Обучение	Использует алгоритмы обратного распространения ошибки для обучения на больших объемах данных	Обучение основано на анализе и классификации данных с использованием различных алгоритмов
Обработка данных	Может обрабатывать большие объемы данных и извлекать сложные закономерности	Обрабатывает данные с использованием статистических методов и алгоритмов
Преимущества	Могут обучаться на неструктурированных данных и находить сложные зависимости	Простота в реализации и интерпретации результатов
Недостатки	Требуют большого количества данных для обучения и вычислительных ресурсов для обработки	Могут быть ограничены в способности обрабатывать сложные данные и находить сложные зависимости
Примеры применения	Распознавание образов, обработка естественного языка, прогнозирование временных рядов	Классификация текстов, регрессионный анализ, кластеризация данных

WebP предлагает значительные преимущества по сравнению с традиционными форматами, такими как JPEG и PNG, особенно в отношении соотношения качества и размера файла [5, 10]. Это делает его популярным выбором для веб-разработчиков и дизайнеров, стремящихся улучшить производительность веб-страниц за счет уменьшения времени загрузки изображений [7, 8].

В исследовании для формирования моделей были применены библиотеки включая PyTorch для работы с нейронными сетями, torchvision для работы с изображениями и датасетами, matplotlib для визуализации результатов, tkinter для создания графического интерфейса, и pandas для обработки и сохранения данных [9]. Определена архитектура для двух типов автокодировщиков. Autoencoder был представлен классом, который содержит такие компонент как энкодер и декодер [1, 11]. Энкодер состоял из последовательности сверточных слоев, которые постепенно уменьшили размерность входного изображения, извлекая ключевые особенности. Декодер применял транспонированные сверточные слои для восстановления соответствующего изображения из сжатого до соответствующих исходных размеров. VAE был сформирован со схожей структурой, но основным отличием является то, что вместо прямого кодирования входных данных в латентное пространство, вариационный автокодировщик кодирует их в распределение, что достигается с помощью двух полносвязных слоев, которые вычисляют среднее значение и логарифм дисперсии для каждой латентной переменной. Затем была применена техника репараметризации для получения выборки из распределения, что и позволило обучать сеть с использованием метода обратного распространения ошибки.

Исследование

Процесс обучения нейронных сетей был инициирован нажатием кнопки «Обучить модели». При этом загружается датасет MNIST,

содержащий изображения рукописных цифр, и последовательно обучаются обе модели.

На рис.1 показан процесс загрузки и распаковки данных из различных источников для набора MNIST. Далее идёт скачивание файлов с метками и изображениями, а также сообщения об ошибках, после чего происходит успешное извлечение файлов в локальную директорию.

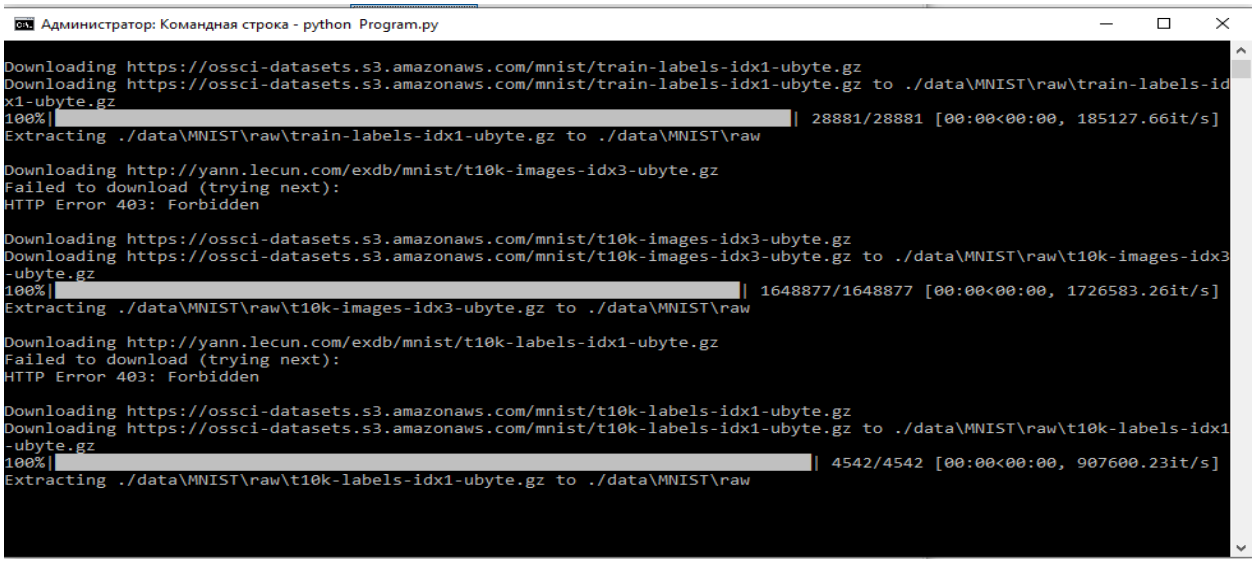

Рис. 1. – Загрузка данных для обучения

После завершения обучения результаты в виде среднеквадратичной ошибки (Mean Squared Error – MSE) и пикового отношения сигнал/шум (peak signal-to-noise ratio – PSNR) добавляются в таблицу на вкладке «Результаты», а графики функций потерь отображаются на вкладке «Графики».

На рис.2 представлен график, на котором сравниваются функции потерь двух методов сжатия данных: AutoEncoder (синяя линия) и VAE (оранжевая линия). По оси X отложены эпохи обучения, а по оси Y – значения функции потерь. Видно, что AutoEncoder демонстрирует стабильно низкие потери, практически близкие к нулю, в то время как VAE начинает с высоких значений (~25000), но затем снижается до уровня ~14000 и стабилизируется.

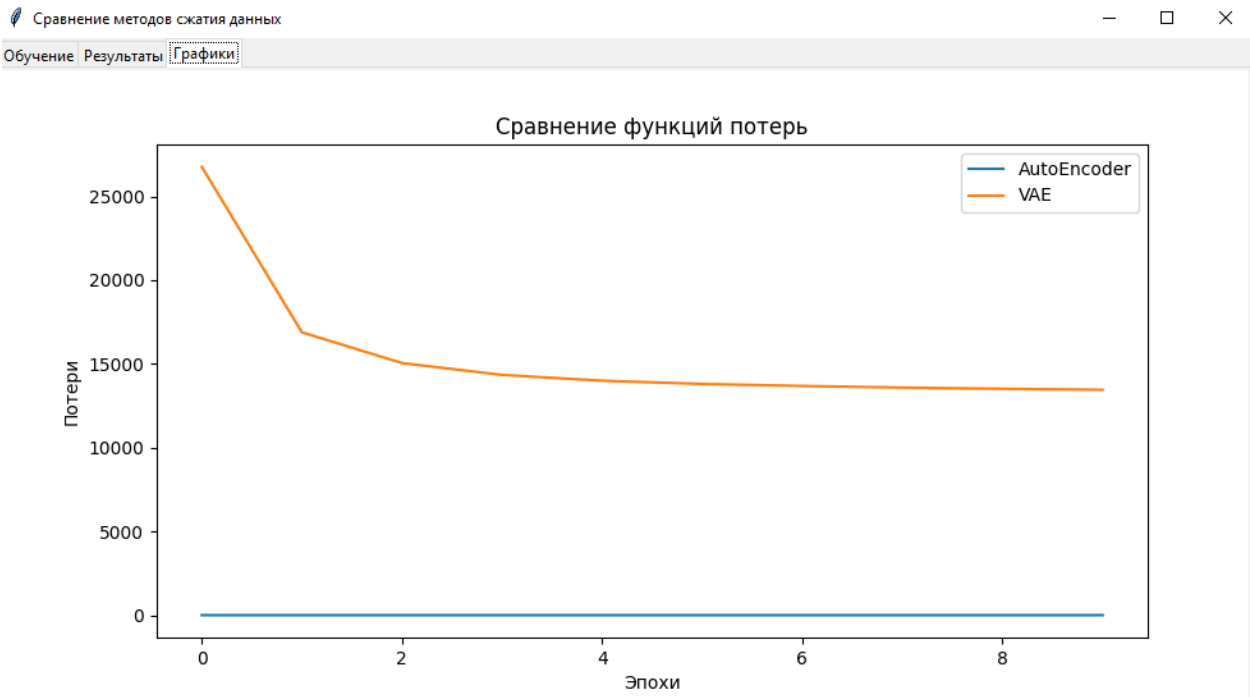

Рис. 2. – Результаты проведенного исследования

Реализованный графический интерфейс программы представлен на рис.3.

Функция потерь для AutoEncoder применяет среднеквадратичную ошибку между входным и реконструированным изображением. Для VAE функция потерь более сложная и состоит из двух компонентов: ошибки реконструкции (бинарная кросс-энтропия) и регуляризационного члена (дивергенция Кульбака-Лейблера, KL-дивергенция), который стимулирует латентное пространство приближаться к стандартному нормальному распределению. AutoEncoder показывает лучшие результаты по обоим метрикам, MSE значительно ниже (0.0027 против 0.0138), что указывает на меньшую ошибку реконструкции. PSNR у AutoEncoder также выше (25.71 против 18.59), что говорит о лучшем качестве восстановленных изображений (рис.4).

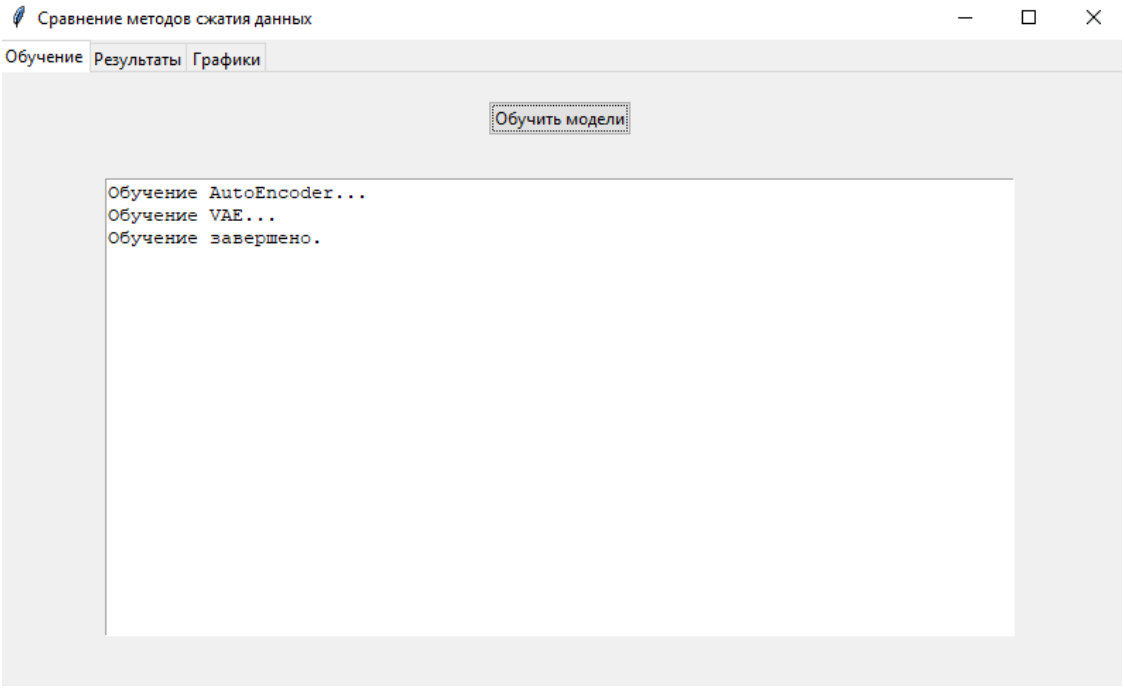

Рис. 3. – Реализация основного интерфейса программы

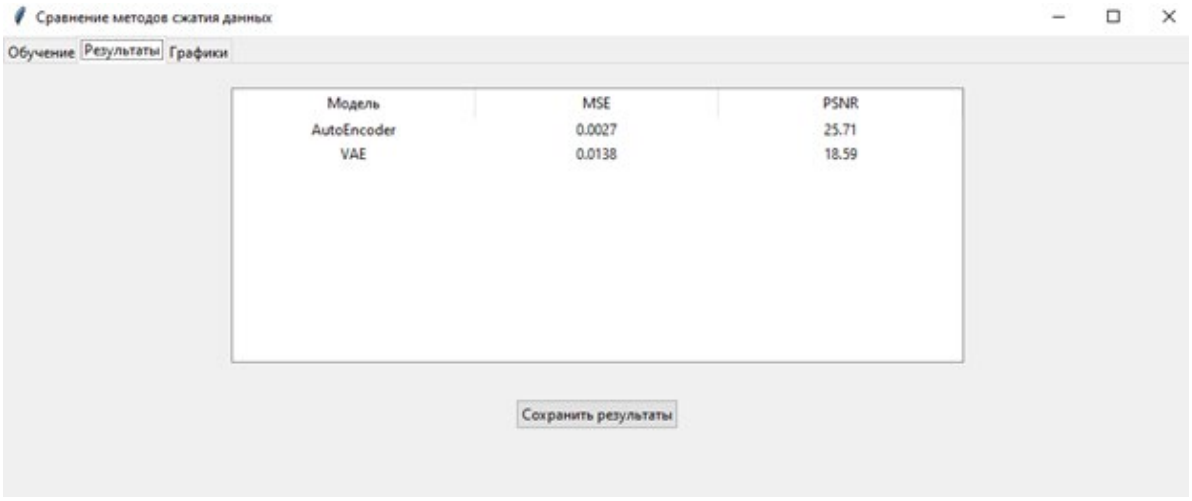

Рис. 4. – Сравнительная характеристика полученных результатов обучения моделей на дата сете MNIST

Процесс обучения моделей реализован в функциях train_autoencoder и train_vae. Функции принимают модель, загрузчик данных, количество эпох и устройство (центральный процессор или графический процессор) в качестве параметров, при нажатии пользователем на кнопку «Обучить модели»

итеративно обучают модели, вычисляя функцию потерь и обновляя веса с помощью оптимизатора Adam.

Для оценки качества сжатия используется функция evaluate_compression, которая вычисляет среднеквадратичную ошибку MSE и пиковое отношение сигнал/шум PSNR для обеих моделей на тестовом наборе данных. Графический интерфейс реализован в классе CompressionGUI. Класс создает окно с тремя вкладками и кнопками для запуска обучения и сохранения результатов, загружается датасет MNIST, обучаются обе модели на основе подгруженных данных, и результаты отображаются в таблице и на графике, также выгружаются в формате csv при нажатии пользователем кнопки «Сохранить результаты».

Заключение

Нейронные сети - очень сильные инструменты для сжатия данных. Они могут делать это гораздо лучше, чем обычные методы. Они могут подстраиваться под детали данных и находить закономерности для создания небольших изображений, видео и звуковых файлов без потери качества. Примером может служить формат WebP, который справляется с этой задачей лучше, чем старый формат JPEG. В результате полученного сравнения двух моделей AutoEncoder и VAE были получены следующие результаты. Для AutoEncoder значение MSE составляет 0,0027, а PSNR – 25,71. Для VAE эти показатели равны 0,0138 и 18,59 соответственно. Превосходство модели AutoEncoder может быть объяснено несколькими факторами. AutoEncoder напрямую оптимизирует ошибку реконструкции, в то время как VAE балансирует между точностью реконструкции и регуляризацией латентного пространства. Но в то же время необходимо отметить, что датасет MNIST, на котором проводилось обучение, содержит относительно простые черно-белые изображения цифр, для которых прямое отображение в «латентное пространство может быть достаточно эффективным». Модель VAE имеет

свои преимущества. Её вероятностная природа позволяет генерировать новые образцы и может обеспечивать лучшую генерализацию на более сложных и разнообразных данных. Построенный по результатам анализа график дает нам дополнительную информацию о поведении моделей во время обучения. График показывает, что функция потерь VAE начинается с гораздо более высоких значений (около 27000) по сравнению с AutoEncoder (менее 5000), соответственно функция потерь VAE включает в себя не только ошибку реконструкции, но и регуляризационный член (KL-дивергенцию), который вносит дополнительный вклад в общую потерю. Сравнивая эти результаты, можно сделать вывод, что для задачи сжатия простых изображений, таких как цифры из датасета MNIST, AutoEncoder может быть более эффективным выбором. VAE демонстрирует стабильное улучшение на протяжении всего процесса обучения, модель продолжает обучаться и показала более лучшие результаты при увеличении количества эпох и на более сложных изображениях.

Литература

1. Антонец Д.В., Русских Н.Е., Минин А.Р., Замятин В.И., Вяткин Ю.В., Раменский В.Е., Штокало Д.Н. Использование условных вариационных автокодировщиков и трансформеров для анализа данных SCRNA-SEQ // Труды 11-й Московской конференции по вычислительной молекулярной биологии. MCCMB’23. Москва: Институт проблем передачи информации РАН, 2023. URL: mccmb.belozersky.msu.ru/2023/thesis/abstracts/72.pdf.

2. Cheng K., Tahir R., Eric L. K., Li M. An analysis of generative adversarial networks and variants for image synthesis on MNIST dataset // Multimedia Tools and Applications. 2020. V. 79. pp. 13725-13752.

3. Талалаев А.А., Тищенко И.П., Фраленко В.П., Хачумов В.М. Анализ эффективности применения искусственных нейронных сетей для решения задач распознавания, сжатия и прогнозирования // Искусственный интеллект и принятие решений. 2008. №2. С. 24-33.

4. Курбатов Г.Р. Методы оценки качества сгенерированных данных // Тенденции развития науки и образования. 2023. №103-8. С. 167–169.

5. Hinton J.E., Salakhoddinov R.R. Data dimensionality reduction with the help of neural networks // Nauka. 2006. V. 313. № 5786. pp. 504-507.

6. Ballé J., Laparra V., Simoncelli E.P. Optimized end-to-end image compression // International Conference on Learning Representation. 2017. URL: arxiv.org/abs/1611.01704.

7. Kingma D.P., Welling M. Autocoding and the variational Bayesian approach // International Conference on Learning Representation. 2014. URL: arxiv.org/abs/1312.6114.

8. Toderici G., Vincent D., Johnston N., Hwang S.J., Minnen D., Shor J., Covell M. Compression of full-size images using recurrent neural networks // Conference on Computer Vision and Pattern Recognition. 2017. URL: arxiv.org/abs/1608.05148.

9. Ballé J., Minnen D., Singh S. Variational image compression with a scaling hyperparameter // International Conference on Learning Representation. 2018. URL: arxiv.org/abs/1802.01436.

10. Theis L., Shi W., Cunningham A., Husar F. Lossy image compression using compressive autoencoders // International Conference on Learning Representation. 2017. URL: arxiv.org/abs/1703.00395.

11. Mentzer F., Agustsson E., Tshannen M. Conditional probability models for deep image compression // Conference on Computer Vision and Pattern Recognition. 2018. URL: arxiv.org/abs/1801.04260.

References

1. Antonets D.V., Russkikh N.E., Minin A.R., Zamyatin V.I., Vyatkin Yu.V., Ramenskiy V.E., Shtokalo D.N. Trudy 11-y Moskovskoy konferentsii po vychislitel'noy molekulyarnoy biologii. MSSMV'23. Moskva: Institut problem peredachi informatsii RAN, 2023. URL: mccmb.belozersky.msu.ru/2023/thesis/abstracts/72.pdf.

2. Cheng K., Tahir R., Eric L. K., Li M. Multimedia Tools and Applications. 2020. V. 79. pp. 13725-13752.

3. Talalaev A.A., Tishchenko I.P., Fralenko V.P., Khachumov V.M. Iskusstvennyy intellekt i prinyatie resheniy. 2008. №2. pp. 24-33.

4. Kurbatov G.R. Tendentsii razvitiya nauki i obrazovaniya. 2023. №103-8. pp. 167–169.

5. Hinton J.E., Salakhoddinov R.R. Nauka. 2006. V. 313. № 5786. pp. 504-507.

6. Ballé J., Laparra V., Simoncelli E.P. International Conference on Learning Representation. 2017. URL: arxiv.org/abs/1611.01704.

7. Kingma D.P., Welling M. International Conference on Learning Representation. 2014. URL: arxiv.org/abs/1312.6114.

8. Toderici G., Vincent D., Johnston N., Hwang S.J., Minnen D., Shor J., Covell M. Conference on Computer Vision and Pattern Recognition. 2017. URL: arxiv.org/abs/1608.05148.

9. Ballé J., Minnen D., Singh S. International Conference on Learning Representation. 2018. URL: arxiv.org/abs/1802.01436.

10. Theis L., Shi W., Cunningham A., Husar F. International Conference on Learning Representation. 2017. URL: arxiv.org/abs/1703.00395.

11. Mentzer F., Agustsson E., Tshannen M. Conference on Computer Vision and Pattern Recognition. 2018. URL: arxiv.org/abs/1801.04260.

Дата поступления: 9.04.25 Дата публикации: 25.05.25