ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И ТЕЛЕКОММУНИКАЦИИ

EDN VTVPZV DOI 10.5281/zenodo.15585989
УДК 004.932.72

ИССЛЕДОВАНИЕ МЕТОДОВ ГЛУБОКОГО ОБУЧЕНИЯ ДЛЯ РАСПОЗНАВАНИЯ РУКОПИСНОГО ТЕКСТА

© 2025 ***Бондаренко В.И., Федоров М.Д.***

Проведен сравнительный анализ четырёх методов глубокого обучения (CNN, LSTM, Transformer, MLP) для распознавания рукописного текста. Проведено обучение и оценка моделей на наборах данных MNIST, EMNIST, IAM и RIMES. Результаты показывают, что CNN достигает наивысшей точности (99.04% на MNIST) и минимальных потерь, делая её наиболее эффективной для данной задачи. Предлагаются рекомендации по дальнейшему развитию моделей распознавания.

Ключевые слова: распознавание рукописного текста, глубокое обучение, CNN, LSTM, Transformer, MLP, MNIST, EMNIST, IAM, RIMES.

Введение. Распознавание рукописного текста является одной из ключевых задач в области искусственного интеллекта и компьютерного зрения. Несмотря на значительные успехи в распознавании печатного текста, проблема распознавания рукописного текста остаётся сложной из-за его высокой вариативности и сложности [1]. Существуют индивидуальные особенности почерка, различия в форме и наклоне букв, а также шумы, возникающие при сканировании, которые затрудняют разработку универсальных систем распознавания. Традиционные методы, основанные на алгоритмах машинного обучения, полагаются на ручное извлечение признаков и требуют значительной предварительной обработки данных. Эти подходы эффективны лишь в ограниченных сценариях, где текст стандартизирован, однако они часто не справляются с более сложными и вариативными данными.

Последние исследования в области глубокого обучения предлагают новые решения для распознавания рукописного текста. Свёрточные нейронные сети (CNN) продемонстрировали высокую эффективность в задачах обработки изображений, позволяя автоматически выделять иерархические признаки в данных [2, 3]. Рекуррентные нейронные сети LSTM зарекомендовали себя как один из наиболее эффективных подходов к распознаванию рукописного текста в автономном режиме благодаря их способности обрабатывать последовательности символов с учетом контекста. Архитектура трансформеров, изначально предложенная для задач обработки естественного языка, также находит применение в задачах распознавания рукописного текста, особенно в обработке последовательностей и контекстов. Многослойные перцептроны (MLP) используются как базовый метод глубокого обучения для классификации изображений, однако они менее приспособлены для выявления пространственных и последовательных признаков.

Несмотря на разнообразие предложенных подходов, остаётся открытым вопрос о том, какой алгоритм наиболее эффективен для распознавания рукописного текста на разных типах данных. Большинство существующих исследований фокусируются на применении одного алгоритма к конкретному набору данных, что затрудняет всестороннее сравнение их эффективности.

Более того, многие из этих работ не дают ясных рекомендаций относительно выбора алгоритма для определённых условий или типов данных, что создаёт пробел в понимании эффективности различных методов в реальных задачах распознавания.

Целью данной работы является проведение сравнительного анализа четырёх современных методов глубокого обучения (CNN, LSTM, Transformer, MLP) на четырёх разных наборах данных: MNIST, EMNIST, IAM и RIMES. Эти наборы данных различаются по сложности и структуре, представляя различные аспекты проблемы распознавания рукописного текста. Проведение такого анализа позволит выявить сильные и слабые стороны каждого метода, определить оптимальные подходы для конкретных типов данных и предложить рекомендации по применению глубокого обучения в задачах распознавания рукописного текста.

Для достижения этой цели необходимо:

– детально исследовать архитектурные особенности и возможности каждого из четырёх алгоритмов глубокого обучения в контексте задачи распознавания рукописного текста;

– провести подготовку и предварительную обработку данных для повышения эффективности обучения;

– обучить и оценить каждую модель на каждом из выбранных наборов данных с использованием метрик точности и потерь;

– провести сравнительный анализ результатов для выявления наиболее эффективного алгоритма для различных типов данных.

Принципы работы алгоритмов. Свёрточные нейронные сети (CNN) – это тип глубоких нейронных сетей, предназначенных для работы с изображениями и данными с сетчатой топологией. Основной принцип работы CNN заключается в использовании свёрточных слоёв, которые автоматически извлекают ключевые признаки из изображения, такие как края, текстуры и формы [2, 3]. Сеть применяет небольшие фильтры, которые скользят по изображению, выполняя операцию свёртки и создавая карты признаков, отражающие различные аспекты входных данных. После этого применяются операции пулинга, уменьшающие размерность данных и повышающие устойчивость к смещениям и искажениям. Наконец, сжатые признаки передаются в полносвязные слои для классификации. Эта архитектура делает CNN эффективными для распознавания образов, включая задачи распознавания рукописного текста, поскольку они способны улавливать пространственные и иерархические структуры в данных.

Рекуррентные сети с долгой кратковременной памятью (LSTM) – это тип рекуррентных нейронных сетей (RNN), специально разработанный для работы с последовательными данными, такими как текст или временные ряды. Принцип работы LSTM заключается в использовании ячеек памяти, которые способны сохранять и передавать информацию на протяжении длительных последовательностей, благодаря механизму управляемых гейтов. Эти гейты (входной, забывающий и выходной) контролируют поток информации, решая, какая информация должна быть сохранена, забыта или передана дальше [4]. Это позволяет LSTM улавливать долгосрочные зависимости в данных и эффективно работать с последовательностями, где контекст играет важную роль. В задачах распознавания рукописного текста LSTM помогает учитывать взаимосвязи между символами, улучшая точность распознавания сложных и неоднозначных последовательностей.

Трансформеры (Transformers) – это архитектура нейронных сетей, разработанная для эффективной обработки последовательных данных, таких как текст. В отличие от рекуррентных сетей, трансформеры обрабатывают всю последовательность одновременно, используя механизм внимания (attention), чтобы определить, какие части данных наиболее важны для текущего шага [5, 6]. Механизм внимания позволяет модели фокусироваться на разных частях входных данных, независимо от их положения в последовательности, улавливая как локальные, так и глобальные зависимости. Архитектура трансформера состоит из многократно повторяемых слоёв энкодера и декодера, которые обрабатывают и преобразуют входные данные. Это позволяет трансформерам эффективно распознавать сложные паттерны и взаимосвязи, делая их особенно полезными для задач, где контекст всей последовательности важен, например, для распознавания рукописного текста с учетом контекста и структуры.

Многослойный перцептрон (MLP) – это базовая архитектура нейронных сетей, состоящая из входного, одного или нескольких скрытых и выходного слоёв. Каждый слой состоит из нейронов, которые связаны с нейронами предыдущего слоя, создавая полносвязную структуру. Принцип работы MLP заключается в том, что входные данные проходят через эти слои, где на каждом шаге к ним применяются линейные преобразования, а затем функции активации, такие как ReLU или сигмоида, чтобы внести нелинейность в модель [7]. Это позволяет MLP учить сложные зависимости в данных. Несмотря на свою простоту, MLP эффективен в задачах классификации и распознавания, однако он не так хорош в извлечении пространственных или последовательных признаков, как CNN или LSTM. В контексте распознавания рукописного текста MLP может использоваться для классификации отдельных символов, но его эффективность снижается на более сложных задачах, требующих учета контекста или пространственных особенностей.

Для данной работы будут использованы четыре основных набора данных для распознавания рукописного текста: MNIST, EMNIST, IAM и RIMES. Каждый из них обладает уникальными характеристиками и предоставляет различные уровни сложности для задач распознавания. Детальное описание каждого набора данных представлено ниже.

Подготовка наборов данных. Для подготовки данных MNIST, содержащих 60,000 изображений рукописных цифр для обучения и 10,000 изображений для тестирования, были использованы файлы в формате IDX, сжатые в формате gzip (рис. 1).

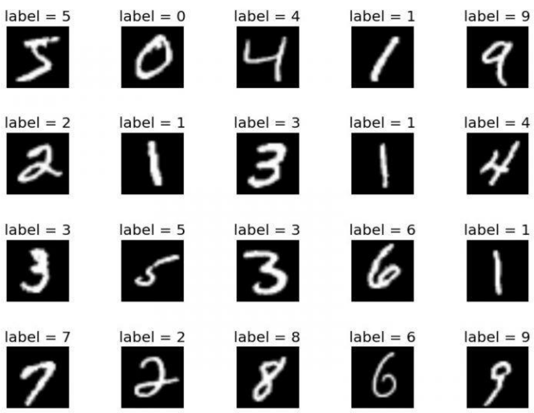

Рис. 1. Рукописные цифры из датасета MNIST

Сначала данные были загружены с использованием функций, которые читали изображения и метки из соответствующих файлов. Изображения были преобразованы в массивы NumPy и нормализованы путём деления на 255, чтобы все значения пикселей находились в диапазоне от 0 до 1. Такая нормализация помогает улучшить производительность моделей глубокого обучения, делая процесс обучения более стабильным. Метки были закодированы с использованием one-hot-кодирования, что представляет каждую цифру в виде вектора, где активен только один элемент, соответствующий правильному классу. После обработки обучающая выборка имела форму (60000,28,28,1), а тестовая – (10000,28,28,1). Дополнительное измерение размера 1 в конце формы используется для обозначения канала изображения, что необходимо для свёрточных нейронных сетей, которые ожидают входные данные в виде четырёхмерных тензоров.

Набор данных EMNIST представляет собой расширение MNIST, включающее как цифры, так и буквы английского алфавита (рис. 2).

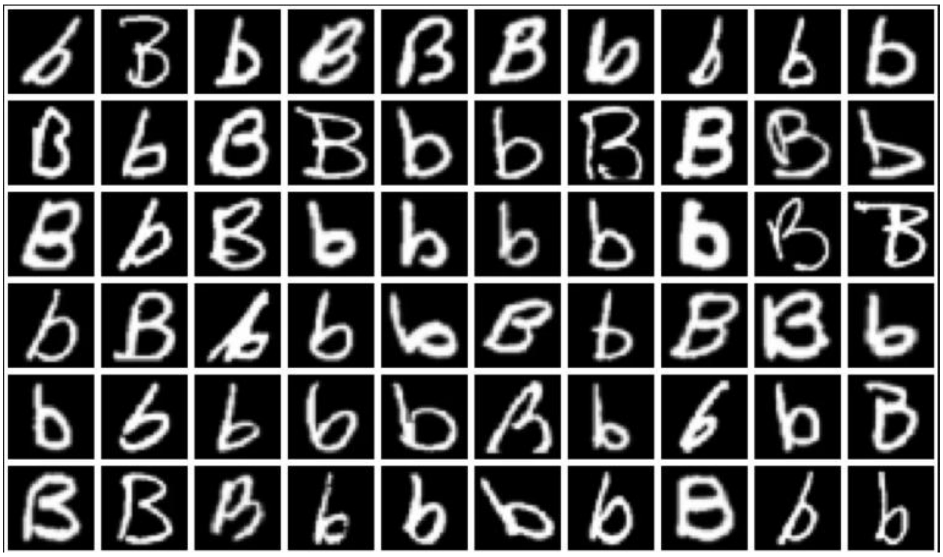

Рис. 2. Рукописные буквы b из датасета EMNIST

Он содержит 47 классов символов и используется для более сложных задач классификации [8]. Для подготовки данных EMNIST изображения и метки были загружены из сжатых IDX-файлов с помощью функций, аналогичных тем, что были использованы для MNIST. Изображения имели размер 28x28 пикселей, как и в случае MNIST, и были нормализованы путём деления на 255, чтобы значения пикселей находились в диапазоне от 0 до 1. После нормализации изображения были преобразованы в формат, совместимый с свёрточными нейронными сетями, путём добавления дополнительного измерения, что привело к форме (num_samples,28,28,1). Метки были также преобразованы в формат one-hot-кодирования для 47 классов, что позволяет моделям эффективно работать с ними. Обработанные данные имели формы (num_samples,28,28,1) для изображений и (num_samples,47) для меток.

Для набора данных IAM, который представляет собой сложный и реалистичный набор рукописных текстов, процесс подготовки данных был несколько сложнее. Изображения IAM содержат слова, строки и целые абзацы рукописного текста различного размера (рис. 3), поэтому они были предварительно обработаны для приведения к единому формату, подходящему для обучения моделей глубокого обучения.

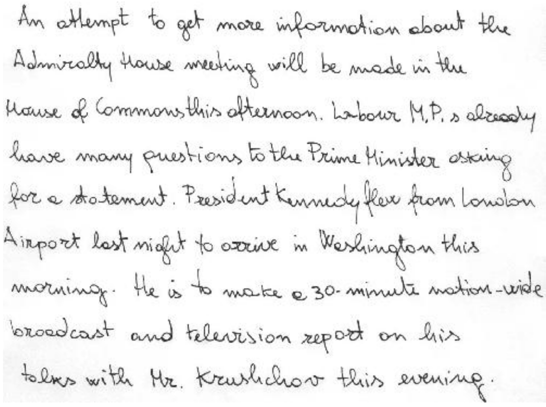
An attempt to get more information about the
Admiralty House meeting will be made in the
House of Commons this afternoon. Labour M.P.s already
have many questions to the Prime Minister asking
for a statement. President Kennedy flew from London
Airport last night to arrive in Washington this
morning. He is to make a 30-minute nation-wide
broadcast and television report on his
talks with Mr. Kruschov this evening.

Рис. 3. Абзац рукописного текста из датасета IAM

Изображения были масштабированы и обрезаны до размера 128х32 пикселей для отдельных слов. Масштабирование выполнялось так, чтобы сохранить пропорции, а пустые области заполнялись нейтральным фоном (белым). После масштабирования изображения были нормализованы путём деления на 255, как и в предыдущих наборах данных. Метки в IAM представлены в виде строк текста, поэтому для их использования в нейронных сетях они были преобразованы в числовые последовательности (с помощью кодировщиков символов).

Набор данных RIMES, содержащий рукописные французские тексты, включает сложные рукописные документы, такие как письма и формы (рис. 4).

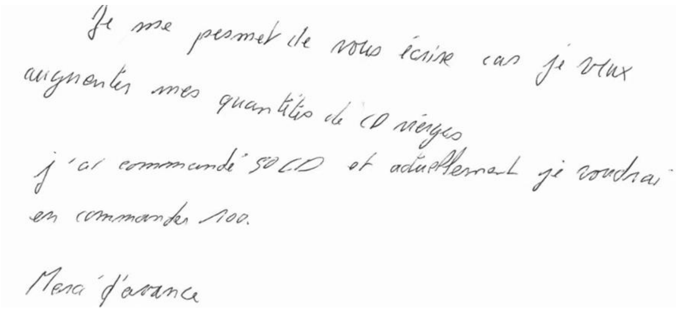
Je me permet de vous écrire car je veux
augmenter mes quantités de CD vierges
j'ai commandé 50 CD et actuellement je voudrai
en commander 100.
Merci d'avance

Рис. 4. Письмо, написанное от руки на французском языке из датасета RIMES

Для подготовки данных RIMES изображения были предварительно обработаны для обеспечения единообразия и совместимости с моделями глубокого обучения. Изображения были масштабированы до фиксированного размера, 128х64 пикселей, чтобы обеспечить одинаковую размерность входных данных. Если изображение было больше целевого размера, оно обрезалось, а если меньше – дополнялось нейтральным фоном. Масштабированные изображения были нормализованы путём деления на 255. Метки в RIMES представлены в виде строк текста, поэтому они также были преобразованы в числовые последовательности, чтобы сделать их пригодными для использования в процессе обучения моделей.

Обучение моделей. Рассмотрим процесс обучения модели с использованием всех алгоритмов на наборе данных MNIST [9]. Обучение каждой модели проводится в течение 5 эпох. Во время каждой эпохи модели обучаются на обучающей выборке и оценивались на тестовой выборке, при этом отслеживались метрики точности и потери. В конце каждой эпохи была выведена текущая точность (accuracy) и величина функции потерь (loss) как для обучающего, так и для тестового набора данных, что можно увидеть на рисунке 5.

```
Epoch 1/5
469/469 ━━━━━━━━━━━━━━━━━━━━ 10s 19ms/step - accuracy: 0.7930 - loss: 0.6547 - val_accuracy: 0.9808 - val_loss: 0.0585
Epoch 2/5
469/469 ━━━━━━━━━━━━━━━━━━━━ 8s 18ms/step - accuracy: 0.9702 - loss: 0.1036 - val_accuracy: 0.9869 - val_loss: 0.0387
Epoch 3/5
469/469 ━━━━━━━━━━━━━━━━━━━━ 8s 17ms/step - accuracy: 0.9776 - loss: 0.0733 - val_accuracy: 0.9890 - val_loss: 0.0343
Epoch 4/5
469/469 ━━━━━━━━━━━━━━━━━━━━ 8s 17ms/step - accuracy: 0.9822 - loss: 0.0594 - val_accuracy: 0.9901 - val_loss: 0.0315
Epoch 5/5
469/469 ━━━━━━━━━━━━━━━━━━━━ 8s 17ms/step - accuracy: 0.9855 - loss: 0.0497 - val_accuracy: 0.9909 - val_loss: 0.0269
```

Рис. 5. Динамика обучения модели CNN по эпохам

Модель CNN обучалась на протяжении 5 эпох, показывая стремительное увеличение точности. В первой эпохе точность модели достигла 79.30% на обучающей выборке и 98.08% на тестовой выборке. К пятой эпохе точность модели на обучающей выборке возросла до 98.55%, а на тестовой – до 99.09%. Потери также существенно уменьшились с каждой эпохой, указывая на улучшение качества модели. Общая точность модели на тестовой выборке составила 99.09%, что свидетельствует о её высокой эффективности в распознавании рукописных цифр.

Общие точности моделей, полученные после 5 эпох обучения, распределились следующим образом: CNN достигла наивысшей точности – 99.04%, MLP также продемонстрировала высокую точность – 97.96%. Модель LSTM немного уступила, показав итоговую точность 97.44%. Модель Transformer, несмотря на заметное улучшение в ходе обучения, оказалась менее точной по сравнению с остальными алгоритмами, достигнув 95.72%.

На основе результатов обучения были построены матрицы ошибок (Confusion Matrices) для каждой модели. На рисунке 6 представлена матрица ошибок для CNN, которая демонстрирует, как модель классифицировала каждую цифру.

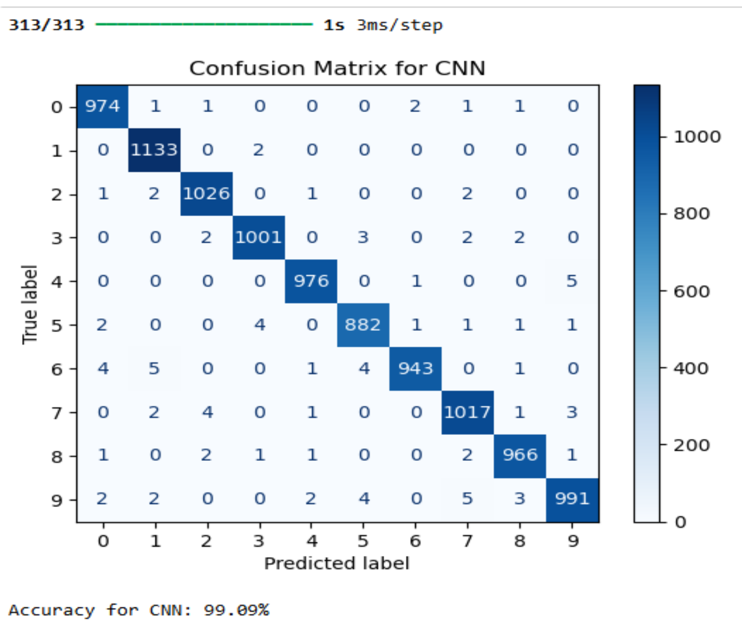

Рис. 6. Матрица ошибок для CNN

Большинство значений находятся на главной диагонали, что указывает на высокую точность классификации. Ошибки, хотя и редкие, указывают на случаи, когда модель спутала одну цифру с другой. Кроме того, были построены графики сравнения точности (рис. 7) и потерь всех моделей (рис. 8) за 5 эпох.

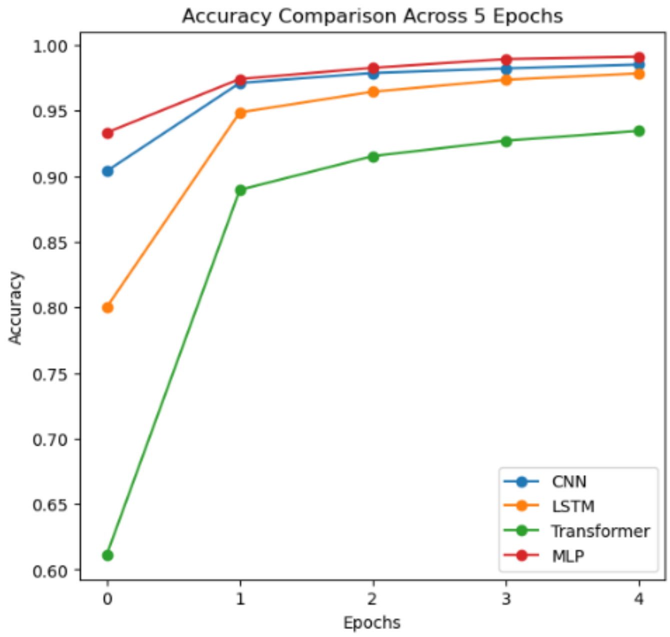

Рис. 7. Сравнение точности обучения различных алгоритмов за 5 эпох

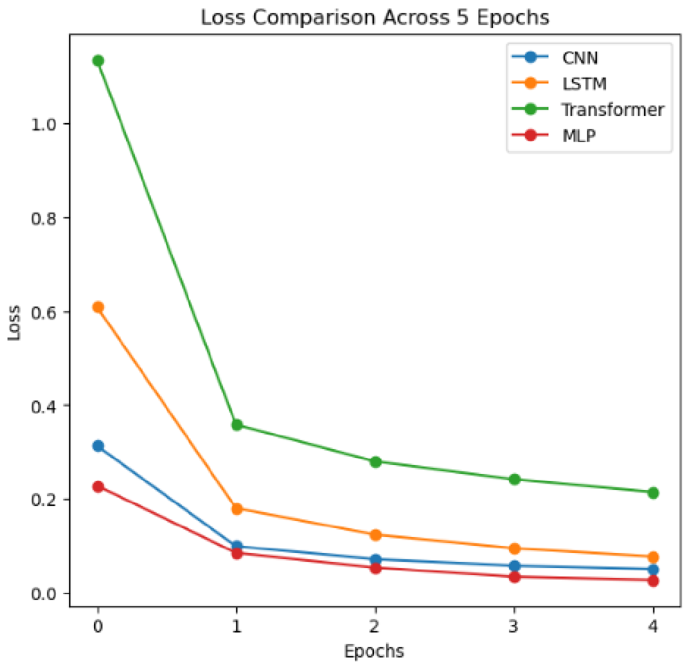

Рис. 8. Сравнение потерь обучения различных алгоритмов за 5 эпох

На графике точности видно, что CNN и MLP достигают наивысшей точности быстрее всех, в то время как LSTM следует за ними. Transformer показывает более медленный рост точности, но также достигает хороших результатов к пятой эпохе. График потерь демонстрирует аналогичную картину: потери уменьшаются быстрее всего у CNN и MLP, что свидетельствует о более эффективном обучении.

Таким образом, свёрточная нейронная сеть (CNN) продемонстрировала наилучшие результаты в распознавании рукописных цифр на наборе данных MNIST, достигнув точности 99.09%. Другие модели, такие как LSTM, Transformer и MLP, также показали хорошие результаты, однако уступили CNN в данной задаче.

Аналогичным образом были обучены и протестированы модели на наборах данных EMNIST, IAM и RIMES для всесторонней оценки эффективности каждого алгоритма в различных сценариях распознавания рукописного текста.

В таблице 1 представлено сравнительное измерение точности и потерь для каждой модели на различных наборах данных, при этом процесс обучения и оценки производился на локальной машине с использованием CPU Intel Core i7 6700K и 16GB RAM.

Таблица 1. Сравнение точности и потерь алгоритмов на наборах данных

Алгоритм	Набор данных	Точность (%)	Потери (%)
CNN	MNIST	99.04	2.69
	EMNIST	86.87	2.82
	IAM	85.34	4.81
	RIMES	87.81	4.06
LSTM	MNIST	97.44	3.52
	EMNIST	84.76	8.52
	IAM	81.57	9.12
	RIMES	83.08	8.52
Transformer	MNIST	95.72	4.23
	EMNIST	83.56	12.91
	IAM	79.46	10.18
	RIMES	81.49	9.51
MLP	MNIST	97.96	3.87
	EMNIST	84.23	7.14
	IAM	75.89	10.67
	RIMES	79.27	9.75

Выводы. Результаты сравнительного анализа четырёх алгоритмов глубокого обучения (CNN, LSTM, Transformer и MLP) на различных наборах данных (MNIST, EMNIST, IAM, RIMES) демонстрируют, что наиболее эффективно в задаче распознавания рукописного текста выступает свёрточная нейронная сеть (CNN). На всех четырёх наборах данных CNN достигает наивысшей точности, особенно выделяясь на наборе данных MNIST с точностью 99.04% и сравнительно низким уровнем потерь — 2.69%. Это указывает на способность CNN эффективно выделять иерархические признаки в изображениях, что делает её наиболее подходящим выбором для задач распознавания рукописных символов.

LSTM также показывает хорошие результаты, особенно на наборе данных MNIST, достигая точности 97.44%. Однако, её производительность на более сложных наборах данных (EMNIST, IAM, RIMES) заметно снижается, с точностью в диапазоне от 81.57% до 84.76% и более высокими потерями, особенно на наборе данных IAM, где потери достигают 9.12%. Это свидетельствует о том, что LSTM, несмотря на свою эффективность в обработке последовательностей, менее приспособлена к задачам распознавания рукописного текста, требующим анализа пространственных признаков.

Трансформерная модель показывает относительно низкую производительность, особенно на более сложных наборах данных (IAM и RIMES), с точностью 79.46% и 81.49% соответственно и высокими потерями (более 10%). Это указывает на то, что в задачах, требующих детального анализа изображений рукописного текста, трансформеры уступают специализированным моделям, таким как CNN.

MLP также уступает по эффективности, особенно на сложных наборах данных. Несмотря на высокую точность на MNIST (97.96%), точность MLP на IAM и RIMES падает до 75.89% и 79.27% соответственно, при этом потери остаются высокими. Это свидетельствует о том, что MLP менее эффективна в выделении сложных признаков из изображений рукописного текста.

В целом, результаты указывают на преимущество свёрточных нейронных сетей (CNN) в задачах распознавания рукописного текста. CNN не только достигает высокой точности на всех наборах данных, но и демонстрирует низкий уровень потерь, что говорит о её стабильности и надёжности. LSTM и MLP могут быть эффективными для менее сложных задач, однако они уступают CNN при работе с более вариативными данными. Трансформерные модели в данном контексте оказываются наименее эффективными, что свидетельствует о необходимости дальнейшей адаптации их архитектуры для задач распознавания рукописного текста.

Перспективы дальнейших исследований заключаются в оптимизации существующих моделей, включая усовершенствование архитектуры трансформеров для обработки рукописного текста, а также в исследовании гибридных моделей, объединяющих преимущества CNN, LSTM и других подходов.

СПИСОК ИСТОЧНИКОВ

1. Wang, Y. Offline Handwritten Text Recognition Using Deep Learning: A Review / Y. Wang, W. Xiao, S. Li // Journal of Physics: Conference Series. – 2021. – 7 p. – DOI 10.1088/1742-6596/1848/1/012015.
2. Mishra, A. Handwritten Text Recognition Using Convolutional Neural Network / A. Mishra, S. Ram, C. Kavyashree // Computer Vision and Pattern Recognition. – 2023. – 6 p. – DOI 10.17148/IARJSET.2021.86148.
3. Перевышин, В. Г. Сравнение точности распознавания рукописных цифр методом k-ближайших соседей (KNN) и сверточной нейронной сетью (CNN) / В. Г. Перевышин, Р. В. Ковальчик // Вестник Донецкого национального университета. Серия Г: Технические науки. – 2023. – № 4. – С. 20-28. – EDN GOXWRE.
4. Sang, D. V. Improving CRNN with EfficientNet-like feature extractor and multi-head attention for text recognition / D. V. Sang, L. T. B. Cuong // Conference: The Tenth International Symposium. – 2019. – P. 285-290. – DOI 10.1145/3368926.3369689.
5. Attention Is All You Need / A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit // 31st Conference on Neural Information Processing Systems. – 2017. – 15 p. – DOI 10.35940/ijrte.D7811.118419.
6. Asha, K. Handwriting Recognition using Deep Learning based Convolutional Neural Network / K. Asha, H. K. Krishnappa // International Journal of Recent Technology and Engineering (IJRTE). – 2019. – V. 8. – № 4. – P. 4826-4828. – DOI 10.35940/ijrte.D7811.118419.
7. Pricope, T. V. A contextual analysis of multi-layer perceptronmodels in classifying hand-written digits and letters:limited resources / T. V. Pricope // Computer Vision and Pattern Recognition. – 2021. – 25 p. – DOI 10.48550/arXiv.2107.01782.
8. Liman, M. Handwritten Character Recognition Using Deep Learning Algorithm with Machine Learning Classifier / M. Liman, A. Josef, G. Kusuma // JOIV International Journal on Informatics Visualization. – 2024. – P. 150-157. – DOI 10.62527/joiv.8.1.1707.
9. Handwritten Digit Recognition with Python & CNN [Электронный ресурс] / TechVidvan: [Сайт]. – https://techvidvan.com/tutorials/handwritten-digit-recognition-with-python-cnn/ (дата обращения 07.09.2024).

Поступила в редакцию 29.01.2025 г., рекомендована к печати 03.03.2025 г.

RESEARCH ON DEEP LEARNING METHODS FOR HANDWRITTEN TEXT RECOGNITION

Bondarenko V.I., Fedorov M.D.

A comparative analysis of four deep learning methods (CNN, LSTM, Transformer, MLP) for handwritten text recognition was conducted. The models were trained and evaluated on the MNIST, EMNIST, IAM, and RIMES datasets. The results show that CNN achieves the highest accuracy (99.04% on MNIST) and the lowest loss, making it the most effective method for this task. Recommendations for further development of recognition models are proposed.

Keywords: handwritten text recognition, deep learning, CNN, LSTM, Transformer, MLP, MNIST, EMNIST, IAM, RIMES.

Бондаренко Виталий Иванович
кандидат технических наук, доцент, доцент кафедры компьютерных технологий ФГБОУ ВО «Донецкий государственный университет»,
Российская Федерация, ДНР, г. Донецк.
E-mail: bondarenko@donnu.ru

Bondarenko Vitalii Ivanovich
Candidate of Technical Sciences, Associate Professor, Associate Professor at Department of Computer Technologies of Donetsk State University,
Russian Federation, DPR, Donetsk.

Федоров Максим Дмитриевич
магистрант кафедры компьютерных технологий ФГБОУ ВО «Донецкий государственный университет»,
Российская Федерация, ДНР, Донецк
E-mail: www.maksim.azot.dn@gmail.com

Fedorov Maksim Dmitrievich
Master student at Department of Computer Technologies of Donetsk State University,
Russian Federation, DPR, Donetsk.