

Верификация автора рукописных документов на основе сиамской сети

Keywords: Сиамские нейронные сети, идентификация и верификация автора, рукописный текст.

Аннотация

В работе представлен метод верификации почерка определённого автора в корпусе рукописных документов на основании малого количества примеров, предложен алгоритм предобработки данных.

Предлагается использование сиамской нейронной сети для сравнения и анализа уникальных характеристик почерка, стиля письма. Такой способ обучения позволяет получать мощные дискриминационные признаки изображения, эмбеддинги, на основе которых можно произвести качественную классификацию автора.

Предложенный подход был применён к задаче верификации возможных автографов Жуковского среди рукописных текстов неизвестных авторов. Подход также был применён к задаче классификации на полностью размеченном датасете IAM.

1. Введение

Глубокие нейронные сети уже давно демонстрируют высокие результаты на задачах классификации изображений. Ярчайший пример - в 2015 году нейросеть ResNet (He et al., 2015) с 152 слоями показала меньшее количество ошибок на конкурсе ImageNet по сравнению с мануальной разметкой. Однако чем больше параметров у модели, тем выше риск переобучения и тем выше потребность в большем количестве данных. Эта проблема встаёт наиболее остро в задаче верификации почерка определённого автора в силу большого количества классов и малого числа представителей каждого класса.

В работах по идентификации автора рукописного текста задача часто сводится к анализу минимальных единиц письменности, графем (Bensefia et al., 2002, Bensefia et al., 2005, Koch et al., 2015). Учитывая специфику предоставленных фотографий, их низкое качество, выделить графемы не получится. Но этого и не требуется: для установления авторства нет необходимости распознавать все графемы рукописного текста. Выдвигается гипотеза, что автора можно узнать и в случае, когда невозможно прочитать написанное. В обычной жизни это делается интуитивно по общему стилю письма: по наклону и размеру текста, по уникальным штрихам, по пробелам между строками, по взаимному расположению строк и т. д. Поэтому в данной работе стиль почерка будет рассматриваться как паттерн изображения, а идентификация будет производиться по фрагменту с несколькими строками.

Актуальность задачи обусловлена тем, что машинный поиск текстов с почерком определенного автора в составе больших баз данных растровых изображений рукописных документов, с одной стороны, значительно расширяет возможности исследователей, в частности литературоведов или историков, в плане выявления в архивных собраниях текстов того или иного лица; с другой стороны, она дает архивным работникам удобный инструмент для автоматической классификации: выявленный программой реестр соответствий почерка

может служить основой для описания рукописи (или ее состава) и внесения соответствующей информации в каталог.

2. Сиамские нейронные сети

Решить главную проблему данных, их нехватку, предполагается с помощью алгоритма однократного обучения — сиамской нейронной сети. Сиамские сети были впервые представлены в начале 1990-х Bromley и LeCun для решения задачи верификации подписи (Bromley et al., 1993b). Она является типом нейронной сети для глубокого обучения, которая использует две или больше идентичных подсети с одинаковой архитектурой. Также они используют одни параметры для обучения.

Сиамские сети особенно полезны в случае классификации с большим количеством классов и с небольшим числом объектов каждого класса. В таких случаях недостаточно примеров каждого класса, чтобы обучить глубокую свёрточную нейронную сеть. К тому же при добавлении новых классов пришлось бы менять архитектуру сети и переобучать. Вместо этого сиамская сеть учится на задачу бинарной классификации пар объектов: принадлежат ли объекты одному классу или нет. Это делает их особенно полезными в задачах, где требуется гибкость и эффективность при ограниченном наборе данных.

Данный вид сетей позволяет получить векторы признаков, эмбеддинги, двух объектов, отражающие их семантическое сходство или различие. Примеры приложений для сиамских сетей: верификация подписи (Bromley et al., 1993a), распознавание лиц (Solomon et al., 2024), идентификация перефразирования предложений (Yin and Schütze, 2015).

Метки класса получаются после обучения простого классификатора на эмбеддингах, в данной работе будет использоваться двухслойная полно связная нейронная сеть. Аналогичный метод классификации символов был использован в (Koch et al., 2015).

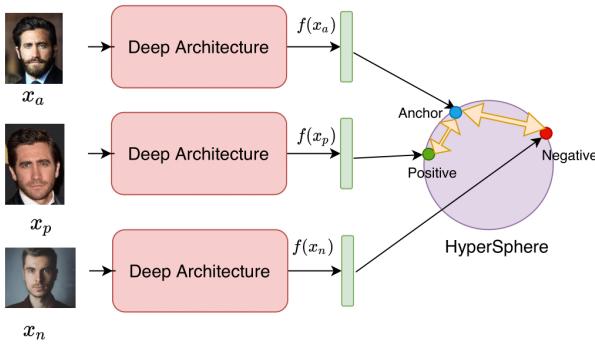


Рис. 1. Обучение сиамской сети с помощью triplet loss
(источник: Shivam Chandhok)

Две наиболее популярные функции потерь для обучения сиамских нейронных сетей — contrastive loss function (Chopra et al., 2005, Hadsell et al., 2006) и triplet loss function (Schroff et al., 2015).

2.1 Contrastive loss function

Contrastive loss использует пару объектов, они могут быть и положительного, и отрицательного класса.

$$L(x, y) = (1 - Z(x, y)) d^2(x, y) + Z(x, y) \max(0, margin - d^2(x, y)) \quad (1)$$

$$Z(x, y) = \begin{cases} 0 & | x, y \text{ из одного класса} \\ 1 & | x, y \text{ из разных классов} \end{cases} \quad (2)$$

$$d(x, y) = \|x - y\|_p \quad (3)$$

Объекты одного класса штрафуются для минимизации расстояние между ними, объекты из разных классов штрафуются, если расстояние меньше отступа *margin*.

2.2 Triplet loss function

Улучшением contrastive loss является triplet loss. В отличии от contrastive loss здесь используется три объекта: объект рассматриваемого класса (*anchor*), с которым будет проводиться сравнение, а также два других объекта: принадлежащий к тому же классу (*positive*), и объект противоположного класса (*negative*).

$$L(a, p, n) = \max\{d(a, p) - d(a, n) + margin, 0\} \quad (4)$$

Функция стремится приблизить объекты одного класса и увеличить расстояние между объектами разных классов. Также функция не штрафует, если уже достигнуто требуемое соотношение расстояний между тремя объектами. *margin* — заранее задаваемый параметр, показывающий, за какую разницу расстояний следует штрафовать.

При обучении модели с triplet loss требуется меньше образцов для сходимости, поскольку сеть обновляется одновременно, используя как похожие, так и непохожие образцы. Поэтому в работе будет использоваться данная функция потерь.

3. Задача верификации почерка автора

Рукописи В.А. Жуковского, великого русского поэта, хранятся во множестве архивов России и зарубежных стран. Между тем архивные описи и каталоги редко достигают исчерпывающей полноты в плане раскрытия содержания отдельных единиц хранения. Большую «темную зону» здесь представляют прежде всего конволюты, подборки рукописных документов, описываемые как единое целое, но включающие в себя отдельные тексты, в том числе разных авторов. Оцифровка рукописей, ведущаяся в данных архивах с разной степенью интенсивности, позволяет в обозримой временной перспективе применить к разбору их содержания разработанные программные методы атрибуции. Более полное описание контекста данной проблемы дано в статье (Kiselev V.S. et al., 2023).

3.1 Постановка задачи

В техническом плане условия задачи следующие: дана небольшая коллекция почерков Жуковского, его автографов (рис. 2) разных периодов жизни — раннего, зрелого и позднего, разной степени аккуратности — идеальный, обычный, неаккуратный. Всего - 25 изображений.

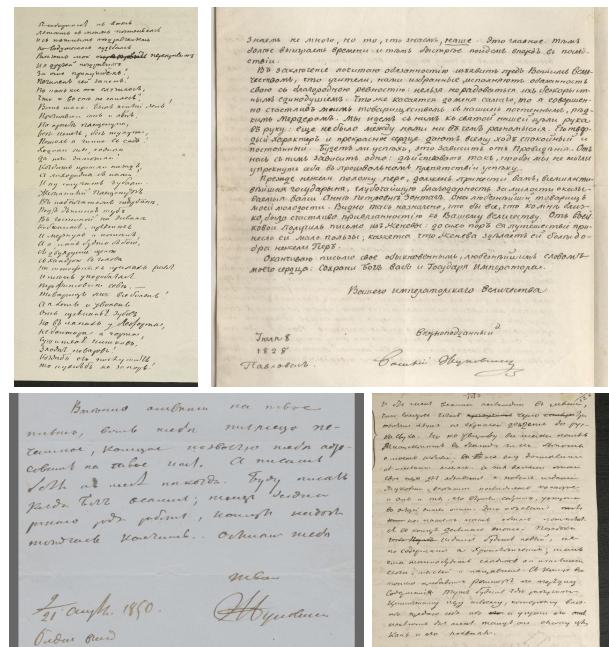


Рис. 2. Autographs

Так же дан набор снимков подборок рукописных документов неизвестных авторов, конволют (рис. 3), всего - 222 изображения. Требуется в подборке изображений определить документы, с высокой степенью вероятности принадлежащих руке Жуковского.

Формальная постановка задачи: для каждого конволюта требуется определить вероятность того, что этот документ является автографом Жуковского. Далее по пороговому значению вероятности пользователь сможет отобрать топ документов для ручной экспертной проверки.

Особенности предоставляемых данных:

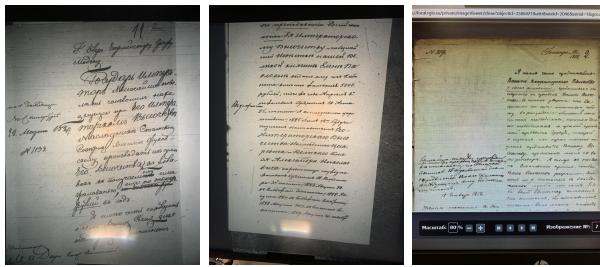


Рис. 3. Convolutes

1. Нехватка размеченных данных. Только у 25 объектов установлен автограф.
2. Автографы и конволюты имеют разный фон.
3. Почерки имеют разный масштаб.
4. Конволюты оцифрованы намного хуже, чем автографы.
5. Изображения из конволютов имеют искажение в виде муарового узора, так как сняты с экрана, возможны блики.

3.2 Предобработка данных

Повысим расширение изображения с помощью фотоусилителя (Picsart). Чтобы предотвратить переобучение на фон, бинаризуем изображения с помощью трансформера DocEnTr (Souibgui et al., 2022). На рисунке 4 приведён пример бинаризации автографов и изображений из конволютов.

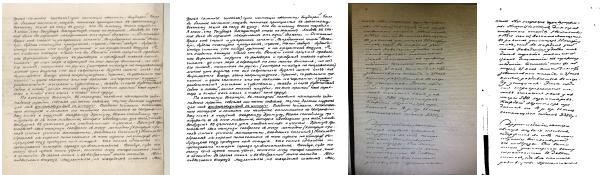


Рис. 4. Binarization

Конволюты сняты примерно в одинаковом масштабе и размер почерков отличается не сильно, чего нельзя сказать про автографы. Поэтому для всех автографов вручную изменим размер так, чтобы во фрагменте высотой 300 пикселей помещалось 7-8 строк.

3.3 Обучающая и валидационная выборка

Предполагается, что среди конволютов не более 2 % автографов Жуковского. Поэтому возьмём все конволюты в качестве отрицательного класса и автографы Жуковского — в качестве положительного.

Для предотвращения утечки данных в валидационную выборку каждая фотография была разрезана пополам. Если высота фотографии больше ширины, то верхняя часть была отнесена в обучающую выборку и нижняя — в валидационную. Если ширина фотографии больше высоты, то левая часть — в обучающую выборку и правая — в валидационную.

Чтобы увеличить количество позитивных и негативных объектов и сбалансировать классы в выборках была

проведена аугментация с помощью случайного вырезания фрагмента 300 x 300 пикселей и случайного преобразования перспективы со степенью искажения 0.3. Для предотвращения попадания белых изображений в выборки аугментация производилась таким образом, чтобы доля белых пикселей была не более 95%. Количество положительных экземпляров увеличилось в 40 раз, количество отрицательных — в 5 раз. Итого: 1000 позитивных, 1110 негативных — в обучающей выборке, столько же в валидационной выборке.

3.4 Генерация вероятности

Общий метод получения вероятности принадлежности к положительному классу:

1. Обучение модели на задачу бинарной классификации (функция потерь — кросс энтропия).
2. Получение предсказания модели для отрицательного класса.
3. Применение функции softmax, получение числа из вероятностного симплекса.
4. Калибровка вероятности.
5. Выделение топа объектов отрицательного класса с наибольшей вероятностью. Далее будем называть такие объекты «подозрительными» изображениями.

3.5 Калибровка вероятности

Предположим, что бинарный классификатор выдаёт некоторую оценку принадлежности объекта к положительному классу. Даже если оценка принадлежит вероятностному симплексу, она может плохо оценивать реальную вероятность того, что объект принадлежит к положительному классу. Как следствие, ответ классификатора плохо интерпретируем.

Назовём классификатор хорошо откалиброванным, если для классификатора a , объекта x класса y выполняется:

$$a(x) \approx P(y|x) = 1 \quad (5)$$

То есть если группе объектов классификатор дал оценку 0.9, то в случае хорошей калибровки в этой группе будет примерно 90% объектов положительного класса. В реальности алгоритм вряд ли большой группе объектов даст идентичную оценку, поэтому все объекты делят на группы по величине вероятности, бины.

Для оценки откалиброванности классификатора строится калибровочная кривая: по оси x откладывается средняя предсказанная вероятность в каждом бине, по оси y — доля объектов в каждом бине, чей класс является положительным. Таким образом, для хорошо откалиброванного классификатора калибровочная кривая является прямой линией.

3.6 Histogram Binning

Применим один из самых простых и универсальных методов калибровки — непараметрический метод гистограммной калибровки, Histogram Binning (Zadrozny and Elkan, 2001, Guo et al., 2017). Решается задача оптимизации по параметрам θ :

$$\theta = \arg \min_{\theta} \sum_{m=1}^M \sum_{i=1}^n [a_i \in B_m] (y_i - \theta_m)^2 \quad (6)$$

где a_i — оценка классификатора
 y_i — истинный класс i -го объекта
 B_1, \dots, B_M — бины

Аналитическое решение задачи: θ_m соответствует среднее значение оценок вероятности положительного класса, попавших в B_m . Обычно в методе используются бины одинаковой ширины. После решения задачи, если оценка попала в i -ый бин, она заменяется на соответствующее значение θ_i .

Минусы метода:

1. Есть гиперпараметр — число бинов.
2. Преобразование вероятности не является непрерывным.
3. Если используются бины равной ширины, то в некоторых бинах может содержаться малое число объектов.

3.7 Метод решения

Применим идею обучения сиамской сети на бинарных изображениях. Возьмём предобученный ResNet18. Последний слой имеет размерность 1000 на выходе, поэтому сиамская сеть будет выучивать 1000-размерный эмбеддинг изображения. Будем обучать последние два слоя, 513000 обучаемых параметров. Построим датасет из 8000 троек изображений обучающей выборки, и датасет из 2000 троек валидационной выборки: в качестве anchor и positive будем брать случайные элементы позитивного класса, в качестве negative — случайный элемент негативного класса.

На графиках 5 приведены значения triplet loss и точность на обучении. Точность считается как доля триплетов, для которых евклидово расстояние между эмбеддингами anchor и positive меньше, чем между anchor и negative. Точность на обучении — 99%, на валидации — 90.75%.

На полученных эмбеддингах обучим классификатор: двуслойную полно связную нейронную сеть с 513 538 параметрами. Точность на обучении — 100%, на валидации — 97.63% (рис. 6).

По матрицам ошибок (рис. 7) видно, что на обучающей выборке модель не ошибается, на валидации — 3% ошибок.

Построим калибровочные кривые с 20 бинами. По графику 8 видно, что на валидационной выборке кривая

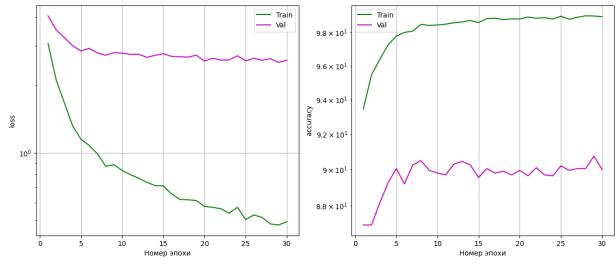


Рис. 5. Triplet loss and accuracy

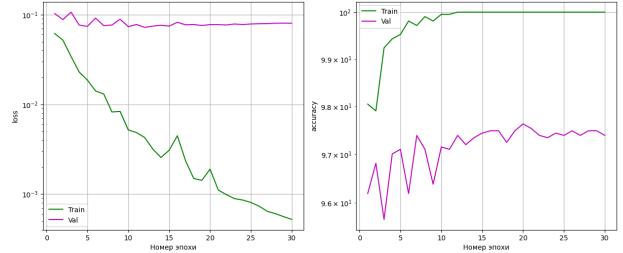


Рис. 6. Loss and accuracy

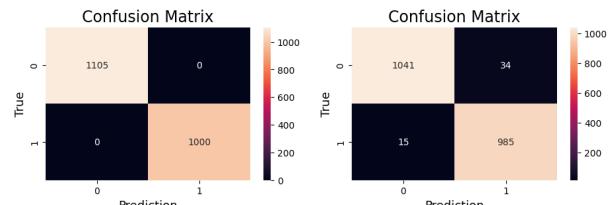


Рис. 7. Матрицы ошибок на обучающей и валидационной выборках

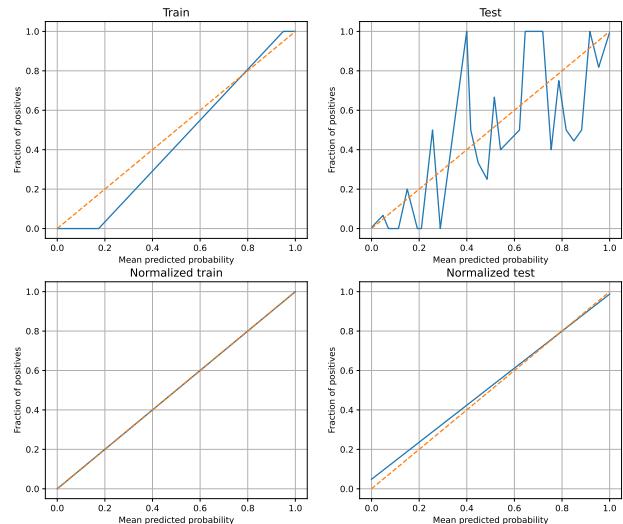


Рис. 8. Histogram Binning

далека от прямой. После калибровки вероятностей с помощью метода Histogram Binning классификатор стал хорошо откалиброванным.

На рисунке 8 приведены распределения вероятности внутри положительного и отрицательного класса до и после калибровки. Распределения получились силь-

но прижаты к краям, есть ярко выраженные выбросы. До калибровки классификатор страдал от «чрезмерной уверенности» предсказания.

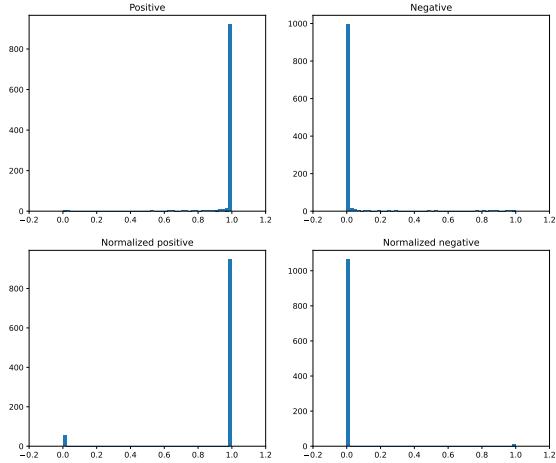


Рис. 9. Распределения вероятности

«Подозрительные» изображения так же получились хорошего качества (рис. 10).

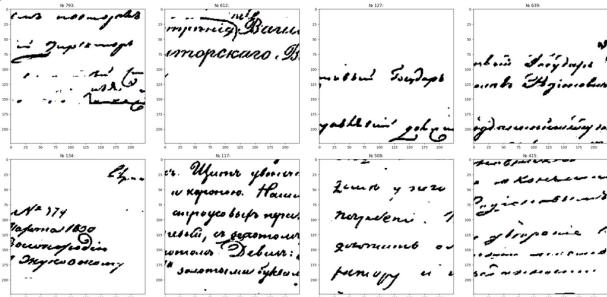


Рис. 10. Топ изображений с максимальной вероятностью

Был проведён аналогичный эксперимент для эмбеддингов сиамской сети меньшей размерности, 128. Качество получилось незначительно хуже: 90.1% - точность сиамской сети, 96.3% - точность классификации эмбеддингов, «подозрительными» были классифицированы в основном те же изображения.

3.8 Результат

Построенная модель принимает на вход фрагмент изображения, но требуется получать вероятность для всего изображения. Применим простую идею: вырежем 10 случайных фрагментов 300 x 300 пикселей из изображения, для каждого фрагмента получим предсказание модели. Итоговой вероятностью изображения будем считать максимум из 10 предсказаний.

4. Задача идентификации почерка автора

Так как в предыдущей задаче слишком мало точной разметки почерка авторов, мы не можем объективно оценить качество результата.

Протестируем предложенный подход на аналогичной задаче классификации почерков на корпусе IAM

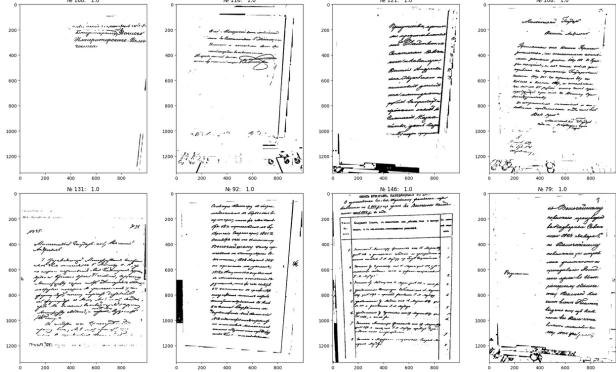


Рис. 11. «Подозрительные» изображения



Рис. 12. IAM dataset

(U. Marti, 2002). Он состоит из рукописных английских предложений, основанных на корпусе Lancaster-Oslo/Bergen (Stig et al., 1978). IAM содержит 1539 картинок с 657 авторами. Благодаря общедоступности, гибкой структуре и большому количеству писателей данные IAM широко используются для идентификации, верификации латинских писателей, распознавания рукописного текста.

Эксперимент будет проводиться на укороченном IAM, так как для подавляющего большинства авторов приведён один пример. Отберём только тех авторов, у которых не менее пяти объектов. Таким образом, в укороченный IAM содержится 397 картинок с 39 авторами. Обрежем верхнюю и нижнюю часть изображения с печатным текстом и подписью автора, чтобы они не влияли на предсказание. Примеры изображений почерка разных авторов приведены на рисунке 12.

Сиамская сеть с размерностью эмбеддинга 1000 обучалась на наборе из 1000 триплетов и достигла точности 100% на валидации (рис. 13). Точность двухслойного классификатора на 39 классов — 98.75% (рис. 14).

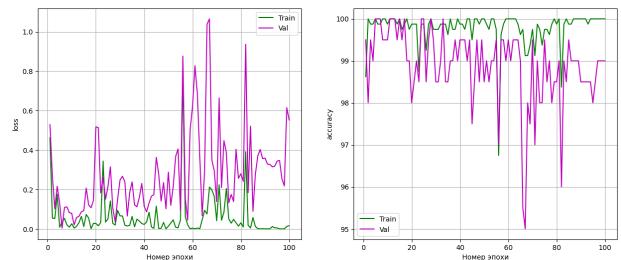


Рис. 13. Triplet loss и accuracy

Такая высокая точность по сравнению с результатом на конволютах объясняется лучшим качеством датасета IAM: в нем нет проблемы разного фона, масштаба написания и расположения текста.

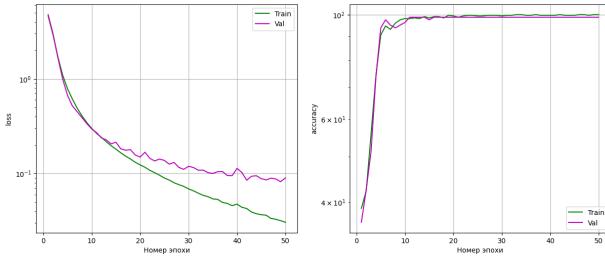


Рис. 14. Loss и accuracy

5. Заключение

В данной работе был предложен подход к верификации почерка определённого автора в корпусе исторических документов на основании малого количества образцов. Были поставлены эксперименты, подтверждающие его эффективность. В будущем предложенный метод может быть улучшен путём перехода от классификации фрагмента изображения с несколькими строками к классификации одной строки рукописного текста, это поможет значительно увеличить выборку и устранить проблему разного масштаба почерка.

Список литературы

- Bensefia, A., Nosary, A., Paquet, T., Heutte, L., 2002. Writer identification by writer's invariants. Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition, IEEE, 274–279.
- Bensefia, A., Paquet, T., Heutte, L., 2005. A writer identification and verification system. Pattern Recognition Letters, 26(13), 2080–2092.
- Bromley, J., Bentz, J., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Sackinger, E., Shah, R., 1993a. Signature Verification using a "Siamese" Time Delay Neural Network. International Journal of Pattern Recognition and Artificial Intelligence, 7, 25.
- Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R., 1993b. Signature verification using a "siamese" time delay neural network. Advances in neural information processing systems, 6.
- Chopra, S., Hadsell, R., LeCun, Y., 2005. Learning a similarity metric discriminatively, with application to face verification. 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), 1, IEEE, 539–546.
- Guo, C., Pleiss, G., Sun, Y., Weinberger, K. Q., 2017. On calibration of modern neural networks.
- Hadsell, R., Chopra, S., LeCun, Y., 2006. Dimensionality reduction by learning an invariant mapping. 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06), 2, IEEE, 1735–1742.
- He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition.
- Kiselev V.S., Lebedeva O.B., T. E. et al., 2023. The problem of machine identification of texts with the handwriting of a specific author as part of large databases of raster images of handwritten documents (based on the experience of identifying letters from Vasily Zhukovsky in office convolutes of the Russian State Historical Archive). Imagologiya i komparativistika – Imagology and Comparative Studies, 20, 247–262.
- Koch, G., Zemel, R., Salakhutdinov, R. et al., 2015. Siamese neural networks for one-shot image recognition. ICML deep learning workshop, 2, Lille.
- Schroff, F., Kalenichenko, D., Philbin, J., 2015. Facenet: A unified embedding for face recognition and clustering. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE.
- Solomon, E., Woubie, A., Emiru, E. S., 2024. Deep learning based face recognition method using siamese network.
- Souibgui, M. A., Biswas, S., Jemni, S. K., Kessentini, Y., Fornés, A., Lladós, J., Pal, U., 2022. Docentr: An end-to-end document image enhancement transformer.
- Stig, J., Leech, G. N., Goodluck, H., 1978. Manual of information to accompany the lancaster-oslo: Bergen corpus of british english, for use with digital computers. (No Title).
- U. Marti, H. B., 2002. The iam-database: An english sentence database for off-line handwriting recognition. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5, IEEE, 39 – 46.
- Yin, W., Schütze, H., 2015. Convolutional neural network for paraphrase identification. Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 901–911.
- Zadrozny, B., Elkan, C., 2001. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. Icml, 1, 609–616.