*СЕНСОРНЫЕ СИСТЕМЫ, 2019, том 33, № 1, с. 65–69*

## РАСПОЗНАВАНИЕ ОБРАЗОВ

*УДК 004.93'11*

# СРАВНЕНИЕ КЛАССИФИЦИРУЮЩЕЙ И МЕТРИЧЕСКОЙ СВЁРТОЧНЫХ СЕТЕЙ НА ПРИМЕРЕ РАСПОЗНАВАНИЯ ПОЛЯ “ПОЛ” ПАСПОРТА ГРАЖДАНИНА РФ

**© 2019 г. А. Н. Чирвоная[1,*], А. Е. Лынченко[2], Ю. С. Чернышова[2,3], А. В. Шешкус[2,3]**

*[1] Национальный исследовательский технологический университет “МИСиС”*
*119049, Москва, Ленинский проспект, д. 4, Россия*

*[2] ООО “Смарт Энджинс Сервис”, 117312 Москва, проспект 60-летия Октября, 9, Россия*

*[3] Федеральное государственное учреждение*
*“Федеральный исследовательский центр “Информатика и управление” Российской академии наук”*
*119333 Москва, Вавилова, д. 44, кор. 2, Россия*

**E-mail: nastyachirvonaya@smartengines.biz*

Поступила в редакцию 16.09.2018 г.

В работе рассматривается использование метрических нейронных сетей в задаче распознавания изображений слов. Подход к распознаванию слов, основанный на распознавании отдельных букв, хорошо изучен, но плохо применим к некоторым видам текста. Действительно, рукописные, написанные арабским языком или имеющие лигатуры тексты трудно сегментируются на буквы. Кроме этого, в тексте могут появляться слипшиеся символы, если изображения сильно зашумлены и/или искажены из-за несовершенства камеры. Все эти проблемы возникают в системах распознавания текста с заданным шаблоном, где набор слов может быть ограничен. В таких случаях разумно распознавать слова целиком, хотя словарь ответов может быть большим и не обязательно известным на этапе обучения. Для решения задачи распознавания изображений слов мы предлагаем использовать метрическую нейронную сеть. В работе приводится сравнение качества распознавания метрической нейронной сети со стандартной классифицирующей на словах, собранных с поля “пол” паспорта гражданина РФ. Параметры всех слоев, кроме последнего, у метрической и классифицирующей сетей были сделаны одинаковыми для обеспечения чистоты эксперимента. Результаты показывают пригодность метрических нейронных сетей для решения задачи распознавания слов. Основными преимуществами предлагаемого метода являются возможность расширения алфавита сети уже после обучения и отсутствие необходимости сегментировать слово на символы.

*Ключевые слова:* распознавание текста, свёрточные нейронные сети, глубокое обучение, сиамские нейронные сети, обучение метрики

**DOI:** 10.1134/S0235009219010049

## ВВЕДЕНИЕ

На протяжении последних десятилетий задачи распознавания образов являются одними из самых актуальных. Одна из таких задач – распознавание текста. Перевод информации на электронные носители значительно ускоряет процесс обработки и передачи, а также способствует более надежному ее хранению. В связи с этим возникла задача автоматического ввода информации в электронные устройства. Для перевода информации в цифровой вид разрабатывается множество систем, в том числе и системы распознавания документов, удостоверяющих личность (Bulatov et al., 2018).

Обычно задачу распознавания текста можно разделить на несколько подзадач: нахождение текста на изображении, сегментация на символы и распознавание символов. Каждая из этих подзадач сама по себе имеет множество способов решения, например, в работе (Чернов и др., 2016) авторы проводят сравнительный анализ двух методов сегментации: основанный на анализе структурных элементов изображения и основанный на применении глубокого обучения. В работе (Venkata Rao et al., 2016) анализируются как различные методы распознавания символов, так и методы их поиска и локализации. Принципиально другим подходом является распознавание слов целиком, без сегментации на отдельные символы (Jaderberg, 2014). При этом в качестве словаря классов может быть использован как полный набор всех возможных слов, так и его часть, но тогда требуется посимвольное распознавание объектов, не указанных в словаре.

**Таблица 1.** Варианты заполнения графы “пол” в паспорте РФ

| | | | | | |
|---|---|---|---|---|---|
| МУЖ | МУЖ. | муж | муж. | Муж. | МУЖСКОЙ |
| ЖЕН | ЖЕН. | жен | жен. | Жен. | ЖЕНСКИЙ |

В настоящее время существует множество различных видов нейронных сетей, решающих разные типы задач. Например, многослойные персептроны состоят из полносвязных слоев, в которых каждый элемент выхода зависит от всех элементов входа. В работе (Лёзин, Соловьёв, 2016) показано применение такого вида сетей для сжатия изображений. Однако с задачами распознавания текста такие сети справляются хуже, чем свёрточные (Lecun et al., 2001) при равном количестве обучающих данных, так как они более склонны к переобучению, что отражено в (Hastie et al., 2009). Автор работы (Прохоров, 2008) приводит сравнение типов сетей при применении их к распознаванию рукописных символов на одном и том же наборе данных. Важной особенностью свёрточных нейронных сетей является то, что за счет применения фильтров к отдельным частям картинки, а не ко всей целиком, они способны учиться выделять признаки вне зависимости от положения объекта на изображении. Именно такие сети мы будем использовать для проведения эксперимента, так как нам важна устойчивость сети к различным сдвигам объекта и другим искажениям изображения, типичным при съемке в неконтролируемых условиях (Арлазаров и др., 2014).

Одной из разновидностей свёрточных сетей являются энкодеры, или кодировщики, или метрические сети. Если обычные классифицирующие сети каждому входному изображению ставят в соответствие вектор размера алфавита сети, где каждый элемент представляет собой оценку соответствующего элемента алфавита, то метрические сети формируют некоторое описание входного объекта, которое сравнивается с описаниями “эталонных” представителей классов. Итоговый ответ сети определяется исходя из степени близости полученного вектора к элементам из набора эталонов. Именно такие сети используются в работе (Liu et al., 2018) для распознавания текста на изображениях. Для обучения метрической сети


**Рис. 1.** Варианты сгенерированных слов из обучающей выборки.

могут использоваться сиамские сети, которые представляют собой пару ветвей с идентичными весами, соединенных одним или несколькими слоями, на которых происходит сравнение векторов признаков (Koch et al., 2015).

В данной работе рассматривается метод распознавания целых слов при помощи свёрточных нейронных сетей при наличии полного словаря и приводится сравнение характеристик работы классифицирующей и метрической сетей на примере распознавания данных поля “пол” в паспорте РФ.

## ПОСТАНОВКА ЗАДАЧИ

Задано множество $X = \{I\}$ изображений поля “пол” в паспорте РФ. Задано множество $A = \{0, 1\}$ классов изображений в зависимости от содержания поля.

Требуется обучить два классификатора вида $C$: $X \to A$, ставящих в соответствие каждому изображению один из классов, согласно его содержимому, и сравнить характеристики их работы. В качестве классификаторов требуется взять классифицирующую и метрическую нейронные сети.

## ОБЗОР ДАННЫХ

Обучение сетей проводилось на синтезированных данных. Вначале были выяснены варианты заполнения графы “пол” в паспорте РФ (в табл. 1), и на их основе составлен словарь. Важно отметить, что одному и тому же классу соответствуют несколько вариантов.

Далее по этому словарю были сгенерированы данные по методу, описанному авторами работы (Chernyshova et al., 2018). Данные представляют собой изображения различных вариантов написания, у каждого из которых была проставлена метка с номером варианта. При генерации использовались различного вида искажения для того, чтобы максимально приблизить данные к натуральным, полученным в неконтролируемых условиях съемки: проективные искажения, смаз и гауссово размытие. Всего было получено 285421 изображений (142515 для женского пола и 142906 для мужского), содержащих различные варианты написания пола, некоторые из которых изображены на рис. 1.

Для обучения классифицирующей сети потребовалось сгруппировать данные по двум классам в соответствии с полом, который они обозначают: по шесть вариантов на каждый. Для обучения метрической сети было сформировано 1500024 пар изображений, которые были помечены в соответствии с тем, отражают ли они один и тот же вариант написания (“0”, если один и тот же вариант, “1” – разные).

Тестовой выборкой послужили натуральные данные – части изображений паспортов, содер-

жащие данные поля “пол”, всего 5310 изображений, поровну каждого класса. Изображения были получены при помощи малоразмерных цифровых камер и путем сканирования.

Для тестовых данных была вручную проведена разметка (“0” – женский пол, “1” – мужской).

## ОБЗОР ИСПОЛЬЗУЕМЫХ МОДЕЛЕЙ ОБУЧЕНИЯ

В качестве классифицирующей была использована глубокая свёрточная сеть, выходной вектор которой состоит из двух элементов и содержит в себе информацию о принадлежности входного объекта одному из классов. Для такой сети функция ошибки рассчитывается как

$$cost = -\sum_{i=0}^{M-1} p_i \ln e_i, \qquad (1)$$

где $M$ – размер алфавита сети, $p_i$ – идеальная оценка $i$-го класса, а $e_i$ – оценка $i$-го класса сетью.

Функция ошибки метрической сети

$$cost = (1 - p)\|e\| - (p * \max(\alpha - \|e\|, 0)), \qquad (2)$$

где $p$ – значение, указывающее на принадлежность входов одному и тому же классу, $e$ – вектор разности выходных векторов ветвей, а $\alpha$ – порог, начиная с которого увеличение расстояния между разными классами не уменьшает итоговую ошибку. Этот параметр вводится для того, чтобы сеть не пыталась в процессе обучения сделать все классы равноудаленными друг от друга, так как это нам не нужно (и не обязательно возможно). Размерность пространства, в которое сеть переводит входные объекты, была выбрана равной 10.

Как можно видеть, основным отличием при фиксированном размере алфавита является метод обучения и, следовательно, свойства ответа сети. Метрическая сеть имеет возможность исходное изображение перевести в пространство, по кластерам которого, в том числе, можно будет делать выводы о похожести различных классов из исходного алфавита. Из ответов классифицирующей сети такого вывода делать нельзя.

## ЭКСПЕРИМЕНТЫ И РЕЗУЛЬТАТЫ

Так как сети принимают на вход изображения строго определенного одного и того же размера, были проанализированы размеры изображений обучающей выборки для классифицирующей сети и найдено оптимальное соотношение сторон. Это соотношение соответствует максимальному значению на гистограмме распределения отношений ширины изображения к его высоте, изображенной на рис. 2. Размер входного изображения должен быть мал в целях экономии ресурсов, но достаточен для того, чтобы изображения были различимы. Так, перед обучением данные сжимались до выбранного размера. Соответственно сдвоенные изображения, на которых обучалась метрическая сеть, имели высоту в 2 раза больше, чем одиночные для классифицирующей.

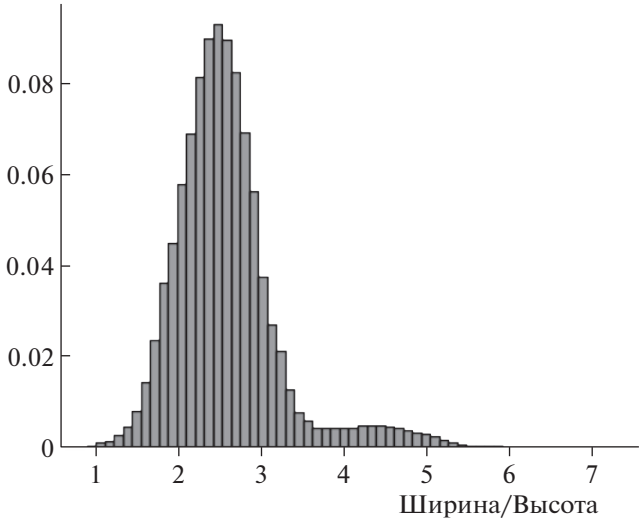


**Рис. 2.** Распределение отношений ширины изображения к его высоте.

В ходе эксперимента было обучено две классифицирующих и две метрических сети, решающих задачу распознавания пола на изображениях паспорта РФ. В качестве начального варианта классифицирующей была использована сеть с архитектурой, описанной в табл. 2. Этой же архитектурой обладала и каждая из ветвей сиамской сети.

Вторым этапом эксперимента было изменение архитектур на более легкие: количество свёрточных слоев было уменьшено, а шаги фильтра на последнем слое были удвоены (архитектура представлена в табл. 3). Так удалось получить близкий по качеству результат, но при этом размер классифицирующей сети уменьшился примерно в

**Таблица 2.** Тяжелая архитектура классифицирующей сети

| № слоя | Тип | Описание |
|---|---|---|
| 1 | Свёрточный | 4 фильтра 3 × 3, без отступов, шаг фильтра 1 × 1 |
| 2 | Свёрточный | 8 фильтров 5 × 5, отступ 2 × 2, шаг фильтра 2 × 2 |
| 3 | Свёрточный | 8 фильтров 3 × 3, отступ 1 × 1, шаг фильтра 1 × 1 |
| 4 | Свёрточный | 12 фильтров 5 × 5, отступ 2 × 2, шаг фильтра 2, 2 |
| 5 | Свёрточный | 12 фильтров 3 × 3, отступ 1 × 1, шаг фильтра 1 × 1 |
| 6 | Свёрточный | 12 фильтров 3 × 3, отступ 1 × 1, шаг фильтра 1 × 1 |
| 7 | Полносвязный | |

**Таблица 3.** Легкая архитектура классифицирующей сети

| № слоя | Тип | Описание |
|---|---|---|
| 1 | Свёрточный | 4 фильтра 3 × 3, без отступов, шаг фильтра 1 × 1 |
| 2 | Свёрточный | 8 фильтров 5 × 5, отступ 2 × 2, шаг фильтра 2 × 2 |
| 3 | Свёрточный | 8 фильтров 3 × 3, отступ 1 × 1, шаг фильтра 2 × 2 |
| 4 | Полносвязный | – |

**Таблица 4.** Качество распознавания рассматриваемых сетей

| Сеть | Тяжелая архитектура | Легкая архитектура |
|---|---|---|
| Классифицирующая | 98.36% | 98.23% |
| Метрическая | 98.05% | 97.86% |

3 раза. Необходимость упрощения архитектуры была обусловлена тем, что данные сети предполагается использовать в том числе и в мобильных приложениях, поэтому занимаемый размер имеет важное значение.

На основании описанной выше тестовой выборки была проверена работа обоих типов сетей с различными архитектурами и произведены замеры качества распознавания. Результаты представлены в табл. 4.

По данным таблицы видно, что классифицирующие сети показывают более высокое качество, чем метрические. Под качеством распознавания подразумевается его точность, т.е. отношение количества корректно классифицированных изображений к размеру тестовой выборки.

## ЗАКЛЮЧЕНИЕ

В работе представлено сравнение характеристик работы двух типов классификаторов, в качестве которых были использованы свёрточные нейронные сети, на примере распознавания пола на изображениях паспорта РФ. В рассмотренной задаче результаты экспериментов показали несущественное преимущество классифицирующей сети в сравнении с метрической, однако метрическая сеть обладает рядом дополнительных свойств. Одним из таких свойств является потенциальная применимость в задачах распознавания с заранее не определенным алфавитом. Другим преимуществом является то, что расстояния в результирующем пространстве сравнимы друг с другом и, в том числе, по позициям центров классов в этом пространстве можно делать выводы об их похожести.

Метрическая сеть, описанная в работе, может быть в дальнейшем использована для распознавания слов даже при условии, что полный словарь неизвестен. Для добавления нового элемента в словарь после обучения требуется по нескольким эталонным примерам при помощи уже обученной сети построить описание класса. Таким образом, становится заметным явное преимущество метрических сетей, которое состоит в возможности изменять алфавит после обучения и оценивать похожесть различных классов по позициям их объектов в результирующем пространстве.

Работа выполнена при финансовой поддержке РФФИ в рамках научных проектов № 17-29-07092 и № 17-29-07093.

## СПИСОК ЛИТЕРАТУРЫ

Арлазаров В.В., Жуковский А.Е., Кривцов В.Е., Николаев Д.П., Полевой Д.В. Анализ особенностей использования стационарных и мобильных малоразмерных цифровых видеокамер для распознавания документов. *Информационные технологии и вычислительные системы*. 2014. № 3. С. 71–81.

Лёзин И.А., Соловьёв А.В. Сжатие изображений с использованием многослойного персептрона. *Известия Самарского научного центра РАН*. 2016. Т. 18. № 4. С. 770–773.

Прохоров В.Г. Использование свёрточных нейронных сетей для распознавания рукописних символов. *Проблеми програмування*. 2008. № 2-3. С. 669-674.

Чернов Т.С., Ильин Д.А., Безматерных П.В., Фараджев И.А., Карпенко С.М. Исследование методов сегментации изображений текстовых блоков документов с помощью алгоритмов структурного анализа и машинного обучения. *Вестник РФФИ*. 2016. № 4 (92). С. 55–71. doi 10.22204/2410-4639-2016-092-04-55-71

Bulatov K., Arlazarov V.V., Chernov T., Slavin O., Nikolaev D.P. Smart IDReader: Document Recognition in Video Stream. *The 14th IAPR International Conference on Document Analysis and Recognition*. 2018. P. 39–44. doi 10.1109/ICDAR.2017.34710.1109/ICDAR.2017.347

Chernyshova Y., Gayer A., Sheshkus A. Generation method of synthetic training data for mobile OCR system. *Proc. SPIE 10696, Tenth International Conference on Machine Vision*. 2018. P. 1–7. doi 10.1117/12.2310119.10.1117/12.2310119

Hastie T., Tibshirani R., Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York. Springer-Verlag, 2009. 745 p.

Jaderberg M., Simonyan K., Vedaldi A., Zisserman A. Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition. *NIPS Deep Learning Workshop*. 2014. P. 1–10.

Koch G., Zemel R., Salakhutdinov R. Siamese Neural Networks for One-shot Image Recognition. *Proceedings of the 32 International Conference on Machine Learning*. 2015. V. 2. 8 p.

*Lecun Y., Bottou L., Bengio Y., Haffner P.* Gradient-based learning applied to document recognition. *Intelligent Signal Processing*. New York. IEEE Press, 2001. P. 306–351.

Liu Y., Wang Z., Jin H., Wassel I. Synthetically supervised feature learinig for scene text recognition. *The European Conference on Computer Vision*. 2018. P. 435–451.

Venkata Rao N., Sastry A.S.C.S., Chakravarthy A.S.N., Kalyanchakravarthi P. Optical character recognition technique algorithms. *Journal of Theoretical and Applied Information Technology*. 2016. V. 83. P. 275–282.

# Comparison of the classifying and similarity metric-based neural networks through the recognition of the filed "gender" in Russian Federation passport

**A. N. Chirvonaya[a,#], A. E. Lynchenko[b], Y. S. Chernyshova[b,c], and A. V. Sheshkus[b,c]**

[a] *National University of Science and Technology "MISIS", 119049 Moscow, Leninsky pr., 4, Russia*

[b] *Smart Engines Limited, 117312 Moscow, pr. 60-letiya Oktyabrya, 9, Russia*

[c] *Federal Research Center "Computer Science and Control" of Russian Academy of Sciences 119333 Moscow, Vavilova str., 44-2, Russia*

[#] *E-mail: nastyachirvonaya@smartengines.biz*

In this paper we consider the applicability of similarity metric classifiers for recognition of words. The approaches based on recognition of single characters are well studied, but they demonstrate poor performance on some kinds of text, especially the ones hard for segmentation, such as handwritten and Arabic texts, or the ones with ligatures. Moreover, if images are strongly noised and/or corrupted because of camera imperfection, touched symbols can appear in the text. All these problems usually occur in recognition systems for text with a predefined pattern, where the set of words is limited. Given this, it is reasonable to recognize whole words, although the dictionary can be huge and unknown while training. In this study we suggest using the similarity metric-based neural networks for word images recognition. We provide the comparison between the similarity metric-based neural network with classifying one on the words collected from "gender" field of Russian National passport. To maintain the experimental integrity, the parameters of all the layers except the last one were the same for both types of networks. The results show the relevance of the similarity metric-based neural networks to word recognition problem solving. The main advantages of the suggested method are the possibility of network alphabet extension after learning and no need for symbol segmentation.

*Key words:* text recognition, convolutional neural networks, deep learning, siamese neural networks, metrics learning

## REFERENCES

Arlazarov V.V., Zhukovskiy A.E., Krivtsov V.E., Nikolaev D.P., Polevoy D.V. Analiz osobennostey ispolzovaniya statsionarnykh i mobilnykh malorazmernykh tsifrovykh video kamer dlya raspoznavaniya dokumentov [The analysis of the features of using stationary and mobile small-size digital video cameras for documents recognition]. *Informatsionnye tekhnologii i vychislitelnye sistemy* [Information technologies and computation systems]. 2014. № 3. P. 71–81. (In Russian).

Lyozin I.A., Solovyov A.V. Performing an image compression by using the multilayer perceptron. *Izvestija Samarskogo nauchnogo centra RAN* [Proceedings of the Samara scientific center of Russian Academy of Sciences]. 2016. V. 18. № 4. P. 770–773. (In Russian).

Prohorov V.G. Ispol'zovanie svjortochnyh nejronnyh setej dlja raspoznavanija rukopisnih simvolov [The using of convolutional neural networks for handwritten symbols recognition]. *Problemi programuvannja* [Programming problems]. 2008. № 2–3. P. 669–674. (In Russian).

Chernov T.S., Il'in D.A., Bezmaternyh P.V., Faradzhev I.A., Karpenko S.M. Research of Segmentation Methods for Images of Document Textual Blocks Based on the Structural Analysis and Machine Learning. *RBRF Information Bulletin*. 2016. № 4 (92). P. 55–71. DOI: 10.22204/2410-4639-2016-092-04-55-71. (In Russian).

Bulatov K., Arlazarov V.V., Chernov T., Slavin O., Nikolaev D.P. Smart IDReader: Document Recognition in Video Stream. *The 14th IAPR International Conference on Document Analysis and Recognition (ICDAR2017)*. 2018. P. 39–44. DOI: 10.1109/ICDAR.2017.347

Chernyshova Y., Gayer A., Sheshkus A. Generation method of synthetic training data for mobile OCR system. *Proc. SPIE 10696, Tenth International Conference on Machine Vision (ICMV 2017)*. 2018. P. 1–7. DOI: 10.1117/12.2310119.

Hastie T., Tibshirani R., Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York. Springer-Verlag, 2009. 745 p.

Jaderberg M., Simonyan K., Vedaldi A., Zisserman A. Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition. *NIPS Deep Learning Workshop*. 2014.

Koch G., Zemel R., Salakhutdinov R. Siamese Neural Networks for One-shot Image Recognition. *Proceedings of the 32 International Conference on Machine Learning*. 2015. V. 2. 8 p.

Lecun Y., Bottou L., Bengio Y., Haffner P. Gradient-based learning applied to document recognition. *Intelligent Signal Processing*. New York. IEEE Press, 2001. P. 306–351.

Liu Y., Wang Z., Jin H., Wassel I. Synthetically supervised feature learinig for scene text recognition. *The European Conference on Computer Vision (ECCV)*. 2018. P. 435–451.

Venkata Rao N., Sastry A.S.C.S., Chakravarthy A.S.N., Kalyanchakravarthi P. Optical character recognition technique algorithms. *Journal of Theoretical and Applied Information Technology*. 2016. V. 83. P. 275–282.