DOI 10.37882/2223-2966.2024.6-2.13

НЕЙРОСЕТЕВОЙ КЛАССИФИКАТОР АВТОМАТИЧЕСКОЙ СИСТЕМЫ ОБРАБОТКИ ДОКУМЕНТОВ

NEURAL NETWORK CLASSIFIER FOR AUTOMATIC DOCUMENT PROCESSING SYSTEM

A. Gorbunov
M. Kuznetsova

Summary. Document classification plays an important role in many areas, such as information retrieval, data mining, etc., where machine learning and deep learning models can be applied. The paper deals with the features of building a neural network classifier for automatic document processing system. In the process of research the DocXClassifier model is presented, which is designed on a convolution-free neural network. The effectiveness of the model has been demonstrated by comparing it with other machine learning algorithms during the classification of scientific articles from the database «Web of Science Dataset».

Keywords: neural network, classification, document, group.

Горбунов Александр Николаевич
Аспирант, Аккредитованное образовательное частное учреждение высшего образования «Московский финансово-юридический университет МФЮА»; старший разработчик, ООО «Клик Групп»
a.gorbunov@click-group.ru
Кузнецова Марина Николаевна
доктор экономических наук, профессор, «Московский финансово-юридический университет МФЮА»
marina_kuzn82@mail.ru

Аннотация. Классификация документов играет важную роль во многих областях, таких как информационный поиск, интеллектуальный анализ данных и т.д., где могут применяться модели машинного и глубокого обучения. В статье рассмотрены особенности построения нейросетевого классификатора автоматической системы обработки документов. В процессе исследования представлена модель DocXClassifier, которая спроектирована на без конволюционной нейронной сети. Эффективность модели была продемонстрирована в ходе сравнения с другими алгоритмами машинного обучения при классификации научных статей из базы данных «Web of Science Dataset».

Ключевые слова: нейронная сеть, классификация, документ, группа.

Полнота и своевременность информации — жизненно важные элементы для современных организаций, будь то крупные межправительственные учреждения или небольшие предприятия. Необходимость получения документов, проверки их достоверности, внесения информации в соответствующие базы данных и т.д., занимает много времени у операторов соответствующих систем. Помимо этого, объем доступной на сегодняшний день информации стал огромным, и тенденция его роста практически экспоненциальная [1]. Такое информационное насыщение на самом деле является недостатком, поскольку традиционные системы регистрации и поиска документов начинают исчерпывать свои возможности. Использование этих систем становится все более сложным для пользователей, которые хотят получить необходимую информацию, а также для тех, кто занимается сопровождением, например, индексированием, классификацией документов и поддержкой тезаурусов. Для решения этой проблемы необходимо реализовать две ключевых задачи:

— упрощение поисковой деятельности, выполняемой неэкспертными пользователями;
— снижение затрат на обслуживание систем классификации документов [2].

В данном контексте очевидным является тот факт, что замена этапов ручной обработки документов автоматизированными системами позволит сократить время их анализа и прохождения, а также даст возможность нивелировать влияние человеческого фактора на точность выполнения работ. Из чего следует, что решение прикладной задачи разработки автоматической системы обработки документов является достаточно актуальной в текущее время широкого внедрения электронного документооборота.

На сегодняшний день для решения этой задачи широко используются различные методы машинного обучения. Как известно, эти методы предполагают применение множества алгоритмов, которые могут очень хорошо справляться с классификацией, но наиболее действенными из них являются нейронные сети. Нейронные сети способны намного быстрее и эффективнее выполнять задачи по классификации документов, по сравнению с традиционными методами машинного обучения. В настоящее время широкое распространение получили различные типы нейронных сетей, такие как конволюционные (CNN) или рекуррентные нейронные сети (RNN). Также в некоторых работах описываются преимущества применения древовидной структуры долговременной памяти (LSTM) [3].

В последние годы глубокое обучение (DL) совершило значительный прорыв в области анализа документов, продемонстрировав исключительную производительность в ряде задач, таких как извлечение ключевой ин-

формации и анализ макетов [4]. Однако, несмотря на эти достижения, остаются две основные проблемы, которые по-прежнему препятствуют безопасному и надежному развертыванию таких систем в реальных сценариях: их неотъемлемая природа «черного ящика» и низкая устойчивость к данным, выходящим за пределы распространения.

Таким образом, вопросы создания нейронных сетей для классификации документов, которые будут простыми в использовании и в тоже время способными продемонстрировать высокую производительность классификации, формируют перспективное направление для научных исследований, что и обусловило выбор темы данной статьи.

Возможности использования самоорганизующихся карт, которые после процесса обучения создают карту пространства документа, рассматривают в своих трудах Алексеев А.А., Зуев Д.С., Катасёв А.С., Кириллов А.Е., Хасьянов А.Ф., Nijia Lu, Guohua Wu, Zhen Zhang, Yitao Zheng, Yizhi Ren.

Над решением задачи выбора подходящих методов для встраивания слов, что играет жизненно важную роль в классификации документов с использованием нейронных сетей, трудятся Кривошеев Н.А., Спицын В.Г., Семенова А.В., Курейчик В.М., Дли М.И., Булыгина О.В., Sachin Dhawan, Rashmi Gupta, Lucas L. Lima, José R. Ferreira Junior, Marcelo C. Oliveira.

Высоко оценивая имеющиеся на сегодняшний день труды и наработки, следует отметить, что в исследуемой предметной плоскости существует еще много сложных вопросов, которые требуют более углубленного анализа. Так, нерешенной остается проблема с выбором оптимального количества признаков, поскольку, если их слишком много, то это вызывает больший риск, делая систему сложной, а также увеличивает время и стоимость классификации. Отдельного внимания заслуживает задача повышения уровня точности систем классификации документов в условиях ограниченности выборки.

Таким образом, цель статьи заключается в рассмотрении особенностей разработки нейросетевого классификатора автоматической системы обработки документов.

Документы — это не что иное, как набор предложений и абзацев. Чтобы обработать эти документы, их нужно сначала преобразовать в подходящий для обработки формат. Текстовые документы — самый простой формат для чтения и обработки, поэтому первым шагом будет преобразование документа в текстовый формат, а затем извлечение текста [5]. Для создания нейросетевого классификатора предлагаем использовать конволюционную нейронную сеть, которая будет проектироваться на базе Keras. Keras — это библиотека нейронных сетей, предоставляющая API, она входит в состав tensor flow, являющейся библиотекой с открытым исходным кодом для проектов машинного обучения.

На рисунке 1 представлена блок-схема предлагаемого классификатора.

Рис. 1. Блок-схема нейросетевого классификатора (составлено автором)

В процессе предварительной обработки и представления документов будут использоваться основные функции очистки, к которым относятся: удаление пустых слов, удаление иностранных символов, удаление знаков препинания, удаление цифр. Среди этих функций очистки целесообразным, по мнению автора, является применение функция токенизации, которая разбивает текстовый поток на слова, предложения, символы или другие значимые элементы, называемые лексемами. Также предполагается использование метода GLOVE (Global Vector for Word Representation). GlOVE — это алгоритм обучения без контроля для создания векторного представления слов. Обучение происходит на основе статистики совпадений слов из корпуса, а представление отображает линейную подструктуру слов [6].

Вектор слов передается в виде матрицы X и определяет мягкое ограничение для каждой пары:

$$w_i w_j + b_i + b_j = \log\left(X_{ij}\right)$$

где i — представляет, как слова появляются в контексте слова j; w_i — вектор для основного слова, w_j — вектор для контекста, а b_i, b_j — смещения. Функция стоимости встраивания GlOVE J представлена в виде:

$$J = \sum_{i=1}^{V}\sum_{j=1}^{V} f\left(X_{ij}\right)(w_i w_j + b_i + b_j - \log X_{ij})^2$$

Вектор слов l, $l \in R^d$ и d — размерность вектора слов. Весь документ представляется как $D \in R^{nd}$ в виде матрицы, n — количество слов в документе. Предложение максимальной длины добавляется там, где это необходимо.

$$l_n = l_1 \oplus l_2 \oplus l_3 \oplus \ldots \oplus l_n$$

где l_n обозначает конкатенацию слов. Свертка включает в себя поле W, которое применяется для формирования признака C. Свертка определяется как:

$$W \in R^{hd}$$

где h — количество слов, которые охватывает свертка, т.е. размер полосы. В данном случае уравнение имеет следующий вид:

$$W \times D_{j:j+h-1} = \sum_{i=j}^{j+h-1}\sum_{k=0}^{d-1} W_i, kD_i, k$$

Предположим, что y — это функция than, тогда y может быть представлено следующим образом:

$$y = f\left(W.X_{i:i+h-1} + b\right)$$

где b — член смещения, а f — нелинейная функция.

В рамках разрабатываемой модели применяется фильтр в различных конфигурациях, чтобы получить карту признаков. Также добавляется член смещения и применяется функция активации. После свертки всего документа формируется окончательная карта признаков C, такая, что:

$$C\left(W\right) = \left[C_1, C_2, C_3, \ldots, C_{n+h+1}\right]$$

Далее модифицируем традиционную архитектуру CNN, заменив глобальное среднее объединение в ConvNeXt механизмом объединения на основе внимания, как показано на рис. 2.

Механизм объединения на основе внимания использует маркер класса запроса для объединения выходных векторов карты признаков модели в виде взвешенной суммы на основе их сходства с вектором обучаемого класса (CLS) размерности d. При этом сходство вычисляется с помощью стандартной операции внимания с масштабированным точечным продуктом:

$$Attn\left(Q, K, V\right) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

где K, Q и V представляют матрицы запросов, ключей и значений слоя внимания, соответственно, а d_k — размерность признака k-й точки внимания.

Благодаря тому, что механизм внимания применяется только один раз, с помощью одной операции softmax, модель, по сути, присваивает важность определенным векторам признаков для каждого конкретного класса. Полученный агрегированный вектор затем добавляется к вектору CLS и обрабатывается сетью с прямой зависимостью. Наконец, для выполнения классификации используется линейная матрица.

Для оценки эффективности предложенного алгоритма было проведено экспериментальное исследование, которое заключалось в сравнении производительности классификатора DocXClassifier с другими алгоритмами машинного обучения, такими как машинный вектор поддержки (SVM), Naive Bayes, линейная классификация.

Для тестирования использовалась база данных, опубликованная в 2018 году «Web of Science Dataset», она

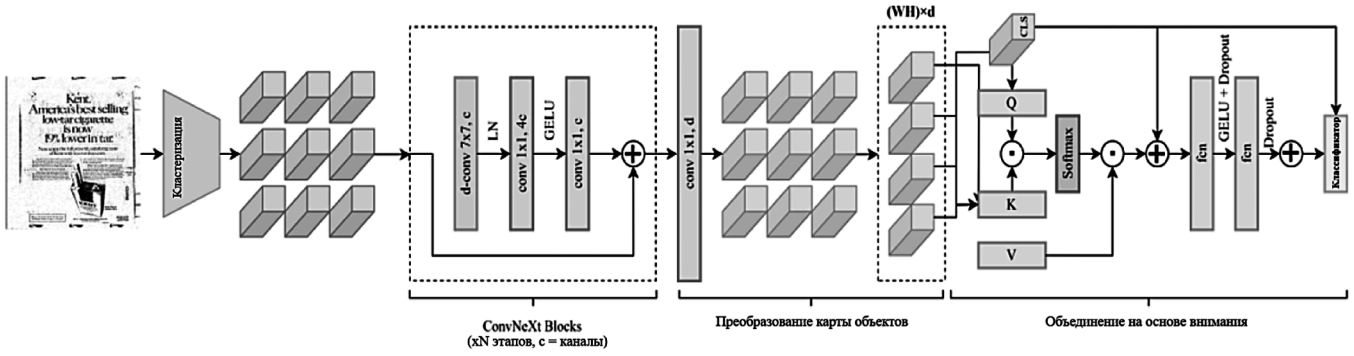

Рис. 2. Полная конфигурация предлагаемой модели DocXClassifier

состоит из 37 316 научных статей. Задача заключалась в классификации статьей в разрезе семи категорий: информатика, психология, медицина, гражданские науки, физика, биохимия и компьютерное зрение. Оценка предложенного метода проводилась с помощью расчета показатель точности (Acc). Он определяется как процент правильных предсказаний. Это происходит независимо от количества классов. Его формула выглядит следующим образом:

$$Acc = \frac{\text{правильно предсказанный класс}}{\text{общее количество классов}} \times 100\%$$

В ходе эксперимента сравнивалась точность алгоритмов между собой в зависимости от типа предварительной обработки, затем были выбраны среди них результаты с наилучшей точностью. Полученные данные представлены в таблице 1.

Таблица 1.

Точность классификации документов при использовании различных алгоритмов[1]

Алгоритм	Тип предварительной обработки			
	CV	WL TF-IDF	NL TF-IDF	C TF-IDF
Naive Bayes	73 %	48 %	66 %	62 %
SVM	48,4 %	48,44 %	50 %	49 %
Линейная классификация	79 %	74 %	69 %	72 %
DocXClassifier	75 %	82 %	70 %	69 %

[1] CV: Счетные векторы
WL TT-IDF: Векторы TF-IDF на уровне слов
NL TF-IDF: Векторы TF-IDF уровня N-грамм
C TF-IDF: TF-IDF-векторы уровня символов

Таблица 1 наглядно показывает, что скорость классификации алгоритмов варьируется в зависимости от используемой предварительной обработки и выбранной архитектуры. Так, для алгоритма DocXClassifier наивысший показатель близкий к 82 % был получен при использовании WL TF-IDF, для алгоритма SVM с предварительной обработкой NL TF-IDF — 50 %, для алгоритма линейной регрессии с предварительной обработкой CV — 79 %, а для алгоритма NB с предварительной обработкой CV — 73 %.

Подводя итоги, отметим, что предложенная в статье модель нейросетевого классификатора DocXClassifier обладает способностью эффективно генерировать карты важности признаков во время тестирования и позволяет получить высокую точность классификации документов (82 %), по сравнению с другими алгоритмами. Модель может быть улучшена с помощью различных предварительно обученных встроенных слов, а также добавления новых категорий группировки, чтобы увеличить область классификации.

ЛИТЕРАТУРА

1. Жалыбин А.А., Маликов А.В. Текстовая классификация документов на основе текстовой сегментации // Перспективы науки. 2021. № 4 (139). С. 187–192.
2. Бартеньев О.В. Оценка эффективности методов токенизации текста // Вестник Московского энергетического института. Вестник МЭИ. 2023. № 6. С. 144–156.
3. Weizhong Zhao, Dandan Fang An effective framework for semistructured document classification via hierarchical attention model // International Journal of Intelligent Systems. 2021. Volume 36, Issue 9. P. 45–49.
4. Кривошеев Н.А. Методы машинного обучения для классификации текстовой информации // Труды Международной конференции по компьютерной графике и зрению «Графикон». 2019. № 29. С. 266–269.
5. Shaobin Huang, Jingyun Sun NeuralConflict: Using neural networks to identify norm conflicts in normative documents // Expert Systems. 2022. Volume 41, Issue 6. P. 78–83.
6. Sheenam Malhotra, Williamjeet Singh A secure neural network-based ranking approach for document searching in cloud data center // Software: Practice and Experience. 2022. Volume 52, Issue 9. P. 76–83.

Журнал «Современная наука: актуальные проблемы теории и практики»