Классификация книг по жанрам на основе текстовых описаний посредством глубокого обучения

П.Л. Николаев

***Аннотация* — В данной статье рассмотрена модель глубокой нейронной сети для классификации книг по жанрам на основе текстовых описаний. При решении подобных задач обычно применяют модели, состоящие из рекуррентных слоев, однако в работе предлагается использовать модель с гибридной архитектурой: нейросеть, состоящую из LSTM и сверточных слоев. В работе приводится структура сети, а также рассматриваются методы улучшения качества ее работы. Обучение и тестирование глубокой нейронной сети проводятся на собственном наборе данных, содержащем информацию о тысячах книг. На основе полученных результатов можно судить о возможности применения обученной модели при решении практических задач. При этом данная модель может быть использована и при классификации текстовых данных в других тематиках.**

***Ключевые слова* — Искусственные нейронные сети, машинное обучение, глубокое обучение, сверточные нейронные сети, рекуррентные нейронные сети, классификация текста.**

I. Введение

При анализе большого количества информации уже невозможно обойтись без интеллектуальных методов, позволяющих автоматически производить обработку данных и извлекать из них необходимые сведения. С помощью данных методов можно значительно упростить и ускорить весь процесс по сравнению с ручной обработкой людьми. В последние годы в различные сферы активно внедряется глубокое обучение (глубокие нейронные сети), при помощи которого можно добиваться высоких результатов в задачах распознавания образов и анализа данных.

Одним из примеров подобного распознавания и анализа является интеллектуальный анализ текстов, одна из подзадач которого – классификация текстовых данных. Ее можно использовать в самых различных областях, в частности, при автоматическом определении жанровой принадлежности литературы. Это может пригодиться при работе с большим количеством книг: например, при создании систем автоматического построения электронных книжных каталогов, или при создании систем инвентаризации книг для библиотек или книжных магазинов.

Целью данной работы является разработка глубокой нейронной сети, способной определять жанровую принадлежность книг по их текстовым описаниям. Для этого в рамках работы решаются следующие задачи:

- создание эффективной модели нейросети для классификации текстовых данных, представляющих собой небольшие описания книг;
- сбор данных для нейросети – описаний книг и информации по жанрам;
- обучение и тестирование нейросети на собранных данных.

II. Обзор существующих работ

В основном работы по рассматриваемой теме посвящены классификации книг на английском языке. В работах [1, 2] рассматривается классификация книг по жанрам исключительно на основе изображений книжных обложек без использования текстов произведений, аннотаций и без указания авторов и названий. Касательно применяемых методов классификации, то в [1] наилучшие результаты показала комбинация предобученной сверточной нейронной сети с методами обработки естественного языка (Stanford NLP Classifier [3] и word2vec [4]), а в [2] – комбинация предобученной сверточной нейросети и рекуррентной LSTM-сети. В обоих случаях сверточные сети распознают само изображение, а методы обработки естественного языка и рекуррентная сеть используются для распознавания текстов на книжных обложках.

В [5] рассматривается классификация книг по жанрам непосредственно по текстовому содержимому. По приведенным в работе данным, наилучшие результаты показаны сверточной нейросетью.

Классификации книг на русском языке посвящена работа [6], в которой в качестве исходных данных выступают полные тексты произведений. В ней также используется сверточная нейронная сеть. При этом в работах [5] и [6] за основу используемых нейросетей берутся сети, рассмотренные в [7], посвященной исследованию методов классификации текстов.

Таким образом, во всех приведенных статьях используются различные методы глубокого обучения, что говорит об их высокой эффективности.

Однако способ классификация книг только по изображениям обложек, даже с учетом размещения на них названий произведений и авторов, применим не во всех случаях, поскольку множество книг, в основном старых, может иметь похожие обложки, в результате их разбиение на категории представляется весьма затруднительным. Также обложки могут быть испорчены, и их распознавание будет весьма

Статья получена 27 сентября 2021.

П.Л. Николаев – старший преподаватель Московского авиационного института (национального исследовательского университета) (e-mail: npavel89@gmail.com).

проблематичным, либо – и вовсе отсутствовать.

Что касается определения жанра по содержимому книг, то у данного подхода есть недостаток, заключающийся в сложности сбора данных, поскольку данный процесс требует проведения значительной и долгой работы по нахождению и систематизации данных. Кроме того, в этом случае для обучения глубоких нейросетей понадобится более мощное оборудование, а сам процесс обучения займет много времени.

В связи с вышеизложенным было принято решение рассмотреть классификацию книг по жанрам на основе небольших текстовых описаний.

III. Модель глубокой сети для классификации текстовых данных

Для решения поставленной задачи была создана модель глубокой нейронной сети, состоящая из разных видов слоев. Структура сети представлена в таблице 1.

Таблица 1. Структура сети.

Слой (тип)	**Выходная форма**
Входной слой	(None, 500)
Embedding()	(None, 500, 300)
SpatialDropout1D(0.3)	(None, 500, 300)
Bidirectional(LSTM(128, dropout=0.2, recurrent_dropout=0.2)	(None, 500, 256)
Dropout(0.2)	(None, 500, 256)
Conv1D(128, 3)+ReLU	(None, 498, 128)
Conv1D(128, 1)+ReLU	(None, 498, 128)
Dropout(0.2)	(None, 498, 128)
GlobalMaxPooling1D()	(None, 128)
Dense(10)+Softmax	(None, 10)

Вначале идет входной слой, на который поступают входные сигналы. Следом за ним следует слой Embedding, используемый для получения векторных представлений токенов (составных частей текста).

Для обработки последовательностей, коими являются текстовые данные, лучше всего подходят рекуррентные нейронные сети. Они обладают обратными связями и способны сохранять предыдущие состояния. Среди рекуррентных сетей можно выделить LSTM-сети (Long Short-Term Memory) или сети долгой краткосрочной памяти. LSTM-слои сохраняют информацию для последующего использования, в результате предотвращается затухание старых сигналов во время обработки [8]. В результате использования сетей с LSTM-слоями можно добиться большей точности при распознавании текста в сравнении с обычными рекуррентными сетями.

В нашей модели сети есть один двунаправленный LSTM-слой с гиперболическим тангенсом и сигмоидой в качестве функций активации. Под двунаправленностью подразумевается обработка текста как в прямом, так и в обратном направлении с последующим объединением результатов обработки. Использование двунаправленных рекуррентных слоев ведет к повышению точности распознавания текста.

Наряду с LSTM-слоем в сети используются и одномерные сверточные слои с функцией активации ReLU. В отличие от двумерных сверточных слоев, лежащих в основе сверточных сетей для распознавания изображений, одномерные слои способны посимвольно обрабатывать текст. Экспериментально было установлено, что улучшения точности распознавания в рамках решаемой задачи можно добиться путем добавления двух одномерных сверточных слоев.

Кроме того, для уменьшения переобучения сети были добавлены слои прореживания SpatialDropout со значением 0,3 и Dropout со значением 0,2. В первом случае в ноль устанавливаются целые карты признаков, а во втором – случайные нейроны. В результате происходит обучение только случайно отобранных частей нейронной сети.

Также на LSTM-слое использовалось рекуррентное прореживание со значением 0,2, означающее долю прореживаемых входных и рекуррентных значений.

В модели есть слой GlobalMaxPooling1D, используемый для уменьшения формы вывода. А после следует блок классификации, включающий только один полносвязный слой из 10 нейронов (по числу классифицируемых жанров, на которые нужно разбить данные). На данном слое используется функция активации softmax.

IV. Набор данных

Для обучения и проверки нейронной сети был собран набор данных, содержащий информацию о 7612 книгах, разбитых на 10 категорий. В таблице 2 представлено количество примеров по каждому жанру.

Таблица 2. Количество примеров в наборе данных по жанрам.

Жанр	**Количество примеров**
Фантастика	828
Фэнтези	912
Детективы	894
Проза	1556
История. Исторические науки	714
Информационные технологии	552
Естественные науки	553
Медицина и здоровье	455
Кулинария	530
Культура. Искусство	618

Для каждой книги в наборе были собраны следующие данные: автор, название и короткое описание. Однако было решено не учитывать имена авторов и названия произведений, поскольку многие авторы пишут в одном и том же жанре, а названия порой мало отражают полную суть книги. В итоге, при обучении и проверке нейронной сети использовались только аннотации в качестве входных данных и названия жанров в качестве выходных.

Собранный датасет был обработан путем удаления

всех лишних символов, за исключением русских букв, пробелов и точки, используемой в качестве разделителя предложений. Также были удалены стоп-слова (местоимения, предлоги, союзы, междометия и частицы), не несущие смысловой нагрузки. Помимо этого, весь текст был приведен к нижнему регистру.

После очистки было произведено разбиение набора данных на три выборки: обучающую (60% данных – 4563 примера), валидационную (10% данных – 761 пример) и тестовую (30% данных – 2288 примеров). При этом разбиение производилось следующим образом: данные были сгруппированы по жанру, затем они были перемешаны внутри каждой группы, а уже после пропорционально выполнялось разбиение на различные выборки в каждой группе. Таким образом, записи, относящиеся к разным жанрам, удалось распределить по выборкам более равномерно.

После всей обработки и разбиения датасета на части было произведено перекодирование текстовых данных в числовые представления.

В случае с входными данными процесс перекодирования выглядел следующим образом. Вначале была произведена токенизация – разбиение текста на составные части (токены), коими могут являться слова, символы или последовательности символом или слов (N-граммы). В нашем случае был создан словарь с максимальным количеством слов, равным 50000. В этом словаре все слова были связаны с целочисленными индексами, основываясь на частоте встречаемости слов. А после этого каждый текст в наборе был преобразован в последовательность целых чисел, т.е. произошла замена слов на целые числа из словаря. Поскольку получившиеся последовательности имели разные длины, то все они были приведены к размерности, равной 300, путем усечения в случае большей длины или же путем дополнения нулями в случае недостаточной длины.

Что касается выходных данных (жанров), то они были перекодированы в последовательность нулей и единиц с размерностью, равной 10 (по количеству уникальных жанров), с помощью унитарного кодирования. В полученной последовательности ноль соответствовал неправильному жанру, а единица – правильному.

V. Обучение сети

Предложенная модель нейронной сети была реализована на языке программирования Python 3.8.5 с применением библиотек глубокого обучения TensorFlow 2.4.1 и Keras 2.4.3. Для обучения использовался компьютер с процессором Intel Core i3-8100 с тактовой частотой 3,6 ГГЦ и объемом оперативной памяти в 16 ГБ. К сожалению, из-за использования рекуррентного прореживания не было возможности использовать видеокарту для ускорения обучения, поскольку данный метод не поддерживался библиотекой cuDNN, способной взаимодействовать с видеокартами, и поверх которой работает TensorFlow.

Для обучения предложенной модели сети в качестве оптимизатора использовался метод Adam с коэффициентом обучения, равным 0,001. В качестве функции потерь была использована перекрестная энтропия, а в качестве метрики для оценивания правильности определения классов – верность (accuracy), определяющая процент правильно классифицированных примеров. Размер мини-выборки был задан равным 64.

Число эпох для обучения было задано равным 5. Все обучение заняло около 1 часа. Наилучшие показатели были достигнуты на 3 эпохе.

В таблице 3 представлены результаты обучения и проверки предложенной сети по лучшим весовым коэффициентам, вычисленным на 3 эпохе.

Таблица 3. Результаты обучения и проверки нейросети.

Выборка	**Точность (accuracy)**	**Потери (loss)**
Обучающая	98.07%	0.08
Валидационная	72.80%	0.84
Тестовая	71.11%	0.93

На рисунке 1 представлена матрица неточностей, полученная после прогона тестовой выборки через обученную нейросеть. Данная матрица показывает количество верно и ошибочно классифицированных примеров по каждому классу (жанру).

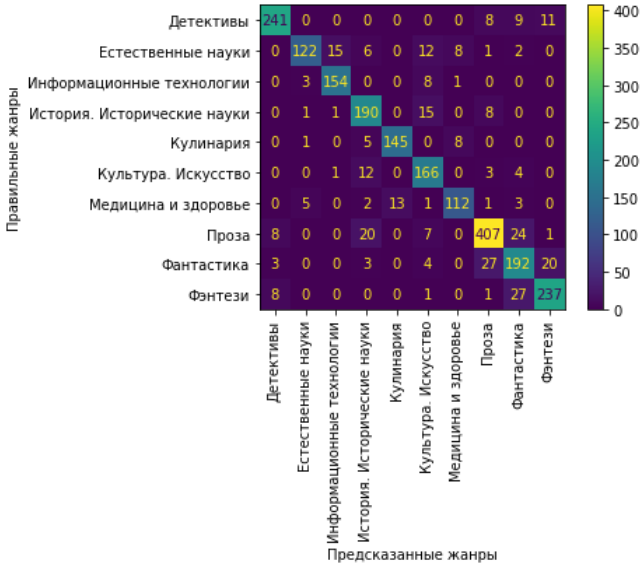

Рисунок 1. Матрица неточностей на тестовой выборке.

VI. Заключение

Таким образом, был рассмотрен метод классификации книг по жанрам на основе только их описаний. Для этого была разработана модель глубокой нейронной сети, состоящая из LSTM и сверточных одномерных слоев. Данная сеть была обучена и проверена на собственном наборе данных.

Показанная на тестовой выборке точность работы сети (около 71%) показывает, что она в достаточной мере способна производить классификацию книг по

жанрам, основываясь лишь на небольших аннотациях. Дальнейшим продолжением работы будет являться улучшение качества работы сети.

Библиография

[1] H. Chiang, Y. Ge, C. Wu. (2015). Classification of Book Genres By Cover and Title [Online]. Available: http://cs229.stanford.edu/proj2015/127_report.pdf

[2] C. Kundu, L. Zheng. (2020). Deep multi-modal networks for book genre classification based on its cover. [Online]. Available: https://arxiv.org/abs/2011.07658v1

[3] Stanford Classifer [Электронный ресурс]. – URL: https://nlp.stanford.edu/software/classifier.shtml

[4] T. Mikolov, K. Chen, G. Corrado, J. Dean, “Efficient Estimation of Word Representations in Vector Space,” in Proceedings of Workshop at ICLR, 2013.

[5] J. Worsham. J. Kalita, “Genre Identification and the Compositional Effect of Genre in Literature,” in Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA, August 20-26, 2018, pp. 1963–1973.

[6] Батраева И.А., Нарцев А.Д., Лезгян А.С. Использование анализа семантической близости слов при решении задачи определения жанровой принадлежности текстов методами глубокого обучения // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2020. №50 С.14-22. DOI: 10.17223/19988605/50/2.

[7] Y. Kim, “Convolutional neural networks for sentence clas-sification,” in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, October 2014, pp. 1746–1751. DOI:10.3115/v1/D14-1181.

[8] Шолле Ф. Глубокое обучение на Python. – СПб.: Питер, 2018. – 400 с.

Book Genre Classification on the Base on Text Description through Deep Learning

P.L. Nikolaev

***Abstract* — This article describes the deep neural network model for books classification by genre on the base of text description. Such problems are usually solved with the use of models consisting of recurrent layers, however, in this work it is proposed to use the model with the hybrid architecture: the neural network consisting of LSTMs and convolutional layers. The paper provides the structure of the network and also discusses methods for improving the quality of its work. Deep neural network training and testing is carried out on its own dataset containing information on thousands of books. We can judge the possibility of using the trained model in solving practical problems on the basis of the results obtained. Moreover, this model can be used for the classification of text data in other topics.**

***Keywords* — Artificial neural networks, machine learning, deep learning, convolutional neural networks, recurrent neural networks, text classification.**

REFERENCES

[1] H. Chiang, Y. Ge, C. Wu. (2015). Classification of Book Genres By Cover and Title [Online]. Available: http://cs229.stanford.edu/proj2015/127_report.pdf

[2] C. Kundu, L. Zheng. (2020). Deep multi-modal networks for book genre classification based on its cover. [Online]. Available: https://arxiv.org/abs/2011.07658v1

[3] Stanford Classifer. [Online]. Available: https://nlp.stanford.edu/software/classifier.shtml

[4] T. Mikolov, K. Chen, G. Corrado, J. Dean, “Efficient Estimation of Word Representations in Vector Space,” in Proceedings of Workshop at ICLR, 2013.

[5] J. Worsham. J. Kalita, “Genre Identification and the Compositional Effect of Genre in Literature,” in Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA, August 20-26, 2018, pp. 1963–1973.

[6] Batraeva I.A., Nartsev A.D., Lezgyan A.S, “Using the analysis of semantic proximity of words in solving the problem of determining the genre of texts within deep learning,” Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitel'naya tekhnika i informatika, 50, pp. 14–22, 2020. DOI: 10.17223/19988605/50/2.

[7] Y. Kim, “Convolutional neural networks for sentence clas-sification,” in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, October 2014, pp. 1746–1751. DOI:10.3115/v1/D14-1181.

[8] F. Chollet, Glubokoe obuchenie na Python in St. Peterburg, Russia: Piter (In Russian), 2018.