Применение сверточной нейронной сети для классификации текста

Л.Е. Сапожникова[1], О.А. Гордеева[1]

[1]Самарский национальный исследовательский университет им. академика С.П. Королева, Московское шоссе 34А, Самара, Россия, 443086

Аннотация. В статье рассматривается один из методов классификации текста с помощью нейронной сети. Сформулирована общая постановка задачи классификации текста, описана архитектура сверточной нейронной сети для решения поставленной задачи, приведены этапы решения и результаты классификации.

1. Введение

В современных условиях непрерывно увеличивающегося объема хранимой и используемой информации возникает ряд задач автоматической обработки текстов, одна из которых – задача классификации текстовых данных, что позволяет разделить тексты по различным тематическим каталогам (категориям, классам). Классифицируются сайты, документы, письма, обращения, новости для распределения их по классам с целью оптимального хранения и использования.

Текстовые классификаторы применяют для распознавания эмоциональной окраски текста при обработке отзывов и комментариев. Классификация текста применяется в борьбе со спамом, а также для персонализации контекстной рекламы на основе анализа активности пользователя в сети и классификации просмотренных сайтов.

Для классификации текстовой информации могут применяться различные методы и технологии. В данной статье речь пойдет об использовании сверточной нейронной сети, аспектах и особенностях ее применения для решения задачи классификации текста, а также о результатах применения данного метода классификации.

2. Постановка задачи классификации текста и методы ее решения

В общем виде задача классификации выглядит следующим образом.

Имеется некоторое множество объектов (текстов) и заранее известное множество классов, с которыми объекты могут быть сопоставлены. Для некоторой части объектов известно, к какому классу они принадлежат – это подмножество будет являться обучающей выборкой, для остальных объектов классы не определены. Следует определить класс принадлежности произвольного объекта (текста) из множества объектов.

Задача классификации может быть формализована следующим образом [1].

Существует множество текстов $D = \{d_1,...d_n\}$. Каждый текст $d_i \in D$ представляет собой последовательность слов $Wd = \{w_1,...,...\}$. Задано конечное множество классов $C = \{c_1,...,c_m\}$. Можно обозначить за $\Phi(d)$ идеальный классификатор, переводящий объект d в его класс c_j. Задача заключается в построении другого классификатора $\tilde{\Phi}(d)$, максимально близкого к

идеальному классификатору, способного классифицировать произвольный объект d.
Решение задачи классификации включает в себя следующие основные этапы:

1. Предварительная обработка текста, включающая токенезацию и векторное представление слов.
2. Построение классификатора.
3. Оценка вероятности ошибочной классификации.

Методы классификации текста разделяют на [2]:

- вероятностные (наивный байесовский классификатор);
- метрические (метод k-ближайших соседей);
- логические (метод деревьев решений);
- линейные (логическая регрессия);
- методы на основе искусственных нейронных сетей.

3. Классификация текста с помощью сверточной нейронной сети

3.1. Общая архитектура сверточной нейронной сети

Сверточные нейронные сети очень эффективно используются для решения задачи классификации текста [3]. Результат классификации представляет собой распределение вероятностей принадлежности текста к заранее известным классам.
Базовая архитектура сверточной нейронной сети состоит из следующих слоев [4].

1. Сверточный слой, который представляет собой набор карт признаков (матриц), у каждой карты есть ядро свертки, представляющее собой фильтр или окно, которое скользит по всей области карты признаков. Набор фильтров определяет размерность новой матрицы. Алгоритм обратного распространения ошибки для сверточных сетей также является сверткой, но с пространственно перевернутыми фильтрами.
2. Субдискретизирующий слой, который уменьшает размер матрицы, на данном слое чаще всего используется метод максимального элемента (max-pooling)
3. Полносвязный слой, в котором каждый нейрон соединен со всеми нейронами на предыдущем уровне, причем каждая связь имеет свой весовой коэффициент.
4. Выходной слой, который связан со всеми нейронами предыдущего слоя. Количество нейронов соответствует количеству распознаваемых классов.

3.2. Применяемая модель сверточной нейронной сети

Рассмотрим модель нейронной сети, применяемую в данной работе. Входными данными являются слова, представленные векторами семантических признаков. В таком представлении близкие по смыслу слова находятся на близком расстоянии в векторном пространстве.
Пусть $x_i \in R^k$ - k-мерный вектор соответствующий i-тому слову в предложении. Тогда предложение длины n можно представить как [3]:

$$x_{1..n} = x_1 \oplus x_2 \oplus ... \oplus x_n, \quad (1)$$

где $\oplus$ - операция конкатенации.
Пусть $x_{i..i+j}$ означает конкатенацию слов $x_i, x_{i+1},..., x_{i+j}$. Операция свертки использует фильтр $w \in R^{hk}$, который применяется к окну из h слов для создания нового признака. Например, признак c_i сгенерируется из окна слов $x_{i..i+h-1}$ как

$$c_i = f\left(w * x_{i..i+h-1} + b\right), \quad (2)$$

где $b \in R$ – шаг смещения, а f - нелинейная функция активации. Этот фильтр применяется к каждому возможному окну слов в предложении $\{x_{1..h}, x_{2..h+1},..., x_{n-h+1..n}\}$, чтобы произвести новую карту признаков

$$c = \{c_1, c_2, ...c_{n-h+1}\}, \quad (3)$$

где $c \in R^{n-h+1}$.

Затем применяется операция объединения набора значений, выбирается максимальное значение $\hat{c} = \max\{c\}$ - наиболее важный признак для каждой свертки.

В процессе работы сеть использует несколько фильтров с разными размерами окон для получения множества признаков. Эти признаки предпоследнего слоя передаются последнему слою, выход из которого является вероятностью распределения признаков по классам.

В качестве функции активации используется функция Leaky ReLU, задаваемая формулой [4]

$$f(x) = 1(x < 0) * (\alpha x) + 1(x \geq 0)(x), \quad (4)$$

где α – константа, имеющая очень небольшое значение.

Данная функция имеет более высокую скорость сходимости по сравнению с другими функциями активации, а также довольно проста в вычислении.

Для регуляризации нейронной сети (для предотвращения переобучения) используется L2-регуляризация [4] в сочетании с дропаут [5].

L2-регуляризация реализуется путем штрафования нейронной сети – увеличения целевой функции (функции потерь). Для каждого веса *w* в сети добавляется функция потерь λw^2, где λ - сила регуляризации.

Регуляризация L2 предотвращает сильное увеличение каких-либо весов и приводит к перераспределению весовых значений. Это заставляет нейронную сеть использовать все нейроны хотя бы в небольшой степени.

Дропаут – это отключение нейронов случайным образом. На каждом этапе обучения отдельные нейроны выпадают из сети, что помогает избежать зависимости между нейтронами во время обучения.

Сочетание регуляризации L2 и дропаут позволяет избежать ситуации, когда сеть показывает отличные результаты на обучающей выборке, но неэффективна при проверке на контрольной выборке.

Архитектура используемой сети представлена на рисунке 1. Первый слой решает задачу векторного представления слов. Далее создаются три сверточных слоя (conv2d_1, conv2d_2, conv2d_3). На рисунке также отображена функция активации Leaky ReLU для каждого слоя, дропаут и субдискретизирующий слой (max-pooling). Далее происходит соединение слоев, и в конце получается полносвязный слой (dense_1). Представленная архитектура построена с использованием системы визуализации TensorBoard [6].

3.3. Формирование входных данных

Тестирование описанного метода классификации проведено на материалах новостного портала РИА-Новости. Данный портал содержит огромное количество публикаций с четко определенной тематикой, что позволило задать категории (классы) для классификации, а также сформировать достаточное количество текстовых фрагментов для обучения сети.

Были определены 8 классов – культура, происшествия, религия, общество, экономика, политика, наука, мир. Для каждого класса получено по 8000 статей из материалов портала. Всего было получено 64 000 новостных статей, относящихся к одному из 8 классов. Для получения данных было разработано приложение на платформе Node.js. Эти данные были подвергнуты предварительной обработке перед началом обучения нейронной сети.

3.4. Предварительная обработка текста

Форматирование входных данных для нейронной сети было выполнено на языке Python 3.6 с использованием библиотеки Keras и Jupyter Notebook.

Максимальный размер текста был ограничен 1000 символами. Тексты большей длины были разбиты на части и отнесены к тому же классу. Для обучения нейронной сети важно одинаковое количество примеров каждого класса, иначе нейронная сеть будет игнорировать семантический смысл текста и будет учитывать априорную вероятность появления статей каждого класса. Во избежание такой ситуации объемы (количество текстов) классов были уравнены с классом наименьшего объема – 11 733 текста. В общей сложности было получено 93864 текста для 8 классов.

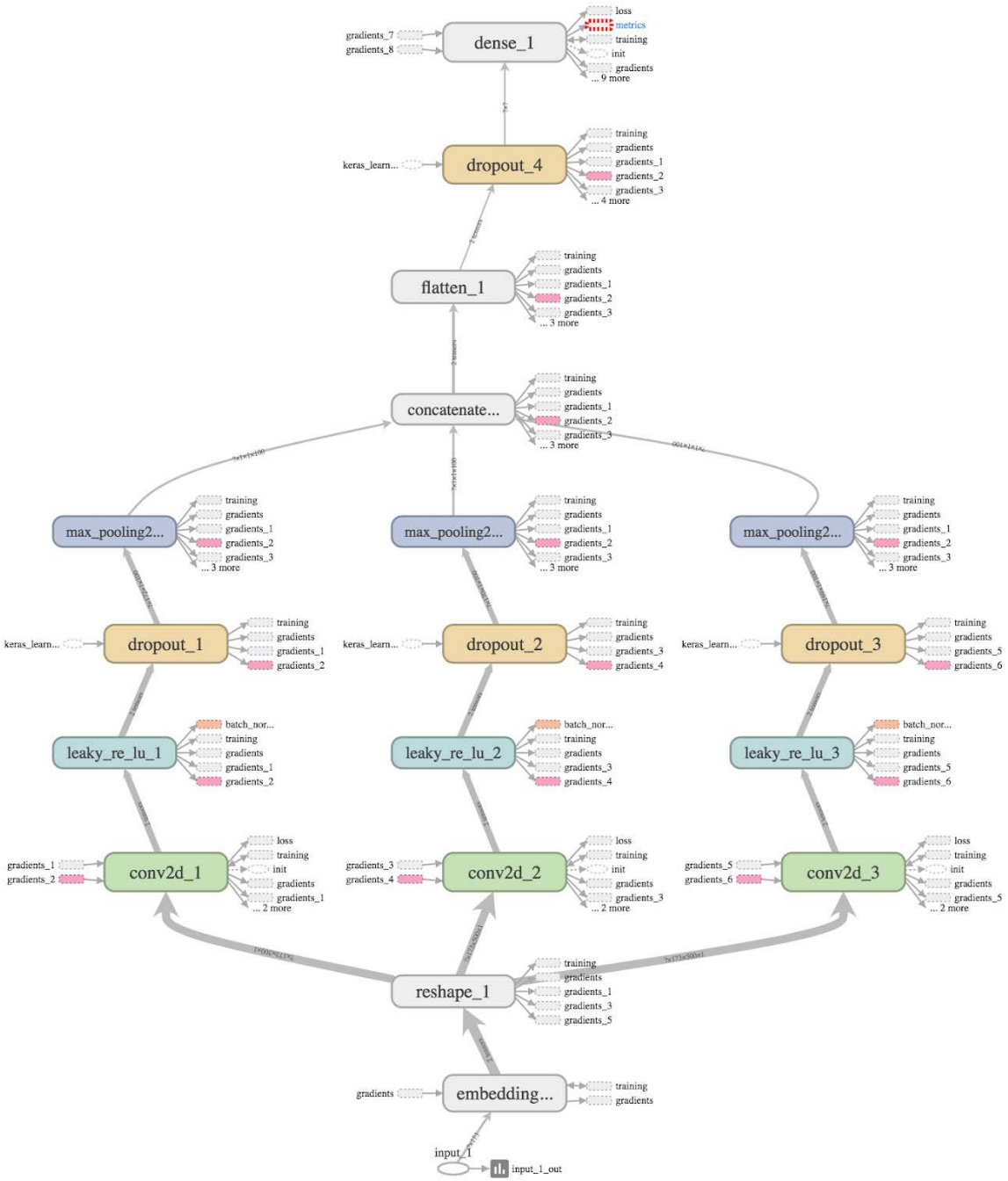

Рисунок 1. Архитектура используемой сверточной сети.

Затем все тексты были разделены на три выборки – обучающую (60% текстов), валидационную (20% текстов) и контрольную (20% текстов).

Далее была проведена токенизация текстов с целью поиска уникальных слов. С помощью библиотеки Keras было найдено 282 972 уникальных токена (слова), которые необходимо перевести в векторные представления признаков для обучения сети.

3.5. Формирование векторов признаков

Векторные представления описывают зависимости между словами, в векторном пространстве похожие слова будут иметь похожие векторы. Томас Миколов [7] получил для векторных представлений значимые синтаксические и семантические закономерности, которые реализуются в существующих сетевых библиотеках и моделях.

Наиболее популярным практическим способом получение вектора признаков является использование языковых моделей word2vec, созданных с использованием нейронных сетей. Эти модели обучаются на очень больших объемах данных естественного языка и не требуют таких предварительных действий, как удаление стоп-слов, стэминг или лемматизация, которые необходимы при классификации другими методами. Сохранение различных словоформ может увеличить точность классификации. Семантически одинаковые слова и синонимы в результате будут иметь близкие векторы значений семантических признаков.
В работе использовалась открытая модель word2vec [8], представляющая слово в виде вектора из 500 значений признаков. Модель является предварительно обученной, предназначена для русского языка. Для начала была проведена серия экспериментов построения векторов для похожих слов, подтверждающая возможность применения самой модели. В результате было установлено, что данная модель векторизации строит близкие вектора признаков для слов, близких по смыслу. В итоге было принято решение использовать обученную модель для перевода слов экспериментальной выборки в вектора признаков без применения стемминга или лемматизации и без удаления стоп-слов.

3.6. Построение, обучение и применение сети
Первый слой созданной сети преобразует поступающие на вход токены в вектора признаков и уже является обученным, Построенная нейронная сеть использует три независимых сверточных слоя с размером фильтров 2, 4, 5. Количество фильтров – 100. В качестве функции активации используется функция Leaky ReLU с коэффициентом $\alpha = 0.01$. Также используется L2 регуляризация совместно с дропаут и субдискретизирующий слой (max-pooling). Сила регуляризации = 0.01, дропаут для сверточных слоев = 0.5, дропаут для полносвязного слоя = 0.6, скорость обучения сети = 0.001. Далее происходит соединение слоев, и в конце получается полносвязный слой.
Обучение сети происходило в 5 эпох, рассчитывались значение функции потерь, точность классификации данных обучающей выборки и валидационной выборки.
В таблице 1 представлены значения точности для обучающей и валидационной выборок на различных эпохах обучения. С каждой эпохой значение точности классификации увеличивается. Важно отметить, что точность классификации обучающей выборки с каждой эпохой будет увеличиваться всегда, более важным является точность классификации валидационной выборки. Необходимо закончить обучение нейронной сети, если точность классификации обучающей выборки растет, а точность классификации валидационной выборки перестает возрастать – это свидетельствует о переобучении сети.

Таблица 1. Точность классификации.

	Обучающая выборка	Валидационная выборка
Epoch 1	72.6	82.27
Epoch 2	82.07	83.46
Epoch 3	85.84	82.91
Epoch 4	88.64	84.05
Epoch 5	90.92	84.34

После 5 эпохи точность классификации валидационной выборки стала равна 0,84 и перестала возрастать, что говорит о начале переобучения сети. Все веса нейронов сохранены для дальнейшего использования в задачах классификации текста. Качество классификации было проверено на контрольной выборке, точность составила 0,84.
В таблице 2 представлены некоторые результаты применения обученной сети для классификации цитат из новостной ленты портала mail.ru.

Таблица 2. Примеры классификации текста.

Текст	Результат классификации	
На некоторых участках порывы ветра достигают такой силы, что препятствуют проведению работ на высоте, — добавил представитель ведомства. Вечером в субботу гроза с порывами ветра до 25 метров в секунду и сильным дождем спровоцировала падение деревьев и обрыв линий электропередачи в нескольких районах Подмосковья. Больше других пострадали север и запад	общество: 0.21 религия: 0.07 происшествия: 0.36 политика: 0.07	культура: 0.05 мир: 0.1 наука: 0.06 экономика: 0.08
Администрация США считает, что открытие Крымского моста отрицательно отразится на людях и полуострове, заявила во вторник в ходе брифинга представитель Госдепартамента США Хизер Нойерт.	общество: 0.03 религия: 0.06 происшествия: 0.05 политика: 0.23	культура: 0.07 мир: 0.41 наука: 0.05 экономика: 0.1
В рамках фестиваля в Каннах состоялась мировая премьера новой части популярной франшизы «Хан Соло: Звездные войны». Ленту представили на Лазурном берегу вне конкурса. Каннские отборщики славятся своей высоколобостью.	общество: 0.04 религия: 0.04 происшествия: 0.01 политика: 0.03	культура: 0.79 мир: 0.03 наука: 0.04 экономика: 0,02
Несколько десятков участников мотопробега в честь запуска автомобильного движения по Крымскому мосту нарушили правила дорожного движения. Они остановились на арочном пролете для того, чтобы сфотографироваться.	общество: 0.41 религия: 0.09 происшествия: 0.1 политика: 0.04	культура: 0.07 мир: 0.07 наука: 0.14 экономика: 0.08

4. Заключение

Задача классификации является актуальным направлением в обработке текстовых данных. Процессы классификации текста реализуются в различных областях: классификаторы по различным признакам (тематика, стилистика, эмоциональная окраска текста), фильтрация спама, контекстная реклама и прочее.

В данной работе было проведено исследование применимости сверточной нейронной сети для решения задачи классификации текста. Сверточная нейронная сеть была обучена с использованием уже обученной нейронной сети для представления слов в виде векторов выделенных признаков, представляющих универсальные семантические значения, которые могут быть использованы для классификации текстов естественного русского языка.

Построенная нейронная сеть способна классифицировать новостные и не только тексты по тематикам и предоставлять распределение вероятности отнесения текста к восьми заранее определенным классам. Точность классификации оценена по контрольной выборке и составила 84%. Для выбранного количества классов такой результат классификации говорит об эффективности применения сверточной нейронной сети для классификации текста.

5. Литература

[1] Епрев, А.С. Автоматическая классификация текстовых документов // Математические структуры и моделирование. – 2010. – № 21. – С. 65-81.

[2] Батура, Т.В. Методы автоматической классификации текстов // Программные продукты и системы. – 2017. – № 1. – С. 85-99.

[3] Kim, Y. Convolutional Neural Networks for Sentence Classification [Электронный ресурс]. – Режим доступа: https://arxiv.org/pdf/1408.5882.pdf (11.11.2018).

[4] Karpathy, A. CS231n: Convolutional Neural Networks for Visual Recognition [Электронный ресурс]. – Режим доступа: http://cs231n.github.io/ (11.11.2018).

[5] Budhiraja, A. Dropout in (Deep) Machine learning [Электронный ресурс]. – Режим доступа: https://medium.com/@amarbudhiraja/https-medium-com-amarbudhiraja-learning-less-to-learn-better-dropout-in-deep-machine-learning-74334da4bfc5 (11.11.2018).

[6] Tensorboard [Электронный ресурс]. – Режим доступа: https://www.tensorflow.org/programmers_guide/summaries_and_tensorboard (11.11.2018).

[7] Mikolov, T. Linguistic Regularities in Continuous Space Word Representations / T. Mikolov, Y. Wen-tau, G. Zweig [Электронный ресурс]. – Режим доступа: https://www.aclweb.org/anthology/N/N13/N13-1090.pdf (11.11.2018).

[8] Обученная модель word2vec [Электронный ресурс]. – Режим доступа: http://panchenko.me /data/dsl-backup/w2v-ru/ (11.11.2018).

Application of convolutional neural network for text classification

L.E. Sapozhnikova[1], O.A. Gordeeva[2]

[1]Samara University, Moskovskoe Shosse 34A, Samara, Russia, 443086

Abstract. In the paper the method of text classification using a convolutional neural network is presented. The problem of text classification is formulated, the architecture of a convolutional neural network for solving the problem is described, the steps of the solution and the results of classification are given.