ПРИНЯТИЕ РЕШЕНИЯ О КЛАССИФИКАЦИИ ТЕКСТА НА ОСНОВЕ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ

MAKING A DECISION ON TEXT CLASSIFICATION BASED ON ARTIFICIAL NEURAL NETWORKS

КОНДРАШОВА Д.А.
НАСЫРОВ Р. В.

KONDRASHOVA D.A.
NASYROV R.V.

*Уфимский государственный
авиационный технический университет*

Ufa state aviation technical university

kovtunov@rambler.ru

kovtunov@rambler.ru

doi: 10.18411/pivsn-31

idsp: pivsn-31

АННОТАЦИЯ

Цель данной статьи - исследовать применение различных модификаций искусственных нейронных сетей к задаче определения принадлежности текста к определенному стилистическому типу, а также сравнить степень эффективности и точности применяемых методов.

Ключевые слова: нейронные сети, анализ текстов, многослойный перцептрон, долгая краткосрочная память.

ABSTRACT

The purpose of this article is to investigate the application of various modifications of artificial neural networks to the problem of determining whether a text belongs to a particular stylistic type, and also to compare the extent of efficiency and accuracy of the methods used.

Keywords: neural networks, multilayer perceptron, MLP, long short term memory, LSTM

ВВЕДЕНИЕ

Задача определения принадлежности текста к тому или иному классу, приобретает все большую актуальность в связи с высокой скоростью развития информационных технологий. Автоматическая классификация текста может быть полезна в различных сферах, в частности в сфере бизнеса, образования, медицины. При помощи автоматической классификации могут быть решены многие задачи технической поддержки, документа оборота и т.д.

Так как за последние годы выявлен ряд задач для решения которых применение искусственных нейронных сетей (ИНС) обладает эффективностью. К таким задачам можно отнести, например, выделение отдельных гитарных аккордов в музыке [1], распознавание цифр [2] и других объектов на изображениях [3]. В связи с этим представляется целесообразным использовать их для решения важной задачи анализа классификации текста.

ЦЕЛЬ ИССЛЕДОВАНИЯ

Целью данной работы является создание и сравнение двух моделей нейронной сети применительно к задаче классификации текста. В данной статье рассмотрена бинарная классификация на 2 класса — «принадлежность к техническому тексту» и «принадлежность к художественному тексту».

ПРЕДЛАГАЕМОЕ РЕШЕНИЕ

Известно, что многослойными перцептронами (MLP) (рис.1) называют нейронные сети прямого распространения [4]. Входной сигнал в таких сетях распространяется в прямом направлении, от слоя к слою. Многослойный перцептрон в общем представлении состоит из следующих элементов:

- множества входных узлов, которые образуют входной слой;
- одного или нескольких скрытых слоев вычислительных нейронов;
- одного выходного слоя нейронов.

Рис. 1 -Архитектура многослойного перцептрона

Частным случаем рекуррентных ИНС является LSTM-сеть[5] - это искусственная нейронная сеть, содержащая LSTM-модули вместо или в дополнение к другим сетевым модулям. LSTM-модуль — это рекуррентный модуль сети, способный запоминать значения как на короткие, так и на длинные промежутки времени. Ключом к данной возможности является то, что LSTM-модуль не использует функцию активации внутри своих рекуррентных компонентов. Таким образом, хранимое значение не размывается во времени, и градиент или штраф не исчезает при использовании метода обратного распространения ошибки во времени (англ. Backpropagation through time) при тренировке сети.

LSTM-модули часто группируются в «блоки», содержащие различные LSTM-модули. Подобное устройство характерно для «глубоких» многослойных нейронных сетей и способствует выполнению параллельных вычислений с применением соответствующего оборудования.

Рис. 2 Простой LSTM-блок с тремя вентилями: входным, выходным и забывания. LSTM-блоки могут обладать большим числом вентиляей

РЕЗУЛЬТАТЫ

Для программной реализации предложенных методов, была использована открытая нейросетевая библиотека keras[6], написанная на языке Python[7]. Данная библиотека представляет собой надстройку над фреймворками Deeplearning4j[8], TensorFlow[9] и Theano[10].

Пример определения модели MLP:

```
model = Sequential()
model.add(Dense(512, input_shape=(num_words,)))
model.add(Activation('relu'))
model.add(Dropout(0.2))
model.add(Dense(num_classes))
model.add(Activation('softmax'))
model.compile(loss='categorical_crossentropy',
optimizer='adam',
metrics=['accuracy'])
```

Пример определения модели LSTM:

```
model = Sequential()
model.add(Embedding(num_words, vocab_size))
model.add(LSTM(32, dropout=0.2, recurrent_dropout=0.2))
model.add(Dense(num_classes, activation='sigmoid'))
model.compile(loss='binary_crossentropy',
optimizer='adam',
metrics=['accuracy'])
```

Для обучения ИНС были выбраны два небольших отрывка из технического[11] и художественного[12] текстов, объемом 500 слов каждый. Обучение ИНС было выполнено за 10 итераций (параметр epochs в примере ниже).

Пример обучения модели MLP:

```
history = model.fit(x_train, y_train, batch_size = 32, epochs = epochs, verbose = 1,
validation_split = 0.1)
```

Пример обучения модели LSTM:

```
history = model.fit(x_train, y_train, batch_size = 32, epochs = epochs,
validation_data=(x_test, y_test)),
```

где x_train - набор данных для обучения, преобразованный в числовой массив;

y_train – числовой массив категорий для обучения (0 – слово соответствует художественному тексту, 1 – слово соответствует техническому тексту);

x_test - набор данных для валидации модели после обучения, преобразованный в числовой массив;

y_test - числовой массив категорий для валидации модели после обучения (0 – слово соответствует художественному тексту, 1 – слово соответствует техническому тексту);

Для оценки качества классификации использовался показатель точности, вычисляемый алгоритмом нейронной сети. Данный показатель является вероятностью того, что случайно выбранные слова из какого-либо класса модель отнесет к тому классу, к которому слово действительно принадлежит.

Пример оценки точности методов:

```
score = model.evaluate(x_train, y_train, batch_size = 32, verbose = 1)
```

В таблице 1 для сравнения, представлены вычисленные значения точности и времени обучения для каждого метода. Если сравнить данные результаты, то можно заметить, что метод LSTM показал более высокую точность, но время обучения в данном случае значительно больше, чем при использовании метода многослойного перцептрона.

Таблица 1

Сравнение методов решения задачи анализа принадлежности текста

Классификатор	Точность	Время обучения
Метод многослойного перцептрона(MLP)	0,8	3 секунды
Метод долгой краткосрочной памяти(LSTM)	0,9	640 секунд

Для проверки результатов работы, на вход ИНС был подан массив слов «Rabbit network messages approximately conversation afterwards»:

```
ynew = model.predict_classes(test)
```

Оба метода определили принадлежность слов «network» и «messages» к техническому тексту, а «Rabbit», «approximately», «conversation», «afterwards» к художественному. Стоит отметить, что слово «approximately» программа классифицировала как относящееся к художественному тексту, несмотря на то, что данное слово использовалось только в обучающих данных из технического текста, что является статистической ошибкой второго рода. Причина такого результата может заключаться как в малом объеме обучающей выборки, так и в слабой классификационной возможности рассматриваемых методов в рамках решаемой задачи.

ОБСУЖДЕНИЕ

В ходе данной работы было произведено сравнение существующих методов обучения с учителем: метода многослойного персептрона(MLP) и метода длинной краткосрочной памяти(LSTM). Несмотря на то, что для обучения нейронной сети были выбраны небольшие тексты, время обучения сети по алгоритму LSTM значительно (в 213 раз) превысило время обучения сети по алгоритму MLP. Существенные дополнительные затраты времени дают увеличение точности всего на 10%.

ЗАКЛЮЧЕНИЯ

Задача классификации текста является одной из наиболее интересных задач в обработке естественного языка. За последние годы методы анализа текста на естественном языке получили достаточно большое развитие. В качестве продолжения исследований по теме данной статьи можно рассмотреть применение векторного представления слова, такие как word2vec [13] или sentence2vec [14]. Эти представления позволяют улучшить результат работы методов классификации в задаче классификации текстов, поэтому они являются перспективными для исследований в области анализа текстов.

1. Бейкер-Левандовски Н., Бенджио Ю., Винсент П. Распознавание аудио аккордов с повторяющимися нейронными сетями // Материалы 14-го Международного общества музыкальной информационно-поисковой конференции, 2013.
2. Лекан У., Боту Л., Беджио У., Хаффнер П. Градиентное обучение, применяемое для распознавания документов // Материалы IEEE, 1998.
3. Крижевский А., Суцкев И., Хинтон Г. Классификация ImageNet с глубокими сверточными нейронными сетями // Достижения в системах обработки нейронной информации, 2012.
4. <http://www.aiportal.ru/articles/neural-networks/multi-perceptron.html>
5. https://ru.wikipedia.org/wiki/%D0%94%D0%BE%D0%BB%D0%B3%D0%B0%D1%8F_%D0%BA%D1%80%D0%B0%D1%82%D0%BA%D0%BE%D1%81%D1%80%D0%BE%D1%87%D0%BD%D0%B0%D1%8F_%D0%BF%D0%B0%D0%BC%D1%8F%D1%82%D1%8C
6. <https://keras.io/>
7. <https://www.python.org/>
8. <https://deeplearning4j.org/>
9. <https://www.tensorflow.org/>
10. <http://www.deeplearning.net/software/theano/>
11. Тюрин С. Ю. Английский язык в сфере информационных и компьютерных технологий, 2012.
12. Кэрролл Л. Alice's Adventures in Wonderland, 1865..
13. Миколов Т., Чен К., Коррадо Г., Дин Дж. Эффективное оценивание представлений слов в векторном пространстве, 2013.
14. Лю Й., Сунь С., Лин Л., Чжао Ю., Ван Х. Вычисление сходства семантического текста с использованием функций, 2015.