# ПРИМЕНЕНИЕ СВЕРТОЧНОЙ НЕЙРОННОЙ СЕТИ ДЛЯ КЛАССИФИКАЦИИ ТЕКСТОВ ПРИ ИЗВЛЕЧЕНИИ ПРИЗНАКОВ НА УРОВНЕ СИМВОЛОВ

В.Г. Журавлев
Научный руководитель: В. Г. Спицын
Томский политехнический университет
vgz2@tpu.ru

**Введение**

Классификация текстов является одной из ключевых тем машинной обработки естественного языка, целью которой является определение принадлежности текстового документа к одному из ранее определённых классов. Задача понимания текста, написанного на естественном языке, предполагает определение явных или неявных признаков отдельных элементов текста, таких как слова, фразы, предложения и параграфы, с целью выделения тех или иных свойств текста [1].

Диапазон исследований, проводимых в рамках темы классификации текстов, варьируется от поиска наилучших текстовых признаков, до выбора наиболее подходящих классификаторов в рамках машинного обучения [2].

В последние годы, всё большую популярность набирает применение свёрточных нейронных сетей, в таких областях, как компьютерное зрение, распознавание голоса, а также обработка текстовой информации [3].

**Описание алгоритма**

Свёрточная нейронная сеть представляет собой технологию глубокого обучения, основанную на работе зрительной коры головного мозга млекопитающих. Идея свёрточных нейронных сетей заключается в применении свёрточных слоёв (англ. convolution) и слоёв подвыборки (англ. subsampling). При этом наиболее важную роль играет операция свёртки, заключающаяся в поэлементном умножении каждого фрагмента изображения на ядро свёртки, суммировании и записи результата в соответствующую позицию выходного изображения (рис. 1). Тем не менее, применение свёрточной нейронной сети продемонстрировало свою эффективность и в других областях, таких как обработка и классификации текстовой информации.

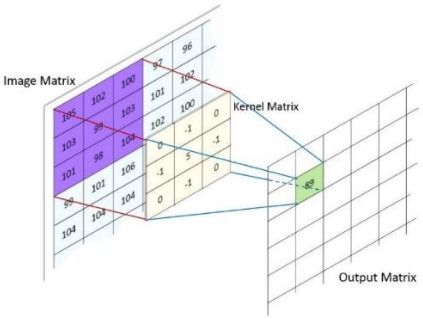


Рис. 1. Операция свёртки изображения

Архитектура свёрточной нейронной сети, описанная в [1], включает в себя шесть слоёв свёртки, три из которых используют слои подвыборки, а также три полносвязных слоя, один из которых является выходным (табл. 1, 2) (рис. 2).

Таблица 1. Свёрточные слои нейронной сети

| Layer | Large Feature | Small Feature | Kernel | Pool |
|---|---|---|---|---|
| 1 | 1024 | 256 | 7 | 3 |
| 2 | 1024 | 256 | 7 | 3 |
| 3 | 1024 | 256 | 3 | - |
| 4 | 1024 | 256 | 3 | - |
| 5 | 1024 | 256 | 3 | - |
| 6 | 1024 | 256 | 3 | 3 |

Таблица 2. Полносвязные слои нейронной сети

| Layer | Output Units Large | Output Unis Small |
|---|---|---|
| 7 | 2048 | 1024 |
| 8 | 2048 | 1024 |
| 9 | 4 | 4 |

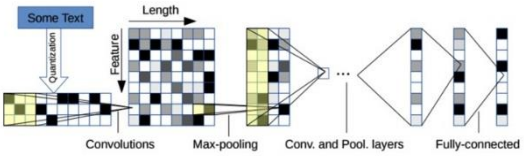


Рис. 2. Визуальное представление модели нейронной сети

Одной из проблем, возникающих в ходе решения задачи классификации текстов, является представление текстовой информации в форме, пригодной для использования нейронной сетью. Извлечение признаков из текста может осуществляться на уровне отдельных слов [4]. При этом могут применяться такие технологии, как one-hot кодировка, предполагающая использование векторов фиксированный длины, а также word2vec – технология от Google, позволяющая получить векторное представление слов на основе статистической информации об их использовании [5].

Однако, это приводит к появлению проблемы «out-of-vocabulary», при которой нейронная сеть не способна корректно обработать слово, которое не вошло в словарь на этапе обучении сети. Решение проблемы можно увидеть в обработке текстовой информации на уровне отдельных символов, при которой каждому символу из заранее определенного набора соответствует целочисленный индекс.

Обучение с учителем предполагает использование того или иного набора данных. В качестве такого набора данных может быть применён AG's News Topic Classification Dataset, который представляет собой коллекцию из более чем одного миллиона газетных заголовков новостей, каждый

из которых может быть отнесён к одному из четырех предопределённых классов (World, Sports, Business, Sci/Tech).

**Тестирование алгоритма поиска**

Для проверки работы приведенной архитектуры нейронной сети был использован фреймворк машинного обучения Keras, основанный на фреймворке Tensorflow от Google. Набор данных, применяемых для обучения свёрточной нейронной сети, включает в себя 120000 обучающих и 7600 проверочных примеров.

В целях улучшения качества обучения был применён оптимизационный алгоритм Adam. Функция потерь представляет собой categorical cross-entropy. В целях ускорения процесса обучения была использования видеокарта NVIDIA GeForce GTX 1050 Ti.

Обучение длилось в течении 10 эпох. В конечном итоге, на одной их эпох обучения удалось добиться точности, равной 89,87% на тестовой выборке. Иными словами, ошибка сети на приведенном наборе данных составила 10,13%, что сравнимо с показателями, приведёнными в [1] (рис. 3).

| Model | AG |
|---|---|
| BoW | 11.19 |
| BoW TFIDF | 10.36 |
| ngrams | 7.96 |
| ngrams TFIDF | **7.64** |
| Bag-of-means | **16.91** |
| LSTM | 13.94 |
| Lg. w2v Conv. | 9.92 |
| Sm. w2v Conv. | 11.35 |
| Lg. w2v Conv. Th. | 9.91 |
| Sm. w2v Conv. Th. | 10.88 |
| Lg. Lk. Conv. | 8.55 |
| Sm. Lk. Conv. | 10.87 |
| Lg. Lk. Conv. Th. | 8.93 |
| Sm. Lk. Conv. Th. | 9.12 |
| Lg. Full Conv. | 9.85 |
| Sm. Full Conv. | 11.59 |
| Lg. Full Conv. Th. | 9.51 |
| Sm. Full Conv. Th. | 10.89 |
| Lg. Conv. | 12.82 |
| Sm. Conv. | 15.65 |
| Lg. Conv. Th. | 13.39 |
| Sm. Conv. Th. | 14.80 |

Рис. 3. Показатели ошибки сети на тестовом наборе данных AG News

На рисунке 4 можно увидеть динамику изменения значения функции потерь, на рисунке 5 – динамику изменения точности (accuracy) на тестовой выборке.

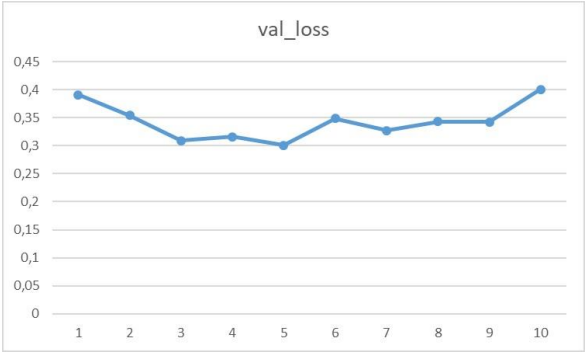


Рис. 4. Изменение значения функции потерь val_loss в ходе обучения


Рис. 5. Изменение значения точности val_acc в ходе обучения

**Заключение**

В результате выполнения работы, был проведён анализ методов векторизации текста, исследованы возможные проблемы и пути их решения.

Осуществлена успешная реализация модели свёрточной нейронной сети для задачи классификации текста на основе набора данных AG's News.

Подтверждено, что подобная архитектура нейронной сети позволяет добиться высокой точности классификации, достигающей 89,87 % на тестовой выборке. Работа выполнена при поддержке гранта РФФИ № 18-08-00977 А.

**Список использованных источников**

1. Zhang, X. Character-level convolutional networks for text classification / Xiang Zhang, Junbo Zhao, Yann LeCun // In Advances in Neural Information Processing Systems. 2015.Feb. 649 - 657 p.
2. LeCun, X. Z. Y. Text understanding from scratch / Xiang Zhang Yann LeCun // Computer Science Department. 2016.
3. Krizhevsky, A. Imagenet classification with deep convolutional neural networks / Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton // NIPS.2012.1106 -1114 p.
4. Kim, Y. Convolutional neural networks for sentence classification / Yoon Kim // IEMNLP. 2014.Sep. 1746 -1751 p.
5. Mikolov, Tomas; et al. "Efficient Estimation of Word Representations in Vector-Space".arXiv:1301.3781.