УДК 004.45 **ТЕХНИЧЕСКИЕ НАУКИ**

Стрелец Андрей Иванович, *магистр*
кафедры «Компьютерные системы и технологии»,
Национальный исследовательский ядерный университет «МИФИ»
Россия, г. Москва
Черникова Елена Андреевна, *магистр*
кафедра «Компьютерные системы и технологии»,
Национальный исследовательский ядерный университет «МИФИ»
Россия, г. Москва
Дороничев Никита Андреевич, *магистр*
кафедры «Компьютерные системы и технологии»,
Национальный исследовательский ядерный университет «МИФИ»
Россия, г. Москва
Сычев Максим Игоревич, *магистр*
кафедры «Компьютерные системы и технологии»,
Национальный исследовательский ядерный университет «МИФИ»
Россия, г. Москва

ПОСТРОЕНИЕ МОДЕЛИ НЕЙРОННОЙ СЕТИ ДЛЯ АНАЛИЗА ТЕКСТОВЫХ ДАННЫХ

Аннотация: Данная статья описывает ключевые моменты построения модели нейронной сети для анализа текстовых данных. Проведена оценки построенной модели.

Ключевые слова: нейронная сеть, машинное обучение, большие данные, анализ текста.

Abstract: This article is about most important parts of neural network for analysis of text data. Article contains evaluation of developed model.

Key words: neural network, machine learning, big data, text analysis.

Введение

В современном мире анализ данных большого объёма является одним из наиболее перспективных направлений в области IT-технологий. Одним из наиболее распространённых форматов представления данных является текст [1]. Текстовые данные, наряду с данными в формате изображений, используются в качестве хранения информации, анализируемой поисковыми системами [2]. Также текстовый формат является наиболее популярным форматом представления данных в социальных сетях и в приложениях для обмена сообщениям. Для анализа текстовых данных используются различные подходы. Нейронные сети на основе машинного обучения являются новаторским подходом в области анализа текстовых данных [3]. В рамках данной статьи рассмотрено построение модели нейронной сети для анализа больших объёмов текста. Также проведён анализ разработанной модели с помощью опробования модели на тестовом наборе данных.

Построение модели

Построение модели включает в себя выбор функции потерь, функции активации для каждого слоя, выбор количества слоёв и количества нейронов в каждом слое, а также выбор оптимизатора.

Оптимизатор определяет то, каким образом будут изменяться весовые коэффициенты сети во время обучения под воздействием функции потерь. Оптимизатор реализует определённый вариант стохастического градиентного спуска. Оптимизатор rmsprop в общем случае является наиболее подходящим для большинства задач. Функция потерь возвращает количественную оценку, которая будет минимизироваться в процессе обучения. Фактически рассматривается как мера успеха в решении стоящей задачи.

Функция активации – это фактор нелинейности нейронной сети. Без него нейронная сеть представляла бы собой совокупность двух линейных операций: скалярного произведения и сложения, однако линейное пространство гипотез

является слишком ограниченным, и использование нескольких слоёв не было бы оправданным, так как стек линейных слоёв всё равно реализует линейную операцию. Нелинейная функция активации даёт доступ к более обширному пространству гипотез. Самой популярной функцией активации является функция relu:

$$f(x) = x^{+} = \max(0, x)$$

Количество слоёв и конфигурация каждого слоя подбираются экспериментальным путём. Для правильного выбора количества слоёв и количества нейронов необходимо пробовать различные конфигурации и проводить оценку качества классификации. Обычно при большем количестве слоёв и нейронов в каждом слое сеть обучается быстрее (за меньшее количество эпох), но и переобучается также быстрее [4]. При меньшем же количестве слоёв наблюдается обратная картина. Однако если две конфигурации дают одинаковый результат, то стоит выбирать ту, у которой меньше слоёв и меньше нейронов в каждом слое, потому что она обладает большей общностью (под общностью имеется в виду качество предсказания на данных, которые нейросеть прежде не видела), так как память такой нейросети меньше, а следовательно, меньше вероятность запоминания нерелевантных шаблонов в данных.

В данной работе количество слоёв было выбрано экспериментальным путём через оценку точности нейросети с каждым количеством слоёв. В качестве оптимизатора был выбран rmsprop, а функции потерь и активации выбирались с учётом того, насколько они подходят для решения конкретной задачи классификации.

Оценка модели

При оценке результатов нейронной сети необходимо помнить главное правило: нейронная сеть не должна работать с тестовыми данными до оценки результатов. Тестовые данные не должны использоваться ни для выбора

параметров, ни для выбора гиперпараметров (параметрами сети обычно называют весовые коэффициенты нейросетей, поэтому количество слоёв и размер слоя обычно называют гиперпараметрами модели). Для настройки параметров и гиперпараметров нужно использовать тренировочную и проверочную выборки. Часто исследователи пренебрегают данным правилом, в результате чего на практике их модели показывают худшие результаты, нежели объявлено изначально. Для оценки построенной модели сформирована тестовая выборка, отсутствующая в наборе данных CIFAR-10, и проверены различные нейросетевые модели, обнаружено падение от 4% до 10% относительно результата на изначальной тестовой выборке.

Поэтому в данной работе для итоговой оценки результатов нейронной сети также использовались тестовые данные, отделённые от обучающей и проверочной выборок на начальном этапе. Тестовые данные не использовались ни для выбора параметров, ни для выбора гиперпараметров сети. Тестовые данные были выбраны из всех доступных размеченных данных путём случайного выбора для увеличения непредвзятости оценки работы нейросети.

Заключение

В данной статье рассмотрена модель для анализа текстовых данных больших объёмов. Проведен анализ разработанной модели на тестовом наборе данных и наборе данных CIFAR-10, являющихся эталонными для подобных нейронных моделей. Разработанная модель может быть использована для анализа текстовых данных большого объёма в поисковых механизмах или для анализа текстовых данных в социальных сетях.

Библиографический список:

1. S. M. Metev and V. P. Veiko, Laser Assisted Microtechnology, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.

2. J. Breckling, Ed., The Analysis of Directional Time Series: Applications to Wind Speed and Direction, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.

3. S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, “A novel ultrathin elevated channel low-temperature poly-Si TFT,” IEEE Electron Device Lett., vol. 20, pp. 569–571, Nov. 1999.

4. O. Sallou and C. Monjeaud., Go-Docker, A batch scheduling system with Docker containers. IEEE International Conference of Cluster Computing, pp. 514-515, 2015.